Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

# Two-Dimensional PCA with F-Norm Minimization

**Qianqian Wang**
State Key Laboratory of ISN, Xidian University
Xi'an China

**Quanxue Gao**
State Key Laboratory of ISN, Xidian University
Xi'an China

## Abstract

Two-dimensional principle component analysis (2DPCA) has been widely used for face image representation and recognition. But it is sensitive to the presence of outliers. To alleviate this problem, we propose a novel robust 2DPCA, namely 2DPCA with F-norm minimization (F-2DPCA), which is intuitive and directly derived from 2DPCA. In F-2DPCA, distance in spatial dimensions (attribute dimensions) is measured in F-norm, while the summation over different data points uses 1-norm. Thus it is robust to outliers and rotational invariant as well. To solve F-2DPCA, we propose a fast iterative algorithm, which has a closed-form solution in each iteration, and prove its convergence. Experimental results on face image databases illustrate its effectiveness and advantages.

## Introduction

Principal component analysis (PCA) (Turk and Pentland 1991), linear discriminant analysis (LDA) (Belhumeur, Hespanha, and Kriegman 1997), locality preserving projection (LPP) (He and Niyogi 2005) and neighborhood preserving embedding (NPE) (He et al. 2005) are four of the most representative methods, where PCA is used to extract the most expressive features, LDA is considered to be capable of extracting the most discriminating features. Different from PCA and LDA, which characterize the global geometric structure, LPP and NPE well preserve the local geometric structure of data.

Applying the aforementioned methods to image recognition, we need to transform each image, which is represented as a matrix, into 1D image vector by concatenating all rows. So, these methods cannot well exploit the spatial structure information that is embedded in pixels of image and important for image representation and recognition (Yang et al. 2004; Zhang et al. 2015; Lu, Plataniotis, and Venetsanopoulos 2008). To handle this problem, many two-dimensional subspace learning methods or tensor methods have been developed (Yang et al. 2004; Lu, Plataniotis, and Venetsanopoulos 2008; Yang et al. 2005). In contrast to the aforementioned methods, two-dimensional subspace learning methods directly extract features from image matrix and fully consider the variation among different rows/columns of an image. The representative two-dimensional methods include two-dimensional PCA (2DPCA) (Yang et al. 2004) and two-dimensional LDA (2DLDA) (Yang et al. 2005). Although their motivations of two-dimensional methods are different, they can be unified within the graph embedding framework (Yan et al. 2005) and measure the similarity between images by using *squared F-norm*. It is commonly known that *squared F-norm* is not robust in the sense that outlying measurements can arbitrarily skew the solution from the desired solution. Thus, these methods are not robust in the presence of outliers (Ke and Kanade 2005; Collins, Dasgupta, and Schapire 2001; Gao et al. 2013).

Recently, $\ell_1$-norm based subspace learning technique is considered to be capable of obtaining the robust projection vectors and has become an active topic in dimensionality reduction. For example, Ke and Kanade (2005) proposed L1-PCA that uses $\ell_1$-norm to measure the reconstruction error. Kwak (2008) used $\ell_1$-norm to measure the variance and proposed PCA-L1 with greedy algorithm. Nie *et al*. (2011) proposed a non-greedy iterative to solve PCA-L1. Motivated by $\ell_1$-norm based PCA, some $\ell_1$-norm based LDA algorithms have been developed, such as LDA-L1 (Zhong and Zhang 2013) and ILDA-L1 (Chen, Yang, and Jin 2014). However, $\ell_1$-norm is not rotational invariant (Ding et al. 2006), which is a fundamental property of Euclidean space with $\ell_2$-norm. It has been emphasized in the context of learning algorithms (Kwak 2014). Based on this content, Ding *et al*. (2006) proposed the rotational invariant $\ell_1$-norm for feature extraction and developed $R_1$-PCA that measures the similarity among data by $R_1$-norm, which is just $\ell_{2,1}$-norm of a matrix. To further analysis robustness of subspace learning technique, Kwak *et al.* extended $\ell_1$-norm to $\ell_p$-norm and proposed $\ell_p$-norm based subspace learning methods (Kwak 2014; Oh and Kwak 2016).

Although the aforementioned methods are robust to outliers, they need to transform 2D image into a vector by concatenating all rows of image. So, these methods cannot well exploit the spatial structure information of data. To handle this problem, Li *et al*. (2010) extended PCA-L1 to 2DPCA-L1 with greedy algorithm. Wang *et al*. (2013) imposed sparse constraint in 2DPCA-L1 and proposed 2DPCAL1-S. Y. Pang *et al*. (2010) proposed $\ell_1$-norm based tensor subspace learning. Wang *et al*. (2015) proposed 2DPCA-L1

with non-greedy algorithm. However, these methods do not have rotational invariance and do not explicitly consider the reconstruction error, which is the real goal of PCA.

