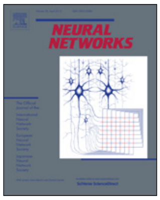
Neural Networks 46 (2013) 190–198

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



2DPCA with L1-norm for simultaneously robust and sparse modelling

Haixian Wang*, Jing Wang

Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, PR China

ARTICLE INFO

Article history:
Received 19 May 2012
Received in revised form 3 April 2013
Accepted 4 June 2013

Keywords:
Dimensionality reduction
2DPCA-L1
Robust modelling
Lasso regularization

ABSTRACT

Robust dimensionality reduction is an important issue in processing multivariate data. Two-dimensional principal component analysis based on L1-norm (2DPCA-L1) is a recently developed technique for robust dimensionality reduction in the image domain. The basis vectors of 2DPCA-L1, however, are still dense. It is beneficial to perform a sparse modelling for the image analysis. In this paper, we propose a new dimensionality reduction method, referred to as 2DPCA-L1 with sparsity (2DPCAL1-S), which effectively combines the robustness of 2DPCA-L1 and the sparsity-inducing lasso regularization. It is a sparse variant of 2DPCA-L1 for unsupervised learning. We elaborately design an iterative algorithm to compute the basis vectors of 2DPCAL1-S. The experiments on image data sets confirm the effectiveness of the proposed approach.

1. Introduction

Dimensionality reduction (DR) is of great importance for multivariate data analysis. For classifying typically high-dimensional patterns in practice, DR can relieve the "curse of dimensionality" effectively (Jain, Duin, & Mao, 2000). Principal component analysis (PCA) (Jolliffe, 1986) is perhaps the most popular DR technique. It seeks a few basis vectors such that the variances of projected samples are maximized. In the domain of image analysis, two-dimensional PCA (2DPCA) (Yang, Zhang, Frangi, & Yang, 2004) is more efficient, due to its direct formulation based on raw two-dimensional images.

Although PCA and 2DPCA have been widely applied in many fields, they are vulnerable at the presence of atypical samples because of the employment of the L2-norm in the variance formulation. As a robust alternative, L1-norm-based approaches were developed. Specifically, the L1-norm-based PCA variants include L1-PCA (Ke & Kanade, 2005), R1-PCA (Ding, Zhou, He, & Zha, 2006), PCA-L1 (Kwak, 2008), and non-greedy PCA-L1 (Nie, Huang, Ding, Luo, & Wang, 2011). Li, Pang, and Yuan (2009) developed the L1-norm-based 2DPCA (2DPCA-L1), which demonstrated encouraging performance for the image analysis.

A limitation of the above methods is that the basis vectors learned are still dense, which makes it difficult to explain the resulting features. It is desirable to select the most relevant or salient elements from a large number of features. To address this issue, sparse modelling has been developed and received increasing attention in the community of pattern classification (Wright et al., 2010). The sparsity was achieved by regularizing objective variables with a lasso penalty term using the L1-norm (Chen, Donoho, & Saunders, 1998; Tibshirani, 1996). Mathematically, the classic PCA approach could be reformulated as a regression-type optimization problem, and then the sparsity-inducing lasso penalty was imposed, resulting in sparse PCA (SPCA) (Zou, Hastie, & Tibshirani, 2006). The sparsity was further generalized to structured version, producing structured sparse PCA (Jenatton, Obozinski, & Bach, 2010). With the graph embedding platform (Yan et al., 2007), various DR approaches were endowed with a unified sparse framework by the L1-norm penalty (Cai, He, & Han, 2007; Wang, 2012; Zhou, Tao, & Wu, 2011). Recently, the robustness of SPCA was improved by the L1-norm maximization (Meng, Zhao, & Xu, 2012).

The sparse modelling for 2DPCA-L1, however, is still not addressed. Note that the L1-norm used in 2DPCA-L1 works as a robust measure of sample dispersion rather than regularizing basis vectors. A common way of enforcing sparsity is to fix the L2-norm and minimize the L1-norm with a length constraint.

In this paper, we limit our attention to the image analysis, and consider extending 2DPCA-L1 with sparsity, referred to as 2DPCAL1-S. On account of the L1-norm used as the lasso penalty in the sparsity-inducing modelling, we propose incorporating the L1-norm lasso penalty, together with the fixed L2-norm, onto the basis vectors of 2DPCA-L1. Consequently, 2DPCAL1-S maximizes the L1-dispersion of samples subject to the elastic net (i.e., L2-norm and L1-norm) (Zou et al., 2006) constraint onto the basis vectors. Formally, we combine the L1-dispersion and the elastic net constraint onto the objective function. As can be seen, we use the L1-norm for both robust and sparse modelling simultaneously. Due to the involvement of the L1-norm in the two aspects, the optimization of 2DPCAL1-S is not straightforward. We design an elegant iterative algorithm to solve 2DPCAL1-S.

* Corresponding author. Tel.: +86 25 83795664; fax: +86 25 83795929.
E-mail addresses: hxwang@seu.edu.cn, haixian_wang@hotmail.com (H. Wang).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.06.002

The remainder of this paper is organized as follows. The conventional 2DPCA-L1 method is briefly reviewed in Section 2. The formulation of 2DPCAL1-S is proposed in Section 3. Section 4 reports experimental results. And Section 5 concludes the paper.

2. Brief review of 2DPCA-L1

The 2DPCA-L1 approach, proposed by Li et al. (2009), finds basis vectors that maximize the dispersion of projected image samples in terms of the L1-norm. Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are a set of training images with size $q \times p$, where n is the number of images. These images are assumed to be mean-centred.

Let $\mathbf{v} \in \mathbb{R}^p$ be the first basis vector of 2DPCA-L1. It maximizes the L1-norm-based dispersion of projected samples

$$g(\mathbf{v}) = \sum_{i=1}^{n} \|\mathbf{X}_i \mathbf{v}\|_1 \tag{1}$$

subject to $\|\mathbf{v}\|_2 = 1$, where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the L1-norm and the L2-norm, respectively. In this paper, for a vector $\mathbf{z} = (z_1, \ldots, z_n)^T$, its Ld-norm is specified as $\|\mathbf{z}\|_d = \left(\sum_{i=1}^{n} |z_i|^d\right)^{1/d}$. Let $\mathbf{x}_{ji} \in \mathbb{R}^p$ be the jth row vector of $\mathbf{X}_i$, i.e.,

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{1i}^T \\ \vdots \\ \mathbf{x}_{qi}^T \end{bmatrix}. \tag{2}$$

Then $g(\mathbf{v})$ can be rewritten as

$$g(\mathbf{v}) = \sum_{i=1}^{n} \sum_{j=1}^{q} |\mathbf{v}^T \mathbf{x}_{ji}|. \tag{3}$$

The computation of $\mathbf{v}$ is implemented by an iterative algorithm as follows. Denote by t the iteration number. The basis vector $\mathbf{v}(t+1)$ at the $(t+1)$th-step is updated according to

$$\mathbf{v}(t+1) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{q} s_{ji}(t) \mathbf{x}_{ji}}{\left\| \sum_{i=1}^{n} \sum_{j=1}^{q} s_{ji}(t) \mathbf{x}_{ji} \right\|_2}, \tag{4}$$

where $s_{ji}(t)$ is defined as

$$s_{ji}(t) = \operatorname{sign}(\mathbf{v}^T(t) \mathbf{x}_{ji}) \tag{5}$$

for $j = 1, \ldots, q$; $i = 1, \ldots, n$, where $\operatorname{sign}(\cdot)$ is the sign function. This iterative procedure was theoretically shown to converge to a local maximum value of $g(\mathbf{v})$ (Li et al., 2009). The reminder basis vectors are computed likewise by using the deflated samples with previously obtained basis vectors.

