Principal Component Analysis based on Nuclear norm Minimization

Jian-Xun Mi\textsuperscript{a,b,}\textsuperscript{*}, Ya-Nan Zhang\textsuperscript{a,b}, Zhihui Lai\textsuperscript{c,d}, Weisheng Li\textsuperscript{a,b}, Lifang Zhou\textsuperscript{a,b,e}, Fujin Zhong\textsuperscript{a}

\textsuperscript{a}Chongqing Key Laboratory of Image cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
\textsuperscript{b}College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
\textsuperscript{c}Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
\textsuperscript{d}Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China
\textsuperscript{e}College of Software, Chongqing Univ. of Posts and Telecommunications, Chongqing 400065, China

\textbf{ARTICLE INFO}

Article history:
Received 30 November 2018
Received in revised form 1 March 2019
Accepted 24 May 2019
Available online 8 June 2019

Keywords:
Principal component analysis (PCA)
Nuclear norm
Robustness
Optimal mean
Low-dimensional representation

\textbf{ABSTRACT}

Principal component analysis (PCA) is a widely used tool for dimensionality reduction and feature extraction in the field of computer vision. Traditional PCA is sensitive to outliers which are common in empirical applications. Therefore, in recent years, massive efforts have been made to improve the robustness of PCA. However, many emerging PCA variants developed in the direction have some weaknesses. First, few of them pay attention to the 2D structure of error matrix. Second, to estimate data mean from sample set with outliers by averaging is usually biased. Third, if some elements of a sample are disturbed, to extract principal components (PCs) by directly projecting data with transformation matrix causes incorrect mapping of sample to its genuine location in low-dimensional feature subspace. To alleviate these problems, we present a novel robust method, called nuclear norm-based on PCA (N-PCA) to take full advantage of the structure information of error image. Meanwhile, it is developed under a novel unified framework of PCA to remedy the bias of computing data mean and the low-dimensional representation of a sample both of which are treated as unknown variables in a single model together with projection matrix. To solve N-PCA, we propose an iterative algorithm, which has a closed-form solution in each iteration. Experimental results on several open databases demonstrate the effectiveness of the proposed method.

1. Introduction

In pattern recognition and machine learning community, feature extraction is one of the most important problems. Principal component analysis (PCA) (Jolliffe, 2011), linear discriminant analysis (LDA) (Belhumeur, Hespanha, & Kriegman, 1997) and Neural Networks (Huang, 1996, 1999, 2004; Huang & Du, 2008; Zhao & Huang, 2007) are three of the classical methods. PCA seeks to learn a transformation matrix to project high-dimensional data into a low-dimensional space to avoid dimensional disaster but at the same time minimizing information loss. It is also an unsupervised method, which is important in practical applications when labeled information is difficult to obtain. LDA is a supervised method. It is intended to find a projection matrix to obtain discriminative information by maximizing the between-class scatter matrix while minimizing the within-class scatter matrix. The conventional L2-norm PCA has been extensively used for many applications, such as pattern recognition and signal procession. In the paper, the application of PCA in face representation and recognition is particularly of our interest.

However, the performance of conventional PCA is limited by dealing with outliers since its covariance matrix is taken from L2-norm and the squared L2-norm exaggerates the effect of outliers. To tackle this problem, plenty of its extended researches have been presented (Jolliffe & Cadima, 2016). These approaches can be broadly divided into two categories in terms of the way to generate the robustness. The first category aims at coping with noise samples of database. Hence, some variants of PCA alleviate the effect of outliers by directly suppressing the noise image, such as rotational invariant L1-norm PCA (R1-PCA) (Ding, Zhou, He, & Zha, 2006), L1-norm PCA (L1-PCA) (Ke & Kanade, 2005) and PCA based on L1-norm maximization (PCA-L1) (Kwak, 2008). L1-PCA applies maximum likelihood estimation to formulate the PCA model, and a heuristic estimate is utilized to solve the L1-norm problem. However, L1-PCA is not invariant to rotation which is an important property in learning algorithm (Ng, 2004). Thus, R1-PCA have been proposed. In R1-PCA, the reconstruction error of each data point is measured in
L2-norm, while the summation of reconstruction error over all the data points uses L1-norm. It not only is rotational invariant but also successfully suppresses the effect of outlying samples by the data points uses L1-norm. It not only is rotational invariant L2-norm, while the summation of reconstruction error overall subspace of a b-dimensional feature space when a < b, which, however, is not true for R1-PCA. Secondly, R1-PCA takes plenty of time to achieve convergence when the input space dimension is very large. Inspired by R1-PCA, many robust methods based on R1-norm have been proposed in Lai, Xu, Yang, Shen, and Zhang (2017), Nie, Huang, Cai, and Ding (2010), Wang, Gao, Gao, and Nie (2016), Wang, Gao, Gao, and Nie (2017a) and Yi, Lai, He, Cheung, and Liu (2017). PCA-L1 employs L1-norm to measure the variance of given data. PCA-L1 (Kwak, 2008), however, this work does not proposed a fine optimize solution of the objective function. Nie, Huang, Ding, Luo, and Wang (2011) proposed a non-greeny iterative algorithm to solve the PCA-L1. Stimulated by PCA-L1, Lu, Zou, Wang, and Wang (2016) proposed a L1-norm-based PCA with adaptive regularization (PCA-L1/AR) which is adaptive to the correlation structure. Tsagkarakis, Markopoulos, Sklavounis, and Pados (2018) presented a L1-norm PCA of complex data to tackle complex-valued data matrices. He, Hu, Zheng, and Kong (2011) presented a rotational-invariant PCA based on maximum correntropy criterion (HQ-PCA), Pang, Li, and Yuan (2010) proposed a L1-norm based on tensor analysis (TPCA-L1). Another category of robust PCA tries to suppress the influence of some contaminated parts in samples, and takes advantage of images as much as possible even if they are partially contaminated. In contrast to the methods belong to the first category, many extended methods of PCA aim at simultaneously weaken the effect of the contaminated parts and improve the role of uncontaminated rests to alleviate the effect of the outliers, such as Modular PCA (Gotumakul & Asari, 2004), Euler PCA (EPICA) (Liwicki, Tzimiropoulos, Zafeiriou, & Pantic, 2013) and weighted PCA (WPCA) (Koren & Carmel, 2004). The Modular PCA divides each image into smaller sub-images, then applies PCA approach to extract features for these sub-images. WPCA uses weight Euclidean distance to improve the robust of the algorithm. EPICA employs a robust dissimilarity measure based on Euler representation to suppress outliers. Both methods ignore the holistic the structure of images.

Applying the above-mentioned methods into image analysis, they all need to convert a two-dimensional image matrix into one-dimensional vector by concatenating all rows or columns of an image prior to feature extraction. Thus, the aforementioned dimensionality reduction methods cannot exploit the spatial structural information well which is embedded in the arrangement of pixels of image and essential for image representation and classification. To tackle this problem, various matrix-based dimensionality reduction methods have been proposed for image representation and classification. For example, Yang, Zhang, Frangi, and Yang (2004) proposed two-dimensional PCA (2-DPCA) which directly calculated covariance matrix from image matrices. 2-DPCA has been widely applied for image classification and representation. But it has a disadvantage that 2-DPCA operates on the row of the matrix and ignores the structure information embedded in the columns of the images. Therefore, Kong et al. (2005) proposed bilateral projection-based 2-DPCA (B2-DPCA) to combine rows and columns information of the matrix. B2-DPCA finds two projection matrices to obtain the row information and column information respectively.

Current matrix form based PCAs such as mentioned 2-DPCA and B2-DPCA are all measured by Euclidean distance. It is commonly known that squared L2-norm is not robust to outliers. Thus, these methods are obvious degradation when there are outliers. To solve this problem, many robust methods, which are based on 2-D image matrix, have been proposed to extract features. For example, Li, Pang, and Yuan (2010) generalize PCA-L1 to L1-norm-based 2-DPCA (2-DPCA-L1), Wang and Wang (2013) proposed 2-DPCA-L1 with sparsity (2DPCAL1-S) which imposed sparse constraint in 2-DPCA-L1. Lai, Xu, Chen, Yang, and Zhang (2014) proposed multilinear sparse PCA (MPCA). Wang and Gao (2017) proposed F-2DPCA which uses F-norm instead of squared F-norm as distance metric.

