Angle Principal Component Analysis

Qianqian Wang **Quanxue Gao*** **Xinbo Gao** **Feiping Nie**
State Key Lab. of Integrated Services Networks, Xidian University
610887187@qq.com qxgao@xidian.edu.cn xd_gxb_pr@163.com 1436883741@qq.com

Abstract

Recently, many ℓ_1-norm based PCA methods have been developed for dimensionality reduction, but they do not explicitly consider the reconstruction error. Moreover, they do not take into account the relationship between reconstruction error and variance of projected data. This reduces the robustness of algorithms. To handle this problem, a novel formulation for PCA, namely angle PCA, is proposed. Angle PCA employs ℓ_2-norm to measure reconstruction error and variance of projected data and maximizes the summation of ratio between variance and reconstruction error of each data. Angle PCA not only is robust to outliers but also retains PCA's desirable property such as rotational invariance. To solve Angle PCA, we propose an iterative algorithm, which has closed-form solution in each iteration. Extensive experiments on several face image databases illustrate that our method is overall superior to the other robust PCA algorithms, such as PCA, PCA-L1 greedy, PCA-L1 nongreedy and HQ-PCA.

1 Introduction

Automated learning of low-dimensional linear models from training data has become a standard paradigm in pattern recognition and machine learning community. Principal component analysis (PCA) [Turk and Pentland, 1991], linear discriminant analysis (LDA) [Belhumeur *et al.*, 1997], locality preserving projection (LPP) [He and Niyogi, 2005] and neighborhood preserving embedding (NPE) [He *et al.*, 2005] are four of the most representative linear models for different tasks. Unlike the other three methods, PCA is also usually employed as a pre-processing step in many models, including LDA, LPP, and NPE. The derived features in PCA are linear combinations of the original features, where the coefficients are from the projection matrix. The optimal projection matrix in PCA is composed of the eigenvectors of the covariance matrix. It has been applied successfully in many applications including face representation and recognition and gene expression data analysis.

Applying PCA to image representation and recognition, we need to transform each image into 1D image vector by concatenating all rows/ columns of image. So, these methods cannot well exploit the spatial structure information that is embedded in pixels [Zhang *et al.*, 2015]. To handle this problem, many two-dimensional subspace learning methods or tensor methods have been developed. The representative two-dimensional methods include two-dimensional PCA (2DPCA) [Yang *et al.*, 2004] and multi-linear PCA [Lu *et al.*, 2008]. These methods minimize the summation of the reconstruction errors of data in the least-squares sense. It is commonly known that least-squares techniques are not robust to outliers.

Recently, ℓ_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 and machine learning. For example, Ke and Kanade [Ke and Kanade, 2005] proposed L1-PCA which uses ℓ_1-norm to measure reconstruction error in the objective function. Kwak [Kwak, 2008] used ℓ_1-norm to measure variance and proposed PCA-L1 with greedy algorithm. Nie *et al.* [Nie *et al.*, 2011] proposed a non-greedy iterative to solve PCA-L1. To well exploit the spatial structure embedded in image, Li *et al.* [Li *et al.*, 2010] and Wang *et al.* [Wang *et al.*, 2015] respectively extended PCA-L1 to 2DPCA-L1. Ju *et al.* [Ju *et al.*, 2015] proposed ℓ_1-norm based 2D probabilistic PCA for dimensionality reduction.

However, ℓ_1-norm is not rotational invariant [Ding *et al.*, 2006], which has been emphasized in the context of learning algorithms [Ng, 2004]. Since data geometric distribution remains unchanged under a rotational transformation of the sample space, the features, which are extracted by criterion function of subspace learning technique, should remain unchanged. It helps avoid performance degradation of subspace learning technique. Based on this content, Half-Quadratic PCA (HQ-PCA) is introduced in [He *et al.*, 2011], which uses the maximum correntropy criterion to extract features. Ding *et al.* [Ding *et al.*, 2006] proposed the rotational invariant ℓ_1-norm for feature extraction and developed R_1-PCA. Wang and Gao [Wang and Gao, 2017] employed F-norm as distance metric and proposed robust 2DPCA with F-norm minimization. These robust PCA algorithms implicitly consider that all the reconstruction errors or variance of data are equally important. This reduces the flexibility of algorithms.

[*] Corresponding author: Q. Gao

In this paper, we propose a novel formula for PCA, namely angle PCA. Angle PCA employs ℓ_2-norm to measure reconstruction error and variance and maximizes the summation of ratio between the variance and reconstruction error of each data. Moreover, Angle PCA assigns a small weight to large reconstruction error. Thus, our proposed method not only is robust to outliers but also well characterizes the relationship between reconstruction error and variance. To solve Angle PCA, we propose a fast iterative algorithm, which has closed-form solution in each iteration. Experiments on several face databases show the effectiveness of our proposed algorithm.