3. 2DPCA-L1 with sparsity

3.1. Basic idea

Sparse modelling has been receiving exploding attention in computer vision and pattern classification (Wright et al., 2010). The obtained basis vectors of 2DPCA-L1, however, are still dense (Li et al., 2009). In other words, the projection procedure involves all the original features. As we know, a typical image usually has a large number of features. There may exist irrelevant or redundant features for classification. It is important to find a few salient features, which correspond to specific parts of the image such as eyes or mouth of a face image. To select a set of representative features, the projection vectors are expected to have very sparse elements with respect to such features. Such sparse projection vectors, if learned correctly, could encode semantic information and thus deliver valuable discriminative information. The sparse modelling has been successfully applied to many classification problems (Wright et al., 2010).

It is desirable to learn sparse basis vectors for the purpose of classification. In light of the advantage of the L1-norm penalty in the sparse modelling (Chen et al., 1998; Tibshirani, 1996), we propose regularizing the basis vectors of 2DPCA-L1 using the L1-norm penalty together with the fixed L2-norm. We refer to the proposed approach as 2DPCAL1-S. It results in sparse basis vectors. Note that the L1-norm used in 2DPCAL1-S takes effect in two different perspectives: measuring dispersion and regularizing basis vectors. Computationally, we elaborately design an iterative algorithm to implement 2DPCAL1-S.

3.2. Objective function

We impose the sparsity-inducing L1-norm penalty, as well as the fixed L2-norm, onto the basis vector $\mathbf{v}$. Specifically, we integrate the elastic net into the objective function. The elastic net generalizes the L1-norm lasso penalty by combining the ridge penalty and can circumvent potential limitations of the lasso (Zou et al., 2006). Consequently, we wish to select a vector $\mathbf{v}$ such that the objective function

$$h(\mathbf{v}) = \sum_{i=1}^{n} \sum_{j=1}^{q} |\mathbf{v}^T \mathbf{x}_{ji}| - \frac{\eta}{2} \|\mathbf{v}\|_2^2 - \gamma \|\mathbf{v}\|_1, \tag{6}$$

is maximized, where η and γ are positive tuning parameters which are usually selected by cross validation. Due to the absolute value operation, it is not a direct issue to solve the optimization problem (6). We thus derive an iterative algorithm for optimization and show its monotonicity in the following two subsections.

3.3. Iterative algorithm

An iterative algorithm for 2DPCAL1-S is formally presented as follows. Let $\mathbf{v}(0)$ be the initial basis vector.

1. Let $t = 0$, and initialize $\mathbf{v}(t)$ as any p-dimensional vector.
2. Compute the quantity $s_{ji}(t)$ as in (5), which results in value 1, 0, or -1 depending on $\mathbf{v}^T(t)\mathbf{x}_{ji}$ larger than zero, equal to zero, or less than zero, respectively.
3. Let

$$\mathbf{y}(t) = \sum_{i=1}^{n} \sum_{j=1}^{q} s_{ji}(t) \mathbf{x}_{ji}, \tag{7}$$

and

$$\mathbf{w}(t) = \left(\frac{|v_1(t)|}{\gamma + \eta |v_1(t)|}, \ldots, \frac{|v_p(t)|}{\gamma + \eta |v_p(t)|} \right)^T, \tag{8}$$

where $v_k(t)$ is the kth entry of $\mathbf{v}(t)$ for $k = 1, \ldots, p$. Then, the basis vector $\mathbf{v}(t)$ is updated as

$$\mathbf{v}(t+1) = \mathbf{y}(t) \circ \mathbf{w}(t), \tag{9}$$

where $\circ$ denotes the element-wise product between two vectors.

4. If the objective function $h(\mathbf{v}(t+1))$ does not grow significantly, then stop the iterative procedure and set $\mathbf{v}^* = \mathbf{v}(t+1)$. Otherwise, set $t \leftarrow t+1$, and go to Step 2.
5. Output $\mathbf{v}^*$ as the basis vector.

The computational complexity of the above algorithm is $O(nqp)$ per iteration. Note that the update formula (9) can be further expanded as

$$\mathbf{v}(t+1) = \begin{pmatrix} |v_1(t)|y_1(t)/(\gamma + \eta|v_1(t)|) \\ \vdots \\ |v_k(t)|y_k(t)/(\gamma + \eta|v_k(t)|) \\ \vdots \\ |v_p(t)|y_p(t)/(\gamma + \eta|v_p(t)|) \end{pmatrix}, \tag{10}$$

where $y_k(t)$ is the kth entry of $\mathbf{y}(t)$ for $k = 1, \dots, p$ (cf. (7)). Comparing the update formula (10) in 2DPCAL1-S with that (4) in 2DPCA-L1, we see that 2DPCAL1-S weights the entries of $\mathbf{y}(t)$ by using the magnitudes of the corresponding entries of $\mathbf{v}(t)$ while 2DPCA-L1 treats the entries of $\mathbf{y}(t)$ equally in each iteration.

3.4. Monotonicity validation

Theorem. *The objective function $h(\mathbf{v}(t))$ increases with each iteration by the above algorithmic procedure.*

Proof. We establish the theorem by showing that $h(\mathbf{v}(t+1)) \geq h(\mathbf{v}(t))$. We start to compute the value of $h(\mathbf{v}(t))$ at iteration t. With the definition of $s_{ji}(t)$, the first term of $h(\mathbf{v}(t))$ can be rewritten as

$$\sum_{i=1}^{n}\sum_{j=1}^{q} |\mathbf{v}^T(t)\mathbf{x}_{ji}| = \sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\mathbf{v}^T(t)\mathbf{x}_{ji}. \tag{11}$$

Taking the sparsity of $\mathbf{v}(t)$ into account, some elements of $\mathbf{v}(t)$ may happen to be zeros. Denote by $\underline{\mathbf{v}}(t)$ the vector that is resulted from removing the zero elements of $\mathbf{v}(t)$, and $\underline{\mathbf{x}}_{ji}$ the vector that is formed by leaving out the elements of $\mathbf{x}_{ji}$ whose indices correspond to the indices of the zero elements of $\mathbf{v}(t)$. For example, suppose $\mathbf{v}(t) = (1,\ 0,\ 2,\ 0,\ 3)^T$ and $\mathbf{x}_{ji} = (4,\ 5,\ 6,\ 7,\ 8)^T$. Then $\underline{\mathbf{v}}(t) = (1,\ 2,\ 3)^T$ and $\underline{\mathbf{x}}_{ji} = (4,\ 6,\ 8)^T$. Therefore, (11) is equal to

$$\sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{v}}^T(t)\underline{\mathbf{x}}_{ji}. \tag{12}$$

