Avoiding Optimal Mean Robust and Sparse BPCA with L1-norm Maximization

*Ganyi Tang, Lili Fan, Jianguo Shi, Jingjing Tan, Guifu Lu**

School of Computer and Information, Anhui Polytechnic University, China
tom365@ahpu.edu.cn, fanlili@ahpu.edu.cn, shijianguo@ahpu.edu.cn, seokgiz@126.com, luguifu_jsj@163.com

Abstract

Recently, the robust PCA/2DPCA methods have achieved great success in subspace learning. Nevertheless, most of them have a basic premise that the average of samples is zero and the optimal mean is the center of the data. Actually, this premise only applies to PCA/2DPCA methods based on L2-norm. The robust PCA/2DPCA method with L1-norm has an optimal mean deviate from zero, and the estimation of the optimal mean leads to an expensive calculation. Another shortcoming of PCA/2DPCA is that it does not pay enough attention to the instinct correlation within the part of data. To tackle these issues, we introduce the maximum variance of samples' difference into Block principal component analysis (BPCA) and propose a robust method for avoiding the optimal mean to extract orthonormal features. BPCA, which is generalized from PCA and 2DPCA, is a general PCA/2DPCA framework specialized in part learning, can makes better use of the partial correlation. However, projection features without sparsity not only have higher computational complexity, but also lack semantic properties. We integrate the elastic network into avoiding optimal mean robust BPCA to perform sparse constraints on projection features. These two BPCA methods (non-sparse and sparse) make the presumption of zero-mean data unnecessary and avoid optimal mean calculation. Experiments on reference benchmark databases indicate the usefulness of the proposed two methods in image classification and image reconstruction.

Keywords: BPCA, Avoiding optimal mean, Sparse modeling, L1-norm, Elastic net

1 Introduction

PCA [1] is a classical subspace learning and feature extraction technique wildly used in the areas of data analysis, pattern recognition, computer vision, etc. It uses pixelwise covariance, where two-dimensional images are expressed as long vectors. 2DPCA, which preserves the 2D spatial structure of images, was firstly developed by Yang and Zhang [2]. Although PCA/2DPCA have been successfully applied to many domains, they are sensitive to outliers as a result of the occupation of the L2-norm in the optimization formulation. In recent years, many L1-norm-based schemes were devised.

L1-PCA [3] and R1-PCA [4] have complex characteristics of finding optimal basis features through linear or quadratic programming, and they are computationally expensive. Kwak proposed an intuitive and simple PCA with L1-norm (PCA-L1) [5], which is invariant to rotations. Li et al. proposed robust 2DPCA with L1-norm referred to as 2DPCA-L1 [6], which avoids transforming 2D images into 1D vectors. For extracting all directional features simultaneously, Nie et al. proposed robust non-greedy version of PCA-L1 referred to as PCAL1-nongreedy [7], and Wang et al. put forward 2DPCAL1-nongreedy [8] subsequently. Since both L2-norm and L1-norm are particular cases of Lp-norm ($0 < p \leq 2$), Kwak et al. naturally proposed Lp-norm-based PCA (PCA-Lp) [9].

Taking the rows of the image matrix as the units of computation, two-dimensional PCA can be rewritten under the umbrella of PCA [10]. Block PCA [11-12] does not compute directly by vector units, but divides each data matrix into several pieces containing some rows and columns, and then expresses blocks-based process in the form of PCA. From this perspective, 2DPCA and PCA are particular cases of BPCA. Replaced L2-norm with L1-norm, Wang et al. proposed a robust L1-norm-based BPCA (BPCA-L1) [13], and Li et al. presented corresponding non-greedy method (BPCA-L1 non-greedy) [14].

The robust PCA techniques usually fixed optimal mean as the average of samples and presume the data are formerly centered. However, this presumption, which practically neglects the processing of mean optimization, is unsubstantial. Actually, it is a theoretical guarantee only in L2-norm-based optimization. Moreover, the outliers often tend to bias the predetermined mean of high-dimensional data and degrade the performance [15-17]. To deal with this issue, Luo et al. [18-19] developed robust PCA/2DPCA with avoiding optimal mean, which maximize the sum of projected differences.

The eigenvectors learned by the aforementioned approaches are still dense. Non-sparse features may have redundant information, which is not only poor in semantics, but also difficult to guarantee performance. It is highly advantageous to find the most relevant or outstanding element from many characteristics. H. Wang et al. proposed a robust and sparse model of 2DPCA with L1-norm (2DPCAL1-S) [20]. And then, J. Wang et al. presented a generalized sparse model of 2DPCA with Lp-norm (G2DPCA) [21].

Inspired by these works, we generalize two novel BPCA

*Corresponding Author: Guifu Lu; E-mail: luguifu_jsj@163.com
DOI: 10.53106/160792642023072404016

models with avoiding optimal mean.

The major merits of our works are list as follows:

1) The methods we proposed are under a unified subspace-learning framework of BPCA, which makes better use of the local correlation of the neighboring pixels. Under the umbrella of BPCA, the L1-norm based PCA/2DPCA with robustness and sparsity can be reformulated.

2) The Avoiding Optimal Mean L1-norm-based BPCA (BPCAL1-AOM) we proposed maximize the metric of samples' difference projection to extract feature vectors and avoid mean optimization quite sensibly.

3) Above BPCAL1-AOM, we further propose a sparse model denoted as BPCAL1S-AOM, by which we acquire sparse features that have superior semantics and help to achieve better performance.

2 Brief Review

In this section, we revisit the methods of L1-norm-based PCA/2DPCA and BPCA.

2.1 PCA-L1 Revisited

Suppose the samples is $X = \{x_1, ..., x_n\} \in R^{d\times n}$, where n is the number of the given data and d is the dimension. Generally, we assume the mean of all x_i (i = 1, 2, …, n) is c ($c = \frac{1}{n}\sum_{i=1}^{n} x_i$). PCA is a dimensionality reduction model, which seeks the optimal linear subspace to solve the following problems:

$$\begin{aligned} &R^* = \arg\min_R \sum_{i=1}^{n} \left\| (x_i - c) - RR^T (x_i - c) \right\|_2^2, \\ &\text{subject to } RR^T = I, \end{aligned} \tag{1}$$

where $\|\cdot\|_2$ denote L2-norm.

Eq. (1) is the Minimum-error Formulation.

Correspondingly, PCA is also solving the dual problem:

$$\begin{aligned} &R^* = \arg\max_R \sum_{i=1}^{n} \left\| R^T (x_i - c) \right\|_2^2 \\ &\text{subject to } RR^T = I. \end{aligned} \tag{2}$$

Eq. (2) is the Maximum Variance Formulation.

The equivalence of problem (1) and problem (2) can be proved directly by Lemma 1 [19].