2 PCA and PCA-L1

2.1 PCA

Assume that we have a set of N sample images $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_N\}$, where $\boldsymbol{x}_i \in \mathbf{R}^m$ denotes the ith training image. Without loss of generality, we assume the data set are centralized, i.e.,$\sum_{i=1}^{N} \boldsymbol{x}_i = 0$. PCA aims to seek projection matrix $\mathbf{W} \in \mathbf{R}^{m \times p}$ by solving the following objective function [Turk and Pentland, 1991].

$$\underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\text{argmin}} \sum_{i=1}^{N} \left\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\right\|_2^2 \tag{1}$$

where $\|\cdot\|_2^2$ denotes the squared ℓ_2-norm, $\mathbf{I}_p$ denotes a p dimensional identity matrix. It is easy to see that, the model (1) is equivalent to the model (2) due to the fact $\sum_{i=1}^{N} \left\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\right\|_2^2 + \sum_{i=1}^{N} \left\|\mathbf{W}^T\boldsymbol{x}_i\right\|_2^2 = \sum_{i=1}^{N} \|\boldsymbol{x}_i\|_2^2$.

$$\underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\text{argmax}} \; tr\left(\mathbf{W}^T\mathbf{S}_t\mathbf{W}\right) = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\text{argmax}} \sum_{i=1}^{N} \left\|\mathbf{W}^T\boldsymbol{x}_i\right\|_2^2 \tag{2}$$

where $\mathbf{S}_t = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T$ is the covariance matrix of data, $tr(\cdot)$ is the trace operator of a matrix,

Solution of Eq. (1) or Eq. (2) is composed of the eigenvectors of $\mathbf{S}_t$ corresponding to the first p largest eigenvalues. As can be seen in the objective functions (1) and (2), PCA employs squared ℓ_2-norm, which is sensitive to outliers, as distance metric in the criterion function. Thus, solution of the objective function (1) is not robust in the sense that outlying measurements skew the solution from the desired solution.

2.2 PCA-L1

To improve robustness of PCA, ℓ_1-norm based PCA technique was proposed by using ℓ_1-norm as distance metric in the criterion function. It has the following two different formulations [Ke and Kanade, 2005; Kwak, 2008].

(a) Reconstruction error based approach. This technique calculates the reconstruction error by using ℓ_1-norm as the distance metric and obtains the principal directions by

$$\underset{\mathbf{W}^T\mathbf{W}=\mathbf{I_p}}{\text{argmin}} \sum_{i=1}^{N} \left\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\right\|_1 \tag{3}$$

(b) Covariance based approach. This technique employs ℓ_1-norm to measure the variance and the principal directions are obtained as

$$\underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\text{argmax}} \left\|\mathbf{W}^T\mathbf{X}\right\|_{L_1} \tag{4}$$

where $\|\cdot\|_1$ denotes 1-norm of a vector, $\|\cdot\|_{L_1}$ denotes ℓ_1-norm of a matrix, which can be defined as follows.

$$\|\mathbf{Q}\|_{L_1} = \sum_{i=1}^{m} \sum_{j=1}^{n} |\mathbf{Q}(i,j)| \tag{5}$$

where $\mathbf{Q}(i,j)$ denotes the element of the ith row jth column of matrix $\mathbf{Q}$.

Compared with traditional PCA, the objective functions (3) and (4) are robust to outliers, but they do not have rotational invariance [Ding *et al.*, 2006], while traditional PCA has this nice property. Moreover, the objective function (3) is not equivalent to the objective function (4) due to the fact $\sum_{i=1}^{N} \left\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\right\|_1 + \sum_{i=1}^{N} \left\|\mathbf{W}^T\boldsymbol{x}_i\right\|_1 \neq \sum_{i=1}^{N} \|\boldsymbol{x}_i\|_1$. It illustrates that solution of Eq. (4) does not minimize the reconstruction error of data, which is true goal of PCA. However, it is difficult to solve Eq. (3). Thus, many algorithms have been proposed to solve the model (4). Finally, both Eq. (3) and Eq. (4) do not take into account the relationship between reconstruction error and variance of projected data.