For the second term of $h(\mathbf{v}(t))$, we have that

$$\|\mathbf{v}(t)\|_2^2 = \|\underline{\mathbf{v}}(t)\|_2^2. \tag{13}$$

In the third term of $h(\mathbf{v}(t))$, $\|\mathbf{v}(t)\|_1$ can be rewritten as

$$\begin{aligned} \|\mathbf{v}(t)\|_1 &= \|\underline{\mathbf{v}}(t)\|_1 \\ &= \frac{1}{2}\underline{\mathbf{v}}^T(t)\underline{\mathbf{U}}(t)\underline{\mathbf{v}}(t) + \frac{1}{2}\|\underline{\mathbf{v}}(t)\|_1, \end{aligned} \tag{14}$$

where $\underline{\mathbf{U}}(t)$ is a diagonal matrix defined as

$$\underline{\mathbf{U}}(t) = \operatorname{diag}\left(|\underline{v}_1(t)|^{-1}, \dots, |\underline{v}_{\underline{p}}(t)|^{-1}\right). \tag{15}$$

Here, $\underline{v}_k(t)$ is the kth entry of $\underline{\mathbf{v}}(t)$ for $k = 1, \dots, \underline{p}$, and $\underline{p}$ is the number of the nonzero elements of $\mathbf{v}(t)$. By substituting (12)–(14) into (6), it follows that

$$\begin{aligned} h(\mathbf{v}(t)) = &\sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{v}}^T(t)\underline{\mathbf{x}}_{ji} - \frac{\eta}{2}\|\underline{\mathbf{v}}(t)\|_2^2 \\ &- \frac{\gamma}{2}\left(\underline{\mathbf{v}}^T(t)\underline{\mathbf{U}}(t)\underline{\mathbf{v}}(t) + \|\underline{\mathbf{v}}(t)\|_1\right). \end{aligned} \tag{16}$$

As a transitional procedure, we introduce a surrogate function given by

$$\begin{aligned} Q(\nu|\underline{\mathbf{v}}(t)) = &\ \nu^T\left(\sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{x}}_{ji}\right) - \frac{\eta}{2}\|\nu\|_2^2 \\ &- \frac{\gamma}{2}\left(\nu^T\underline{\mathbf{U}}(t)\nu + \|\underline{\mathbf{v}}(t)\|_1\right), \end{aligned} \tag{17}$$

where ν is a vector of $\underline{p}$-dimensional variable. We emphasize that Q is a function of ν while $\underline{\mathbf{v}}(t)$ is fixed at iteration t. With ν as the independent argument, we maximize the function Q. Differentiating $Q(\nu|\underline{\mathbf{v}}(t))$ with respect to ν and setting it to zero read

$$\frac{\partial Q(\nu|\underline{\mathbf{v}}(t))}{\partial \nu} = \sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{x}}_{ji} - \eta\nu - \gamma\underline{\mathbf{U}}(t)\nu = 0, \tag{18}$$

which implies that

$$\nu = \left(\eta\mathbf{I}_{\underline{p}} + \gamma\underline{\mathbf{U}}(t)\right)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{x}}_{ji}\right), \tag{19}$$

where $\mathbf{I}_{\underline{p}}$ denotes the $\underline{p}$-dimensional identity matrix. Let $\underline{\mathbf{v}}(t+1) = \nu$, i.e., $\underline{\mathbf{v}}(t+1)$ is the maximum point of Q. Noting again that Q is a function of ν, we have that

$$Q(\underline{\mathbf{v}}(t+1)|\underline{\mathbf{v}}(t)) \geq Q(\underline{\mathbf{v}}(t)|\underline{\mathbf{v}}(t)), \tag{20}$$

i.e.,

$$\begin{aligned} &\underline{\mathbf{v}}^T(t+1)\left(\sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{x}}_{ji}\right) - \frac{\eta}{2}\|\underline{\mathbf{v}}(t+1)\|_2^2 \\ &\quad - \frac{\gamma}{2}\left(\underline{\mathbf{v}}^T(t+1)\underline{\mathbf{U}}(t)\underline{\mathbf{v}}(t+1) + \|\underline{\mathbf{v}}(t)\|_1\right) \\ &\geq \underline{\mathbf{v}}^T(t)\left(\sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{x}}_{ji}\right) - \frac{\eta}{2}\|\underline{\mathbf{v}}(t)\|_2^2 \\ &\quad - \frac{\gamma}{2}\left(\underline{\mathbf{v}}^T(t)\underline{\mathbf{U}}(t)\underline{\mathbf{v}}(t) + \|\underline{\mathbf{v}}(t)\|_1\right). \end{aligned} \tag{21}$$

As will be seen, the purpose of the above inequality is to bridge the values between $h(\mathbf{v}(t+1))$ and $h(\mathbf{v}(t))$.

By (16), the right hand of (21) is $h(\mathbf{v}(t))$. We proceed to consider the left hand of (21). For this purpose, we define the updated p-dimensional vector $\mathbf{v}(t+1)$ at the $(t+1)$th iteration as follows. It is formed by inserting zero elements into $\underline{\mathbf{v}}(t+1)$ such that the indices of the inserted zero elements of $\mathbf{v}(t+1)$ are identical with the indices of the zero elements of $\mathbf{v}(t)$. Continuing the example at the beginning of the proof, if $\underline{\mathbf{v}}(t+1) = (9,\ 9,\ 9)^T$, then $\mathbf{v}(t+1) = (9,\ 0,\ 9,\ 0,\ 9)^T$. With this designation of $\mathbf{v}(t+1)$, the first term of the left hand of (21) can be rewritten as

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{v}}^T(t+1)\underline{\mathbf{x}}_{ji} &= \sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\mathbf{v}^T(t+1)\mathbf{x}_{ji} \\ &\leq \sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t+1)\mathbf{v}^T(t+1)\mathbf{x}_{ji} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{q} |\mathbf{v}^T(t+1)\mathbf{x}_{ji}|. \end{aligned} \tag{22}$$

The inequality holds based on the following observation: $s_{ji}(t+1)\mathbf{v}^T(t+1)\mathbf{x}_{ji}$ is always nonnegative due to the definition of $s_{ji}(t+1)$ while $s_{ji}(t)\mathbf{v}^T(t+1)\mathbf{x}_{ji}$ could possibly be negative. For the second term of the left hand of (21), we have that

$$\|\underline{\mathbf{v}}(t+1)\|_2^2 = \|\mathbf{v}(t+1)\|_2^2. \tag{23}$$

For the third term of the left hand of (21), we have that

$$\begin{aligned} &\underline{\mathbf{v}}^T(t+1)\underline{\mathbf{U}}(t)\underline{\mathbf{v}}(t+1) + \|\underline{\mathbf{v}}(t)\|_1 \\ &\quad = \sum_{k=1}^{\underline{p}} \frac{\underline{v}_k^2(t+1)}{|\underline{v}_k(t)|} + \|\underline{\mathbf{v}}(t)\|_1 \end{aligned}$$

$$\begin{aligned}&\geq \sum_{k=1}^{p} \frac{\underline{v}_k^2(t+1)}{|\underline{v}_k(t+1)|} + \|\underline{\mathbf{v}}(t+1)\|_1 \\ &= 2\|\underline{\mathbf{v}}(t+1)\|_1 \\ &= 2\|\mathbf{v}(t+1)\|_1, \end{aligned} \tag{24}$$

where $\underline{v}_k(t+1)$ is the kth entry of $\underline{\mathbf{v}}(t+1)$. The inequality holds due to the following lemma. Note that if some entries $\underline{v}_j(t+1)$ are zeros, we can only consider the nonzero entries. By some operations, we see that the result of (24) still holds.

