IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 13, 2020, accepted December 26, 2020, date of publication January 6, 2021, date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049535

Robust 2DPCA by Tℓ_1 Criterion Maximization for Image Recognition

XIANGFEI YANG[1], WENSI WANG[2], (Member, IEEE), LIMING LIU[1], YUANHAI SHAO[3], LITING ZHANG[2], AND NAIYANG DENG[4]

[1]School of Statistics, Capital University of Economics and Business, Beijing 100070, China
[2]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[3]School of Management, Hainan University, Haikou 570228, China
[4]College of Science, China Agricultural University, Beijing 100083, China

Corresponding authors: Wensi Wang (wensi.wang@bjut.edu.cn) and Liming Liu (llm5609@163.com)

This work was supported by the Beijing Nova Star Project under Grant K1043001201801.

ABSTRACT Two-dimensional principal component analysis (2DPCA) has been widely used to extract image features. As opposed to PCA, 2DPCA directly treats 2D matrices to extract image features instead of transforming 2D matrices into vectors. However, the classical 2DPCA based on F-norm square is sensitive to noise. To handle this problem, 2DPCAs based on ℓ_1-norm, ℓ_p-norm, and other norms have been studied. In this paper, as a further development, 2DPCA based on Tℓ_1 criterion is proposed, referred as 2DPCA-Tℓ_1. Notice that, different from some norms used before, Tℓ_1 criterion is bounded and Lipschitz-continuous. So it can be expected that our 2DPCA-Tℓ_1 should be more robust. In fact, the experimental results have shown that its performance is superior to that of classical 2DPCA, 2DPCA-L1, 2DPCAL1-S, N-2-DPCA, G2DPCA, and Angle-2DPCA.

INDEX TERMS Two-dimensional principal component analysis (2DPCA), Tℓ_1 criterion, robust, dimensionality reduction, feature extraction.

I. INTRODUCTION

Principal component analysis (PCA) [1], [2] is a popular dimensionality reduction and feature extraction method. It has been widely used in the fields of image recognition and computer vision. However, classical PCA is sensitive to outliers and noise. Thus, many improved versions are proposed, e.g., L1-PCA [3], R1-PCA [4], PCA-L1 [5], PCA-Lp [6], kernel PCA [7], [8], and low-rank PCA [9]. PCA aims to search for several principal components resulting in a projection matrix, such that the dimensionality reduction is realized. In addition to PCA, linear discriminant analysis (LDA) [10], [11] and locally preserving projection (LPP) [12] are also the representative dimensionality reduction methods. The former extracts the most discriminating features, the latter, as the linear approximation of locally linear embedding (LLE) [13], characterizes the local geometric structure.

However, when the above-mentioned methods are applied to extract features from images, we have to transform the image matrices (2D matrices) into high-dimensional image vectors (1D vectors) by concatenating all columns of image matrices, resulting in damage to the spatial structure embedded in pixels of the image. To tackle this issue, a new kind of PCA called 2DPCA is proposed, which directly deals with the 2D image matrices rather than 1D vectors. In addition to retaining spatial structure information, another advantage is that its covariance matrix is much smaller than that of PCA because the covariance matrix is computed directly using the original image matrices, resulting in being evaluated with much less time consuming and higher accuracy. Just like PCA, its model can be constructed by either maximizing the dispersion or minimizing the reconstruction error. And the corresponding optimization problem can be solved by either greedy strategy or non-greedy strategy.

2DPCA proposed by Yang *et al*. [14] is the early one to deal with 2D matrices directly. Since F-norm square of matrix is employed, it is sensitive to outliers and noise. It is well known that ℓ_1-norm is more robust than F-norm square. Therefore, some ℓ_1-norm-based 2DPCAs have been studied. More precisely, 2DPCA-L1 [15] was proposed as a generalization of PCA-L1 [5]. Then Wang *et al*. [16] proposed its non-greedy version. Based on 2DPCA-L1, 2DPCAL1-S [17] was proposed, aiming at improving both the robustness and sparseness simultaneously. G2DPCA [18] was a further

The associate editor coordinating the review of this manuscript and approving it for publication was Varuna De Silva.

study, where the general ℓ_p-norm was introduced, and the parameter p is restricted by $p \geq 1$ and $p > 0$ in the objective function and the constraint respectively. 2DPCA, 2DPCA-L1, and 2DPCAL1-S are all the special cases of G2DPCA.