Although the aforementioned methods are all based on 2-D matrices, 2-DPCA and 2-DPCA-L1 can be equivalent to PCA and PCA-L1 due to the fact that L2-norm and L1-norm are essentially one-dimension vector norm (Gao, 2007; Kong et al., 2005; Mashhooni & Jahromi, 2013). To further enrich the family of PCAs, nuclear norm has been put forward as the distance metric. It can effectively exploit the spatial structure information to improve the robustness since it is a two-order matrix norm. For instance, Gu, Shao, Li, and Fu (2012) proposed Schatten 1-norm PCA which uses nuclear norm as the distance metric to calculate the variance. However, Schatten 1-norm PCA needs an additional constraint on the desired projection matrix, which requires the projection matrix is full rank. In more general cases, the projection matrix is a column full rank matrix for dimensionality reduction problem. So Schatten 1-norm PCA method uses an approximate algorithm. To get rid of this constraint, Zhang, Yang, Qian, and Xu (2015) extended 2DPCA to nuclear norm-based 2-DPCA (N2-DPCA) which utilizes nuclear norm to measure the reconstruction error. But there is a disadvantage of 2D projection based PCA which is each pixel on an image related closely to its near neighbors not all pixels along its row or column. Therefore, the category of 2D based PCAs can be further improved.

In the paper, we propose a novel PCA based on nuclear norm which can help improve the robustness by coping the contaminated features in image samples. Usually, there can be different occlusions on images such as sunglasses on face or other noises of low-rank (Yang et al., 2017). For this case, L2-norm based methods are not robust and L1-norm methods show some robustness but ignore spatial structure of images. Recently, studies on structure have drawn many interests from researchers (Gui, Sun, Ji, Tao, & Tan, 2017; Liu et al., 2013; Zhao, Shkolnisky, & Singer, 2016). Usually, the low-rank function is difficult to solve since it is a non-convex function (Fornasier, Raufut, & Ward, 2011). Fornasier et al. (2011) use nuclear norm minimization instead of minimizing rank due to nuclear norm is the convex envelope of matrix rank (He, Sun, Tan, and Zheng, 2011). Thus, we use nuclear norm as a convex surrogate of the rank in models. To solve N-PCA, we propose an iterative algorithm, which has a closed-form solution in each iteration. Moreover, we use Singular Value Thresholding (SVT) (Cai, Candès, & Shen, 2010) and Alternating Direction Method of Multipliers (ADMM) (Boyd, Parikh, Chu, Peleato, Eckstein, et al, 2011) algorithm in each iteration. In summary, our method has several merits over many previous PCAs, which are given as follows:

1. N-PCA provides a robust approach to conduct mean subtraction. Traditional way of subtracting mean from data in PCA is inaccurate when outliers exist in data since the calculation of the data mean by its center of is biased. N-PCA treats the mean as an optimization variable in its model and the estimation of the data mean exhibits robustness against noises.

2. We propose an approach to estimate the genuine location of an image in the low dimensional feature subspace. Unlike, conventional PCAs, low dimensional representation of a sample is computed by directly projection, and in this way outlier feature can significantly affect the result.

3. N-PCA is a 2D based PCA in terms of the way to measure the reconstruction error by nuclear norm, which exploits the spatial structure of images to compute robust projection vectors against existing outliers on the images.
The remainder of the paper is organized as follows. Section 2 reviews the related works including L2-PCA, R1-PCA and PCA-L1. In Section 3, we introduce the model (N-PCA) and algorithm along with many analyses on them. In Section 4, we conduct with experiments on five public face databases and compare experimental results with state-of-the-art methods. The conclusion is drawn in Section 5.

2. Related works

First of all, a general scenario of PCA is defined. Given m sample images $X = [x_1, x_2, \ldots, x_m]$ in $R^{n \times m}$, where n denotes the original image space dimensionality. Without loss of generality, we assumed that the samples have zero mean, which is calculated by the L2-norm, i.e., $\sum_{i=1}^{m} (x_i - 1/m \sum_{j=1}^{m} x_j) = 0$. PCA aims to find a projection matrix $W \in R^{d \times r}$, where $r \ll m$. And $W$ is normalized to prevent the variance being infinity, i.e., $W^T W = I$, where $I$ is an identity matrix.

2.1. Different derivations for L2-norm PCA

PCA has two different formulations. One is the minimization of the reconstruction error which means each sample should lose least information, measured by L2-norm, when we compare it with its counterpart from a low dimension projection subspace. Such motivation is formulized by following objective function:

$$\min_{W^T W = I} \sum_{i=1}^{m} \|x_i - W W^T x_i\|_2^2$$  \hspace{1cm} (1)

where $\|\cdot\|_2$ denotes the squared L2-norm.

Another formulation is defined by employing L2-norm to measure the variance. To maximize the variance, the projection points of all samples in the low dimensional projection space should be separated as far as possible. The objective function can be expressed as:

$$\max_{W^T W = I} \sum_{i=1}^{m} \|W^T x_i\|_2^2$$  \hspace{1cm} (2)

Then, through some simple algebraic operations, we can obtain

$$\min_{W^T W = I} \sum_{i=1}^{m} \|x_i - W W^T x_i\|_2^2$$

$$= \min_{W^T W = I} \sum_{i=1}^{m} \text{tr}(x_i^2 - 2x_i^T W W^T x_i + W W^T W W^T x_i)$$

$$= \max_{W^T W = I} \sum_{i=1}^{m} \text{tr}(W^T x_i x_i^T W)$$

$$= \max_{W^T W = I} \sum_{i=1}^{m} \|W^T x_i\|_2^2$$

where $\text{tr}(\cdot)$ is the trace operator of a matrix. It is easy to see that Eq. (1) is equivalent to Eq. (2). That is, Eqs. (1) and (2) exact same optimal solution. The solution of Eq. (1) or (2) is composed of the eigenvectors of the covariance matrix $S_r = \sum_{i=1}^{m} x_i x_i^T$ corresponding to the first $r$ largest eigenvalues. But the L2-norm PCA has a flaw that it is susceptible to outliers since outlier skew the solution from the desired solution (Kwak, 2008).

2.2. R1-PCA

To handle this problem, many approaches have been proposed, one representative work is rotational invariant L1-norm PCA (R1-PCA) (Ding et al., 2006). In R1-norm, distance in spatial dimensional is measured in L2-norm, while the summation over different data points in the form of L1-norm. R1-norm is defined as:

$$\|X\|_{R1} = \max_{m} \sum_{i=1}^{n} \|x_i - W W^T x_i\|_1$$

R1-PCA combines the merits of L2-PCA and L1-PCA. It not only is rotational invariant, but also suppresses the effect of outliers. The projection matrix is solved by following objective function

$$\min_{W^T W = I} \sum_{i=1}^{m} \|x_i - W W^T x_i\|_{R1}$$ \hspace{1cm} (3)

However, there are several drawbacks on R1-PCA: (1) It is dependent on the dimension of a subspace to be found (i.e., the optimal solution $W$ when $r = a$ may not be in a subspace of $W$ when $r = b$ ($a < b$)); (2) It is computationally expensive for high-dimensional data. Thus, it is only limited to small-scale data.

2.3. PCA-L1

It has been well known that L1-norm is more robust to outliers due to the fact that L1-norm can suppress the outliers in the database (Kwak, 2008). Therefore, many L1-norm based PCA approaches have been proposed, the most representative one is:

$$\max_{W^T W = I} \sum_{i=1}^{m} \|W^T x_i\|_1$$ \hspace{1cm} (4)

where $\|\cdot\|_1$ denotes L1-norm of a vector. This approach calculates the variance by using L1-norm as the distance metric. Another PCA using L1-norm is presented in Ke and Kanade (2005), which employs L1-norm to measure the reconstruction error and the objective function is expressed as

$$\min_{W^T W = I} \sum_{i=1}^{m} \|x_i - W W^T x_i\|_1$$ \hspace{1cm} (5)

It should point out that the objective function (4) is not equivalent to the objective function (5) which is a crucial goal of PCA while regarding it as dimensionality reduction approach, due to the fact $\|x_i - W W^T x_i\|_1 + \|W W^T x_i\|_1 \neq \|x_i\|_1$. Compared to PCA-L1, L1-PCA is computationally more expensive. Moreover, the objective function (5) is not invariant to rotations and invariant to rotation is an important property in learning algorithm. Although the objective function (4) and (5) are different since they are not dual to each other, the two solutions are equally important in image analysis (Gao, Ma, Liu, Gao, & Nie, 2018; Wang et al., 2017a).