To handle these problems, we propose a robust 2DPCA with F-norm minimization, namely F-2DPCA for feature extraction. In F-2DPCA, distance in spatial dimensions (attribute dimensions) is measured in F-norm, while the summation over different data points uses $\ell_1$-norm. Furthermore, we solve F-2DPCA by non-greedy iterative algorithm, which has a closed-form solution in each iteration. Finally, we prove the convergence of our proposed algorithm. Compared with $\ell_1$-norm based 2DPCA methods, our approach has the following advantages. Fist, F-2DPCA not only is robust to outliers but also has rotational invariance, which has been emphasized in the context of learning algorithms; Second, our proposed non-greedy algorithm has a local solution and best minimizes the objective function value. Third, our approach (solution) relates to image covariance matrix.

## 2DPCA and 2DPCA-L1

Denote by $\mathbf{A}_i \in \mathbf{R}^{m\times n}(i = 1, 2, ..., N)$ the $N$ training images, and $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_k] \in \mathbf{R}^{n\times k}$ the projection matrix. Without loss of generality, we assume the data set are centralized, i.e., $\sum_{i=1}^{N} \mathbf{A}_i = 0$. 2DPCA aims to seek a projection matrix by (Yang et al. 2004):

$$\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}_k} \sum_{i=1}^{N} \|\mathbf{A}_i\mathbf{V}\|_F^2 \tag{1}$$

where $tr(\cdot)$ is the trace operator of a matrix, $\mathbf{I}_k \in \mathbf{R}^{k\times k}$ is an identity matrix, and $\|\cdot\|_F^2$ denotes the squared F-norm. It is easy to see that, the objective function (1) is totally equivalent to the objective function (2) due to the fact $\sum_{i=1}^{N} \left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F^2 + \sum_{i=1}^{N} \|\mathbf{A}_i\mathbf{V}\|_F^2 = \sum_{i=1}^{N} \|\mathbf{A}_i\|_F^2$.

$$\min_{\mathbf{V}^T\mathbf{V}=\mathbf{I}_k} \sum_{i=1}^{N} \left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F^2 \tag{2}$$

The solution of the objective function (1) or (2) is composed of the eigenvectors of the image covariance matrix $\mathbf{S}_t = \sum_{i=1}^{N} (\mathbf{A}_i)^T\mathbf{A}_i$ corresponding to the first $k$ largest eigenvalues. We can see that squared large distance will remarkably dominate the solution of the objective function (1) or (2). Thus, the objective function (1) or (2) is not robust in the sense that outlying measurements can skew the solution from the desired solution. To handle this problem, 2DPCA-L1 was proposed (Li, Pang, and Yuan 2010; Wang et al. 2015). It aims to find the projection matrix by solving the following objective function.

$$\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}_k} \sum_{i=1}^{N} \|\mathbf{A}_i\mathbf{V}\|_{L_1} \tag{3}$$

where $\|\cdot\|_{L_1}$ denotes the $\ell_1$-*norm* of a matrix, which is defined as follows:

$$\|\mathbf{D}\|_{L_1} = \sum_{i=1}^{m}\sum_{j=1}^{n} |\mathbf{D}(i,j)|$$

$\mathbf{D}(i,j)$ denotes the element of the $i$-th row $j$-th column of matrix $\mathbf{D}$.

Compared with traditional 2DPCA, $\ell_1$-norm based 2DPCA technique is robust, but it has several shortcomings as follows. Traditional 2DPCA has rotational invariance, while $\ell_1$-norm based 2DPCA does not have this property. Given an arbitrary rotation matrix $\mathbf{\Gamma}$( $\mathbf{\Gamma}\mathbf{\Gamma}^T = \mathbf{I}$), in general, we have $\|\mathbf{\Gamma}\mathbf{A}_i\mathbf{V}\|_{L_1} \neq \|\mathbf{A}_i\mathbf{V}\|_{L_1}$. Moreover, it is not clear whether $\ell_1$-norm based PCA (i.e., solution) relates to the covariance matrix. Finally, the objective function (3) does not explicitly consider the reconstruction error, which is the real goal of PCA, due to the fact $\sum_{i=1}^{N} \left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_{L_1} + \sum_{i=1}^{N} \|\mathbf{A}_i\mathbf{V}\|_{L_1} \neq \sum_{i=1}^{N} \|\mathbf{A}_i\|_{L_1}$. To handle these problems, we propose a robust 2DPCA with F-norm minimization in the following section.