Besides, to be more robust, 2DPCAs based on other norms have also been proposed. R-2DPCA [19], Angle-2DPCA [20], F-norm 2DPCA [21], and OMF-2DPCA [22] employ F-norm as the distance metric instead of F-norm square, resulting in both robustness and rotational invariance. R_1-2DPCA [23] is based on R_1-norm and helps encode discriminant information. N-2DPCA [24] uses nuclear norm to measure the reconstruction error. Its robustness comes from the fact that nuclear norm is essentially the convex envelope of the matrix rank. Moreover, 2DPCA-Sp [25] is based on Schatten p-norm ($0 < p < \infty$) to maximize the dispersion, and GC-2DPCA [26] is based on $\ell_{2,p}$-norm ($0 < p \leq 2$) to minimize the reconstruction error. Both of them are regarded as a framework of 2DPCA.

Different from using the norms mentioned above, the Tℓ_1 criterion is used to construct our 2DPCA-Tℓ_1. The Tℓ_1 criterion looks like the ℓ_p-norm with $p \in (0, 1)$. However, they are markedly different since Tℓ_1 criterion has two properties: boundedness and Lipschitz-continuity, where Lipschitz-continuity measures relative changes in the objective function with respect to the input. These two properties make the Tℓ_1 criterion to be a suitable distance metric for PCA, particularly for robustness, due to its stronger suppression of noise. Thus we employ Tℓ_1 criterion as a distance metric to formulate the optimization problem. For solving the optimization problem, a modified gradient ascent method is designed. This leads to our 2DPCA-Tℓ_1. 2DPCA-Tℓ_1 has two major advantages:

- Because the distance metric Tℓ_1 criterion has the stronger suppression effect to noise, 2DPCA-Tℓ_1 is robust to noise.
- Compared with PCAs, the spatial structure information is preserved.

The experimental results on real datasets have shown the effectiveness of our 2DPCA-Tℓ_1.

The rest of this paper is organized as follows. In Section II, we present briefly related works including 2DPCA, 2DPCA-L1, 2DPCAL1-S, G2DPCA, N2DPCA and Angle-2DPCA. In Section III, our Tℓ_1-criterion-based 2DPCA is described in detail. Its performance is compared with the related works in Section IV. Finally, the conclusion follows in Section V.

II. RELATED WORKS

Suppose that there are N training image matrices $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, where $\mathbf{X}_i \in \mathbb{R}^{m\times n}$, $i = 1, \ldots, N$, m and n stand for the image height and width, respectively. Without loss of generality, assume that the image matrices have been centralized, i.e., $\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i = 0$. For a given $d > 0$, our task is to find a projection matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^{n\times d}$, where $\mathbf{w}_i \in \mathbb{R}^n$ is the ith projection vector (principal component), $i = 1, \ldots, d$. Then, the corresponding low-dimensional representation $\mathbf{Y}_i \in \mathbf{R}^{m\times d}$ of image $\mathbf{X}_i$ is given by

$$\mathbf{Y}_i = \mathbf{X}_i\mathbf{W} = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{im} \end{pmatrix}\mathbf{W}, \quad i = 1, \ldots, N,$$

where $\mathbf{X}_{ij} \in \mathbb{R}^{1\times n}$ is the jth row of $\mathbf{X}_i$, $j = 1, \ldots, m$.

Herein, for finding the projection matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d]$, the approach we are interested in is maximizing the dispersion by greedy strategy. So the key step is to find the first projection vector $\mathbf{w}_1$ by constructing and solving an optimization problem because the followed $\mathbf{w}_2, \ldots, \mathbf{w}_d$ can be obtained one by one similarly. In the following review of the related works, we are only concerned with the first projection vector $\mathbf{w}_1$ and the corresponding optimization problems with the single vector variable $\mathbf{w}$.

A. 2DPCA

Remind the early 2DPCA [14] dealing with the 2D matrices directly, its key point is to solve the following optimization problem with the projection vector $\mathbf{w}$

$$\begin{aligned} &\max_{\mathbf{w}} \sum_{i=1}^{N} ||\mathbf{X}_i\mathbf{w}||_2^2 \\ &\text{s.t. } ||\mathbf{w}||_2^2 = 1. \end{aligned} \tag{1}$$

Its solution, the projection vector $\mathbf{w}$, could be obtained by calculating the eigen decomposition of the image covariance matrix

$$\mathbf{S} = \sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{X}_i \in \mathbb{R}^{n\times n},$$

and selecting the eigenvector with the largest eigenvalue. As can be seen in problem (1), ℓ_2-norm square is employed as the metric. Its sensitivity to outliers and noise leads to the following improvement.