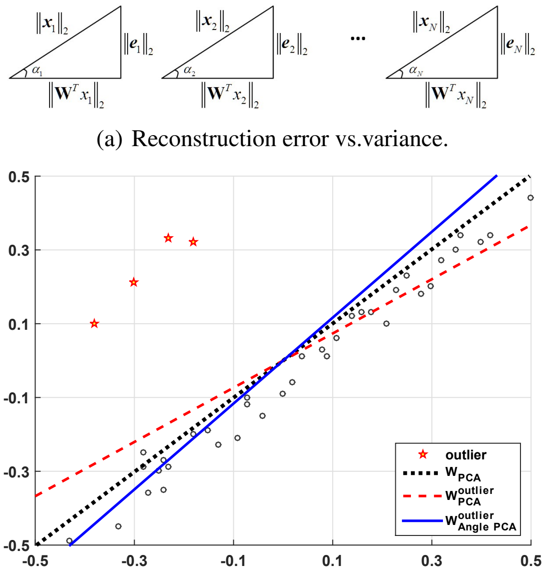
3 Angle PCA

3.1 Motivation and Objective Function

The objective function (1) shows that PCA minimizes the summation of each squared reconstruction error $\|\boldsymbol{e}_i\|_2^2$, where $\boldsymbol{e}_i = \boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i$, and implicitly considers that the reconstruction errors of all data points are equally important. Thus, the squared large reconstruction errors, i.e. distances will dominate the objective function (1). In the presence of outliers, which deviate significantly from the rest data, the real projection solution will remarkably deviate from the ideal solution because the outliers have large reconstruction errors under the ideal projection vectors. This results in sensitivity of PCA to outliers. Figure1(b) shows some clean data points and outliers, which are constructed by Matlab function, and plots the optimal projection vector of PCA and optimal projection vector of our method, where $\mathbf{W}_{\text{PCA}}$ refers to the projection vector of PCA without outliers, $\mathbf{W}_{\text{PCA}}^{\text{outlier}}$ and $\mathbf{W}_{\text{AnglePCA}}^{\text{outlier}}$ respectively refer to the projection vector of PCA and our method when the training data include outliers. Figure1(b) illustrates that, $\mathbf{W}_{\text{PCA}}^{\text{outlier}}$ remarkably deviates from the direction of $\mathbf{W}_{\text{PCA}}$, while our method is close to the ideal directions. Thus, to achieve robustness, the contribution of distance metric to the criterion function (1) should reduce the effect of outliers. Moreover, we hope the low-dimensional representations are not uniquely determined up to an orthogonal transformation. Finally, it is easy to see that, compared with squared Euclidean distance, ℓ_2-norm not only can weaken the effect of large distance but also has rotational invariance, however, in ℓ_2-norm, we have $\sum_{i=1}^{N} \left\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\right\|_2 + \sum_{i=1}^{N} \left\|\mathbf{W}^T\boldsymbol{x}_i\right\|_2 \neq \sum_{i=1}^{N} \|\boldsymbol{x}_i\|_2$.

Combining the aforementioned analysis, we propose a novel formulation for PCA. It not only integrates the relationship between variance and reconstruction error in the objective function but also is robustness to outliers. Specifically, we aim to seek the projection matrix $\mathbf{W}$ by the model (6).

$$\underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\arg\max} \sum_{i=1}^{N} \mathbf{F}_i(\mathbf{W}) \tag{6}$$



(a) Reconstruction error vs.variance.

(b) Projection vectors of PCA and Angle PCA on artificial dataset with/without outliers.

Figure 1: Diagram of Angle PCA

where $\mathbf{F}_i(\mathbf{W}) = \dfrac{\mathbf{H}_i(\mathbf{W})}{\mathbf{M}_i(\mathbf{W})}$, $\mathbf{H}_i(\mathbf{W}) = \|\mathbf{W}^T\boldsymbol{x}_i\|_2$, $\mathbf{M}_i(\mathbf{W}) = \|\boldsymbol{e}_i\|_2$.

It is easy to see that $\mathbf{F}_i(\mathbf{W})$ is the cotangent value of the angle between reconstruction error and variance of the ith data (See Figure 1(a)), i.e.,$\mathbf{F}_i(\mathbf{W}) = \cot \alpha_i$. Thus, the objective function (6) is called angle PCA (Angle PCA).

3.2 Algorithm

Before solving the objective function (6), we first introduce the following theorem [Guo *et al.*, 2003; Wang *et al.*, 2007].
Theorem 1. Suppose $\mathbf{A}_w$ and $\mathbf{B}_w$ are related to $\mathbf{W}$, then the optimal solution of Eq. (7)

$$\underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\arg\max} \frac{\|\mathbf{A}_w\|_2}{\|\mathbf{B}_w\|_2} \tag{7}$$

can be obtained by iteratively solving the model (8).

$$\underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p,\lambda_w}{\arg\max} \quad \|\mathbf{A}_w\|_2 - \lambda_w\|\mathbf{B}_w\|_2 \tag{8}$$

where λ_w relates to $\mathbf{W}$.

Now, we consider how to solve the objective function (6). Motivated by Theorem 1, the optimal solution of the objective function (6) can be obtained by iteratively solving the objective function (9)

$$G(\mathbf{W},\lambda_i) = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\arg\max} \sum\nolimits_{i=1}^{N} \{\mathbf{H}_i(\mathbf{W}) - \lambda_i\mathbf{M}_i(\mathbf{W})\} \tag{9}$$

where λ_i relates to $\mathbf{H}_i(\mathbf{W})$ and $\mathbf{M}_i(\mathbf{W})$.