## 2DPCA with F-norm minimization

### Motivation and Objective function

2DPCA uses the *squared F-norm* to measure the similarity among images in the objective function. It is well known that *squared F-norm* is not robust in the sense that outlying measurements can arbitrarily skew the solution from the desired solution. This results in sensitivity of 2DPCA. To handle this problem, the contribution of distance metric to the criterion function (2) should reduce the effect of large distance. Moreover, we hope to obtain a robust low-dimensional subspace that is not uniquely determined up to an orthogonal transformations. Compared with *squared F-norm*, *F-norm* not only can weaken the effect of large distance but also has rotational invariance. Thus, an intuitive and reasonable way is to use *F-norm* instead of *squared F-norm*, i.e.,

$$\left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F^2 \rightarrow \left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F \tag{4}$$

Substituting Eq. (4) into the objective function (2), we have

$$\operatorname*{argmin}_{\mathbf{V}^T\mathbf{V}=\mathbf{I}_k} \sum_{i=1}^{N} \left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F \tag{5}$$

The objective function (5) is called 2DPCA with F-norm minimization (F-2DPCA). In the objective function (5), distance in spatial dimensions (attribute dimensions) is measured in F-norm, while the summation over different data points uses $\ell_1$-norm. Compared with 2DPCA, our proposed method can further weaken the effect of large distance, and compared with 2DPCA-L1, our proposed method has rotational invariance due to the fact $\|\mathbf{\Gamma}\mathbf{A}_i\mathbf{V}\|_F = \|\mathbf{A}_i\mathbf{V}\|_F$.

Now we consider how to solve the objective function (5). By simple algebra, we have

$$\begin{aligned}\sum_{i=1}^{N} \left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F &= \sum_{i=1}^{N} \frac{\left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F^2}{\left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F} \\ &= \sum_{i=1}^{N} \frac{tr(\mathbf{A}_i{}^T\mathbf{A}_i) - tr(\mathbf{V}^T\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V})}{\left\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\right\|_F}\end{aligned} \tag{6}$$

Substituting Eq. (6) into Eq. (5), and by simple algebra, the objective function (5) becomes

$$\underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}_k}{\operatorname{argmin}} \sum_{i=1}^{N} \Big( tr(\mathbf{A}_i{}^T\mathbf{A}_i) - tr(\mathbf{V}^T\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V}) \Big) \boldsymbol{d}_i \quad (7)$$

where $\boldsymbol{d}_i = \dfrac{1}{\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\|_F}$ , in order to avoid being divided by 0, $\boldsymbol{d}_i$ is defined as follows.

$$\boldsymbol{d}_i = \frac{1}{\|\mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T\|_F + \gamma} \quad (8)$$

where $\gamma > 0$ is a small constant.

In the objective function (7), we have two unknown variables $\mathbf{V}$ and $\boldsymbol{d}_i$ which relate to $\mathbf{V}$ . Thus, it has no closed-form solution and is difficult to directly solve the solution of the objective function (7). An algorithm can be developed for alternatively updating $\mathbf{V}$ (while fixing $\boldsymbol{d}_i$) and $\boldsymbol{d}_i$ (while fixing $\mathbf{V}$). To be specific, in the $(t+1)$-th iteration, when $\boldsymbol{d}_i{}^{(t)}$ is known, we can update $\mathbf{V}$ by minimizing the objective function (7). In this case, the first term in the objective function (7) becomes constant. Thus, the objective function (7) is converted to solve the following objective function:

$$\underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}_k}{\operatorname{argmax}} tr(\mathbf{V}^T\mathbf{H}\mathbf{V}) \quad (9)$$

where $\mathbf{H} = \sum_{i=1}^{N} \mathbf{A}_i{}^T\boldsymbol{d}_i\mathbf{A}_i$, which is the weighted image covariance matrix.

According to the matrix theory, the column vectors of the optimal projection matrix $\mathbf{V}$ of Eq. (9) are composed of the eigenvectors of $\mathbf{H} = \sum_{i=1}^{N} \mathbf{A}_i{}^T\boldsymbol{d}_i\mathbf{A}_i$ corresponding to the $k$ largest eigenvalues. After that, we can calculate $\boldsymbol{d}_i$ by Eq. (8). This iterative procedure is repeated until convergence, which is proved in the subsequent subsection. Eq. (9) illustrates that solution of our proposed method relates to the weighted image covariance matrix. We summarize the pseudo code of solving the objective function (5), i.e., F-2DPCA in *Algorithm* 1.