3. Nuclear norm-based PCA

In this section, we present the novel principal component subspace learning method, including the motivation, its objective function, the according optimization approach and algorithm, along with computational complexity and many analyses.

3.1. Motivation and objective function

In the community of machine learning, many nuclear norm-based methods have been proposed recently (Gu et al., 2012; Yang et al., 2017; Zhang et al., 2015), which have the following merits.
(1) In the presence of outliers, the distance metric affects the robustness of algorithms. From Wang and Gao (2017), we know that the square L2-norm (Euclidean distance) exaggerates the effect of outliers while the nuclear norm emphasizes the importance of small variations in distance metric. For example, it is known that the variations among face images, which are sampled from the same person under illumination changes and block occlusion, are always larger than image variations from different people. For example, we selected four images from the Extended Yale B database. Fig. 1(a), (b) and (c) are from same person, where Fig. 1(a) is a template image, Fig. 1(b) and (c) are lighting and block occlusion images, respectively. Fig. 1(d) belongs to another person. It is easy to see in Table 1, the distance $d(a, d)$ between images $a$ and $d$ is smaller than $d(a, b)$ and $d(a, c)$ when the Euclidean distance is employed as the distance measure. It is the classification error. However, nuclear norm can exactly classify the images since the $d(a, d)$ is larger than $d(a, b)$ and $d(a, c)$. Thus, nuclear norm is more robust than L2-norm.

(2) The structured noise (e.g. block occlusion) causes that the error image is of low rank (Yang et al., 2017). For example, Fig. 2 shows three images from the Extended Yale B database. Fig. 2(a) is a template image. Fig. 2(b) is the case that Fig. 2(a) is added with a block occlusion. The error image, which represents the difference between Fig. 2 (a) and (b), is shown in Fig. 2(c). It is clear that the rank of the error image is low from Table 2. However, the low-rank function is difficult to optimize since it is a non-convex function. We use nuclear norm as a convex surrogate of the rank in models. Thus, it is more reliable than the L1 or L2 norm that unclear norm is used as distance metric (Yang et al., 2017).

(3) Nuclear norm makes full use of the structure information of the error image. L2-norm and L1-norm both are vector-based element-wise norm. They are always assumed that pixels are independent. That means, the results of the L2-norm and L1-norm of a vector are not changed when we randomly exchange the location of the elements of the vector. Therefore, they cannot well exploit the correlation between pixels in pixels of an image which is important for image representation and recognition. However, the nuclear norm of a matrix is the sum of singular values of the matrix. The operation is intended for coping with the information on the entire error image matrix rather than handling single pixels independently. If the position of the elements is exchanged, the value of the nuclear norm of the matrix will change. Therefore, it can efficiently exploit the spatial structure information.

(4) L2-norm provides an optimal characterization for errors with the Gaussian distribution (Jolliffe, 2011), and the L1-norm provides an optimal characterization with the Laplace distribution (Gao, 2008). They make a hypothesis that the error image satisfies a restricted distribution. Fig. 3(a) is the image fitted by different distributions. Fig. 3(b) shows that singular values of the error matrix fitted by different distributions. From Fig. 3(a), we know that the error image does not comply with Gaussian and Laplace distributions. Thus L2-norm and L1-norm cannot effectively describe the reconstruction error. Nuclear norm is the sum of all singular values of the matrix. It is equivalent to L1-norm of the singular value vector. Fig. 3(b) demonstrates that singular values of error matrix fit Laplace distribution. That means, nuclear norm is more suitable for the structural noise than L2-norm and L1-norm.

Thus, we characterize the reconstruction error matrix using the nuclear norm. The objective function of N-PCA is defined as:

$$\min_{W^{T}W=I} \sum_{i=1}^{m} \|g^{-1}(WW^{T}x_{i} - x_{i})\|_{*}.$$  \hspace{1cm} (6)

where $\| \cdot \|_{*}$ denotes nuclear norm, $g^{-1}(\cdot)$ denotes that a transformed vector is converted back into its original matrix form of image.

To correctly conduct PCA, Eq. (6) should assume that the samples have zero mean. Usually sample mean is estimated by the average or the empirical mean and samples can be easily centered by subtracting mean. (Notice that for traditional PCA, if we add data mean as a variable into Eq. (1), the result is average over all the samples.) However, in the other hand the objective function (6) neglects the optimal mean calculation problem and in the other hand the estimation of data mean by averaging samples leads to biased results when there are outliers (Luo et al., 2017; Nie, Yuan, & Huang, 2014; Wang, Gao, Gao, & Nie, 2017b). If the data mean is not accuracy, it is impossible to extract correct principal

---

**Table 1**

<table>
<thead>
<tr>
<th>Distance metric</th>
<th>$d(a, b)$</th>
<th>$d(a, c)$</th>
<th>$d(a, d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>0.271</td>
<td>0.298</td>
<td>0.1936</td>
</tr>
<tr>
<td>Nuclear norm</td>
<td>0.722</td>
<td>0.444</td>
<td>1.174</td>
</tr>
</tbody>
</table>

Comparison of distance between images using different norms.

---

**Table 2**

<table>
<thead>
<tr>
<th>Rank</th>
<th>2(a)</th>
<th>2(b)</th>
<th>2(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64</td>
<td>64</td>
<td>20</td>
</tr>
</tbody>
</table>

Comparison of rank from different images.

---

**Fig. 1.** Four face images having different class labels in the Extended Yale B database.

**Fig. 2.** Example images. (a) Template image. (b) Template image is added with a black block occlusion. (c) Error image.

**Fig. 3.** Comparison of distribution between images using different norms. (a) Gaussian distribution. (b) Laplace distribution. (c) Nuclear norm.

---

### References

components with proper projection vectors. To tackle this issue, we add the data mean as a variable into (5). By jointly optimizing mean variable and projection vectors, a lower minimum is expected, which means the model is likely to produce a better result despite of noises. By the way, according to our subsequent derivation, the computed sample mean is not trivially equal to sample average. Therefore, the corrected objective function is rewritten as follows:

$$
\min_{W^{T}W=d, b} \sum_{i=1}^{m} \| g^{-1}(WW^{T}(x_{i} - b) - (x_{i} - b)) \|_{*} 
$$

where \( x_{i}, i = 1, \ldots, m \) is an unprocessed sample, and \( b \in \mathbb{R}^{n} \) is an unknown mean vector.

Remind that the main goal of the study is to help PCA to deal with contaminated images some of whose pixels are disturbed. When one uses PCA to extract principal components as features to proceed his own task, such as recognition and clustering, to compute \( y_{i}, \) i.e., the representation of a sample in the low-dimensional subspace, is crucial, which can be calculated by simple algebra for traditional and other PCAs, i.e., \( y_{i} = WW^{T}(x_{i} - b) \). But we should remind that some elements are contaminated in \( x_{i} \). Therefore, even if a correct projection matrix \( W \) is available, the genuine \( y_{i} \) cannot be produced by straightforwardly multiplying \( W \) with a noisy \( x_{i} \). Nevertheless, in consideration of reconstruction, a good \( x_{i} \) can be reconstructed by a true \( y_{i} \). Inspired by this, we directly add \( y_{i} \) as a new variable into the model which replaces \( WW^{T}(x_{i} - b) \) to estimate the genuine location of an image in the low dimensional feature subspace. Notice that we do not need to worry about the reversed action that to measure the difference between a noisy \( x_{i} \) with its clear reconstructed counterpart because of the virtue of nuclear norm which suppresses the structure noise on \( x_{i} \). Here, the further remedied objective function is given as follows:

$$
\min_{W^{T}W=d, b, y} \sum_{i=1}^{m} \| g^{-1}(Wy_{i} - (x_{i} - b)) \|_{*} 
$$

where \( y_{i}, i = 1, \ldots, m \) is the representation of \( x_{i} \) in the low dimensional subspace.