By simple algebra, we have

$$\begin{aligned} &\sum\nolimits_{i=1}^{N} \{\mathbf{H}_i(\mathbf{W}) - \lambda_i\mathbf{M}_i(\mathbf{W})\} \\ &\quad = \sum\nolimits_{i=1}^{N} \left\{ \frac{\left\|\mathbf{W}^T\boldsymbol{x}_i\right\|_2^2}{\|\mathbf{W}^T\boldsymbol{x}_i\|_2} - \lambda_i \frac{\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\|_2^2}{\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\|_2} \right\} \\ &\quad = tr(\mathbf{W}^T\mathbf{Z}\mathbf{W}) - tr(\mathbf{X}\mathbf{D}\mathbf{X}^T) \end{aligned} \tag{10}$$

where $\mathbf{Z} = \sum_{i=1}^{N} \boldsymbol{x}_i(\dfrac{1}{\|\mathbf{W}^T\boldsymbol{x}_i\|_2} + \dfrac{\lambda_i}{\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\|_2})\boldsymbol{x}_i^T$, and $\mathbf{D}$ is a diagonal matrix whose diagonal elements are $d_i = \dfrac{\lambda_i}{\|\boldsymbol{x}_i - \mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\|_2}$.

```
Algorithm 1: Angle PCA
Input: $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_N\} \in \mathbf{R}^{m\times N}$, $p$ .
Initialize: $\mathbf{W}^1 \in \mathbf{R}^{m\times p}$ which satisfies $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, $k = 1$, $\varepsilon = 10^{-1}$, $J(\mathbf{W}^0) = 1$, and $J(\mathbf{W}) = \sum_{i=1}^{N} \mathbf{F}_i(\mathbf{W})$.
while $J(\mathbf{W}^k) - J(\mathbf{W}^{k-1}) \geq \varepsilon$ do
    1.For all training samples, calculate
      $\lambda_i^k = \dfrac{\left\|(\mathbf{W}^{k-1})^T\boldsymbol{x}_i\right\|_2}{\left\|\boldsymbol{x}_i - \mathbf{W}^{k-1}(\mathbf{W}^{k-1})^T\boldsymbol{x}_i\right\|_2}$.
    2.Calculate $\mathbf{Z}^k$ and $\mathbf{D}^k$by Eq. (10).
    3.Solve $\mathbf{W}^k$ by Eq.(12).
    4. If $J(\mathbf{W}^k) \geq J(\mathbf{W}^{k-1})$, go to step 6, else go to step 5.
    5. Find $\mathbf{W}^k$ that satisfies $J(\mathbf{W}^k) \geq J(\mathbf{W}^{k-1})$ by using sub-gradient method with Armigo line search [Liu et al., 2017]. If there is no solution, go to step 7, else go to step 6.
    6.Update $k \leftarrow k + 1$.
end while
7. Output $\mathbf{W}^k \in \mathbf{R}^{m\times p}$.
```

Combining Eq. (10) and Eq. (9), the model (6) finally becomes

$$G(\mathbf{W},\mathbf{Z},\mathbf{D}) = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\arg\max} \left(tr(\mathbf{W}^T\mathbf{Z}\mathbf{W}) - tr(\mathbf{X}\mathbf{D}\mathbf{X}^T)\right) \tag{11}$$

In the objective function (11), we have three unknown variables $\mathbf{W}$, $\mathbf{Z}$ and $\mathbf{D}$, where $\mathbf{Z}$ and $\mathbf{D}$ relate to $\mathbf{W}$. Thus, it has no closed-form solution and is difficult to directly solve the solution of the model (11). An algorithm can be developed for alternatively updating $\mathbf{W}$ (while fixing $\mathbf{Z}$ and $\mathbf{D}$), $\mathbf{Z}$ (while fixing $\mathbf{W}$ and $\mathbf{D}$) and $\mathbf{D}$ (while fixing $\mathbf{W}$ and $\mathbf{Z}$). To be specific, in the kth iteration, when $\mathbf{Z}$ and $\mathbf{D}$ are known, we can update $\mathbf{W}$ by maximizing the objective function (11). In this case, the second term in the objective function (11) becomes constant. In other words, we update $\mathbf{W}$ by solving the objective function (12).

$$\mathbf{W}^* = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}_p}{\arg\max}\ tr\left(\mathbf{W}^T\mathbf{Z}\mathbf{W}\right) \tag{12}$$

According to the matrix theory, solution $\mathbf{W}^*$ of the objective function (12) are composed of the eigenvectors of $\mathbf{Z}$ corresponding to the p largest eigenvalues. After that, we calculate $\mathbf{Z}$ and $\mathbf{D}$ by the updated $\mathbf{W}$. *Algorithm* 1 lists the pseudo code of solving the objective function (9), i.e., Angle PCA.

3.3 Convergence Analysis

Theorem 2. *Algorithm* 1 will converge to a stationary point of the objective function (6).