**Algorithm 1: F -2DPCA**

**Input:** $\mathbf{A}_i \in \mathbf{R}^{m\times n}(\, i = 1, \cdots, N)$, $k$ , where $\mathbf{A}$ is centralized, $\gamma = 0.00001$. Initialize $\mathbf{V}^{(t)} \in \mathbf{R}^{m\times k}$ which satisfies $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, $t = 1$.

**while** not converge **do**

1. For all training samples, calculate $\boldsymbol{d}^{(t)}(i = 1, \cdots, N)$ by Eq. (8).
2. Calculate $\mathbf{H}^{(t)}$ according to Eq. (9), i.e., $\mathbf{H}^{(t)} = \sum_{i=1}^{N} \mathbf{A}_i{}^T\boldsymbol{d}_i{}^{(t)}\mathbf{A}_i$ .
3. Solve $\mathbf{V}^{(t+1)} = \underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}_k}{\operatorname{argmax}} tr(\mathbf{V}^T\mathbf{H}^{(t)}\mathbf{V})$: the column of the optimal solution $\mathbf{V}^{(t+1)}$ are the eigenvectors of $\mathbf{H}^{(t)}$ corresponding to the $k$ largest eigenvalues.
4. Update $t \leftarrow t + 1$.

**end while**

**Output**:$\mathbf{V}^{(t+1)} \in \mathbf{R}^{m\times k}$

## Convergence analysis

**Theorem 1:** In each iteration of *Algorithm* 1, we have:

$$\begin{aligned} &\sum_{i=1}^{N} \left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T \right\|_F \\ &\le \sum_{i=1}^{N} \left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F \end{aligned} \quad (10)$$

i.e., *Algorithm* 1 monotonically decreases the objective function value of F-2DPCA.

**Proof:** For each iteration t, according to step 3 in *Algorithm* 1, we have the following inequality

$$\begin{aligned} &\sum_{i=1}^{N} \frac{tr((\mathbf{V}^{(t+1)})^T\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V}^{(t+1)})}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F} \\ &\ge \sum_{i=1}^{N} \frac{tr((\mathbf{V}^{(t)})^T\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V}^{(t)})}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F} \end{aligned} \quad (11)$$

Multiplying -1 and adding $\sum_{i=1}^{N} \dfrac{tr(\mathbf{A}_i{}^T\mathbf{A}_i)}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F}$ on both sides of Eq. (11), and by simple algebra, the Eq. (11) becomes

$$\begin{aligned} &\sum_{i=1}^{N} \frac{tr(\mathbf{A}_i{}^T\mathbf{A}_i) - tr((\mathbf{V}^{(t+1)})^T\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V}^{(t+1)})}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F} \\ &\le \sum_{i=1}^{N} \frac{tr(\mathbf{A}_i{}^T\mathbf{A}_i) - tr((\mathbf{V}^{(t)})^T\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V}^{(t)})}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F} \end{aligned} \quad (12)$$

According to $\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}\mathbf{V}^T \right\|_F = tr(\mathbf{A}_i{}^T\mathbf{A}_i) - tr(\mathbf{V}^T\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V})$, Eq. (12) becomes

$$\begin{aligned} &\sum_{i=1}^{N} \frac{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T \right\|_F^2}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F} \\ &\le \sum_{i=1}^{N} \frac{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F^2}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F} \end{aligned} \quad (13)$$

According to inequality $a^2 + b^2 \ge 2ab \Rightarrow \dfrac{b^2}{a} \ge 2b - a$ , we have

$$\begin{aligned} &2\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T \right\|_F - \left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F \\ &\le \frac{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T \right\|_F^2}{\left\| \mathbf{A}_i - \mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T \right\|_F} \end{aligned} \quad (14)$$

Eq. (14) holds for each index $i$, thus we can rewrite (14) as

$$\begin{aligned}\sum_{i=1}^{N}(2\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T\Big\|_F \\ -\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T\Big\|_F) \\ \leq\sum_{i=1}^{N}\frac{\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T\Big\|_F^2}{\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T\Big\|_F}\end{aligned} \tag{15}$$

Combining Eq. (13) and Eq. (15) yields

$$\begin{aligned}\sum_{i=1}^{N}(2\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T\Big\|_F \\ -\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T\Big\|_F) \\ \leq\sum_{i=1}^{N}\frac{\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T\Big\|_F^2}{\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T\Big\|_F}\end{aligned} \tag{16}$$

By simple algebra, Eq. (16) becomes

$$\begin{aligned}\sum_{i=1}^{N}\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t+1)}(\mathbf{V}^{(t+1)})^T\Big\|_F \\ \leq\sum_{i=1}^{N}\Big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}^{(t)}(\mathbf{V}^{(t)})^T\Big\|_F\end{aligned} \tag{17}$$

Eq. (17) shows that the *Algorithm* 1 monotonically decreases the objective function value of F-2DPCA in each iteration.