B. 2DPCA-L1

2DPCA-L1 [15] is formulated by replacing ℓ_2-norm square with ℓ_1-norm in the objective function of the problem (1). Thus, its optimization problem is as follows

$$\begin{aligned} &\max_{\mathbf{w}} \sum_{i=1}^{N} ||\mathbf{X}_i\mathbf{w}||_1 \\ &\text{s.t. } ||\mathbf{w}||_2^2 = 1. \end{aligned} \tag{2}$$

Since ℓ_1-norm is employed as the distance metric, 2DPCA-L1 is more robust. Due to the fact that problem (2) does not exist closed-form solution, an iterative algorithm is necessary.

C. G2DPCA

Corresponding to problems (1) and (2), G2DPCA constructs its optimization problem as follows.

$$\max_{\mathbf{w}} \sum_{i=1}^{N} ||\mathbf{X}_i\mathbf{w}||_s^s$$
$$\text{s.t. } ||\mathbf{w}||_p^p = 1, \tag{3}$$

where $||.||_s$ and $||.||_p$ stand respectively for s-norm ($s \geq 1$) and p-norm ($p > 0$). Obviously, it is a general formulation and both the above 2DPCA and 2DPCA-L1 are its special cases. Different from 2DPCA-L1, it may be not only robust but also sparse.

In addition to the aforementioned three methods, the other related methods include 2DPCAL1-S, N2DPCA, and Angle-2DPCA, which are also compared with 2DPCA-Tℓ_1 in our numerical experiments later.

III. Tℓ_1-CRITERION-BASED 2DPCA

In this section, we first introduce Tℓ_1 criterion as a distance metric, which is based on the transformed ℓ_1 (Tℓ_1) penalty function [27]–[32]. For a vector $\mathbf{z} = [z_1, \ldots, z_n]^T \in \mathbb{R}^n$, its T$\ell_1$ criterion is defined as

$$TL1_a(\mathbf{z}) = \sum_{i=1}^{n} \rho_a(z_i), \tag{4}$$

where $\rho_a(t)$ is the operator of the component:

$$\rho_a(t) = \frac{(a+1)|t|}{a+|t|},$$

and $a > 0$ is a positive shape parameter.

Let us compare Tℓ_1 criterion with some relevant norms. Remind that the ℓ_p-norm of a vector $\mathbf{z} = [z_1, \ldots, z_n]^T \in \mathbb{R}^n$ is defined as

$$||\mathbf{z}||_p = \left(\sum_{i=1}^{n} \mu_p(z_i)\right)^{1/p}, \quad 0 < p < 1,$$

where $\mu_p(t)$ is its operator of component $\mu_p(t) = |t|^p$. The ℓ_1-norm and ℓ_0-norm are respectively defined as

$$||\mathbf{z}||_1 = \sum_{i=1}^{n} \mu_1(z_i)$$

and

$$||\mathbf{z}||_0 = \sum_{i=1}^{n} \mu_0(z_i)$$

with the operators of component $\mu_1(t) = |t|$ and $\mu_0(t) = |t|^0$, respectively.

Notice that a norm should satisfy the following three properties: the positive definiteness, the triangle inequality and the absolutely homogeneity. The ℓ_p-norm ($0 < p < 1$) only satisfies the positive definiteness and the absolutely homogeneity. Similarly, the Tℓ_1 criterion only satisfies the positive definiteness and the triangle inequality. So strictly speaking, Tℓ_1 criterion is not a norm. However, this should not prevent it to be a distance metric.

In order to compare Tℓ_1 criterion with the ℓ_p-norm ($0 < p < 1$), examine their operators of component $\rho_a(t)$ and $\mu_p(t)$. For any fixed t, with the change of parameter a, we have

$$\lim_{a\to 0^+} \rho_a(t) = \mu_0(t), \ \lim_{a\to\infty} \rho_a(t) = \mu_1(t),$$

which shows that Tℓ_1 criterion interpolates ℓ_0- and ℓ_1-norm. It seems that Tℓ_1 criterion with the parameter $a \in (0, \infty)$ is similar to ℓ_p-norm with a parameter $p \in (0, 1)$, but they have significant difference. In fact, investigate their operators of component $\rho_a(t)$ and $\mu_p(t)$ first. Obviously, both $\rho_a(t)$ and $\mu_p(t)$ are even functions. So, we only consider the case where $t > 0$. In this case, for any fixed parameter $a(a > 0)$ and parameter $p(0 < p < 1)$, it is easy to see that