Lemma 1.

$$\left\| y - RR^T y \right\|_2^2 + \left\| R^T y \right\|_2^2 = \left\| y \right\|_2^2. \tag{3}$$

Proof. We represented the L2-norm by the trace, the left side of Eq. (3) is

$$\begin{aligned} &\left\| y - RR^T y \right\|_2^2 + \left\| R^T y \right\|_2^2 \\ &= tr((y - RR^T y)(y - RR^T y)^T) \\ &\quad + tr(R^T y y^T R) \\ &= tr(yy^T - yy^T RR^T - RR^T yy^T \\ &\quad + RR^T yy^T RR^T) + tr(R^T yy^T R). \end{aligned} \tag{4}$$

By the formulations as follows

$$\begin{aligned} &tr(U + V) = tr(U) + tr(V), \\ &tr(UV) = tr(VU). \end{aligned} \tag{5}$$

Eq. (4) is converted to

$$\begin{aligned} &tr(yy^T) - tr(yy^T RR^T) - tr(yy^T RR^T) \\ &\quad + tr(RR^T yy^T RR^T) + tr(R^T yy^T R) \\ &= tr(yy^T) = \left\| y \right\|_2^2. \end{aligned} \tag{6}$$

The proof is completed.

Classical PCA is sensitive to outliers since L2-norm exaggerates the effect of noise. Kwad replaced L2-norm with L1-norm in maximum variance formulation and proposed a simple and intuitive robust PCA, which is referred to as PCA-L1 [5]:

$$\begin{aligned} &r^* = \arg\max_r \sum_{i=1}^{n} \left| r^T (x_i - c) \right|, \\ &\text{subject to } rr^T = 1. \end{aligned} \tag{7}$$

PCA-L1 seeks the optimal feature r^* through an iterative procedure to solve the nonlinear optimization upon the objective function containing an absolute value. Define a formulation of the projection vector r (t + 1) at the (t + 1)th iteration as

$$r(t+1) = \frac{\sum_{i=1}^{n} p_i(t)(x_i - c)}{\left\| \sum_{i=1}^{n} p_i(t)(x_i - c) \right\|_2}, \tag{8}$$

where $p_i(t)$ is polarity function, which defined as

$$p_i(t) = \begin{cases} 1, & r^T(t)(x_i - c) \geq 0 \\ -1, & r^T(t)(x_i - c) < 0 \end{cases}. \tag{9}$$

2.2 2 DPCA-L1 Revisited

Let $X = \{X_1, ..., X_n\}$ be the n sample matrices without mean-centered. The dimension of X_i = (1, 2,..., n) is $h \times w$. The purpose of 2DPCA-L1 is to tackle the following maximum optimization problem

$$r^* = \arg\max_r \sum_{i=1}^{n} \|(X_i - C)r\|_1 = \arg\max_r \sum_{i=1}^{n}\sum_{j=1}^{h} \left|r^T(x_{ij} - c_j)\right|, \quad \text{subject to } rr^T = 1, \tag{10}$$

where $c_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and x_{ij} is the jth row transpose of the ith image matrix. i.e.,

$$X_i = \begin{bmatrix} x_{i1}^T \\ \vdots \\ x_{ih}^T \end{bmatrix}. \tag{11}$$

The iterative formulation of $r(t+1)$ is

$$r(t+1) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{h} p_{ij}(t)(x_{ij} - c_j)}{\left\|\sum_{i=1}^{n}\sum_{j=1}^{h} p_{ij}(t)(x_{ij} - c_j)\right\|_2}, \tag{12}$$

where $p_{ij}(t)$ is polarity function defined as

$$p_{ij}(t) = \begin{cases} 1, & r(t)^T(x_{ij} - c_j) \geq 0 \\ -1, & r(t)^T(x_{ij} - c_j) < 0 \end{cases}. \tag{13}$$

2.3 BPCA-L1 Revisited

Eq. (8) and Eq. (12) show that 2DPCA-L1 can be reformulated as the form of PCA-L1. If we generalize row vectors to general pixels blocks, we can extend 2DPCA-L1 to BPCA-L1.

Let $X = \{X_1, X_2, ..., X_n\}$ denotes n image matrices without mean-centered. BPCA-L1 method separates each image into m small blocks, which have the same number of pixels in a small block, i.e. $X_i = \{b_1^{(i)}, b_2^{(i)}, \cdots, b_m^{(i)}\}$. When the block $b_k^{(i)}$ transforms to a vector $x_k^{(i)}$, we get the vectorization version of X_i, which is denoted as $X_i^v = \{x_1^{(v)}, x_2^{(v)}, \cdots x_m^{(v)}\}$. The vectorization is illustrated as Figure 1.

BPCA-L1 tends to seek a projection feature r^* to maximize the L1-norm-based variance, i.e.,

$$r^* = \arg\max_r \sum_{i=1}^{n}\sum_{j=1}^{m} \left|r^T(x_j^{(i)} - c_j)\right|, \quad \text{subject to } r^T r = 1, \tag{14}$$

where $c_j = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}$.

To address this issue, the iterative formulation of $r(t+1)$ is defined as

$$r(t+1) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(t)(x_j^{(i)} - c_j)}{\left\|\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(t)(x_j^{(i)} - c_j)\right\|_2}, \tag{15}$$

where $p_{ij}(t)$ is polarity function, i.e.,

$$p_{ij}(t) = \begin{cases} 1, & r(t)^T(x_j^{(i)} - c_j) \geq 0 \\ -1, & r(t)^T(x_j^{(i)} - c_j) < 0 \end{cases}. \tag{16}$$

Compared with Eq. (8) & Eq. (9) of PCA-L1, Eq. (12) & Eq. (13) of 2DPCA-L1 and Eq. (15) & Eq. (16) of BPCA-L1, we can see that BPCA-L1 is a general framework. Both PCA-L1 and 2DPCA-L1 are special cases of BPCA-L1. If each matrix is only one block, then BPCA-L1 degenerates into PCA-L1. If each row of the matrix is a block, BPCA-L1 becomes 2DPCA-L1.

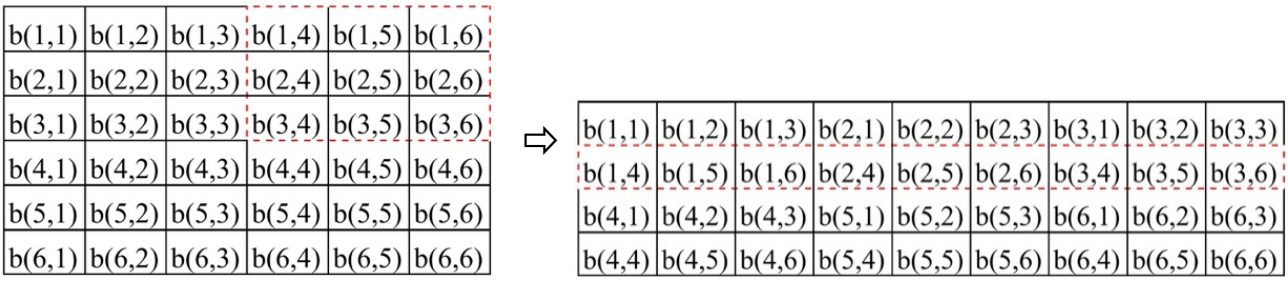

Figure 1. Transforming original image matrix to 3×3 block based vectored representation

3 Problem Formulation

In Eq. (7), Eq. (10), Eq. (14), sample mean is assumed to be the optimal mean of x_i. However, this assumption is incorrect in robust L1-norm-based BPCA methods. Lots of research work, such as optimal mean methods [15-17] or avoiding optimal mean methods [18-19] have attracted extensive attention. Inspired by these works, we generalize the BPCA framework with avoiding optimal mean.