**Theorem 2:** *Algorithm* 1 will converge to a local solution of the objective function (5).

**Proof:** The Lagrangian function of the objective function (5) is

$$L(\mathbf{V})=\sum_{i=1}^{N}\big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}\mathbf{V}^T\big\|_F-tr\left(\mathbf{\Lambda}^T(\mathbf{V}^T\mathbf{V}-\mathbf{I})\right) \tag{18}$$

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

$$\frac{\partial L}{\partial\mathbf{V}}=\sum_{i=1}^{N}(\big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}\mathbf{V}^T\big\|_F^{-1})\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V}-\mathbf{V}\mathbf{\Lambda}^T=0 \tag{19}$$

By simple algebra, we have

$$\sum_{i=1}^{N}(\big\|\mathbf{A}_i-\mathbf{A}_i\mathbf{V}\mathbf{V}^T\big\|_F^{-1})\mathbf{A}_i{}^T\mathbf{A}_i\mathbf{V}=\mathbf{V}\mathbf{\Lambda}^T \tag{20}$$

According to step 3 in *Algorithm* 1, we find the optimal solution of the objective function (9). Thus the converged solution of *Algorithm* 1 satisfies the KKT condition of the objective function (9). The Lagrangian function of Eq. (9) is

$$L_2(\mathbf{V})=tr(\mathbf{V}^T\sum_{i=1}^{N}\mathbf{A}_i{}^T\boldsymbol{d}_i\mathbf{A}_i\mathbf{V})-tr(\mathbf{\Lambda}^T(\mathbf{V}^T\mathbf{V}-\mathbf{I})) \tag{21}$$

Taking the derivative w.r.t. $\mathbf{V}$ and setting it to zero, we get the KKT condition of Eq. (9) as follows

$$\sum_{i=1}^{N}\mathbf{A}_i{}^T\mathbf{A}_i\boldsymbol{d}_i\mathbf{V}-\mathbf{V}\mathbf{\Lambda}^T=0 \tag{22}$$

Eq. (22) is formally similar to Eq. (20). The main difference between Eq. (22) and Eq. (20) is that $\boldsymbol{d}_i$ is known in each iteration in *Algorithm* 1. Suppose we obtain the optimal solution $\mathbf{V}^*$ in the $(t+1)$-th, thus, we have $\mathbf{V}_{t+1}=\mathbf{V}^*=\mathbf{V}_t$. According to the definition of $\boldsymbol{d}_i$, we can see that Eq. (22) is the same as Eq. (20) in this case. It means that the converged solution of *Algorithm* 1 satisfies the KKT condition of Eq. (5), i.e.

$$\left.\frac{\partial L}{\partial\mathbf{V}}\right|_{\mathbf{V}=\mathbf{V}^*}=0 \tag{23}$$

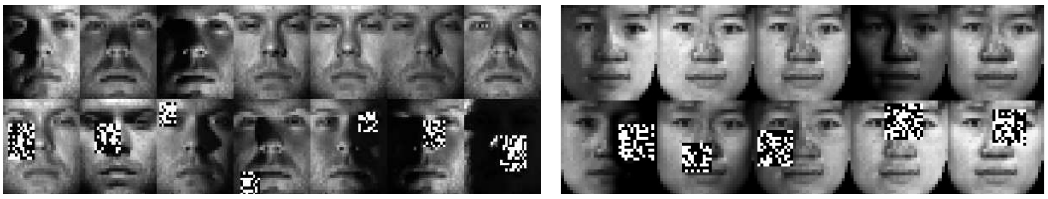
Combining **Theorem 1** and Eq. (23), we have that the converged solution of *Algorithm* 1 is a local solution of Eq. (5).

## Experimental results

We validate our approach in three face databases (Extended Yale B, AR and PIE) and compare it with 2DPCA (Yang et al. 2004), 2DPCA-L1 (Li, Pang, and Yuan 2010), 2DPCA-L1 non-greedy (Wang et al. 2015), 2DPCAL1-S (Wang and Wang 2013) and N-2DPCA (Zhang et al. 2015). In our experiments, we use 1-nearest neighbor (1NN) for classification. We set the number of projection vectors as 25 in the Extended Yale B and CMU PIE databases, 30 in the AR database.