We illustrate the robustness of our model by an example in Fig. 4. Fig. 4(a) shows a template face image \( x_{1} \) with the size of 64 \( \times \) 64 from the Extended Yale B database. We test the image with black block occlusion. Fig. 4(b) shows an image \( x_{2} \) with the occlusion of a block size, 20 \( \times \) 10. Fig. 4(c) and (d) are two reconstructed images of \( x_{2} \) by different approaches, where Fig. 4(c) refers to the reconstruction point of low-dimensional coordinate obtained by traditional directly projection \( WW^{T}(x_{2} - b) \), and Fig. 4(d) is \( y \) the result of ours. From Fig. 4, it is obvious that Fig. 4(d) is much similar to template image \( x_{1} \) than Fig. 4(c). That means, \( y \) estimated by our model can automatically reduce the outlying features on samples. Therefore, as variable into PCA model, the result is more suitable to estimate the genuine location of an image in the low dimensional feature space.

### 3.2. The general PCA framework and its connected to traditional PCA

In the subsection, we first compare our model with traditional PCA. For most previous PCAs, data is pre-processed by subtracting empirical mean or average and the low dimensional projection of a sample is calculated by directly projecting using projection vector. In the proposed PCA, mean and low dimensional representation of data are modeled as new variables. Actually, this methodology of building PCA model can be seen as a general framework. Here, we demonstrate how this framework is applied to the traditional PCA. The objective function in the Euclidean space can be written as:

$$
\min_{W^{T}W=d, b, Y} \| X - b2 - WY \|_{F}^{2}
$$

where \( Y = [y_{1}, \ldots, y_{m}], X = [x_{1}, \ldots, x_{m}] \) and 1 = [11, \ldots, 1m] (1 is a raw vector and every element is equal to one). The Lagrangian function \( L \) of Eq. (9) is defined as

$$
L = \| X - b1 - WW^{T} \|_{F}^{2} + tr(Z^{T}(WW^{T} - I_{r}))
$$

First, \( Y \) can be easily obtained and has the closed solution as follows

$$
Y = W^{T}(X - b1)
$$

Then, substituting Eq. (11) into Eq. (10), the Lagrangian function can be rewritten as

$$
L = \| X - b1 - WW^{T}(X - b1) \|_{F}^{2} + tr(Z^{T}(WW^{T} - I_{r}))
$$

Taking the derivative of Eq. (12) with respective to \( b \) and setting it to zero. With some simple algebra calculation, we can obtain:

$$
0 = (X - WW^{T}(X - b1))1_{T} - 11_{T}b
$$

$$
= (I_{r} - WW^{T})X1_{T} - (I_{r} - WW^{T})b11_{T}
$$

$$
= (I_{r} - WW^{T})(X1_{T} - b11_{T})
$$
Let $M$ be the sample space. Suppose $W^\perp$ is the orthogonal complement of $W$, i.e., $M = W^\perp \otimes W$. For the vector $X1^T - b1^T$, we can obtain

$$X1^T - b1^T = W\alpha + W^\perp \beta$$

Adding (13) and (14), we have $W^\perp \beta = 0$. Then Eq. (14) becomes:

$$b = 1/m(X1^T + W\alpha)$$

where $\alpha$ is a $r$-dimension column vector. From Nie et al. (2014), we know that the result of substituting Eq. (15) into Eq. (12) is the same as the traditional PCA. That means, in the proposed framework, we can simply calculate the representation of samples in the low dimension subspace such that $Y = W^T (X - b1)$ and the mean of the data such that $b = 1/m \sum_{i=1}^{m} X_i$ when the square L2-norm is employed as the distance metric. Therefore, the objective function (9) is equivalent to the objective function (1) in the Euclidean space. It indicates the proposed framework is valid in the Euclidean space.

Although N-PCA is also an implementation under this general PCA framework, according to the results in the following part, by using the nuclear norm as the distance metric, N-PCA possesses its particular approach to compute mean and low dimensional representation. The detail of solving the $Y$ and $b$ is given in Section 3.3 and Appendix.

### 3.3. Optimization algorithm for N-PCA

The alternating direction method of multipliers (ADMM) and the Singular Value Thresholding (SVT) algorithms have been widely used to minimize the nuclear norm problems (Hansson, Liu, & Vandenberghe, 2012; Nie et al., 2014). In this section, we present the details of using an iterative algorithm to solve the objective function (8).

Let $g(e_i) =Wy_i - (x_i - b)$, the objective function (8) can be rewritten as

$$\begin{align*}
\min_{W, y_i, b, \alpha} & \sum_{i=1}^{m} \|e_i\|_f \\
\text{s.t.} & \quad g(e_i) = Wy_i - (x_i - b) \quad i = 1 \ldots m \\
& \quad W^T W = I
\end{align*}$$

where $g(\cdot)$ denotes that original matrix is converted into its vector form of an image. For convenience, instead of $W^T W = I$, $W \leftarrow (WW^T)^{-\frac{1}{2}} W$ is established to guarantee the orthogonality constraint of projection matrix (Hyvärinen, Karhunen, & Oja, 2004). Eq. (16) can be defined as an unconstrained optimization problem by using the augmented Lagrange method. The augmented Lagrangian function $L_\alpha$ is defined by

$$L_\alpha = \sum_{i=1}^{m} (\|e_i\|_f + z_i^T (Wy_i - (x_i - b) - g(e_i)))$$

where $u > 0$ is a penalty parameter, $z_i$ is the Lagrange multiplier, and Eq. (17) is the standard Lagrange function when $u = 0$. For convenience, the augmented Lagrange function can be converted into a different form since the last two items in Eq. (17) can be expressed as:

$$z_i^T(Wy_i - (x_i - b) - g(e_i)) + u/2 \|Wy_i - (x_i - b) - g(e_i)\|_2^2$$

$$= z_i^T(Wy_i - (x_i - b) - g(e_i)) + u/2 \|Wy_i - (x_i - b) - g(e_i)\|_2^2$$

$$\times (Wy_i - x_i + b - g(e_i)) = u/2 \|Wy_i - (x_i - b) - g(e_i) + 1/u z_i\|_2^2$$

$$- 1/(2u) \|z_i\|_2^2$$

The augmented Lagrange function can be rewritten as

$$L_u = \sum_{i=1}^{m} (\|e_i\|_f + u/2 \|Wy_i - (x_i - b) - g(e_i) - 1/u z_i\|_2^2)$$

$$\quad - 1/(2u) \|z_i\|_2^2$$

Solving with joint variables $\gamma_i, e_i, z_i, i = 1 \ldots m, W$ and $b$ is difficult. Therefore, an iterative method is used to solve Eq. (19). The pseudo code is summarized in algorithm 1.

In algorithm 1, we need to optimize Eqs. (20), (21) and (22). Because they are all solved by ADMM algorithm, the detail of solving Eq. (20) is only given as follows (other are presented in the Appendix).

Given $W^k$ which satisfies $(W^k)^T W^k = I$ and $y^k_i, i = 1 \ldots m$, calculate $b$ by ADMM algorithm. ADMM has the following iterations to solve Eq. (20).