3.1 Avoiding Optimal Mean BPCA-L1

Let $X = \{X_1, \ldots, X_n\}$ denote n image matrices without mean-centered. Separate each image into m small blocks with the same number of pixels and get vectorization version of X_i, which is denoted as X_i^v, i.e., $X_i^v = \{x_1^{(i)}, \cdots, x_m^{(i)}\}$.

The sample mean of $X_i^v (i = 1,2,\cdots,n)$ is $C = \frac{1}{n}\sum_i^n X_i^v$, that is to say the kth column of C is $c_k = \frac{1}{n}\sum_{i=1}^n x_k^i$.

Theorem 1. The solution R^* of BPCA which minimizes the reconstruction error based on Euclidean distance, i.e.,

$$R^* = \arg\min_R \sum_{i=1}^n \left\|(X_i^v - A) - RR^T(X_i^v - A)\right\|_2^2, \quad \text{subject to } R^T R = I, \tag{17}$$

is likewise the solving strategy of the following issue

$$R^* = \arg\max_R \sum_{i,j} \left\|R^T(X_i^v - X_j^v)\right\|_2^2, \quad \text{subject to } R^T R = I. \tag{18}$$

Proof.

In Eq. (18), Let $F = \sum_{i=1}^n \left\|(X_i^v - A) - RR^T(X_i^v - A)\right\|_2^2$, then expand the L2-norm of the matrix in terms of the column vectors, we get

$$F = \sum_i^n \sum_k^m \left\|(x_k^{(i)} - a_k) - RR^T(x_k^{(i)} - a_k)\right\|_2^2. \tag{19}$$

F reaches a local maximum only if $\frac{\partial F}{\partial a_k} = 0$, that is

$$\frac{\partial F}{\partial a_k} = \frac{\partial \sum_{i=1}^n \left\|(I - RR^T)(x_k^{(i)} - a_k)\right\|_2^2}{\partial a_k} = \sum_{i=1}^n -2(I - RR^T)^2(x_k^{(i)} - a_k) = 0, \tag{20}$$

namely, the optimal mean is the mean of samples, i.e.,

$$a_k = \frac{1}{n}\sum_{i=1}^n x_k^{(i)}. \tag{21}$$

In PCA/2DPCA, variance maximization and reconstruction error minimization are dual problems. So, Eq. (17) can be converted to the following expression:

$$R^* = \arg\max_R G = \arg\max_R \sum_{i=1}^n \left\|R^T(X_i^v - A)\right\|_2^2, \quad \text{subject to } R^T R = I. \tag{22}$$

In Eq. (22), we expand the L2-norm of the matrix in terms of the column vectors, then

$$G = \sum_i^n \sum_k^m \left\|R^T(x_k^{(i)} - a_k)\right\|_2^2. \tag{23}$$

We denote the square L2-norm in G as trace, that is

$$\begin{aligned} G &= \sum_{i=1}^n \sum_{k=1}^m tr((x_k^{(i)} - a_k)^T RR^T(x_k^{(i)} - a_k)) \\ &= \sum_{i=1}^n \sum_{k=1}^m tr(x_k^{(i)T} RR^T x_k^{(i)}) \\ &\quad - \sum_{i=1}^n \sum_{k=1}^m tr(a_k^T RR^T x_k^{(i)}) \\ &\quad - \sum_{i=1}^n \sum_{k=1}^m tr(x_k^{(i)T} RR^T a_k) \\ &\quad + \sum_{i=1}^n \sum_{k=1}^m tr(a_k^T RR^T a_k). \end{aligned} \tag{24}$$

Let's substitute $a_k = \frac{1}{n}\sum_{j=1}^n x_k^{(j)}$ into Eq. (24) and reformulate it as

$$\begin{aligned} G &= \sum_{i=1}^n \sum_{k=1}^m tr(x_k^{(i)T} RR^T x_k^{(i)}) \\ &\quad - \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^m \sum_{j=1}^n tr(x_k^{(j)T} RR^T x_k^{(i)}) \\ &\quad - \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^m \sum_{j=1}^n tr(x_k^{(i)T} RR^T x_k^{(j)}) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^m \sum_{j=1}^n tr(x_k^{(i)T} RR^T x_k^{(j)}). \end{aligned} \tag{25}$$

The trace of a scalar is itself, so

$$\begin{aligned} G &= \sum_{k=1}^m \sum_{i=1}^n x_k^{(i)T} RR^T x_k^{(i)} \\ &\quad - \frac{1}{n}\sum_{k=1}^m \sum_{i=1}^n \sum_{j=1}^n x_k^{(i)T} RR^T x_k^{(j)}. \end{aligned} \tag{26}$$

In Eq. (18), let $H = \sum_{i,j} \left\| R^T (X_i^v - X_j^v) \right\|_2^2$. Expand the square L2-norm of the matrix in H, wet get

$$H = \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| R^T (x_k^{(i)} - x_k^{(j)}) \right\|_2^2. \tag{27}$$

Denote the square L2-norm in H as trace, i.e.

$$\begin{aligned} H = & \\ & \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} tr((x_k^{(i)T} - x_k^{(j)T}) R R^T (x_k^{(i)} - x_k^{(j)})) \\ = & \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} tr(x_k^{(i)T} R R^T x_k^{(i)}) \\ & - \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} tr(x_k^{(i)T} R R^T x_k^{(j)}) \\ & - \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} tr(x_k^{(j)T} R R^T x_k^{(i)}) \\ & + \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} tr(x_k^{(j)T} R R^T x_k^{(j)}). \end{aligned} \tag{28}$$

It is easy to turn into

$$\begin{aligned} H = & 2n \sum_{k=1}^{m} \sum_{i=1}^{n} x_k^{(i)T} R R^T x_k^{(i)} \\ & - 2 \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} x_k^{(i)T} R R^T x_k^{(j)}. \end{aligned} \tag{29}$$

In keeping with Eq. (26) and Eq. (29), G and H have the same optimal solution obviously.

The proof is completed.