$$\rho_a''(t) = \frac{-2a(a+1)}{(a+t)^3} < 0, \ \rho_a'(t) = \frac{a(a+1)}{(a+t)^2} > 0,$$
$$\lim_{t\to 0} \rho_a'(t) = 1 + a^{-1}, \quad \lim_{t\to\infty} \rho_a(t) = a + 1, \tag{5}$$

and

$$\mu_p'(t) = pt^{p-1} > 0, \ \lim_{t\to 0} \mu_p'(t) = \infty, \ \lim_{t\to\infty} \mu_p(t) = \infty. \tag{6}$$

This means that, on the one hand, $\rho_a(t)$ is bounded and Lipschitz-continuous; on the other hand, $\mu_p(t)$ is unbounded and not Lipschitz-continuous. Thus we conclude that Tℓ_1 criterion should have a stronger suppression effect to noise and better continuity than the ℓ_p-norm ($0 < p < 1$) in theory.

A. MODEL

Motivated by the advantages of Tℓ_1 criterion, we employ Tℓ_1 criterion as the distance metric and construct our Tℓ_1-criterion-based 2DPCA called 2DPCA-Tℓ_1. Corresponding to problems (1), (2), and (3), our optimization problem is as follows

$$\max_{\mathbf{w}} \sum_{i=1}^{N} TL1_a(\mathbf{X}_i\mathbf{w})$$
$$\text{s.t. } ||\mathbf{w}||_2^2 = 1. \tag{7}$$

Since $\mathbf{X}_i\mathbf{w} = \begin{pmatrix} \mathbf{X}_{i1}\mathbf{w} \\ \mathbf{X}_{i2}\mathbf{w} \\ \vdots \\ \mathbf{X}_{im}\mathbf{w} \end{pmatrix} \in \mathbb{R}^{m\times 1}$ and $TL1_a(\mathbf{X}_i\mathbf{w}) = \sum_{j=1}^{m} \frac{(a+1)|\mathbf{X}_{ij}\mathbf{w}|}{a+|\mathbf{X}_{ij}\mathbf{w}|}$, problem (7) can be reformulated as

$$\max_{\mathbf{w}} f(\mathbf{w}) = \sum_{i=1}^{N}\sum_{j=1}^{m} \frac{(a+1)|\mathbf{X}_{ij}\mathbf{w}|}{a+|\mathbf{X}_{ij}\mathbf{w}|}$$
$$\text{s.t. } ||\mathbf{w}||_2^2 = 1, \tag{8}$$

where $\mathbf{X}_{ij} \in \mathbb{R}^{1\times n}$ is the jth row of $\mathbf{X}_i$, $i = 1, \ldots, N$, $j = 1, \ldots, m$.

The optimization problem (8) can be used to find the first projection vector $\mathbf{w}_1$, and the followed projection vectors $\mathbf{w}_2, \ldots, \mathbf{w}_d$ as well.

B. ALGORITHM

Our Tℓ_1-criterion-based 2DPCA can be described by two parts: solving the problem (8); and searching for the projection vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d$ by greedy strategy.

1) GRADIENT ASCENT ALGORITHM FOR SOLVING (8)

Since problem (8) is non-convex and non-smooth, the traditional optimization techniques [33], [34] could not be used directly. Thus a modified gradient ascent algorithm is designed. The search direction of the steepest ascent algorithm at $\mathbf{w}$ is the subgradient

$$\partial f(\mathbf{w}) = \sum_{i=1}^{N}\sum_{j=1}^{m} \frac{a(a+1)sign(\mathbf{X}_{ij}\mathbf{w})\mathbf{X}_{ij}^T}{(a+|\mathbf{X}_{ij}\mathbf{w}|)^2}, \tag{9}$$

where

$$sign(\lambda) = \begin{cases} 1, & \lambda > 0, \\ 0, & \lambda = 0, \\ -1, & \lambda < 0. \end{cases}$$

Noticing the unit sphere constraint $||\mathbf{w}||_2^2 = 1$, project $\partial f(\mathbf{w})$ onto the tangent plane to this sphere at $\mathbf{w}$ and get

$$\mathbf{g}(\mathbf{w}) = \partial f(\mathbf{w}) - \langle \partial f(\mathbf{w}), \mathbf{w}\rangle \mathbf{w}.$$