Lemma (Jenatton et al., 2010). The L1-norm of any vector $\mathbf{z}$ has the following variational equality:

$$\|\mathbf{z}\|_1 = \min_{\boldsymbol{\zeta}\in\mathbb{R}_+^n} \frac{1}{2}\sum_{i=1}^{n}\frac{z_i^2}{\zeta_i} + \frac{1}{2}\|\boldsymbol{\zeta}\|_1, \tag{25}$$

and the minimum value is uniquely achieved at the situation $\zeta_i = |z_i|$ for $i = 1, \ldots, n$, where z_i and ζ_i are the ith entries of $\mathbf{z}$ and ζ, respectively.

Combining (22)–(24), we have that

$$\begin{aligned} h(\mathbf{v}(t+1)) &= \sum_{i=1}^{n}\sum_{j=1}^{q} |\mathbf{v}^T(t+1)\mathbf{x}_{ji}| \\ &\quad - \frac{\eta}{2}\|\mathbf{v}(t+1)\|_2^2 - \gamma\|\mathbf{v}(t+1)\|_1 \\ &\geq \underline{\mathbf{v}}^T(t+1)\left(\sum_{i=1}^{n}\sum_{j=1}^{q} s_{ji}(t)\underline{\mathbf{x}}_{ji}\right) - \frac{\eta}{2}\|\underline{\mathbf{v}}(t+1)\|_2^2 \\ &\quad - \frac{\gamma}{2}\left(\underline{\mathbf{v}}^T(t+1)\underline{\mathbf{U}}(t)\underline{\mathbf{v}}(t+1) + \|\underline{\mathbf{v}}(t)\|_1\right). \end{aligned} \tag{26}$$

Combining (16), (21) and (26), we obtain that

$$h(\mathbf{v}(t+1)) \geq h(\mathbf{v}(t)). \tag{27}$$

Note that the expression of $\mathbf{v}(t+1)$ given in the proof is equivalent to (9). With this update procedure, the monotonicity of the objective function is theoretically guaranteed. The proof of the theorem is thus completed. □

3.5. Multiple basis vectors

We compute the first basis vector $\mathbf{v}_1$ by the algorithm outlined in Section 3.3. Then, we use the deflation technique to extract the remaining basis vectors. Specifically, the τth $(1 < \tau \leq p_0)$ basis vector $\mathbf{v}_\tau$ is computed by using the deflated samples

$$\mathbf{x}_{ji}^{\text{deflated}} = \mathbf{x}_{ji} - \sum_{l=1}^{\tau-1}\mathbf{v}_l(\mathbf{v}_l^T\mathbf{x}_{ji}), \tag{28}$$

where $\mathbf{v}_l$ are normalized to have unit length. That is, the information contained in the previously obtained basis vectors is deducted.

The algorithmic procedure of 2DPCAL1-S is formally summarized in Table 1.

4. Experiments

In order to evaluate the proposed 2DPCAL1-S algorithm, we compare its performances of image classification and reconstruction with four unsupervised learning algorithms: PCA, PCA-L1, 2DPCA, and 2DPCA-L1. Two benchmark face databases FERET and AR are used in our experiments.

In the experiments, the initial components of PCA-L1 are set as the corresponding components of PCA. The initial components of 2DPCA-L1 and 2DPCAL1-S are set as the corresponding components of 2DPCA.

Fig. 1. Sample images of the FERET face database.

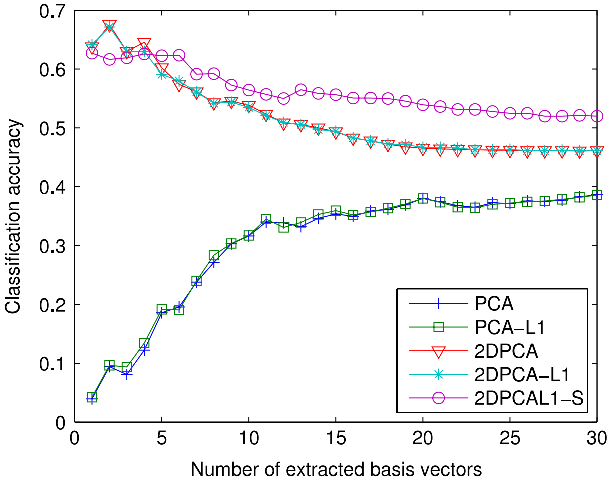


Fig. 2. Classification accuracy of the five algorithms on the FERET face database.

There are two tuning parameters in 2DPCAL1-S. It seems difficult to choose a pair of optimal parameters. However, by the updated expression of $\mathbf{v}(t+1)$, we see that the projection vector essentially relates to the ratio η/λ. Without loss of generality, we consider selecting the optimal ratio in our experiments.

4.1. FERET face database

The first experiment is conducted on a subset of the FERET face database. We use 1400 images of 200 individuals, where each individual has seven images, which show varying expressions and view angles. The image size is 80 by 80. Some sample images are illustrated in Fig. 1. For computational convenience, the images are further resized into 30 by 30.

Ten-fold cross-validation (CV) strategy is adopted for performance evaluation. That is, all images are randomly separated into ten folds, in which nine folds are used for training and the remaining one fold is for testing. This procedure is repeated ten times, and the average classification rate is reported. For 2DPCAL1-S, in each CV repetition, the ratio η/λ is determined on the training data by again a ten-fold CV. Different values of η/λ are tried. Specifically, $\log_{10}(\eta/\lambda)$, denoted as ρ, ranges from -3 to 3 with a step of 1. The value of ρ corresponding to the maximal average classification rate is chosen to classify the testing data.

The five algorithms mentioned above are applied to extract features, followed by the nearest neighbour classifier. Fig. 2 shows the classification accuracy of the five algorithms. It tells us that 2DPCAL1-S outperforms 2DPCA and 2DPCA-L1. This result suggests that introducing the L1-norm regularization term into 2DPCA-L1 could improve the classification performance.