We replace the L2-norm of Eq. (18) in **Theorem 1** with L1-norm, then get the following problem formulation of Avoiding Optimal Mean BPCA-L1 which we refer to as BPCAL1-AOM.

$$\begin{aligned} & R^* = \arg\max_R \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| R^T (x_k^{(i)} - x_k^{(j)}) \right\|_1, \\ & \text{subject to } R^T R = I. \end{aligned} \tag{30}$$

Note that the robust L1-norm-based BPCA methods often incorrectly uses the sample mean as the optimal mean, which affects the model performance. Problem formulation of Eq. (30) automatically avoids computing the optimal mean based on the L1-norm and makes assumptions on the central data unnecessary.

3.2 Avoiding Optimal Mean BPCA-L1 with Sparsity

Features extracted from images may contain redundant or irrelevant elements. Generally, a few outstanding features are very significant for image recognition, and these features often correspond to some key areas in the image, like mouth or eyes in the face. It makes sense to look for the most outstanding or significant elements from a great many of features. Sparsity usually contributes to the efficient execution of the algorithm, and sparse representation often corresponds to the local linear structure of the data. Sparse feature is efficient for classification and easy to interpret. Features extractions with the constraints of L1-norm, which is a common method for sparsity. Elastic net linearly combines L1-norm-measured lasso penalty with L2-norm-measured ridge penalty, which can overcome some limitations of lasso [20, 22]. In this methodology, we add constraints of L1-norm to variance and elastic net to feature vectors themselves to make features more robust and sparse.

The objective function is designed as

$$\begin{aligned} & G(R) \\ & = \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| R^T (x_k^{(i)} - x_k^{(j)}) \right\|_1 - \frac{\eta}{2} \|R\|_2^2 - \lambda \|R\|_1, \end{aligned} \tag{31}$$

where R denotes projection matrix and $R^T R = I$, η and λ are regulable coefficients of sparsity.

The problem formulation of Avoiding Optimal Mean BPCA-L1 with sparsity (BPCAL1S-AOM) is described as

$$\begin{aligned} R^* = \arg\min_R (& \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| R^T (x_k^{(i)} - x_k^{(j)}) \right\|_1 \\ & - \frac{\eta}{2} \|R\|_2^2 - \lambda \|R\|_1), \\ & \text{subject to } R^T R = I. \end{aligned} \tag{32}$$

4 Problem Solution

The optimizations of Eq. (30) and Eq. (32) are difficult because L1-norm is nonlinear. In this section, we propose greedy strategies to address the multi-feature extraction problems.

4.1 Extracting Features by BPCAL1-AOM

Now, we develop a novel features extraction algorithm for BPCAL1-AOM to search the optimal feature r^* that maximizes the variance of Eq. (30).

Algorithm BPCAL1-AOM for d ($d > 0$) features extraction

1. Let $\tau = 1$. τ is a counter for number of features
2. Find the optimal projection vector r_τ
   1) Let $t = 0$. t is a counter for iterations
   2) Initialization:
      Generate $r(t)$ randomly subject to $r(t)^T r(t) = 1$
   3) Let $p_{ijk}(t) = \begin{cases} 1, & r(t)^T (x_k^{(i)} - x_k^{(j)}) \geq 0 \\ -1, & r(t)^T (x_k^{(i)} - x_k^{(j)}) < 0 \end{cases}$
   4) Let $r(t+1) = \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ijk}(t)(x_k^{(i)} - x_k^{(j)})$
   5) Normalization: $r(t+1) = r(t+1) / \|r(t+1)\|_2$
   6) Convergence test:
      if $\|r(t+1) - r(t)\|_2 > \varepsilon$ then go to step 3)

else $r_\tau \leftarrow r$, exit the iteration and go to 3.

7) $t = t + 1$

3. get deflated samples for greedy strategy:

For all $i \in \{1, 2, ..., n\}$ do

$$X_i^v(\tau+1) = X_i^v(\tau) - r_\tau(r_\tau^T X_i^v(\tau))$$

4. If $\tau < d$, let $\tau = \tau + 1$ and go to step 2

The convergence of the algorithm of BPCAL1-AOM can be validated.

Theorem 2. In the above procedure of BPCAL1-AOM, the objective function holds non-decreasing, that is

$$\begin{aligned}&\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|r(t+1)^T(x_k^{(i)}-x_k^{(j)})\right| \\ &\geq \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|r(t)^T(x_k^{(i)}-x_k^{(j)})\right|.\end{aligned} \tag{33}$$

Proof.

In the right of inequality in Theorem 2,

$$\begin{aligned}&\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|r(t)^T(x_k^{(i)}-x_k^{(j)})\right| \\ &=\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}r(t)^T p_{ijk}(t)(x_k^{(i)}-x_k^{(j)}).\end{aligned}$$

In the left of inequality in Theorem 2,

$$\begin{aligned}&\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|r(t+1)^T(x_k^{(i)}-x_k^{(j)}\right| \\ &=\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t+1)r(t+1)^T(x_k^{(i)}-x_k^{(j)}) \\ &\geq\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)r(t+1)^T(x_k^{(i)}-x_k^{(j)}) \\ &=r(t+1)^T(\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)(x_k^{(i)}-x_k^{(j)})).\end{aligned}$$

In the above procedure of BPCAL1-AOM algorithm

$$r(t+1)=\frac{\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)(x_k^{(i)}-x_k^{(j)})}{\left\|\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)(x_k^{(i)}-x_k^{(j)})\right\|_2},$$

Namely,

$r\,(t+1)$ and $\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)(x_k^{(i)}-x_k^{(j)})$ are parallel.

It means that

$$\begin{aligned}&r(t+1)^T(\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)(x_k^{(i)}-x_k^{(j)})) \\ &\geq r(t)^T(\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)(x_k^{(i)}-x_k^{(j)})).\end{aligned}$$

So, the proof is completed.

Because $G(r(t))$ is non-decreasing and the number of samples is limited, the iterative process of BPCAL1-AOM is convergent.

Theorem 3. In BPCAL1-AOM, the orthonormality of the features is guaranteed.

Proof.

By multiplying r_τ^T to the deflated formulation of samples in BPCAL1-AOM, we get

$$\begin{aligned}&r_\tau^T X_i^v(\tau+1) \\ &=r_\tau^T X_i^v(\tau)-r_\tau^T r_\tau(r_\tau^T X_i^v(\tau)) \\ &=r_\tau^T X_i^v(\tau)-r_\tau^T X_i^v(\tau)=0.\end{aligned}$$

That is to say, r_τ is orthogonal to $X_i^v(\tau+1)$.

On the other hand, $r_{\tau+1}$ is a linear representation of the samples of $X_i^v(\tau+1), i=1,2,\cdots n$. That means $r_{\tau+1}$ is parallel to $X_i^v(\tau+1)$. So, $r_{\tau+1}$ is orthogonal to r_τ.

In BPCAL1-AOM procedure, each r_τ is normalized obviously. The proof is completed.

4.2 Extracting Features by BPCAL1S-AOM

To address the problem of Eq. (32), we proposed a novel robust and sparse BPCA-L1 method with avoiding optimize mean, which is referred to as BPCAL1S-AOM.