Obviously, $\mathbf{g}(\mathbf{w})$ is perpendicular to $\mathbf{w}$

$$\langle \mathbf{g}(\mathbf{w}), \quad \mathbf{w}\rangle = 0, \tag{10}$$

and

$$\langle \mathbf{g}(\mathbf{w}), \quad f(\mathbf{w})\rangle \geq 0. \tag{11}$$

The direction $\mathbf{g}(\mathbf{w})$ can be considered as the steepest ascent one among the ones satisfying the constraint as much as possible. This is a reasonable search direction if the line search is applied. However, in order to keep the constraint strictly a curve search is constructed. It is easy to see that $\mathbf{g}(\mathbf{w})$ lies on the plane π spanned by $\mathbf{w}$ and $\partial f(\mathbf{w})$. Furthermore, due to (10), the intersection of the plane and the unit sphere yields a great circle

$$\mathbf{w}cos\theta + \mathbf{g}_0(\mathbf{w})sin\theta, \tag{12}$$

where $\mathbf{g}_0(\mathbf{w}) = \mathbf{g}(\mathbf{w})/||\mathbf{g}(\mathbf{w})||_2$ is the unit vector along the tangent of the great circle on the plane π. The schematic is plotted in Fig. 1. Obviously, for the great circle, when $\theta \to 0_+$, we have

$$f(\mathbf{w}cos\theta + \mathbf{g}_0(\mathbf{w})sin\theta) - f(\mathbf{w}) \approx \langle \mathbf{g}_0(\mathbf{w}), \partial f(\mathbf{w})\rangle sin\theta \geq 0.$$

Note that $\langle \mathbf{g}_0(\mathbf{w}), \partial f(\mathbf{w})\rangle$ can be considered as the increasing rate of the objective when moving from $\mathbf{w}$ along the great circle (12). Let $\mathbf{h}_0(\mathbf{w})$ be the unit vector along the tangent of any smooth curve at $\mathbf{w}$. It is not difficult to see that

$$\langle \mathbf{g}_0(\mathbf{w}), \partial f(\mathbf{w})\rangle \geq \langle \mathbf{h}_0(\mathbf{w}), \partial f(\mathbf{w})\rangle.$$

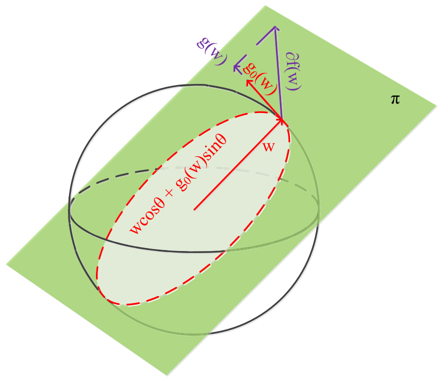


FIGURE 1. Schematic of search direction. The plane π is spanned by $\mathbf{w}$ and $\partial f(\mathbf{w})$. Project $\partial f(\mathbf{w})$ onto the tangent plane of the unit sphere at $\mathbf{w}$ as $\mathbf{g}(\mathbf{w})$ and normalize $\mathbf{g}(\mathbf{w})$ as $\mathbf{g}_0(\mathbf{w})$. $\mathbf{w}$ and $\mathbf{g}_0(\mathbf{w})$ yields the red circle $\mathbf{w}cos\theta + \mathbf{g}_0(\mathbf{w})sin\theta$.

This means that moving along the great circle (12) is the path of the steepest ascent on the unit sphere. So the search along the above circle (12) is selected. More exactly, $\theta = 0$ corresponds to the initial point $\mathbf{w}$, and increasing θ corresponds to moving on the circle. The Armijo-type rule is used to adjust its value here. Note that the basic idea of the algorithm is followed from [35]. Its reasonability and efficiency have also been discussed there.