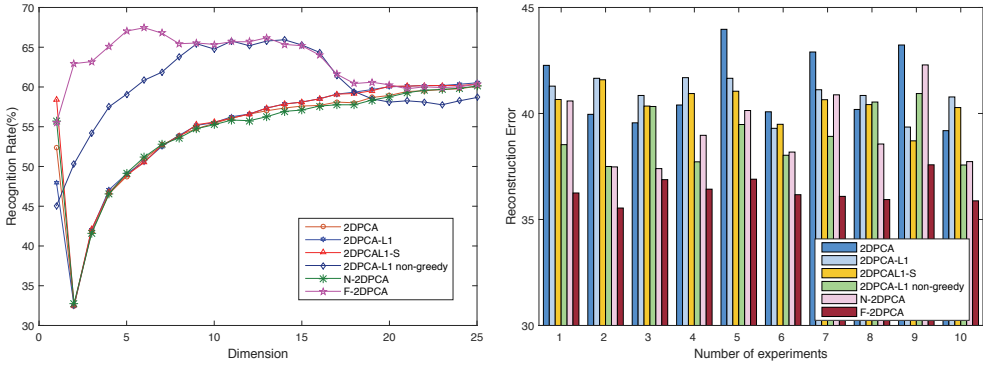
The Extended Yale B database (Georghiades, Belhumeur, and Kriegman 2001) consists of 2144 frontal-face pictures of 38 individuals with different illuminations. There are 64 pictures for each person except 60 for 11th and 13th, 59 for 12th, 62 for 15th and 63 for 14th, 16th and 17th. Figure 1(a) shows some samples of one person in the Extended Yale B database. In the experiments, each image was normalized to $32\times32$ pixels. 14 images of each individual were randomly selected and noised by black and white dots with random distribution. The location of noise is random and ratio of the pixels of noise to number of image pixels is intervenient 0.05 to 0.15. We randomly select 32 images, which include 7 noisy images, per person for training, and the remaining images for testing. 2DPCA, 2DPCA-L1, 2DPCA-L1 non-greedy, 2DPCAL1-S, N-2DPCA and our approach are used to extract features, respectively. We repeat this process 10 times.

In the AR database (Martinez 1998), the pictures of 120 individuals were taken in two sessions. Each session contains 13 color images, which include 6 images with occlusions and 7 full facial images with different facial expressions and lighting conditions. We manually cropped the face



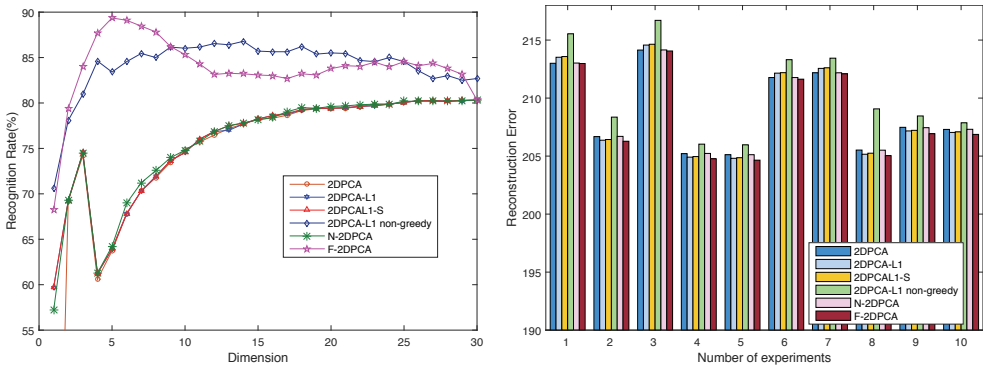
(a) (b)

Figure 1: (a) Some samples of one person in the Extended Yale B database. (b) Some samples of one person in the CMU PIE database. (The second row is noised samples.)




(a) (b)

Figure 2: (a) Classification accuracy *vs.* the number of projection vectors. (b) The optimal Reconstruction Error of six approaches under ten experiments on the Extended Yale B database.