Proof: The Lagrangian function of the problem (9) is

$$L=\sum\nolimits_{i=1}^{N}\left\{\mathbf{H}_i(\mathbf{W})-\lambda_i\mathbf{M}_i(\mathbf{W})\right\}-tr\left(\Lambda\left(\mathbf{W}^T\mathbf{W}-\mathbf{I}_p\right)\right) \tag{13}$$

where the Lagrangian multiplies Λ is a diagonal matrix for enforcing the orthonormal constrains $\mathbf{W}^T\mathbf{W}=\mathbf{I}_p$. In the kth iteration, when λ_i $(i=1,\cdots,N)$ are known, the KKT condition of the problem (9) specifies that the gradient of L must be zero, i.e.,

$$\frac{\partial L}{\partial \mathbf{W}}=\mathbf{ZW}-\mathbf{W}\Lambda=0 \tag{14}$$

By simple algebra, we have

$$\mathbf{ZW}=\mathbf{W}\Lambda \tag{15}$$

According to step 3 in *Algorithm* 1, we find the optimal solution of the objective function (12). Thus the converged solution of *Algorithm* 1 satisfies the KKT condition of the problem (12). The Lagrangian function of Eq. (12) is

$$L_2(\mathbf{W})=tr(\mathbf{W}^T\mathbf{ZW})-tr(\mathbf{\Lambda}^T(\mathbf{W}^T\mathbf{W}-\mathbf{I})) \tag{16}$$

Taking the derivative w.r.t. $\mathbf{W}$ and setting it to zero, we get the KKT condition of (12) as follows.

$$\mathbf{ZW}-\mathbf{W\Lambda}=0 \tag{17}$$

Note that, in (17), matrix $\mathbf{Z}$ relates to $\mathbf{W}^{k-1}$. Suppose we obtain the local solution $\mathbf{W}^*$ in the kth, thus, we have $\mathbf{W}^*=\mathbf{W}^k=\mathbf{W}^{k-1}$. In this case, Eq. (17) is just the same as Eq. (15). It means that the converged solution of *Algorithm* 1 satisfies the KKT condition of Eq. (9), i.e.,

$$\left.\frac{\partial L}{\partial \mathbf{W}}\right|_{\mathbf{W}=\mathbf{W}^*}=0 \tag{18}$$

Eq. (18) illustrates that the converged solution of *Algorithm* 1 is at least a stationary point of the model (9). Moreover, according to Theorem 1, we have that solution of the objective function (6) can be approximately obtained by solving Eq. (9). Thus, the converged solution of *Algorithm* 1 is also a stationary point of the model (6), i.e., Angle PCA. ■

Step 4 and step 5 in *Algorithm* 1 illustrate that the objective function value of angle PCA is non-decreasing in each iteration. Combining the theorems 1 and 2, we have that *Algorithm* 1 may converge to a local solution of Angle PCA in some cases.

3.4 Rotational Invariance

We show that our proposed approach has nice rotational invariance. Rotational invariance means that the low-dimensional representations remain unchanged under a rotational transformation of the sample space. Given an arbitrary rotation matrix $\mathbf{\Gamma}$ ($\mathbf{\Gamma}^T\mathbf{\Gamma}=\mathbf{I}$). Then, the rotational transformation of $\mathbf{W}$ and $\boldsymbol{x}_i$ is defined as follows:

$$\tilde{\mathbf{W}}=\mathbf{\Gamma W},\qquad \tilde{\boldsymbol{x}}_i=\mathbf{\Gamma}\boldsymbol{x}_i \tag{19}$$

For each term in the objective function (6), we have

$$\begin{aligned}\frac{\left\|\mathbf{W}^T\boldsymbol{x}_i\right\|_2}{\left\|\boldsymbol{x}_i-\mathbf{W}\mathbf{W}^T\boldsymbol{x}_i\right\|_2}&=\frac{\left\|\boldsymbol{u}_i\right\|_2}{\left\|(\mathbf{\Gamma}^T\mathbf{\Gamma})(\boldsymbol{x}_i-\mathbf{W}\boldsymbol{u}_i)\right\|_2}\\&=\frac{\left\|\boldsymbol{u}_i\right\|_2}{\left\|\mathbf{\Gamma}\boldsymbol{x}_i-\mathbf{\Gamma W}\boldsymbol{u}_i\right\|_2}\\&=\frac{\left\|\boldsymbol{u}_i\right\|_2}{\left\|\tilde{\boldsymbol{x}}_i-\tilde{\mathbf{W}}\boldsymbol{u}_i\right\|_2}\end{aligned} \tag{20}$$

where $\boldsymbol{u}_i=\mathbf{W}^T\boldsymbol{x}_i$ is the low-dimensional representation of $\boldsymbol{x}_i$. The above equation illustrates that, if $\mathbf{W}$ is the solution of the objective function of Angle PCA, then $\tilde{\mathbf{W}}$ is the projection matrix of Angle PCA under a rotational transformation $\mathbf{\Gamma}$. Thus, we have

$$\tilde{\mathbf{W}}^T\tilde{\boldsymbol{x}}_i=\mathbf{W}^T\mathbf{\Gamma}^T\mathbf{\Gamma}\boldsymbol{x}_i=\mathbf{W}^T\boldsymbol{x}_i=\boldsymbol{u}_i \tag{21}$$

Hence, the low-dimensional representation of $\boldsymbol{x}_i$, remains unchanged in the rotational transformation.