As we all know, it is extremely hard to obtain multiple sparse features simultaneously. Therefore, we calculate one optimal feature and deflate samples to extract the others greedily.

From Eq. (31), we get the objective function for seeking one optimal vector as

$$\begin{aligned}G(r(t))&=\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|r(t)^T(x_k^{(i)}-x_k^{(j)})\right| \\ &\quad-\frac{\eta}{2}\|r(t)\|_2^2-\lambda\|r(t)\|_1 \\ &=r(t)^T\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(x_k^{(i)}-x_k^{(j)}) \\ &\quad-\frac{\eta}{2}\|r(t)\|_2^2-\lambda\|r(t)\|_1 \\ &\text{subject to } r(t)^T r(t)=1,\end{aligned} \tag{34}$$

where p_{ijk} is polarity function as follows

$$p_{ijk}(t)=\begin{cases}1, & r(t)^T(x_k^{(i)}-x_k^{(j)})\geq 0\\ -1, & r(t)^T(x_k^{(i)}-x_k^{(j)})<0\end{cases}. \tag{35}$$

We try to construct appropriate iterative formulation of $r(t+1)$ for non-decreasing, i.e., $G(r(t+1)) \geq G(r(t))$. Suppose $r(t)$ is sparse, it contains some zero elements. Removed zero elements from $r(t)$ and referred to as $\bar{r}(t)$. Accordingly, took out the elements from $x_k^{(i)}$ at the same indices and denoted as $\bar{x}_k^{(i)}$. For instance, if $r(t) = (8, 0, 5, 0, 4)^T$ and $x_k^{(i)} = (51, 52, 53, 54, 55)^T$, then $\bar{r}(t) = (8, 5, 4)^T$ and $\bar{x}_k^{(i)} = (51, 53, 55)^T$.

Now, we remove the zero from sparse feature and rewrite Eq. (33) as

$$\begin{aligned} G(r(t)) &= \bar{r}(t)^T \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(\bar{x}_k^{(i)} - \bar{x}_k^{(j)}) \\ &\quad -\frac{\eta}{2}\|\bar{r}(t)\|_2^2 - \lambda\|\bar{r}(t)\|_1. \end{aligned} \tag{36}$$

Eq. (36) above can be turned into

$$\begin{aligned} G(r(t)) &= \bar{r}^T \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(\bar{x}_k^{(i)} - x_k^{(j)}) \\ &\quad -\frac{\eta}{2}\|\bar{r}(t)\|_2^2 \\ &\quad -\frac{\lambda}{2}(\bar{r}(t)^T \bar{D}(t)\bar{r}(t) + \|\bar{r}(t)\|_1. \end{aligned} \tag{37}$$

where $\bar{D}(t) = \mathrm{diag}(|\bar{r}_1(t)|^{-1}, \cdots, |\bar{r}_{\bar{h}}(t)|^{-1})$.

Following the classical approach, we construct an surrogate function of Eq. (37) [20] as

$$\begin{aligned} S(w|\bar{r}(t)) &= r^T \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(\bar{x}_k^{(i)} - \bar{x}_k^{(j)}) \\ &\quad -\frac{\eta}{2}\|r\|_2^2 \\ &\quad -\frac{\lambda}{2}(r^T \bar{D}(t) r + \|\bar{r}(t)\|_1, \end{aligned} \tag{38}$$

where $S(r \mid \bar{r}(t))$ is a function of r, and $\bar{r}(t)$ is fixed.

$S(r \mid \bar{r}(t))$ reaches a local maximum only if $\frac{\partial S(r \mid \bar{r}(t))}{\partial r} = 0$, that is

$$\begin{aligned} &\frac{\partial S(r \mid \bar{r}(t))}{\partial r} \\ &= \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(\bar{x}_k^{(i)} - \bar{x}_k^{(j)}) - \eta r - \lambda \bar{D}(t) r = 0. \end{aligned} \tag{39}$$

which means that

$$\begin{aligned} \bar{r}^* &= \arg\max S\left(r \mid \bar{r}(t)\right) \\ &= \left(\eta I + \lambda \bar{D}(t)\right)^{-1} \left(\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(\bar{x}_k^{(i)} - \bar{x}_k^{(j)}) \right). \end{aligned} \tag{40}$$

Let $\bar{r}(t+1) = \bar{r}^*$, therefore,

$$\begin{aligned} &S(\bar{r}(t+1) \mid \bar{r}(t)) \\ &\geq S(\bar{r}(t) \mid \bar{r}(t)) = G(r(t)). \end{aligned} \tag{41}$$

Return the zeros to $\bar{r}(t+1), \bar{x}_k^{(i)}, \bar{x}_k^{(j)}$ and get

$$\begin{aligned} &r(t+1) \\ &= \left(\eta I + \lambda D(t)\right)^{-1} \left(\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(x_k^{(i)} - x_k^{(j)}) \right). \end{aligned} \tag{42}$$

Evidently, $(\eta I + \lambda D(t))^{-1}$ is a diagonal matrix, therefor, rewrite Eq. (42) as

$$r(t+1) = a(t) \circ b(t), \tag{43}$$

where "$\circ$" indicates the element-wise product and

$$a(t) = \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(x_k^{(i)} - x_k^{(j)}). \tag{44}$$

$$b(t) = \begin{bmatrix} \frac{|r_1(t)|}{\lambda + \eta|r_1(t)|} \\ \frac{|r_2(t)|}{\lambda + \eta|r_2(t)|} \\ \cdots \\ \frac{|r_h(t)|}{\lambda + \eta|r_h(t)|} \end{bmatrix}. \tag{45}$$

In Eq. (45), $r_p(t)$ is the pth entry of $r(t)$.

We formally present an iterative algorithm for BPCAL1S-AOM as follows.

Algorithm BPCAL1S-AOM for $d(d > 0)$ features extraction

1. Let $\tau = 1$. τ is a counter for number of features
2. Find the optimal projection vector r_τ
   1) Let $t = 0$, t is a counter for iterations
   2) Initialization:
      Generate $r(t)$ randomly subject to $r(t)^T r(t) = 1$
   3) Let $p_{ijk}(t) = \begin{cases} 1, & r(t)^T(x_k^{(i)} - x_k^{(j)}) \geq 0 \\ -1, & r(t)^T(x_k^{(i)} - x_k^{(j)}) < 0 \end{cases}$
   4) Let $a(t) = \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(x_k^{(i)} - x_k^{(j)})$

5) Let

$$b(t)=\begin{bmatrix}\dfrac{|r_1(t)|}{\lambda+\eta|r_1(t)|}\\ \dfrac{|r_2(t)|}{\lambda+\eta|r_2(t)|}\\ \cdots\\ \dfrac{|r_h(t)|}{\lambda+\eta|r_h(t)|}\end{bmatrix}$$

6) Let $r(t+1)=a(t)\circ b(t)$

7) Normalization: $r(t+1)=r(t+1)/\|r(t+1)\|_2$

8) Convergence test:

if $\|r(t+1)-r(t)\|_2>\varepsilon$ then go to step 3)

else $r_\tau \leftarrow r(t+1)$, exit the iteration and go to 3.