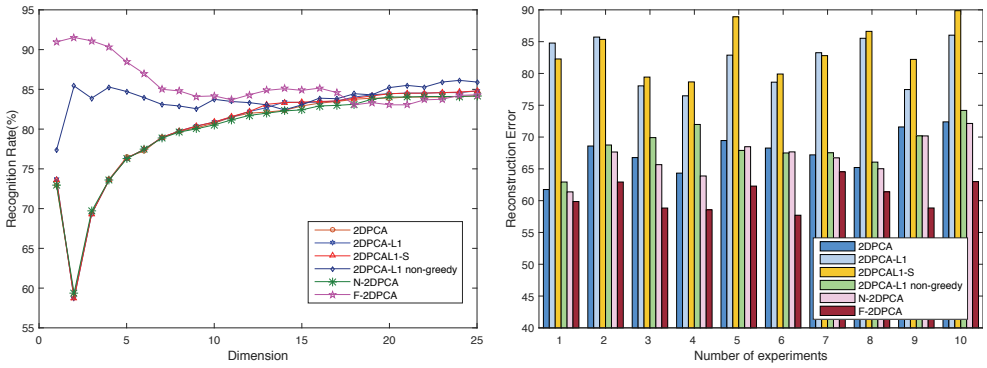

(a) (b)

Figure 3: (a) Classification accuracy *vs.* the number of projection vectors. (b) The optimal Reconstruction Error of six approaches under ten experiments on the AR database.

portion of the image and then normalized it to $50 \times 40$ pixels. In the experiments, we randomly select 13 images per person for training and the remaining images for testing, and then repeat this process 10 times.

The CMU PIE database (Sim, Baker, and Bsat 2002) consists of 2856 frontal-face images of 68 individuals with different illuminations. In the experiments, each image was normalized to $32 \times 32$ pixels, we randomly selected 10 images and added the same noise as that in the Extended Yale B database. Figure 1(b) shows some samples of one person in the CMU PIE database. We randomly select 21 images, which include 16 without noisy images, per person for training and the remaining images for testing, and then repeat this process 10 times.

Tables 1 and 2 list the average recognition accuracy, running time and the corresponding standard deviation of each method on the Extended Yale B, AR, and CMU PIE databases, respectively. Figures 2, 3, and 4 plot the classification curve versus the number of projection vectors and the reconstruction error of six approaches under ten




(a) (b)

Figure 4: (a) Classification accuracy *vs.* the number of projection vectors. (b) The optimal Reconstruction Error of six approaches under ten experiments on the CMU PIE database.

experiments on the Extended Yale B, AR and CMU PIE databases, respectively. Figure 5 shows the convergence curve of our method on the Extended Yale B, AR, and CMU PIE databases.

(1) 2DPCA is overall inferior to the other five approaches. The main reason is that 2DPCA is not robust to outliers such as illumination and occlusion. 2DPCA-L1, 2DPCA-L1 non-greedy and 2DPCAL1-S are not remarkably better than 2DPCA. This is probably because that they do not explicitly consider the reconstruction error. N-2DPCA is not better. The reason is that in classification stage, we use Euclidean distance to measure similarity between data rather than nuclear norm as in (Zhang et al. 2015).

Table 1: The average classification accuracy (%) and the corresponding standard deviation on the Extended Yale B, AR and CMU PIE databases.

| Methods | Experiments | | |
|---|---|---|---|
| | Extended Yale B | AR | CMU PIE |
| 2DPCA | 59.92±0.42 | 80.40±0.88 | 85.39±0.73 |
| 2DPCA-L1 | 60.33±0.38 | 80.39±0.88 | 85.71±0.77 |
| 2DPCA-L1 non-greedy | 66.63±1.01 | 87.49±1.47 | 86.34±0.71 |
| N-2DPCA | 59.99±0.57 | 80.37±0.91 | 85.39±0.73 |
| 2DPCAL1-S | 60.37±0.54 | 80.38±0.85 | 85.91±0.69 |
| F-2DPCA | **67.35±0.95** | **89.19±0.70** | **91.60±0.74** |

(2) F-2DPCA is superior to the other five approaches. Compared with 2DPCA, which uses squared F-norm to measure similarity, F-norm is robust to outliers. Compared with the other $\ell_1$-norm approaches, F-2DPCA is intuitive and directly derived from 2DPCA. Moreover, F-2DPCA retains 2DPCA's desirable properties. For example, F-2DPCA considers the reconstruction error and the solution of F-2DPCA relates to the image covariance matrix. Another reason may be that F-2DPCA best optimizes the objective function. Figure 5 and table 2 illustrate that our proposed algorithm is fast and convergent. This is consistent with our theory analysis in the Convergence analysis section.