In step1, given $g(e_i) = g(e_i^{k-1})$ and $z_i = z_i^{k-1}$ update $b$ by

$$b_i^{k+1} = \arg \min_{\gamma_i} l_{\alpha_i}(e_i, b, z_i)$$

$$= \arg \min_{b} u/2 \|W^T y_i^k - (X - b1 + g(E) - 1/u Z)\|_2^2$$

where $g(E) = [g(e_1), \ldots, g(e_m)]$ and $Z = [z_1, \ldots, z_m]$. $b_i^{k+1}$ can be obtained by taking the gradient of Eq. (23) with respective to $b$ and setting it to zero.

$$b_i^{k+1} = 1/m(X - W^k y_i^k + g(E) - 1/u Z)^T$$

In step2, given $b = b_i^{k+1}$ and $z_i = z_i^{k-1}$ update $e_i$ by

$$e_i^{k+1} = \arg \min \|e_i\|_f$$

$$\quad + 1/2 \|e_i - g^{-1}(W^k y_i^k - (x_i - b_i^{k+1}) + 1/u Z)\|_2^2$$

Eq. (25) can be computed via the singular value thresholding algorithm (Cai et al., 2010). Consider the singular value decomposition of a matrix $Q \in R^{p \times q}$ of rank $r$. $Q$ can be expressed as
3.4. Computational complexity

Given $m$ training images with the size of $p \times q$. The computational complexity of $b$, $Y$ and $w$ are $O(k_1(min(mp^2q, mpq^2)))$, $O(k_2(min(mp^2q, mpq^2)))$ and $O(k_3(min(mp^2q, mpq^2)))$, where $k_1$, $k_2$ and $k_3$ are the iteration times, respectively. Thus, the computational complexity of the algorithm is $O(K(max(k_1, k_2, k_3) \times [min(mp^2q, mpq^2)]))$, where $K$ is the iteration times which optimize the objective function (8). Empirically, our algorithm stops within 50 iterations (i.e., $k \leq 50$).

3.5. Convergence analysis

In this section, we prove that the algorithm, which is proposed to solve objective function (8), has a local optimal solution. Here, we first prove the ADMM algorithm which optimizes $b$ has an optimal solution and the details are given as follows. The convergence of ADMM method has been investigated extensively (Kontogiorgis & Meyer, 1998; Lions & Mercier, 1979; Yuan & Yang, 2009).

Theorem 2. When penalty parameter is some constant greater than zero, the sequence $\{g(E^k), b^k, Z^k\}$ generated by ADMM algorithm converges to a saddle point $(g(E^\infty), b^\infty, Z^\infty)$ of the Lagrangian function.

Proof. It is commonly known that the objective function (8) is closed, proper and convex. That means, there are $g(E)$ and $b$ to minimize the augmented Lagrangian function.

Assumption 1. The unaugmented Lagrangian function $L$ has a saddle point. Let $(g(E^k), b^k, Z^k)$ be the saddle point, $(g(E^\infty), b^\infty, Z^\infty)$ meet $L(g(E^k), b^k, Z^k) \leq L(g(E^\infty), b^\infty, Z^\infty) \leq L(g(E), b, Z)$ for all $g(E), b, Z$.

Under Assumption 1, if the ADMM iterates algorithm satisfies: Residual convergence, Objective convergence and Dual variable convergence (Boyd et al., 2011), the ADMM algorithm has good convergence. Let $q^{k_i} = \|e_i^{k_i}\|_\ast, q^{k_i} = \|e_i^{k_i}\|_\ast$ and $r^{k_i} = W_yi - g(e_i^{k_i}) - (x_i - b^{k_i}).$

Since $(g(E^k), b^k, Z^k)$ is a saddle point of the Lagrangian function $L$, we have $L(g(E^k), b^k, Z^k) \leq L(g(E^{k+1}), b^{k+1}, Z^{k+1})$. Using $g(E^k) = W_yi - (X - b^{k_1})$, the above equation can be rewritten as

$$\sum_{i=1}^{m}(q_i^k - q_i^{k+1}) \leq tr((Z^k)^T R^{k+1})$$

where $R = \{r_1, \ldots, r_m\}$. Meanwhile, we know $b^{k+1} = arg min_w L_w(g(e_i^{k+1}), b^{k+1}, z^{k+1})$. The optimality condition is

$$0 \in \partial arg min_w L_w(g(e_i^{k+1}), b^{k+1}, z^{k+1})$$

$$= u(W_yi - X) = u((g(E^{k+1}) - Z^k)^T + ub^{k+1})$$

Since $Z^{k+1} = Z^k + ub^{k+1}$, we have $0 \in Z^{k+1} - u(g(E^{k+1}) - g(E^k))$. This implies that $b^{k+1}$ minimizes $(Z^{k+1} - u(g(E^{k+1}) - g(E^k)))^T b$. A similar argument shows that $g(e_i^{k+1})$ minimizes $\|e_i\|_\ast - (z_i^{k+1})^T g(e_i)$. It follows that

$$\sum_{i=1}^{m}(z_i^{k+1} - u(g(e_i^{k+1}) - g(e_i))^T b^{k+1}$$

$$\leq \sum_{i=1}^{m}(z_i^{k+1} - u(g(e_i^{k+1}) - g(e_i))^T b$$

$$\sum_{i=1}^{m}(\|e_i^{k+1}\|_\ast - (z_i^{k+1})^T g(e_i^{k+1})) \leq \sum_{i=1}^{m}(\|e_i\|_\ast - (z_i^{k+1})^T g(e_i))$$

Adding above two inequalities and rearranging, we have

$$\sum_{i=1}^{m}(q_i^{k+1} - q_i^k) \leq \sum_{i=1}^{m}(-z_i^{k+1})^T r_i^{k+1} - u(g(e_i^{k+1}) - g(e_i^k))^T$$

$$(-r_i^{k+1} + (g(e_i^{k+1}) - g(e_i^k)))$$
Adding (30) and (34), by some techniques, then we obtain
\[
\sum_{i=1}^{m} \left( (1/u)\|z_i^k - z_i^k\|_2^2 + u\|g(e_i^1) - g(e_i^1)\|_2^2 \right)
- \left( (1/u)\|z_i^{k+1} - z_i^k\|_2^2 + u\|g(e_i^1) - g(e_i^1)\|_2^2 \right)
\geq \sum_{i=1}^{m} \left( u\|e_i^{k+1} + g(e_i^{k+1}) - g(e_i^1)\|_2 \right)
\geq u\|R^{k+1}\|_2 + \|g(E^{k+1}) - g(E^{k})\|_2
\]
Let \(V^{k_1} = (1/u)\|Z^{k_1} - Z^k\|_2^2 + u\|g(E^{k_1}) - g(E^{k})\|_2^2\), we can obtain \(V^{k_1} - V^{k_1+1} \geq u\|R^{k+1}\|_2 + \|g(E^{k+1}) - g(E^{k})\|_2^2\). Since \(V^{k_1}\) is non-negative, it is easy to obtain that
\[
u \sum_{k=0}^{\infty} \left( \|R^{k+1}\|_2^2 + \|g(E^{k+1}) - g(E^{k})\|_2^2 \right) \leq V^0
\]
Because Eq. (36) is an infinite sum and there is an upper bound \(V^{k_1+1} \leq V^k \leq V^0\), the \(R^{k_1} \rightarrow 0\) and \(g(E^{k_1}) \rightarrow g(E^{k})\) as \(k_1 \rightarrow \infty\). In addition, from He and Yang (1998), we obtain that \(Z^{k_1} \rightarrow Z^k\) as \(k_1 \rightarrow \infty\). Therefore, the algorithm of optimizing \(b\) is convergent. By the same token, the algorithm which optimize \(W\) and \(Y\) have good convergence too.

**Theorem 3.** The algorithm which solves the objective function (8) will converge to a local solution \((W^*, Y^*, b^*)\).

**Proof.** We already know that the algorithms of Eqs. (20), (21) and (22) have good convergence. Next, we prove Eq. (8) is convergent. From the above analysis, we can see that any two variables are given from \(b^1, Y^1, W^1\), the remaining variable must exist an optimal solution to meet the constrain condition \(WY - g(E) = X - b1 = 0\). It is obvious that we can obtain
\[
W^*Y^* - g(E) - (X - b^*)^1 = 0
\]
Similarly, from Theorem 2 we can know that \(\nabla L(W^*, y^*, b^*) = 0\), where \(\nabla(\cdot)\) denotes gradient. In addition, \(\frac{\partial g}{\partial y_i} = 0, i = 1, \ldots, m\) is true since Eq. (37) holds. So the algorithm, which solved Eq. (8), satisfies the Karush–Kuhn–Tucker (KKT) condition (Boyd & Vandenberghe, 2004). The \((b^*, Y^*, W^*)\) is at least a local solution of the objective function (8).