4 Experimental Results

We validate our approach on three face databases (Extended Yale B, CMU PIE and AR) and compared it with traditional PCA [Turk and Pentland, 1991] and recently proposed robust PCA methods including PCA-L1 greedy [Kwak, 2008], PCA-L1 non-greedy [Nie *et al.*, 2011], and HQ-PCA [He *et al.*, 2011]. In our experiments, we use 1-nearest neighbor (1NN) for classification and set the number of projection vectors from 10 to 200. We estimate the performance by both recognition accuracy which is obtained by 1NN classifier and reconstruction error. In our experiments, we calculate the reconstruction error of clean data by

$$\textit{\textbf{error}}=\frac{1}{N}\sum\nolimits_{i=1}^{N}\left\|\boldsymbol{x}_i^{\,clean}-\mathbf{W}\mathbf{W}^T\boldsymbol{x}_i^{\,clean}\right\|_2 \tag{22}$$

where N is the number of training data, $\mathbf{W}\in R^{m\times p}$ is the projection matrix. $\boldsymbol{x}_i^{\,clean}$ is the i-th clean training sample.

4.1 Experiments on The Extended Yale B Database

The Extended Yale B database [Georghiades *et al.*, 2001] consists of 2144 frontal-face pictures of 38 individuals with different illuminations. There are 64 pictures for each person except 60 for the 11th and 13th, 59 for the 12th, 62 for the 15th and 63 for the 14th, 16th and 17th. In the experiments, each image was normalized to 32×32 pixels [He and Niyogi, 2005]. 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. PCA, PCA-L1 greedy, PCA-L1 nongreedy, HQ-PCA and our approach Angle PCA are used to extract features, respectively. We repeat this process 10 times.

Table 1 lists the average reconstruction error, recognition accuracy and the corresponding standard deviation of each

Table 1: The average reconstruction error and recognition accuracy of each method on the Extended Yale B database.

Methods	Error	Accuracy
PCA	323.34±2.76	55.98±0.55
PCA-L1 greedy	277.83±2.26	57.90±0.43
PCA-L1 nongreedy	337.92±3.08	55.22±0.5
HQ-PCA	310.57±2.76	57.05±0.49
Angle PCA	**216.99±1.62**	**59.12±0.35**

Table 2: The average reconstruction error and recognition accuracy of each method on the CMU PIE database.

Methods	Error	Accuracy
PCA	204.94±2.17	84.17±0.78
PCA-L1 greedy	143.16±0.88	85.34±0.7
PCA-L1 nongreedy	205.39±2.00	84.10±0.75
HQ-PCA	185.05±1.34	84.62±0.75
Angle PCA	**62.72±0.98**	**86.57±0.67**

method on the Extended Yale B database. Figure 3 (a),(b) respectively plots the reconstruction error and recognition accuracy *vs.* number of projection vectors on the Extended Yale B database. Table 1 and Figure 3 (a),(b) show that our method Angle PCA obviously outperforms the other four methods both in terms of reconstruction error and recognition accuracy. The reason is that Angle PCA employs ℓ_2-norm as distance metric to measure reconstruction error and variance, which is robust to outliers and has rotational invariance, in the criterion function. Moreover, Angle PCA explicitly considers the relationship between reconstruction error and variance of projected data, which is important for data representation and classification.

4.2 Experiments on The CMU PIE Database

The CMU PIE database [Sim *et al.*, 2002] consists of 2856 frontal-face images of 68 individuals with different illuminations. In the experiments, each image was normalized to 32×32 pixels, we randomly selected 10 images and added the same noise as that in the Extended Yale B database. We randomly selected 21 images, which include 16 noise-free images, per person for training and the remaining images for testing. This process is repeated 10 times.

Table 2 lists the average reconstruction error, recognition accuracy and the corresponding standard deviation of each method on the CMU PIE databases. Figure 3 (c),(d) plots the curve of reconstruction error and recognition accuracy vs. dimension on the CMU PIE databases, respectively. Table 3 and Figure 3 (c),(d) show that our proposed method is superior to the other methods under the reconstruction error and recognition accuracy. Traditional PCA is inferior to PCA-L1 and HQ-PCA. These conclusions are consistent with that in the Extended Yale B database. The main reason may be that large distance dominates the criterion function in PCA, and the variations between images, which have the same class label with different illumination, is larger than the change of face identity.