Algorithm 1 Algorithm for Solving (8)

Input: The image matrices $\mathbf{X}_i \in \mathbb{R}^{m\times n}$, $i = 1, \ldots, N$, the parameter a of Tℓ_1 criterion.
Output: The projection vector $\mathbf{w}$.
Initialize: $\mathbf{w}(0) \in \mathbb{R}^{n\times 1}$ satisfying $\mathbf{w}(0)^T\mathbf{w}(0) = 1$, $\theta(0) \in (0, \pi/2]$.
$t \leftarrow 0$.
Repeat:
Compute the subgradient $\partial f(\mathbf{w}(t))$ by (9);
Project $\partial f(\mathbf{w}(t))$ onto the tangent plane of the unit sphere at $\mathbf{w}(t)$ as $\mathbf{g}(\mathbf{w}(t)) = \partial f(\mathbf{w}(t)) - \langle \partial f(\mathbf{w}(t)), \mathbf{w}(t)\rangle \mathbf{w}(t)$;
Normalize $\mathbf{g}(\mathbf{w}(t))$ as $\mathbf{g}_0(\mathbf{w}(t)) = \mathbf{g}(\mathbf{w}(t))/||\mathbf{g}(\mathbf{w}(t))||_2$;
Update $\mathbf{w}(t+1) = \mathbf{w}(t)\cos\theta(t) + \mathbf{g}_0(\mathbf{w}(t))\sin\theta(t)$,
Repeat:
$\theta(t) \leftarrow \theta(t)/2$,
Until $f(\mathbf{w}(t+1)) \geq f(\mathbf{w}(t))$;
Update $\theta(t+1) = \min(2\theta(t), \pi/2)$;
Until convergence

2) GREEDY STRATEGY

By implementing Algorithm 1, we can obtain the first projection vector $\mathbf{w}_1$ directly. To get more than one projection vector, greedy search is applied here. Suppose the first $j-1$ orthonormal projection vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}$ have been obtained, to compute $\mathbf{w}_j$ for $j > 1$, we use the deflation technique to extract it, the training samples have to be updated

$$\mathbf{X}_i^j = \mathbf{X}_i^{j-1} - \mathbf{X}_i^{j-1}\mathbf{w}_{j-1}\mathbf{w}_{j-1}^T, \tag{13}$$

with $\mathbf{X}_i^0 = \mathbf{X}_i \in \mathbb{R}^{m\times n}$ and $\mathbf{w}_0 = \mathbf{0} \in \mathbb{R}^n$, $i = 1, \ldots, N, j = 1, \ldots, d$. (13) means that $\mathbf{X}_i^j$ are computed such that the information contained in the previously obtained projection vectors is deducted. Obviously, $\mathbf{w}_j$ is a unit vector according to Algorithm 1. And as proved in [36], $\mathbf{w}_j$ is orthogonal to $\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}$. Thus $\mathbf{W}$ is an orthonormal matrix, i.e. $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. Algorithm 2 shows the details of the greedy strategy.

Algorithm 2 2DPCA-Tℓ_1

Input: The image matrices $\mathbf{X}_i \in \mathbb{R}^{m\times n}$, $i = 1, \ldots, N$, the parameter a of Tℓ_1 criterion, and the number d of projection vectors.
Output: The projection matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d]$.
Initialize: $\mathbf{W} \leftarrow \varnothing$, $\mathbf{w}_0 = \mathbf{0} \in \mathbb{R}^n$, $\{\mathbf{X}_i^0 \leftarrow \mathbf{X}_i\}_{i=1}^N$.
$j \leftarrow 1$.
Repeat:
Compute $\{\mathbf{X}_i^j\}_{i=1}^N$ according to (13);
Apply Algorithm 1 to $\{\mathbf{X}_i^j\}_{i=1}^N$ and get $\mathbf{w}_j$;
Update $\mathbf{W} \leftarrow [\mathbf{W}, \mathbf{w}_j]$;
Until $j = d$

IV. EXPERIMENTS

In this section, we evaluate the performance of 2DPCA-Tℓ_1 on three human face databases Yale [37], ORL [38], Jaffe [39] and one object database COIL-20 [40], where the block noise with black and white dots is added to examine the robustness. We compare our method with classical 2DPCA [14], 2DPCA-L1 [15], 2DPCAL1-S [17], N-2-DPCA [24], G2DPCA [18] and Angle-2DPCA [20] in the task of 2D image dimensionality reduction and classification, where 1-nearest neighbor (1-NN) is used for classifying. Among the above methods, 2DPCA-Tℓ_1, 2DPCAL1-S and G2DPCA depend on the selection of parameters. For 2DPCA-Tℓ_1, we tune a from $\{100, 50, 10, 1, 0.5, 0.1, 0.05, 0.01, 0.001\}$. For 2DPCAL1-S, it has a positive tuning parameter λ. In [17], λ is in the range of [0.001, 1000], so we search the optimal λ from $\{0.001, 0.02, 1, 10, 200, 500, 1000\}$. For G2DPCA, it depends on two parameters $s \geq 1$ and $p > 0$. Since all other methods have the ℓ_2-norm-based constraints, to be fair, we set $p = 2$ and $s = \{1.1, 1.3, 1.5, 1.7, 1.9\}$ in the following experiments. For 2DPCA-Tℓ_1, 2DPCAL1-S and G2DPCA, we employ the parameters with the best classification performance as their final ones, respectively. All the experiments are performed in MATLAB R2017a.