We investigate how the classification performance of 2DPCAL1-S depends on ρ. Fig. 3(a) shows the classification accuracy of 2DPCAL1-S with different values of ρ. By comparison with the parameter ρ determined by the CV, we see that a constant value, say $\rho = 0$ here, could yield a competitive (or the same) classification accuracy. It could be observed that when $\rho \geq 0$ the classification accuracy tends to decrease with the increasing number of extracted features while when $\rho < 0$ the classification accuracy

Table 1
Algorithmic procedure of 2DPCAL1-S with multiple basis vectors.

```
Input: training samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of size $q \times p$, number of basis vectors $p_0$ (usually $p_0 < p$), and regularization parameter $\gamma$.
Output: projection matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{p_0}) \in \mathbb{R}^{p \times p_0}$ of $p_0$ basis vectors.

1. Set $\tau = 1$.
2. Compute the basis vector $\mathbf{v}_\tau$:
  (a) Use any $p$-dimensional vector as the initial vector $\mathbf{v}(0)$, and set $t = 0$.
  (b) Define the sign function
          $s_{ji}(t) = \mathrm{sign}(\mathbf{v}^T(t)\mathbf{x}_{ji})$,
      where $\mathbf{x}_{ji}$ is the $j$th row vector of $\mathbf{X}_i$.
  (c) Update $\mathbf{v}(t)$ by
          $\mathbf{v}(t+1) = \mathbf{y}(t) \circ \mathbf{w}(t)$,
      where
          $\mathbf{y}(t) = \sum_{i=1}^{n} \sum_{j=1}^{q} s_{ji}(t)\mathbf{x}_{ji}$,
          $\mathbf{w}(t) = \left(\frac{|v_1(t)|}{\gamma+\eta|v_1(t)|}, \ldots, \frac{|v_p(t)|}{\gamma+\eta|v_p(t)|}\right)^T$.
  (d) If the objective function
          $\sum_{i=1}^{n} \sum_{j=1}^{q} |\mathbf{v}^T(t+1)\mathbf{x}_{ji}| - \frac{\eta}{2}\|\mathbf{v}(t+1)\|_2^2 - \gamma\|\mathbf{v}(t+1)\|_1$
      does not grow significantly, then exit the inner loop and set
          $\mathbf{v}_\tau = \mathbf{v}(t+1)$.
      Otherwise, set $t \leftarrow t+1$, and go to Step 2(b).
3. Using the obtained basis vectors $\mathbf{v}_1, \ldots, \mathbf{v}_\tau$, deflate $\mathbf{x}_i$ as
          $\mathbf{x}_{ji}^{\text{deflated}} = \mathbf{x}_{ji} - \sum_{l=1}^{\tau} \mathbf{v}_l(\mathbf{v}_l^T\mathbf{x}_{ji})$,
  where $\mathbf{v}_l$ are normalized to have unit length.
4. If $\tau < p_0$, then let $\tau \leftarrow \tau + 1$ and return to Step 2(a), wherein the deflated samples are used. Otherwise, stop the run and output the basis vectors $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{p_0})$,
where the basis vectors are normalized.
```

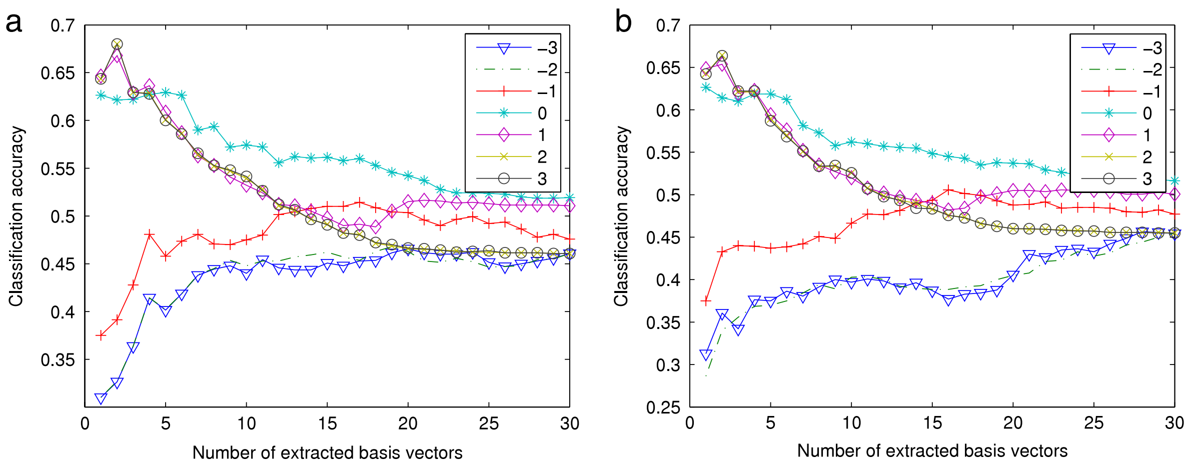


Fig. 3. Classification accuracy of 2DPCAL1-S with varying ρ on the FERET face database. Each curve corresponds to a value of ρ. (a) Use the basis vectors of 2DPCA as initialization. (b) Use random initialization.

tends to increase with the increasing number of extracted features. In general, a positive ρ leads to better results than a negative one. These trends are due to different weights of the L1-norm and the L2-norm imposed on the projection vector $\mathbf{v}$ in the regularization term.

In our experiments, we use the basis vectors of 2DPCA as the initial basis vectors of 2DPCAL1-S. It may be helpful to compare the output of the 2DPCAL1-S algorithm under different initializations. We find that different initializations lead to slightly different results. For example, as compared with Fig. 3(a), the classification accuracy of 2DPCAL1-S using random initialization on the FERET face database is shown in Fig. 3(b). It demonstrates that the algorithm of 2DPCAL1-S finds a (local) maximum.

It is observed that the recognition accuracy of the 2DPCA-based methods reaches the maximum value at the point of using two, three, or four basis vectors and declines with larger numbers of basis vectors. It suggests that the first few basis vectors extract sufficient features for classification. The subsequent basis vectors may produce redundant features and noise which deteriorate the classification accuracy. Note that the classification accuracy varies abruptly at some points. The reason may be that, for the 2DPCA-based methods, one basis vector produces many features, which are equal to the row size of the image. So, the classification accuracy is sensitive to the number of basis vectors, which leads to the classification peak.

It is well known that PCA can also be defined from the perspective of minimizing the reconstruction error with a few principal components (Jolliffe, 1986). The reconstruction error measures the expressive capacity of the principal components. We thus proceed to consider the task of image reconstruction. To investigate the robustness, the database is intentionally contaminated. The polluted data set is designed as follows. 20% of the total 1400 images are randomly selected and occluded with a rectangular noise. The rectangle is randomly located, and its size is at least 20 by 20, in which the noise consists of random black and white dots.

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ be the m ($m = 1120$) unoccluded images, $\bar{\mathbf{X}}$ the mean of all the 1400 images, and $\mathbf{V}$ the projection matrix obtained by 2DPCA, 2DPCA-L1, or 2DPCAL1-S. Then, on the polluted database, the average reconstruction error of the three methods is defined by using the clean images, given by

$$\frac{1}{m}\sum_{i=1}^{m} \|\mathbf{Y}_i - ((\mathbf{Y}_i - \bar{\mathbf{X}})\mathbf{V}\mathbf{V}^T + \bar{\mathbf{X}})\|_F, \qquad (29)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The average reconstruction error of PCA or PCA-L1 can be likewise defined.

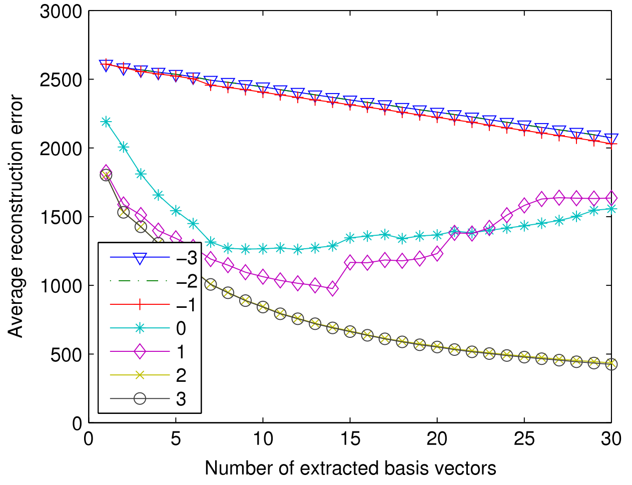


Fig. 4. Average reconstruction errors of 2DPCAL1-S with various values of ρ on the polluted FERET face database.