3. get deflated samples for greedy strategy:

For all $i \in \{1, 2, ..., n\}$ do

$$X_i^v(\tau+1)=X_i^v(\tau)-r_\tau(r_\tau^T X_i^v(\tau))$$

4. If $\tau<d$, let $\tau=\tau+1$ and go to step 2

We can futher validate the convergence of BPCAL1S-AOM.

Theorem 4. In the procedure of BPCAL1S-AOM, the objective function holds non-decreasing:

$$G(r(t+1))\geq G(r(t)). \tag{46}$$

Proof.

Search optimal $\overline{r}^*$ in surrogate function guarantee that

$$\begin{aligned}&S(\overline{r}(t+1)\,|\,\overline{r}(t))\\ &\geq S(\overline{r}(t)\,|\,\overline{r}(t))=G(r(t)).\end{aligned} \tag{47}$$

Eq. (46) will be true if $G(r(t+1))\geq S(\overline{r}(t+1)\,|\,\overline{r}(t))$.

Note that in Eq. (41),

$$\begin{aligned}&S(\overline{r}(t+1)\,|\,\overline{r}(t))\\ &=\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\overline{r}(t+1)^T p_{ijk}(t)(\overline{x}_k^{(i)}-\overline{x}_k^{(j)})\\ &\quad-\frac{\eta}{2}\left\|\overline{r}(t+1)\right\|_2^2\\ &\quad-\frac{\lambda}{2}\left(\overline{r}(t+1)^T\overline{D}(t)\overline{r}(t+1)+\left\|\overline{r}(t)\right\|_1\right).\end{aligned} \tag{48}$$

So, we rewrite

$$\begin{aligned}&G(r(t+1))\\ &=\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}r(t+1)^T p_{ijk}(t+1)(x_k^{(i)}-x_k^{(j)})\\ &\quad-\frac{\eta}{2}\left\|r(t+1)\right\|_2^2-\lambda\left\|r(t+1)\right\|_1.\end{aligned} \tag{49}$$

In Eq. (48) and Eq. (49), the 1st term holds the inequality:

$$\begin{aligned}&\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t+1)r(t+1)^T(x_k^{(i)}-x_k^{(j)})\\ &\geq\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ijk}(t)r(t+1)^T(x_k^{(i)}-x_k^{(j)}).\end{aligned} \tag{50}$$

The 2nd term holds equality

$$-\frac{\eta}{2}\left\|r(t+1)\right\|_2^2=-\frac{\eta}{2}\left\|\overline{r}(t+1)\right\|_2^2. \tag{51}$$

In 3rd term of $S(\overline{r}(t+1)\,|\,\overline{r}(t))$, there is

$$\begin{aligned}&\overline{r}(t+1)^T\overline{D}(t)\overline{r}(t+1)+\left\|\overline{r}(t)\right\|_1\\ &=\sum_q\frac{\overline{r}_q(t+1)^2}{\left|\overline{r}_q(t)\right|}+\left\|\overline{r}(t)\right\|_1\\ &\geq\sum_q\frac{\overline{r}_q(t+1)^2}{\left|\overline{r}_q(t+1)\right|}+\left\|\overline{r}(t+1)\right\|_1\\ &=2\left\|\overline{r}(t+1)\right\|_1=2\left\|r(t+1)\right\|_1.\end{aligned} \tag{52}$$

The inequality in (52) guaranteed by Lemma 2 [20, 23].

Lemma 2. Any vector r holds the variational equality as:

$$2\left\|r\right\|_1=\min_{\varsigma\in\mathbb{R}_+^h}\left(\sum_q\frac{r_q}{\varsigma_q}+\left\|\varsigma\right\|_1\right), \tag{53}$$

and it reaches uniquely the minimum value while $\varsigma_q=\left|r_q\right|$ for q = 1, 2, ..., h. Follow (52), the 3rd term of $G(r(t+1))$ holds the inequality:

$$\begin{aligned}&-\lambda\left\|r(t+1)\right\|_1\\ &\geq-\frac{\lambda}{2}\left(\overline{r}(t+1)^T\overline{D}(t)\overline{r}(t+1)+\left\|\overline{r}(t)\right\|_1\right).\end{aligned} \tag{54}$$

Combining (50), (51) and (54), we get

$$G(r(t+1))\geq S(\overline{r}(t+1)\,|\,\overline{r}(t)). \tag{55}$$

The proof is completed.

Obviously, the number of samples is limited, the iterative process of BPCAL1S-AOM is convergent.

5 Experiments

For evaluating the performance of BPCAL1-AOM and BPCAL1S-AOM, we designed experimental schemes of classification and reconstruction upon three benchmark image sets: ORL, Yale and Feret. In BPCAL1S-AOM, there are seemed two joint tunable parameters of sparsity, η and λ. However, $r(t+1)$ in Eq. (42) will be normalized, that means,

Eq. (42) is equivalent to

$$r(t+1) = \left(I + \frac{\eta}{\lambda} D(t)\right)^{-1} \left(\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ijk}(t)(x_k^{(i)} - x_k^{(j)})\right). \quad (56)$$

So the sparse feature extraction only depends on the radio of η to λ. In experiment involve sparsity, we define $\rho = \log_{10} \eta/\lambda$ and try to find the optimal value of ρ.

5.1 Classification

We applied BPCAL1-AOM and BPCAL1S-AOM to ORL and Yale in images classification and investigated the dependence of BPCAL1S-AOM on ρ, then compared them with 2DPCA, 2DPCA-L1, BPCA and BPCAL1-AOM.

Left of Figure 2 illustrate the great effect of ρ. In $\rho \in [-2, -1, 0, 1, 2]$, we find that $\rho - 1$ is a quite good value upon ORL. Left of Figure 3 shows a similar situation upon Yale but a better optional value of ρ is -2.

Figure 2 and Figure 3 depict the change of classification accuracies as the number of features increases. In general, the classification accuracies increase along with the number of features and reaches the maximum while the number is 5~10, then slightly declines and holds roughly the same. It means that the first serval features are adequate for recognition and surplus may deteriorate the performance.

Comparing with other algorithms in classification experiments, we chose $\rho = 1$ upon ORL and $\rho = -2$ upon Yale. Right of Figure 2 and Figure 3 show that the performance of BPCAL1-AOM and BPCAL1S-AOM are better than other methods. Compared with BPCAL1-AOM, Sparse features extracted by BPCAL1S-AOM not only have better semantics, but also contribute to better performance.