### 3.6. Discussion

In this subsection, we discuss and analyze how our model works. To facilitate the analysis, we refer to the traditional PCA.

By simple algebra, we rewrite the solution of traditional PCA as follows:
\[y_i = W^T(x_i - b), i = 1, \ldots, m\]
\[b = 1/m(X - W^TY)^T\]
\[W = (X - b1)^T(Y^TY)^{-1}\]

As the comparison, we exhibit the optimization of \(b, y_i, i = 1, \ldots, m\) and \(W\) of N-PCA computed in Eqs. (24), (A.2) and (B.2) respectively as follows:
\[y_i = W^T(x_i - b + g(e_i) - 1/u\zeta), i = 1, \ldots, m\]
\[b = 1/m(X - W^T + g(E) - 1/u\zeta)^T\]
\[W = (X - b1 + g(E) - 1/u\zeta)^T(Y^TY)^{-1}\]

Compared with Eq. (38), there are two extra items \(g(E)\) and \(-1/u\zeta\) for each equation in Eq. (39). Next, we analyze the significance of each item. Suppose all samples are covered by a black block occlusion which is of low-rank. First, we consider the variable \(g(e_i)\). Remember nuclear norm is employed as the distance measure to measure residual map \(e_i, i = 1, \ldots, m\). From Eq. (28), we know that the \(e_i, i = 1, \ldots, m\) denotes a low-rank error matrix, \(g(e_i)\) is mainly to capture the low rank block residual as our aim. As we see in Eq. (39), all samples are associated with a \(g(e_i)\) as a compensation for the occlusion. Then, we discuss the other term \(-1/u\zeta\) which compensates the rest residual excluding the low-rank block occlusion. To see more detail, we take the algorithm of solving \(W\) as an example, the updating of \(Z^{k+1} = Z^{k1} + u(W^{k+1}y^{k+1}T - (X_i - b^k1) - g(e_i^{k+1}))\) in Eq. (B.5), where the \(W^{k+1}y^{k+1}T - (X_i - b^k1)\) is the reconstruction error of a sample. From Eq. (B.5), we consider that \(1/u\zeta^{k+1}\) is composed of three parts, including the low-rank error which is the difference between \(g(e_i^{k+1})\) and the low-rank error part of \(W^{k+1}y^{k+1}T - (X_i - b^k1)\), the reconstruction error apart from the low rank error of \(W^{k+1}y^{k+1}T - (X_i - b^k1)\) and \(1/u\zeta^{k1}\). \(1/u\zeta^{k1}\) is the sum of the reconstruction errors in the first \(k_1\) iterations, which aims to obtain \(g(e_i^{k+1})\) and \(-1/u\zeta^{k1}\) since the \(g(e_i^{k+1})\) is the cumulative sum of the first \(k_1 + 1\) reconstruction errors in Eq. (B.4).

To visualize the analysis, we present an illustration in Fig. 5. Fig. 5(a) shows a template image from the Extended Yale B database. In Fig. 5(b), the template image is added with a black block occlusion as a sample. Fig. 5(c) shows the reconstruction images of \(y^{k+1}\), the reconstruction error images and the images produced by \(g(e_i^{k+1})\) from left to right respectively. From Fig. 5(c), the reconstruction error images (middle column) indicate not only the low-rank error, but also the reconstruction error apart from the low rank error of \(W^{k+1}y^{k+1}T - (X_i - b^k1)\). In addition, the \(g(e_i^{k+1})\) (right column) is different from the low-rank error part of \(W^{k+1}y^{k+1}T - (X_i - b^k1)\) (middle column). One can see that the noise parts of the reconstruction images and reconstruction error images are continuously eliminated with each iteration. It means that the \(-1/u\zeta^{k+1}\) can remove the errors caused by reconstruction error and \(g(e_i^{k+1})\) in Eq. (39).

To show the effectiveness of N-PCA, the reconstruction images of the same occluded image, which are produced by \(y^{k+1}\), \((W^{k+1}y^{k+1}T - (X_i - b^k1))\) (i.e., direct projection) and traditional PCA, are shown in Fig. 5(d). It is clear that the robustness of algorithm can be improved by removing optimal mean and estimating the genuine location of an image in the low-dimensional feature space. The reason is that the \(g(E)\) and \(-1/u\zeta\) play an important role in the updating of \(y^{k+1}\) and \(b^{k1}\) to compensate the contaminated part of the image. Therefore, the \(g(E)\) and \(-1/u\zeta\) are very useful for image representation and classification.

### 3.7. Feature extraction for a testing sample

Given a testing sample image \(x^{test}\) in \(R^n\), where \(n\) denotes the original space dimensionality of testing sample. First, we obtain the \(b^*, W^*, y^*\) and \(g(e_i^*)\) by substituting training set into the N-PCA model. The mean vector \(b^*\) and the low-dimensional feature space \(W^*\) are not changed when the model is used to extract features of the testing sample. In contrast, the genuine location of an image in the low-dimensional feature space \(y^*\) and the error \(g(e_i^*)\) correspond to a sample \(x_i\). That means, we need to optimal the \(y^{test}\) and \(g(e^{test})\) for the testing sample \(x^{test}\). The model can be expressed as
\[
\min_{y^{test}} \|g^{-1}(W^*y^{test} - (x^{test} - b^*))\|
\]
\[
\text{s.t. } g(e^{test}) = W^*y^{test} - (x^{test} - b^*), i = 1, \ldots, m
\]
Fig. 5. Illustration of effectiveness of N-PCA. (a) Template image from Extended Yale B database, (b) The template image is added with a black block occlusion, (c) The first column is the reconstruction images of $y_{r+1}^{m}$; the second column is the reconstruction error images; the third column is the images produced by $g(x^{k+1})$. From top to bottom, they denote the procedure from first iteration to the last iteration. (d) The reconstruction images of occluded image using $y_{r+1}^{m}$, $(W^{r+1}x^{k+1} + b^{r+1})$ and traditional PCA, respectively.

The ADMM optimization approach, which is utilized in solving the objective function (21), is used to solve Eq. (41). According to the character of ADMM and the convexity of objective function (41), we know that the algorithm of solving Eq. (41) is convergent. Therefore, $(y_{r+1}^{m})^*$, $b^*$ and $W^*$ are the genuine location of an image in the low-dimensional feature space, mean vector and low-dimensional feature space respectively of the testing sample $x^{test}$.

4. Experiments

In the section, the proposed N-PCA is evaluated on five well-known face image databases including the Extended Yale B, the AR, the ORL, the CMU PIE, and the Georgia Tech (GT) database and compared with the other feature extraction methods, including PCA (Jolliffe, 2011), R1-PCA (Ding et al., 2006), PCA-L1 greedy (Kwak, 2008), PCA-L1 nongreedy (Nie et al., 2011), Modular PCA (Gottumukkal & Asari, 2004), Angle PCA (Wang et al., 2017a), HQ-PCA (He, Hu et al., 2011), L2,p-PCA (Wang et al., 2016), Schatten 1-norm PCA (Gu et al., 2012) and N2-DPCA (Zhang et al., 2015). The parameters of PCA, R1-PCA, PCA-L1 greedy, PCA-L1 nongreedy, Modular PCA, Angle PCA, HQ-PCA, L2,p-PCA, Schatten 1-norm PCA and N2-DPCA follow the suggestion in Ding et al. (2006), Gottumukkal and Asari (2004), Gu et al. (2012), He, Hu et al. (2011), Jolliffe (2011), Kwak (2008), Nie et al. (2011), Wang et al. (2016), Wang et al. (2017a), and Zhang et al. (2015), respectively. The sample mean $b$ of all methods is the same as that of PCA but HQ-PCA and N-PCA, HQ-PCA and N-PCA can estimate the sample mean during optimization. As pointed out in Boyd et al. (2011), the algorithm of N-PCA is sensitive with respect to the parameter $u$. In our experiments, we estimate the parameter $u$ by $u = (cpq)/m$, where $c$ is a constant, $p$ and $q$ are the number of rows and columns of an image matrix, and $m$ denotes the number of samples.