Fig. 4 shows the average reconstruction errors of 2DPCAL1-S with varying values of ρ on the polluted FERET face database. From the figure, we see that the average reconstruction error tends to decrease when ρ increases. The curves corresponding to $\rho = 2$ and $\rho = 3$ are very close to each other. Larger ρ has also been tried. The average reconstruction error, however, has little variation when ρ is larger than two.

We now set ρ as three in 2DCPAL1-S and compare its reconstruction performance with other four algorithms. Fig. 5 shows the reconstructed images of the five algorithms (i.e., PCA, PCA-L1, 2DPCA, 2DPCA-L1, and 2DPCAL1-S), wherein the first ten projection vectors are used for the image reconstruction. Fig. 6 shows the average reconstruction error of the five algorithms. In the figure, the two curves corresponding to 2DPCA-L1 and 2DPCAL1-S overlap, which means that they obtain nearly the same average reconstruction error. When the feature number is larger than twelve, PCA-L1 outperforms PCA. When the feature number is larger than eight, 2DPCA-L1 and 2DPCAL1-S outperform 2DPCA. We emphasize that, given the learned projection vectors, the reconstruction error delineated in Fig. 6 is defined on the clean images. Fig. 6 is not necessarily reflected in Fig. 5(b) which illustrates the polluted images.

Note that the mechanisms of the PCA and the 2DPCA-based methods are different in the sense that the former process full images (converted into vectors) while the latter accommodate each row vectors of the images. As a result, each basis vector of the PCA-based methods yields one feature for an image while each basis vector of the 2DPCA-based methods produces many features (equalling the number of row vectors). So, it is not very meaningful to compare the performance of the PCA-based methods with that of the 2DPCA-based methods using the same number of basis vectors.

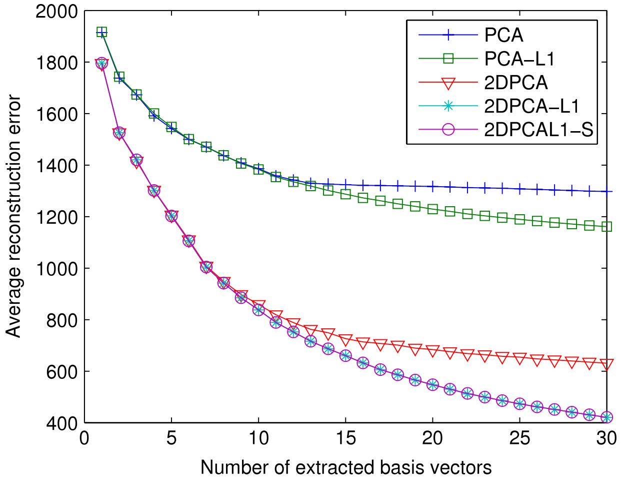


Fig. 6. Average reconstruction errors of the five different algorithms on the polluted FERET face database .

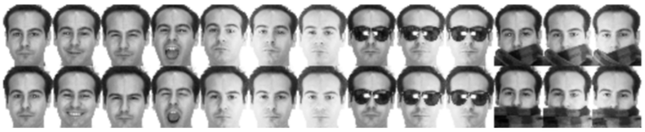


Fig. 7. Sample images of the AR face database.

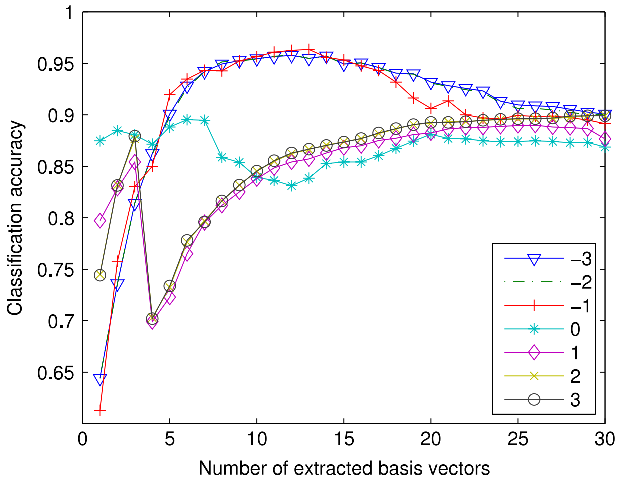


Fig. 8. Classification accuracy of 2DPCAL1-S with varying ρ on the AR face database.

4.2. AR face database

The AR face database contains 3120 images of 120 subjects, i.e., 26 images per subject. The images were taken with different facial

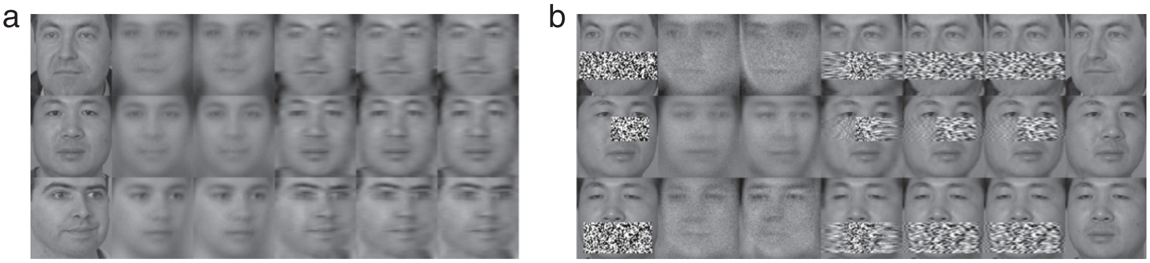


Fig. 5. The reconstructed images on the polluted FERET face database. The first column shows the faces from the database; the following five columns show the reconstructed faces by using the first ten projection vectors produced by PCA, PCA-L1, 2DPCA, 2DPCA-L1, and 2DPCAL1-S, respectively. (a) Images without occlusion. (b) Images with occlusion. The last column in (b) shows the original unoccluded faces, which are used for visual comparison.

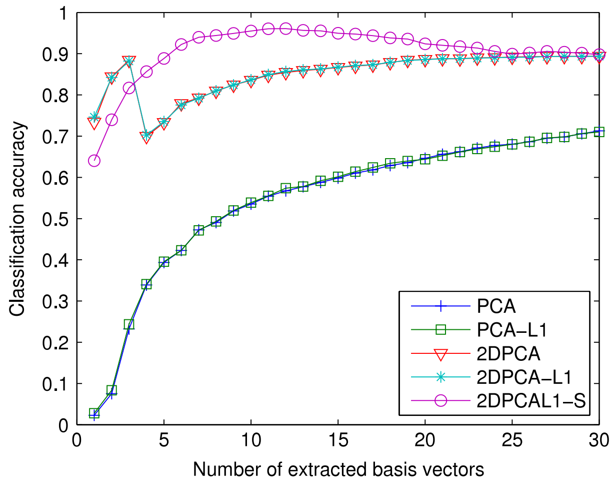


Fig. 9. Classification accuracy of the five algorithms on the AR face database.