The Yale face database consists of 165 grayscale images of 15 individuals under different lighting conditions and facial expressions, these facial expressions include happy, normal, sad, sleepy, surprised, and wink. Each individual has 11 images. Each image in Yale database is reshaped into 32×32 pixels. 6 images of each individual are randomly selected for training, the others for testing. For these training images, the $i \times i$ ($i = 16, 20, 23$) block noise with black and white dots is added to them, and the location of this block is random for each image. Fig. 2(a) shows some original and noised images from this database. Our method and the aforementioned six methods are employed to extract low-dimensional representations, respectively. Then 1-NN is used for classification. This process is repeated ten times.

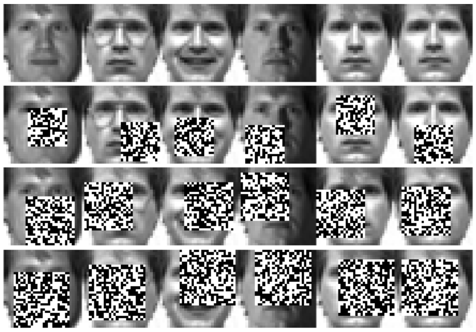
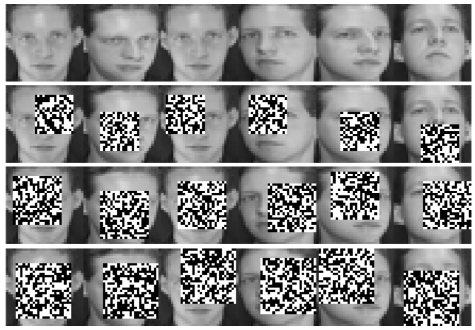
The ORL face database contains 40 individuals, each individual contains 10 images. For some individuals, the images are taken at different times, varying facial expressions and lighting conditions. Here we resize each image to 32×32 pixels. Then we randomly select 6 images per person for training, adding randomly $i \times i$ ($i = 16, 20, 23$) black and white noise to them, and the rest for testing. Some images with and without noise are shown in Fig. 2(b). 2DPCA, 2DPCA-L1, 2DPCAL1-S, N-2-DPCA, G2DPCA, Angle-2DPCA and our 2DPCA-Tℓ_1 are respectively to extract features, and 1-NN is used for classification. We repeat this process 10 times.

The Jaffe database contains 213 images of 7 facial expressions posed by 10 Japanese female individuals. Each image is resized to 32×32 pixels. We randomly choose 70% of each individual's images for training and the remainders for testing. Like Yale and ORL database, the same noise is added to the training images. Some original and noised images from Jaffe database are shown in Fig. 2(c). 2DPCA, 2DPCA-L1, 2DPCAL1-S, N-2-DPCA, G2DPCA, Angle-2DPCA and 2DPCA-Tℓ_1 are applied to extract features. Based on the extracted features, we compute the 1-NN classification accuracy. We do this process ten times to evaluate performance of each method.

Columbia Object Image Library (COIL-20) consists of 20 objects. While each object is rotated through 360 degrees on a turntable, its images are taken at pose intervals of 5 degrees with a color camera. So each object has 72 images and COIL-20 database contains 1440 images. To reduce the computational time, we transform color images into grayscale images and crop them to 64×64 pixels. 46 images of each object are randomly selected for training, the remainders for testing. We add $i \times i$ ($i = 24, 36, 48$) block noise with black and white dots to all training images. Several samples are shown in Fig. 2(d). Then we employ 2DPCA, 2DPCA-L1, 2DPCAL1-S, N-2-DPCA, G2DPCA, Angle-2DPCA and our 2DPCA-Tℓ_1 to extract low-dimensional features and compute classification accuracy by 1-NN. This process is repeated ten times to evaluate performance of each method.