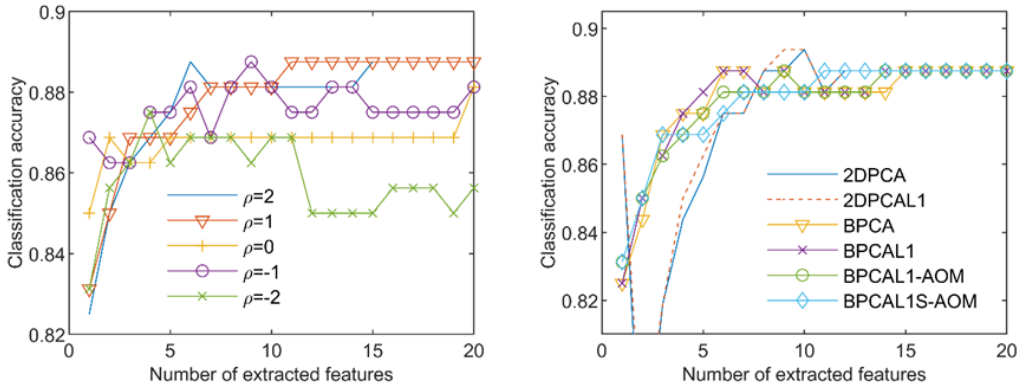

Figure 2. Classification accuracies of BPCAL1S-AOM upon ORL with various ρ and comparing with 2DPCA, 2DPCA-L1, BPCA, BPCAL1, BPCAL1-AOM

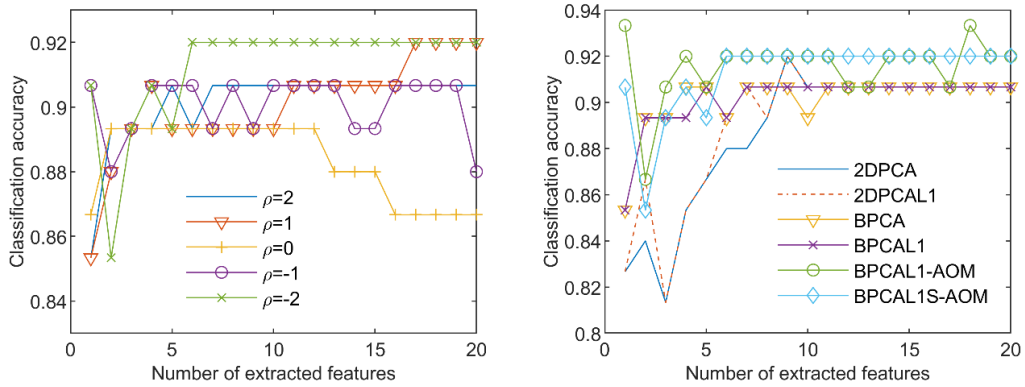

Figure 3. Classification accuracies of BPCAL1S-AOM upon Yale with various ρ and comparing with 2DPCA, 2DPCA-L1, BPCA, BPCAL1, BPCAL1-AOM

5.2 Reconstruction Error

In the experiments, we randomly selected 20 percent of all images and added random rectangle noises. Figure 4 and Figure 5 demontrates the average reconstruction error in experiments upon ORL and Yale database. The experimental results indicate that the value of sparse parameter ρ has a great impact on the algorithm performance. We can select appropriate ρ value in practice to obtain the sparse features with excellent performance. In the comparison experiment with other methods, we let $\rho = 3$ and observed that all reconstruction errors decrease with the increase of the feature number. The performance of BPCAL1-AOM and BPCAL1S-AOM are better than other algorithms.

5.3 Reconstruction of Image

This experiment is upon the Feret database. Figure 6 illustrates the original images with or without noise and the reconstructed ones using 5 features. Three origin images are at column 1. The following six columns depict the rebuilding versions using the first 5 features, which produced respectively by 2DPCA, 2DPCA-L1, BPCA, BPCA-L1, BPCAL1-AOM, BPCAL1S-AOM.

6 Conclusion

We propose two avoiding optimal mean Block PCA methods which are denoted as BPCAL1-AOM and BPCAL1S-AOM respectively. These two robust methods take full advantage of the partial association among neighboring pixels. The extracted sparse features are semantic and effective. These virtues come from block-based computation, sparsity constraints of elastic net and the utilization of L1-norm. On the other hand, these two L1-norm-based robust methods automatically avoid calculating the optimal mean without assuming zero average. The proposed approachs exert on several image application problems upon ORL, Yale and Feret. Experiments above demonstrate the efficacy of BPCAL1-AOM and BPCAL1S-AOM.

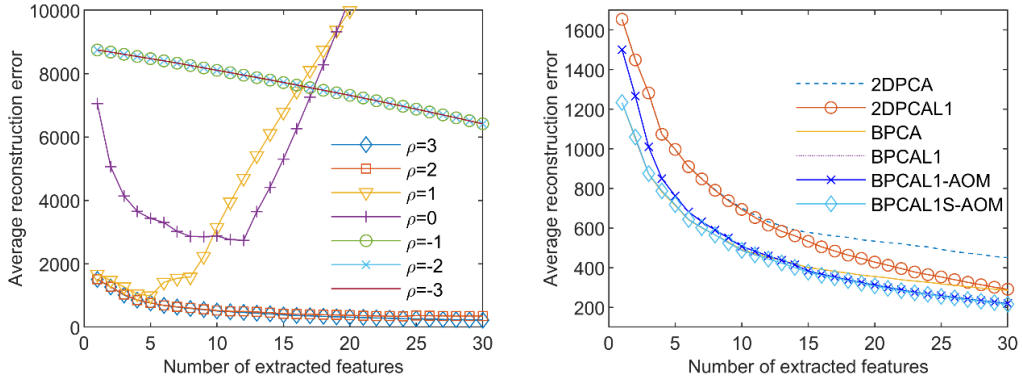

Figure 4. Average reconstruction errors of BPCAL1S-AOM upon ORL with various ρ and comparing with 2DPCA, 2DPCA-L1, BPCA, BPCAL1, BPCAL1-AOM

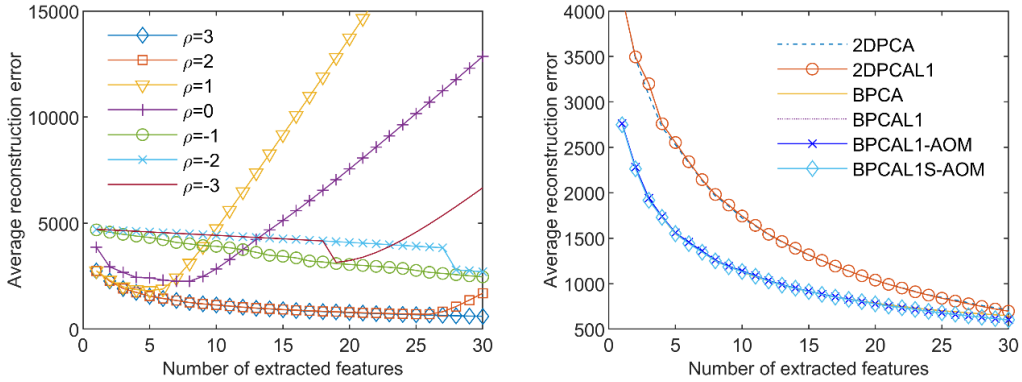

Figure 5. Average reconstruction errors of BPCAL1S-AOM upon Yale with various ρ and comparing with 2DPCA, 2DPCA-L1, BPCA, BPCAL1, BPCAL1-AOM