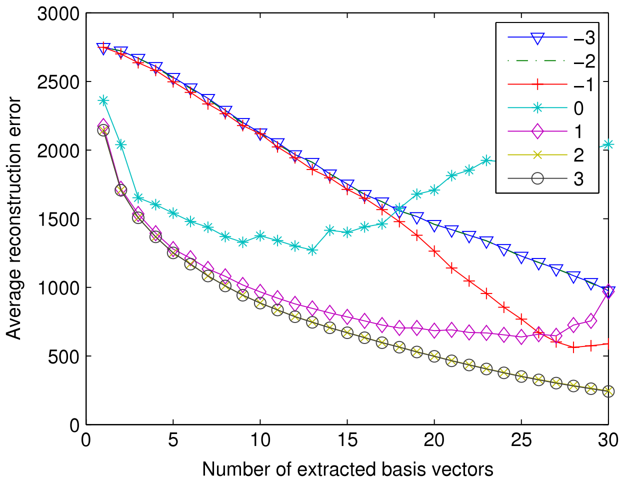


Fig. 10. Average reconstruction errors of 2DPCAL1-S with various values of ρ on the AR face database with the natural occlusion.

expressions and illuminations, and some images were occluded with black sunglasses or towels, as shown in Fig. 7. The image size is 50 by 40.

The classification rates of 2DPCAL1-S with different values of ρ are demonstrated in Fig. 8, where the ten-fold CV scheme is employed. It can be observed that the maximal classification accuracy is generally achieved when taking ρ as three. It also tells us that $\rho > 0$ and $\rho < 0$ lead to different trends of classification accuracy with respect to the number of extracted features.

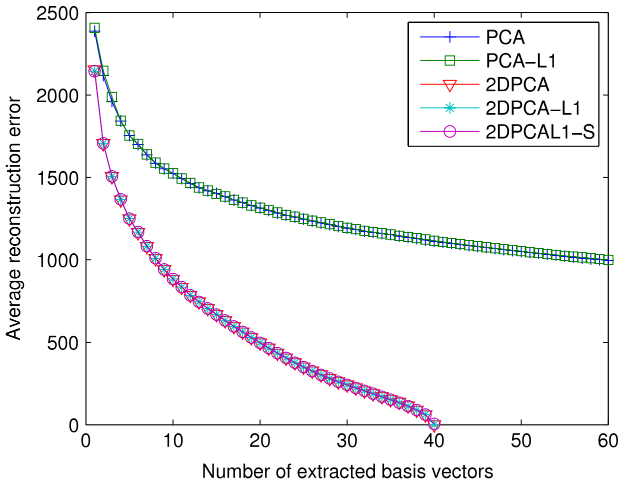


Fig. 12. Average reconstruction errors of the five different algorithms on the AR face database with the natural occlusion.

Based on Fig. 8, we fix ρ to be -3 in 2DPCAL1-S. The classification performance of the five algorithms is shown in Fig. 9. We see that 2DPCAL1-S achieves improved performance.

In the task of face reconstruction, we first learn the basis vectors on all AR faces, and then calculate the reconstruction error via (29), where the faces with black sunglasses or towels are regarded as outlying samples. Fig. 10 shows the average reconstruction errors of 2DPCAL1-S with varying ρ on the AR face database with the natural occlusion. The error tends to decrease when ρ increases. We choose ρ to be three in 2DPCAL1-S and compare its reconstruction performance with the other four algorithms, as shown in Figs. 11 and 12. It could be observed that PCA and PCA-L1 obtain nearly the same reconstruction error. Also, 2DPCA, 2DPCA-L1 and 2DPCAL1-S obtain nearly the same result. The reason may be that the influence of the natural occlusion is not enough to demonstrate the superiority of the L1-norm-based approaches. We thus consider polluting the faces artificially. Specifically, we exclude the naturally occluded images in the experiment. Instead, 20% images are randomly selected from the clean images and occluded with a rectangle of random black and white dots. In this case, Fig. 13 shows the reconstructed images of the five algorithms on the artificially polluted AR face database. The reconstruction errors of the five algorithms are shown in Fig. 14, where the value of ρ in 2DPCAL1-S is likewise selected as three. Again, 2DPCAL1-S and 2DPCA-L1 obtain nearly the same average reconstruction error. When the number of features is larger than eight, the two algorithms are superior to 2DPCA. The classification accuracy of the five algorithms with the ten-fold CV is shown in Fig. 15, where the value of ρ in 2DPCAL1-S is selected as -3. We see that a satisfactory classification accuracy is obtained on this artificially polluted AR face database, and 2DPCAL1-S demonstrates its superiority.

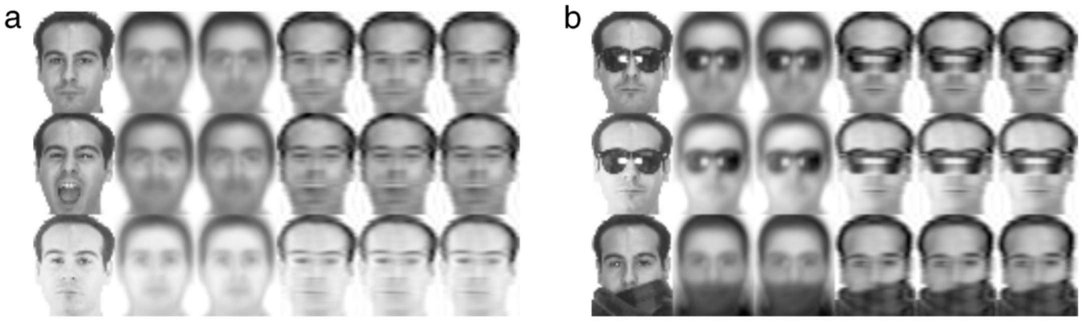


Fig. 11. The reconstructed images on the naturally polluted AR face database. The first column shows the faces from the database; the following five columns show the reconstructed faces by using the first ten projection vectors produced by PCA, PCA-L1, 2DPCA, 2DPCA-L1, and 2DPCAL1-S, respectively. (a) Images without occlusion. (b) Images with natural occlusion.