4.3 Experiments on The AR Database

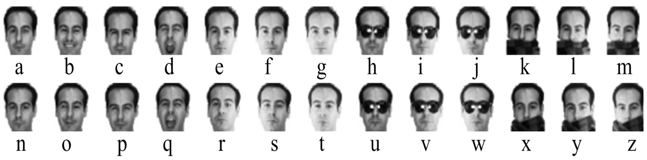
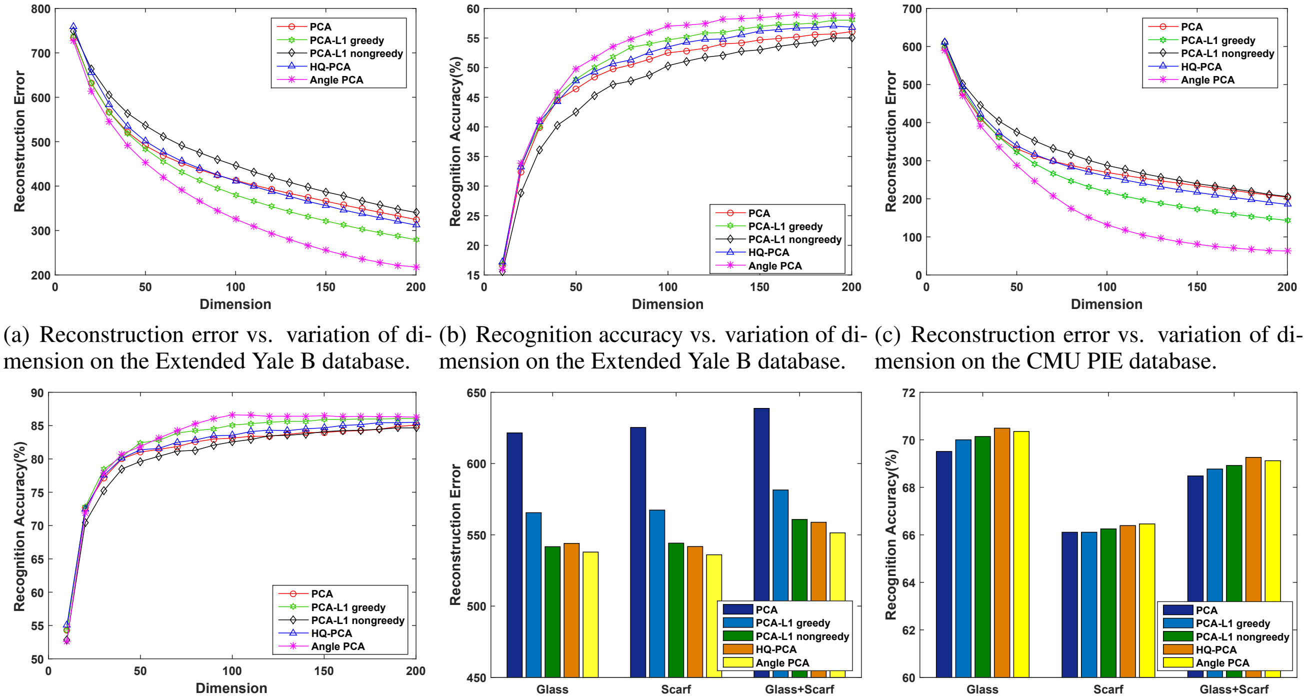


Figure 2: Some samples of one person in the AR database

The AR database [Martinez, 1998] contains over 4000 color face image of 126 people (70 men and 56women), including frontal views of faces with different facial expressions, lighting conditions and occlusions such as glasses and scarfs. The pictures of 120 individuals were taken in two sessions (separated by two weeks). 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 portion of the image and then normalized it to 50×40 pixels. Figure 2 shows some normalized images of one person, where images (a) to (m) are from session 1, and the other images are from session 2.

In this database, we do the following three group experiments. In the first group experiment, we select 9 images, which include images (a)-(h) and image k from the first session for training, and the remaining images for testing. In the second group experiment, we select images from (a) to (h) per person for training, and 12 images, which include images from (i) to (j) and images (n) to (w), per person for testing. In the third group experiment, we select images from (a) to (g) and image (k) per person for training, the remaining images, which do not include images with glasses, for testing. Figure 3 (e),(f) shows the reconstruction error and recognition accuracy of each method. Figure 3 (e) shows that our approach Angle PCA has the minimum reconstruction error and is remarkably superior to the other robust PCA methods. PCA is remarkably inferior to the other four approaches. This is because PCA characterizes the similarity between data by squared ℓ_2-norm, which is very sensitive to outliers, while ℓ_1-norm based methods and our approach are robust to outliers. Figure 3 (f) illustrates that our approach Angle PCA is superior to PCA-L1 greedy, PCA-L1 non-greedy and PCA for image classification. The reason is that our method not only is robust to outliers but also has rotational invariance. Moreover, non-greedy PCA-L1 greedy and PCA-L1 non-greedy do not minimize the reconstruction error, which is real goal of PCA. As can be seen in Figure 3 (f) , Angle PCA is overall inferior to HQ-PCA for image classification. The reason may be that Angle PCA employs the fixed mean value. Moreover, it is commonly known that PCA extracts the most expressive features rather discriminant features, thus, it is reconstruction error that has been widely used to evaluate performance of PCA.

Table 3 shows the average running time and the corresponding standard deviation of each method on three databas-



(a) Reconstruction error vs. variation of dimension on the Extended Yale B database.

(b) Recognition accuracy vs. variation of dimension on the Extended Yale B database.

(c) Reconstruction error vs. variation of dimension on the CMU PIE database.

(d) Recognition accuracy vs. variation of dimension on the CMU PIE database.

(e) The reconstruction error under three experiments on the AR database.

(f) Top recognition accuracy under three experiments on the AR database.

Figure 3: The reconstruction error and recognition accuracy results on three databases.

Table 3: The average running time and the corresponding standard deviation of each method on three databases.