To be fair, we assess the performance by both reconstruction error and recognition accuracy which is obtained by applying PCAs to classification tasks using a nearest neighbor (NN) classifier. In our experiments, the reconstruction error is calculated by

$$\text{error} = \frac{1}{m} \|X^{\text{clean}} - b - W^{\text{noise}}\|_F$$

where $X^{\text{clean}} \in \mathbb{R}^{n \times m}$ is the clean training sample, and $Y^{\text{noise}} \in \mathbb{R}^{n \times m}$ is the low-dimensional representation of $m$ training sample with the noise.

4.1. Experiments on the extended Yale B database

In the Extended Yale B database (Lee, Ho, & Kriegman, 2005), there are 38 individuals. It includes 2144 images under different illuminations. The 31 individuals have 64 images while the 11th, 12th, 13th, 14th, 15th, 16th and 17th classes respectively have 60, 59, 60, 63, 62, 63, 63 images. Here, we used 1984 images from 31 individuals under different illuminations and every image is resized to $64 \times 64$ pixels. Fig. 6 shows some samples in the Extended Yale B database and the corresponding noised images. In the experiment, we evaluated the robustness of all methods to noise caused by black block occlusion (Zhang et al., 2015). Here, we randomly selected half of the images for training and the rest images for testing. Then all samples were imposed by the black block occlusion with the size of $20 \times 10$. PCA, PCA-L1 greedy, PCA-L1 nongreedy, R1-PCA, Modular PCA, Angle PCA, HQ-PCA, L2,p-PCA and our method were used to extract features respectively and the dimensions of features were within range from 10 to 200. We repeated this process 10 times.

Fig. 7(a) shows that the reconstruction error of all methods with varying features number, while Fig. 7(b) shows the recognition accuracy of all methods with varying features number. From Fig. 7, we can see that the recognition accuracies rise along with the increase of projection dimensions for all methods. Our method achieves better performance than the other eight methods for image reconstruction and classification in most cases, which obviously indicates that our method exhibits better tolerance for block occlusion. Especially, N-PCA is the only one whose accuracy exceeds 60%.

In the second experiment, we evaluated the robustness of all methods to noise caused by different illuminations. For per class, the images of Subsets 1 and 2 were used for training and the images of Subsets 3 and 4 for testing. Fig. 8(a) shows that the reconstruction error of all methods with varying features number,
Fig. 6. Some samples in the Extended Yale B database. The second row is the corresponding noised images.

Fig. 7. Reconstruction error and recognition accuracy of different methods with varying feature number on the Extended Yale B database. (a) Reconstruction error. (b) Recognition accuracy with a NN classifier.

Fig. 8. Reconstruction error and recognition accuracy of different methods with varying feature number on the Extended Yale B database. (a) Reconstruction error. (b) Recognition accuracy with a NN classifier.

while Fig. 8(b) shows the recognition accuracy of all methods with varying features number. From Fig. 8, we can see that our method is remarkably superior to the other methods for image reconstruction and classification in most cases.

4.2. Experiments on the AR database

AR face database (Martinez, 1998) contains over 4000 color images corresponding to 126 people’s faces (70 mans and 56
women). The same images were taken in two sessions. Each person is separated by two weeks. Each section contains 13 images, including frontal views of faces with different facial expressions, lighting conditions and occlusions. In the experiment, we selected 2600 images of 100 people (50 men and 50 women) as a gallery. Each image was cropped to $64 \times 64$. Fig. 9 shows images of one person from the AR face database and the corresponding noised images. For each individual, we selected Section 1 for training, and Section 2 for testing. Then the images without occlusions of two sessions were imposed by the black block occlusion with the size of $20 \times 10$. PCA, PCA-L1 greedy, PCA-L1 nongreedy, R1-PCA, Modular PCA, Angle PCA, HQ-PCA, $L_2,p$-PCA and our method were used to extract features respectively, and the dimensions of features were within range from 10 to 120. We repeated this process 10 times.

Fig. 10(a) shows the reconstruction error of all methods with varying feature number, while Fig. 10(b) plots the recognition accuracy with varying features number. From Fig. 10, we can see that our method has the advantage of performance in all cases. Our method has the top reconstruction accuracy and low reconstruction error. And it outperformed the second best over 2% on the AR database.

4.3. Experiments on the CMU PIE database

The CMU PIE database (Sim, Baker, & Bs, 2002) includes 68 subjects. Each people were taken across 13 different poses, under 43 different lighting conditions, and with four different expressions. We chose one subset (Pose C9) for experiment. Each image was resized to $64 \times 64$ pixels. In the first experiment, half of the images were randomly selected for training (i.e., 12 images per subject), and the rest images for testing. All images were imposed by the black block occlusion with the size of $20 \times 10$. PCA, PCA-L1, R1-PCA, Modular PCA, Angle PCA, HQ-PCA, $L_2,p$-PCA and our method were used to extract features respectively, and the dimensions of features were within range from 10 to 120. We repeated this process 10 times.

Fig. 11(a) and (b) plot the reconstruction error and recognition accuracy vs. dimension of all methods on the CMU PIE database. Fig. 11 shows that our method is remarkably superior than the other eight methods. It has the top reconstruction accuracy and minimal reconstruction error in all cases. And it outperforms the second best over 5% on the CMU PIE database.

In the second experiment, we evaluated the robustness to noise caused by block occlusion with different occlusion rates. For per person, half of the images were used for training and the rest images for testing. All training samples and all testing samples were added with black block occlusion and the occlusion rates from 5% to 30%. Fig. 12 shows occluded images with occlusion rates from 5% to 30%. Fig. 13 shows the recognition accuracy of all methods with different occlusion rates. It can be seen that our method outperforms other methods in all cases.

4.4. Experiments on the ORL databases

The ORL face database (Belhumeur et al., 1997) contains 400 images of 40 distinct subjects. Each people has 10 images with sorts of variations such as facial expressions, varying illuminations and facial detail (glasses or not). In this database, we randomly selected 5 images from each subject for training and the remaining images for testing. All images were resized to $56 \times 46$. Then all training samples and testing samples were imposed by the black block occlusion with the size of $10 \times 10$. PCA, PCA-L1 greedy, PCA-L1 nongreedy, R1-PCA, Modular PCA, Angle PCA, HQ-PCA, $L_2,p$-PCA and our method were used to extract features respectively, and the dimensions of features were within range from 10 to 100. We repeated this process 10 times.

Fig. 14(a) shows the reconstruction error of all methods with varying feature number, while Fig. 14(b) plots the recognition accuracy with varying features number. From Fig. 14, we can
Fig. 11. Reconstruction error and recognition accuracy of different methods with varying feature number on the CMU PIE database. (a) Reconstruction error. (b) Recognition accuracy with a NN classifier.

Fig. 12. Occluded images with different occlusion rate.

Fig. 13. Recognition accuracy of all methods with different occlusion rates on the CMU PIE database with a NN classifier.

4.5. Experiments on the Georgia tech (GT) database

The GT database (Nefian, 2013) has 750 images, which are samples from 50 peoples taken in different time. Each people have 15 images under different facial expression, cluttered backgrounds and lighting conditions. In our experiments, images were processed to an order of $64 \times 64$. For each subject, we selected 10 images for training and the remaining images for testing. Then all training samples and all testing samples were imposed by the black block occlusion with the size of $20 \times 10$. The features were extracted by PCA, PCA-L1 greedy, PCA-L1 nongreedy, R1-PCA, Modular PCA, Angle PCA, HQ-PCA, $L_p$-PCA and our method respectively, and the dimension of features were within range from 10 to 100. We repeated this process 10 times.

Fig. 16(a) shows the reconstruction error of all methods with varying features number, while Fig. 16(b) plots the recognition accuracy with varying features number. From Fig. 16, we can see that the top reconstruction accuracy of our method is superior to other eight methods. The recognition accuracy of our method reaches 45.28% which is over other algorithm 6%.