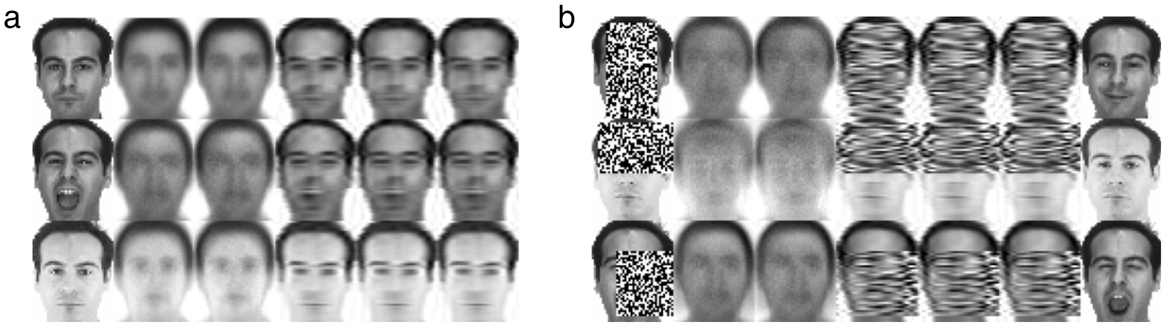


Fig. 13. The reconstructed images on the artificially polluted AR face database. The first column shows the faces from the database, the following five columns show the reconstructed faces by using the first ten projection vectors produced by PCA, PCA-L1, 2DPCA, 2DPCA-L1, and 2DPCAL1-S, respectively. (a) Images without occlusion. (b) Images with artificially occluded occlusion. The last column in (b) shows the original unoccluded faces, which are used for visual comparison.

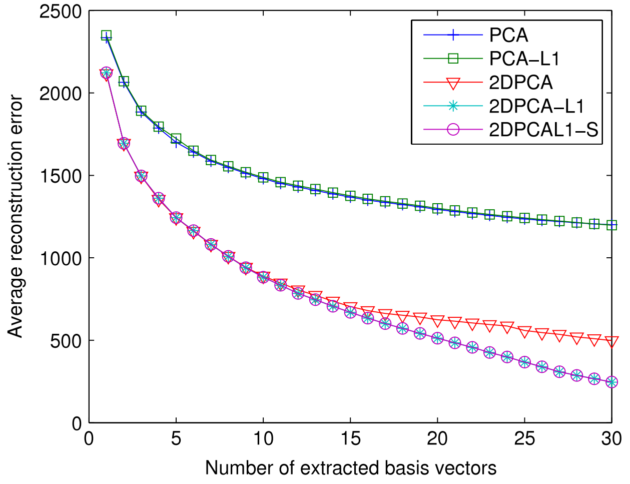


Fig. 14. Average reconstruction errors of the five different algorithms on the artificially polluted AR face database.

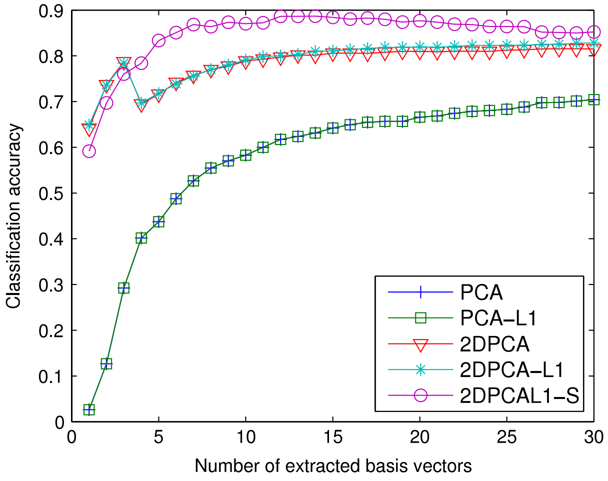


Fig. 15. Classification accuracy of the five algorithms on the artificially polluted AR face database.

4.3. Sparsity of 2DPCAL1-S

The sparsity of the projection matrix of 2DPCAL1-S is investigated in this subsection. In the task of face reconstruction, the average reconstruction error of 2DPCAL1-S decreases with the increasing value of ρ. When ρ is very large, the average reconstruction error of 2DPCAL1-S is low. At the same time, the projection matrix of 2DCPAL1-S is not sparse.

In the task of face classification, the appropriate value of ρ is small, and the projection matrix of 2DCPAL1-S is very sparse. Since the ten-fold CV is employed in classification, a projection matrix is obtained each time. For the FERET faces, there are ten projection matrices with a size 30 × 30. And for AR faces, there are ten projection matrices with a size 40 × 30. Table 2 shows the average cardinality of the first eight projection vectors. For FERET faces, some pixels of an image are linearly combined as a feature. For AR faces, the projection vector is very sparse, which results in a single feature selection. In both the cases, most pixels of an image are ignored, but we still obtain satisfactory classification accuracy, which means that the selected features contain valuable discriminative information.

Table 2
Average sparsity of 2DPCAL1-S.

Database	ρ	Dimension	Projection vector							
			1	2	3	4	5	6	7	8
FERET	0	30	23	6	4	5	4	5	6	7
AR	−3	40	1	2	1	1	1	1	1	1

5. Conclusion

A new subspace learning method, called 2DPCAL1-S, is developed for image analysis in this paper. It uses the L1-norm for both robust and sparse modelling. The role of the L1-norm is two-fold. One is the robust measurement of the dispersion of samples, as in 2DPCA-L1. The other is to introduce penalty, resulting in the sparse projection vectors. 2DPCAL1-S utilizes the feature extraction and the feature selection simultaneously and robustly. Computationally, an iterative algorithm is designed, the monotonicity of which is theoretically guaranteed. The effectiveness of 2DPCAL1-S on image classification and reconstruction is experimentally demonstrated.

The optimal parameter ρ, however, may depend on the data set at hand. It is hard to be determined analytically, which needs further investigation and is our future work.

Acknowledgements

The authors would like to thank the anonymous referees for the constructive recommendations, which greatly improve the paper. This work was supported in part by the National Natural Science Foundation of China under Grants 61075009 and 31130025, in part by the Natural Science Foundation of Jiangsu Province under Grant BK2011595, in part by the Program for New Century Excellent Talents in University of China under Grant NCET-12-0115, and in part by the Qing Lan Project of Jiangsu Province.

References

Cai, D., He, X., & Han, J. (2007). Spectral regression: a unified approach for sparse subspace learning. In *Proceedings of the seventh IEEE international conference on data mining* (pp. 73–82).

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*, 33–61.

Ding, C., Zhou, D., He, X., & Zha, H. (2006). R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization. In *Proceeding of international conference on machine learning* (pp. 281–288).

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

Jenatton, R., Obozinski, G., & Bach, F. (2010). Structured sparse principal component analysis. In *Proceedings of the 13th international conference on artificial intelligence and statistics*.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Ke, Q., & Kanade, T. (2005). Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceeding of IEEE international conference on computer vision and pattern recognition* (pp. 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.

Li, X., Pang, Y., & Yuan, Y. (2009). L1-norm-based 2DPCA. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *40*(4), 1170–1175.

Meng, D., Zhao, Q., & Xu, Z. (2012). Improve robustness of sparse PCA by L1-norm maximization. *Pattern Recognition*, *45*, 487–497.

Nie, F., Huang, H., Ding, C., Luo, D., & Wang, H. (2011). Robust principal component analysis with non-greedy L1-norm maximization. In *International joint conference on artificial intelligence* (pp. 1433–1438).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*, *58*, 267–288.

Wang, H. (2012). Structured sparse linear graph embedding. *Neural Networks*, *27*, 38–44.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., & Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, *98*(6), 1031–1044.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(1), 40–51.

Yang, J., Zhang, D., Frangi, A. F., & 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), 1–7.

Zhou, T., Tao, D., & Wu, X. (2011). Manifold elastic net: a unified framework for sparse dimension reduction. *Data Mining and Knowledge Discovery*, *22*, 340–371.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.