Figure 6. Image reconstructions: First column: original. Second to seven column: 2DPCA, 2DPCA-L1, BPCA, BPCA-L1, BPCAL1-AOM, BPCAL1S-AOM

Acknowledgment

This research is supported by the National Natural Science Foundation of China (Grant No. 61976005), the Key Project of Natural Science Research of Higher Education Institution of Anhui Province of China (Grant No. KJ2020A0363, KJ2020A0361), the Anhui Natural Science Foundation (Grant No. 1908085MF183) , the Key State Laboratory for Novel Software Technology (Nanjing University) Research Program (Grant No. KFKT2019B23), the Safety-Critical Software Key Laboratory Research Program (Grant No. NJ2018014) and the Training Program for Young and Middle-aged Top Talents of Anhui Polytechnic University (Grant No. 201812).

References

[1] I. Jolliffe, *Principal Component Analysis*, Springer, 2004.

[2] J. Yang, D. Zhang, A. F. Frangi, J.-Y. Yang, Two-dimensional PCA: A New Approach to Appearance-based Face Representation and Recognition, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 1, pp. 131-137, January, 2004.

[3] Q. Ke, T. Kanade, Robust L1 Norm Factorization in the Presence of Outliers and Missing Data by Alternative Convex Programming, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, USA, 2005, pp. 739-746.

[4] C. Ding, D. Zhou, X. He, H. Zha, R1-PCA: Rotational Invariant L1-norm Principal Component Analysis for Robust Subspace Factorization, *the 23rd International Conference on Machine Learning (ICML'06)*, Pittsburgh, USA, 2006, pp. 281-288.

[5] N. Kwak, Principal Component Analysis Based on L1-Norm Maximization, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 9, pp. 1672-1680, September, 2008.

[6] X. Li, Y. Pang, Y. Yuan, L1-Norm-Based 2DPCA, *IEEE Transactions on Systems Man and Cybernetics*, Vol. 40, No. 4, pp. 1170-1175, August, 2010.

[7] F. Nie, H. Huang, C. Ding, D. Luo, H. Wang, Robust Principal Component Analysis with Non-Greedy L1-Norm Maximization, *The Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11)*, Barcelona, Spain, 2011, pp. 1433-1438

[8] R. Wang, F. Nie, X. Yang, F. Gao, M. Yao, Robust 2DPCA With Non-greedy L1-Norm Maximization for Image Analysis, *IEEE Transactions on Cybernetics*, Vol. 45, No. 5, pp. 1108-1112, May, 2015.

[9] N. Kwak, Principal Component Analysis by Lp-Norm Maximization, *IEEE Transactions on Cybernetics*, Vol. 44, No. 5, pp. 594-609, May, 2014.

[10] Q.-X. Gao, Is Two-dimensional PCA Equivalent to A Special Case of Modular PCA? *Pattern Recognition Letters*, Vol. 28, No. 10, pp. 1250-1251, July, 2007.

[11] R. Gottumukkal, V. K. Asari, An Improved Face Recognition Technique Based on Modular PCA Approach, *Pattern Recognition Letters*, Vol. 25, No. 4, pp. 429-436, March, 2004.

[12] C. Kim, C.-H. Choi, Image Covariance-based Subspace Method for Face Recognition, *Pattern Recognition*, Vol. 40, No. 5, pp. 1592-1604, May, 2007.

[13] H. Wang, Block Principal Component Analysis with L1-norm for Image Analysis, *Pattern Recognition Letters*, Vol. 33, No. 5, pp. 537-542, April, 2012.

[14] B. N. Li, Q. Yu, R. Wang, K. Xiang, M. Wang, X. Li, Block Principal Component Analysis with Nongreedy L1-norm Maximization, *IEEE Transactions on Cybernetics*, Vol. 46, No. 11, pp. 2543-2547, November, 2016.

[15] F. Nie, J. Yuan, H. Huang, Optimal Mean Robust Principal Component Analysis, *the 31st International Conference on Machine Learning (ICML'14)*, Beijing, China, 2014, pp. 1062-1070.

[16] J. Oh, N. Kwak, Generalized Mean for Robust Principal Component Analysis, *Pattern Recognition*, Vol. 54, pp. 116-127, June, 2016.

[17] Q. Wang, Q. Gao, X. Gao, F. Nie, Optimal Mean Two-dimensional Principal Component Analysis with F-norm Minimization, *Pattern Recognition*, Vol. 68, pp. 286-294, August, 2017.

[18] M. Luo, F. Nie, X. Chang, Y. Yang, A. Hauptmann, Q. Zheng, Avoiding Optimal Mean Robust PCA/2DPCA with Non-greedy L1-norm Maximization, *Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*, New York, USA, 2016, pp. 1802-1808.

[19] M. Luo, F. Nie, X. Chang, A. Hauptmann, Q. Zheng, Avoiding Optimal Mean L2, 1-norm Maximization-based Robust PCA for Reconstruction, *Neural Computation*, Vol. 29, No. 4, pp. 1124-1150, April 2017.

[20] H. Wang, J. Wang, 2DPCA with L1-norm for Simultaneously Robust and Sparse Modelling, *Neural Networks*, Vol. 46, pp. 190-198, October, 2013.

[21] J. Wang, Generalized 2-D Principal Component Analysis by Lp-Norm for Image Analysis, *IEEE Transactions on Cybernetics*, Vol. 46, No. 3, pp. 792-803, March, 2016.

[22] H. Zou, T. Hastie, R. Tibshirani, Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, Vol. 15, No.2, pp. 265-286, June, 2006.

[23] R. Jenatton, G. Obozinski, F. Bach, Structured Sparse Principal Component Analysis, *the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Chia Laguna Resort, Italy, 2010, pp. 366-373.

Biographies

Ganyi Tang received the M. S. degree in 2006 from Southwest University, Chongqing, China. He is an associate professor in the School of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui, China. His research interests include machine learning, pattern recognition and computational intelligence.

Lili Fan received M.S. degree in 2007 from Xi'an Shiyou University, Shaanxi, China. Since 2007, she has been teaching in the School of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui, China. Her research interests include machine learning, pattern recognition.

Jianguo Shi received M.S. degree in Hefei University of Technology, Anhui. He is an associate professor in the School of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui, China. His research interests include pattern recognition and data mining.

Jingjing Tan is an undergraduate student in the School of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui, China. Her research interests include machine learning and intelligence algorithms.

Guifu Lu received the Ph.D. degree in 2012 from Nanjing University of Science and Technology, China. He is a professor in the School of Computer Science and Information, Anhui Polytechnic University, Wuhu, Anhui, China. His research interests include computer vision, digital image processing and pattern recognition.