Methods	Extended Yale B	AR	CMU PIE
PCA	0.01±0.14	4.10±0.05	0.16±0.02
PCA-L1 greedy	23.33±0.51	103.16±1.33	29.44±0.62
PCA-L1 nongreedy	2.72±0.06	7.12±0.27	2.64±0.06
HQ-PCA	3.02±0.10	5.30±0.09	3.43±0.13
Angle PCA	**5.65±0.09**	**27.81±0.81**	**5.79±0.16**

es. To be fair, we set the number of projection vectors 400, and set the number of iteration 20. From Table 3, we can see the running time of our method is similar as PCA-L1 nongreedy, and HQ-PCA. Besides our method is faster than PCA-L1 greedy. Figure 4 shows the convergence curve of our proposed method vs. number of iteration on three databases. It can be seen that our method can monotonically increase the value of the objective function (6) in each iteration. Moreover, in each iteration, our proposed algorithm has a closed-form solution. Thus, these experiments illustrate that our proposed algorithm overall has a local solution.

5 Conclusions

We propose a novel formula for PCA, called Angle PCA, for data representation and classification. Angle PCA employs ℓ_2-norm as distance metric to measure reconstruction error and variance and aims to seek projection matrix that maximizes the summation of the ratio between variance and reconstruction error of each data. It explicitly takes into account the relationship between reconstruction error and variance. To solve Angle PCA, we propose an iterative algorithm, which is fast and has closed-form solution in each iteration. Experiments on the several databases illustrate that Angle PCA is superior to the other methods.

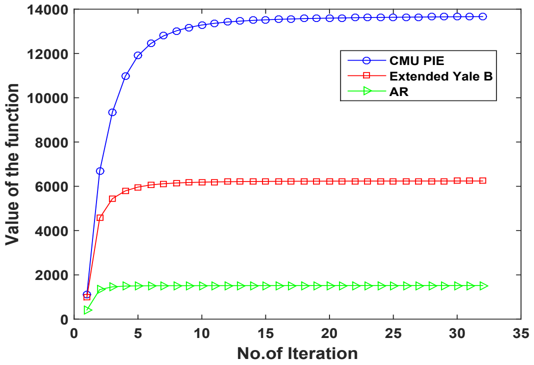


Figure 4: Convergence curve of our method.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61271296, Innovation Fund for Postgraduate's International Cooperation and Exchanges of Xidian University.

References

[Belhumeur *et al.*, 1997] Peter N. Belhumeur, Joao P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[Ding *et al.*, 2006] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization. In *International Conference on Machine Learning*, pages 281–288, 2006.

[Georghiades *et al.*, 2001] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. 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, 2001.

[Guo *et al.*, 2003] Yuefei Guo, Shijin Li, Jingyu Yang, Tingting Shu, and Lide Wu. A generalized foleycsammon transform based on generalized fisher discriminant criterion and its application to face recognition. *Pattern Recognition Letters*, 24(1C3):147–158, 2003.

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

[He *et al.*, 2005] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong Jiang Zhang. Neighborhood preserving embedding. In *The Tenth IEEE International Conference on Computer Vision*, pages 1208–1213, 2005.

[He *et al.*, 2011] Ran He, Baogang Hu, Weishi Zheng, and Xiangwei Kong. Robust principal component analysis based on maximum correntropy criterion. *IEEE Transactions on Image Processing*, 20(6):1485–1494, 2011.

[Ju *et al.*, 2015] Fujiao Ju, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin. Image outlier detection and feature extraction via l1-norm-based 2d probabilistic pca. *IEEE Transactions on Image Processing*, 24(12):4834–4846, 2015.

[Ke and Kanade, 2005] Qifa Ke and Takeo Kanade. 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, pages 739–746, 2005.

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

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

[Liu *et al.*, 2017] Yang Liu, Quanxue Gao, Xinbo Gao, Feiping Nie, and Yunsong Li. A non-greedy algorithm for l1-norm lda. *IEEE Transactions on Image Processing*, 26(2):684–695, 2017.

[Lu *et al.*, 2008] Haiping Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. Mpca: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, 19(1):18–39, 2008.

[Martinez, 1998] Aleix M Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[Ng, 2004] Andrew Y Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78, 2004.

[Nie *et al.*, 2011] Feiping Nie, Heng Huang, Chris H. Q. Ding, Dijun Luo, and Hua Wang. Robust principal component analysis with non-greedy ℓ1-norm maximization. In *Proceedings of the International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain*, pages 1433–1438, 2011.

[Sim *et al.*, 2002] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–51, 2002.

[Turk and Pentland, 1991] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[Wang and Gao, 2017] Qianqian Wang and Quanxue Gao. Two-dimensional pca with f-norm minimization. In *The Thirty-First AAAI Conference on Artificial Intelligence*, pages 180–186, 2017.

[Wang *et al.*, 2007] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

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

[Yang *et al.*, 2004] Jian Yang, David Zhang, Alejandro F. Frangi, and Jingyu Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131–137, 2004.

[Zhang *et al.*, 2015] Fanlong Zhang, Jian Yang, Jianjun Qian, and Yong Xu. Nuclear norm-based 2-dpca for extracting features from images. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2247–2260, 2015.