Table 2: The running time and the corresponding standard deviation on the Extended Yale B, AR and CMU PIE databases.

| Methods | Experiments | | |
|---|---|---|---|
| | Extended Yale B | AR | CMU PIE |
| 2DPCA | 0.01±0.00 | 0.04±0.00 | 0.01±0.00 |
| 2DPCA-L1 | 7.07±0.7 | 11.30±1.37 | 9.20±1.27 |
| L1-2DPCA non-greedy | 7.05±0.19 | 14.71±0.10 | 7.97±0.09 |
| N-2DPCA | 12.69±1.38 | 24.36±5.16 | 15.41±3.15 |
| 2DPCAL1-S | 6.66±0.48 | 11.73±0.79 | 8.33±0.71 |
| F-2DPCA | **3.35±0.08** | **7.20±1.32** | **4.70±0.21** |

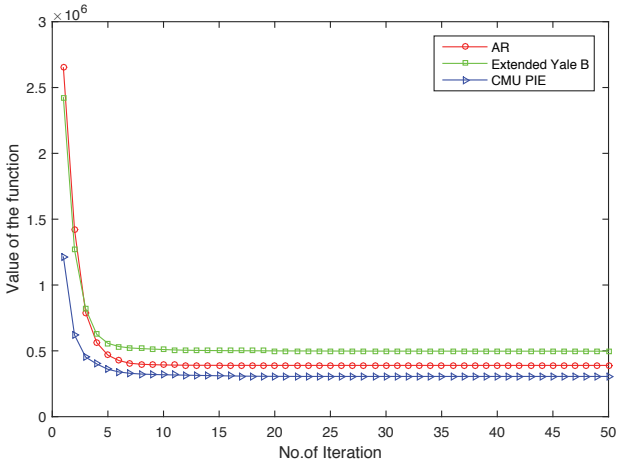



Figure 5: Convergence curve of our method on three databases.

## Conclusions

We present a robust unsupervised dimensionality reduction method, namely F-2DPCA. F-2DPCA uses *F-norm* instead of *squared F-norm* as distance metric to measure the reconstruction error in the criterion function. Compared with $\ell_1$-norm, F-norm of matrix not only has rotational invariance but also retains 2DPCA's desirable properties such as rotational invariance. Moreover, our method explicitly takes into account the reconstruction error while $\ell_1$-norm based 2DPCA technique does not. To solve F-2DPCA, we present a fast iterative algorithm, which has a closed-form solution in each iteration and convergence. Experimental results on several face image databases illustrate the effectiveness and advantages of our proposed method.

## Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61271296, China Postdoctoral Science Foundation (Grant 2012M521747), the 111 Project of China (B08038), and Fundamental Research Funds for the Central Universities of China under Grant BDY21.

## References

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

Chen, X.; Yang, J.; and Jin, Z. 2014. An improved linear discriminant analysis with l1-norm for robust feature extraction. In *International Conference on Pattern Recognition*, 1585–1590.

Collins, M.; Dasgupta, S.; and Schapire, R. E. 2001. A generalization of principal components analysis to the exponential family. In *Proceedings of Advances in Neural Information Processing Systems*, 617–624.

Ding, C.; Zhou, D.; He, X.; and Zha, H. 2006. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, 281–288. ACM.

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

Georghiades, A. S.; Belhumeur, P. N.; and Kriegman, D. J. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(6):643–660.

He, X., and Niyogi, P. 2005. Locality preserving projections. In *Proceedings of Advances in Neural Information Processing Systems*, 186–197.

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

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

Kwak, N. 2008. Principal component analysis based on l1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(9):1672–1680.

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

Li, X.; Pang, Y.; and Yuan, Y. 2010. L1-norm-based 2dpca. *IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics* 40(4):1170–1175.

Lu, H.; Plataniotis, K. N.; and Venetsanopoulos, A. N. 2008. Mpca: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks* 19(1):18–39.

Martinez, A. M. 1998. The ar face database. *CVC Technical Report* 24.

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

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

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

Sim, T.; Baker, S.; and Bsat, M. 2002. The cmu pose, illumination, and expression (pie) database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 46 – 51.

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

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

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

Yan, S.; Xu, D.; Zhang, B.; and Zhang, H.-J. 2005. Graph embedding: a general framework for dimensionality reduction. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 830–837.

Yang, J.; Zhang, D.; Frangi, A. F.; and Yang, J. Y. 2004. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(1):131–7.

Yang, J.; Zhang, D.; Yong, X.; and Yang, J. Y. 2005. Two-dimensional discriminant transform for face recognition. *Pattern Recognition* 38(7):1125–1129.

Zhang, F.; Yang, J.; Qian, J.; and Xu, Y. 2015. Nuclear norm-based 2-dpca for extracting features from images. *IEEE Transactions on Neural Networks and Learning Systems* 26(10):2247–2260.

Zhong, F., and Zhang, J. 2013. Linear discriminant analysis based on l1-norm maximization. *IEEE Transactions on Image Processing* 22(8):3018–27.