4.6. Convergence of N-PCA

Theoretical analysis in Section 3.5 indicates that N-PCA is convergent. Fig. 17 shows some convergence curves of N-PCA on the Extended Yale B, AR, CMU PIE, ORL and GT databases, respectively. It reveals that the proposed algorithm has good convergence on the Extended Yale B, AR, CMU PIE, ORL and GT databases. It is consistent with our theory analysis in Section 3.5.
4.7 Discussions on experimental results

In this section, N-PCA is compared with the other feature extraction methods, including PCA, PCA-L1 greedy, PCA-L1 non-greedy, R1-PCA, Modular PCA, Angle PCA, HQ-PCA, $L_2, p$-PCA, Schatten 1-norm PCA and N2-DPCA. And the top recognition accuracy and reconstruction error of each method on the Extended Yale B, AR, CMU PIE, ORL and GT databases is illustrated in Tables 3 and 4, respectively.

From the above-mentioned experimental results, we can know the following.

1. On the Extended Yale B, AR and CMU PIE databases, Modular PCA may be superior to PCA, PCA-L1 greedy, PCA-L1 non-greedy, R1-PCA, Angle PCA, HQ-PCA, $L_2, p$-PCA and N-PCA when the number of the projection is small. The reason can be explained as follows: For Modular PCA, it divides each image into smaller sub-images, before applying ordinary PCA approach to extract features from these sub-images. The number of features which are extracted by Modular PCA approach is more than other methods (i.e., If an image is divided into four sub-images, Modular PCA approach extracts four times as many features as other methods). In the other words, the first few PCs extracted by Modular PCA are more information. However, holistic appearance based PCAs outperform Modular PCA when most useful information of data is extracted as the number of projection vector is high. The main reason may be that sub-images could not retain the global information of the image. Once it is lost, the accuracy of such method may deteriorate.

2. N2-DPCA and N-PCA achieve better performance than the other eight methods in image reconstruction task and classification problem. This is probably because that N2-DPCA and N-PCA utilize the spatial structure information which is embedded in...
pixels of image and the structure information is important for image representation and recognition. Moreover, in our experiment, all images are added with a black block noise. The structured noise (e.g., block occlusion) causes low rank error image. Nuclear norm is the convex envelope of the matrix rank. So nuclear norm is more suitable as distance metric for the structure noise. Another reason may be that N-PCA and N2-DPCA take full advantage of the reconstruction error which is an essential characteristic of PCA. In contrast, the objective function of PCA-L1 greedy, PCA-L1 non-greedy and Schatten 1-norm PCA are to maximize the variance which is not equivalent to minimizing the reconstruction error of data.

(3) From Table 3, we can note that our method is remarkably superior to the N2-DPCA and Schatten 1-norm PCA for the image classification with a NN classifier. This is probably because that N2-DPCA and Schatten 1-norm PCA neglect problem of the bias of empirical mean when images are occluded with block which does not follow Gaussian distribution. Our method is to approximate the true mean of given data matrix. Moreover, as a distance metric, nuclear norm can weaken the impact of outlier on an image.

At the meantime, our method pays attention to estimate the genuine location of images in the low dimensional feature subspace which is also neglected in N2-DPCA and Schatten 1-norm PCA.

5. Conclusions

In this paper, a novel robust principal component analysis method, i.e. N-PCA, is presented by employing nuclear norm to capture the structure information of 2D reconstruction error. Regarding the problems that structured noises are involved in the sample set, we model the N-PCA under a new framework which minimizes the reconstruction error by jointly optimizing mean and low-dimensional principal component besides projection matrix which is the only variable in most previous PCA methods. An iterative algorithm is proposed to solve N-PCA effectively along with analyses of convergence and computation complexity. We provide an insight into the mechanism that how N-PCA is capable of dealing with occlusion block on images. Experience results on five databases illustrate that the performance of

![Fig. 16. Reconstruction error and recognition accuracy of different methods with varying feature number on the GT database. (a) Reconstruction error. (b) Recognition accuracy with a NN classifier.](image)

![Fig. 17. Value of the function versus iteration times on five databases.](image)
the proposed method outperforms several state-of-the-art robust PCA methods. In the future work, we will focus on reducing the computational cost of the algorithm used to solve the N-PCA model.

Acknowledgments

This work was supported by the Natural Science Foundation of Chongqing (under Grant Nos. cstc2018jcyjAX0532), the National Natural Science Foundation of China (under Grant Nos. 61472055, 61573248, 61802267[kong] and 61732011) and the Shenzhen Municipal Science and Technology Innovation Council (under Grant Nos. JCYJ20180305124834854).

Appendix A. The solution of Eq. (21)

Given \( W^k \) which satisfies \( (W^k)^TW^k = I \) and \( b^{k-1} \) from Eq. (20), calculate \( y_i, i = 1, \ldots, m \) by ADMM algorithm. ADMM is composed of the following iterations.

In step1, given \( g(e_i) = g(e_i^k) \) and \( z_i = z_i^k \) update \( y_i, i = 1, \ldots, m \) by

\[
y_i^{k+1} = \arg \min L_u(e_i, y_i, z_i) = u/2tr(y_i^T(W^k)^TW^ky_i - 2y_i^T(W^k)^T)\times(e_i - b^{k-1} + g(e_i) - 1/uz_i + \text{const})
\]  
(A.1)

where \( y_i^{k+1} \) can be obtained by the gradient of Eq. (A.1).

\[
y_i^{k+1} = (W^k)^T(x_i - b^{k-1} + g(e_i) - 1/uz_i)
\]  
(A.2)

In step2, given \( y_i = y_i^{k+1} \) and \( z_i = z_i^{k+1} \) update \( e_i \) by

\[
e_i^{k+1} = \arg \min\{1/uz_i\}
\]

\[
e_i^{k+1} = D_i/u(g^{-1}(W^ky_i^{k+1} - (x_i - b^{k-1} + 1/uz_i))
\]  
(A.4)

In step3, given \( e_i = g(e_i^{k+1}) \) and \( y_i = y_i^{k+1} \), update \( z_i \) by

\[
z_i^{k+1} = z_i^{k} + u(W^ky_i^{k+1} - (x_i - b^{k-1}))
\]  
(A.5)

This iterative procedure is repeated until the sequence \( \{g(E^k), Y^k, Z^k\} \) is generated by ADMM algorithm converges to a saddle point \( (g(E^k), Y^k, Z^k) \).

Appendix B. The solution of Eq. (22)

Given \( b^{k+1} \) from Eq. (20) and \( y_i^{k+1}, i = 1, \ldots, m \) from Eq. (21), calculate \( W \) by ADMM algorithm. ADMM is composed of the following iterations.

In step1, given \( g(e_i) = g(e_i^k) \) and \( z_i = z_i^k \) update \( W \) by

\[
W_{k+1} = \arg \min L_u(Z, W, z_i)
\]

\[
= \arg \min u/2||z_i||_F^2 - \langle X - b^{k+1} - g(E) - 1/uz_i, 0-z_i \rangle_2^2
\]  
(B.1)

We can get the solution of \( W \) by Eq. (B.1).

\[
W^{k+1} = (X - b^{k+1} + g(E) - 1/uz_i)(Y^{k+1}(Y^{k+1}Y^{k+1})^{-1})^{-1}
\]  
(B.2)

In step2, given \( W = W^{k+1} \) and \( z_i = z_i^{k+1} \) update \( e_i \) by

\[
e_i^{k+1} = \arg \min\{1/uz_i\}
\]

\[
e_i^{k+1} = D_i/u(g^{-1}(W^ky_i^{k+1} - (x_i - b^{k-1} + 1/uz_i))
\]  
(B.3)

From Theorem 1, the optimal solution of (B.3) can be expressed as

\[
e_i^{k+1} = D_i/u(g^{-1}(W^ky_i^{k+1} - (x_i - b^{k-1} + 1/uz_i))
\]  
(B.4)

In step3, given \( e_i = g(e_i^{k+1}) \) and \( W = W^{k+1} \), update \( z_i \) by

\[
z_i^{k+1} = z_i^k + u(W^{k+1}y_i^{k+1} - (x_i - b^{k+1}))
\]  
(B.5)

This iterative procedure is repeated until the sequence \( \{g(E^k), W^k, Z^k\} \) is generated by ADMM algorithm converges to a saddle point \( (g(E^k), W^k, Z^k) \).
References