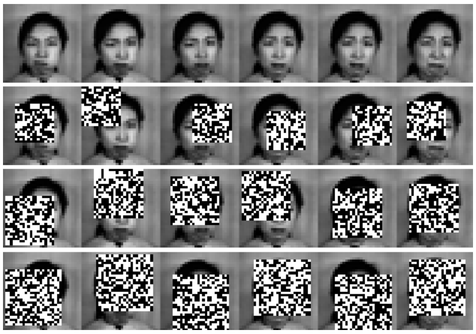
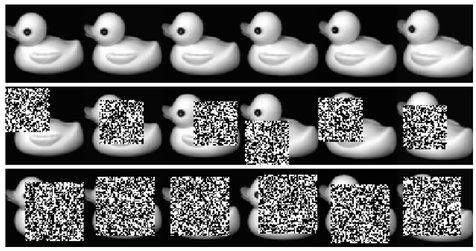
A. PARAMETER SELECTION

2DPCA-Tℓ_1 has a parameter a required to be optimal. a controls the shape of Tℓ_1 criterion. In order to find optimal a, for every a, we compute the corresponding average classification accuracy with the different dimensions of reduced space on each database. Based on the performance of the average classification accuracy, we choose the parameter a with the best performance as the optimal parameter. Tables 1, 2, 3, and 4 list the optimal parameters a of Yale, ORL, Jaffe, and COIL-20 database under different noise intensities, respectively. For 2DPCAL1-S and G2DPCA, they also have parameters to be

(a) The first row is the original images. The second row is 16×16 noised images. The third row is 20×20 noised images. The fourth row is 23×23 noised images

(b) The first row is the original images. The second row is 16×16 noised images. The third row is 20×20 noised images. The fourth row is 23×23 noised images

(c) The first row is the original images. The second row is 16×16 noised images. The third row is 20×20 noised images. The fourth row is 23×23 noised images

(d) The first row is the original images. The second row is 36×36 noised images. The third row is 48×48 noised images.

FIGURE 2. Sample images from four databases. (a). Yale. (b). ORL. (c). Jaffe. (d). COIL-20.

TABLE 1. Optimal parameters on Yale database under different noise intensities for 2DPCAL1-S, G2DPCA, and 2DPCA-Tℓ_1.

Method	Optimal parameters		
	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	$\lambda = 10$	$s = 1.1$	$a = 50$
With 16×16 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.05$
With 20×20 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$
With 23×23 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$

TABLE 2. Optimal parameters on ORL database under different noise intensities for 2DPCAL1-S, G2DPCA, and 2DPCA-Tℓ_1.

Method	Optimal parameters		
	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	$\lambda = 0.001$	$s = 1.1$	$a = 0.05$
With 16×16 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$
With 20×20 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$
With 23×23 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$

optimized and the method of selecting optimal parameters is similar to 2DPCA-Tℓ_1. Their optimal parameters are also respectively listed in Tables 1-4.

From Tables 1-4, it can be seen that the optimal parameter a of 2DPCA-Tℓ_1 is relatively small, especially for the data with noise. The reason may be that the upper bound of Tℓ_1 criterion is also small when a is small, making 2DPCA-Tℓ_1 of stronger robustness to noise than other 2DPCAs. Empirically, the value of parameter a is between 0.01 and 1 for noised data in most cases.

TABLE 3. Optimal parameters on Jaffe database under different noise intensities for 2DPCAL1-S, G2DPCA, and 2DPCA-Tℓ_1.

Method	Optimal parameters		
	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	$\lambda = 0.02$	$s = 1.1$	$a = 100$
With 16×16 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$
With 20×20 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$
With 23×23 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.05$

TABLE 4. Optimal parameters on COIL-20 database under different noise intensities for 2DPCAL1-S, G2DPCA, and 2DPCA-Tℓ_1.

Method	Optimal parameters		
	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	$\lambda = 10$	$s = 1.1$	$a = 0.001$
With 36×36 block noise	$\lambda = 0.001$	$s = 1.1$	$a = 0.01$
With 48×48 block noise	$\lambda = 0.001$	$s = 1.5$	$a = 0.05$

B. CLASSIFICATION COMPARISON

In this subsection, we compare the performance of our 2DPCA-Tℓ_1 with classical 2DPCA, 2DPCA-L1, 2DPCAL1-S, N-2-DPCA, G2DPCA, and Angle-2DPCA on Yale, ORL, Jaffe, and COIL-20 database.

Under the optimal parameters of Tables 1-4, Figs. 3, 4, 5, and 6 plot the average classification accuracy curves versus the dimension of reduced space on Yale, ORL, Jaffe and COIL-20 database, respectively. Tables 5, 6, 7, and 8 list the average classification accuracy of each method under the optimal dimension (i.e., the dimension corresponding to

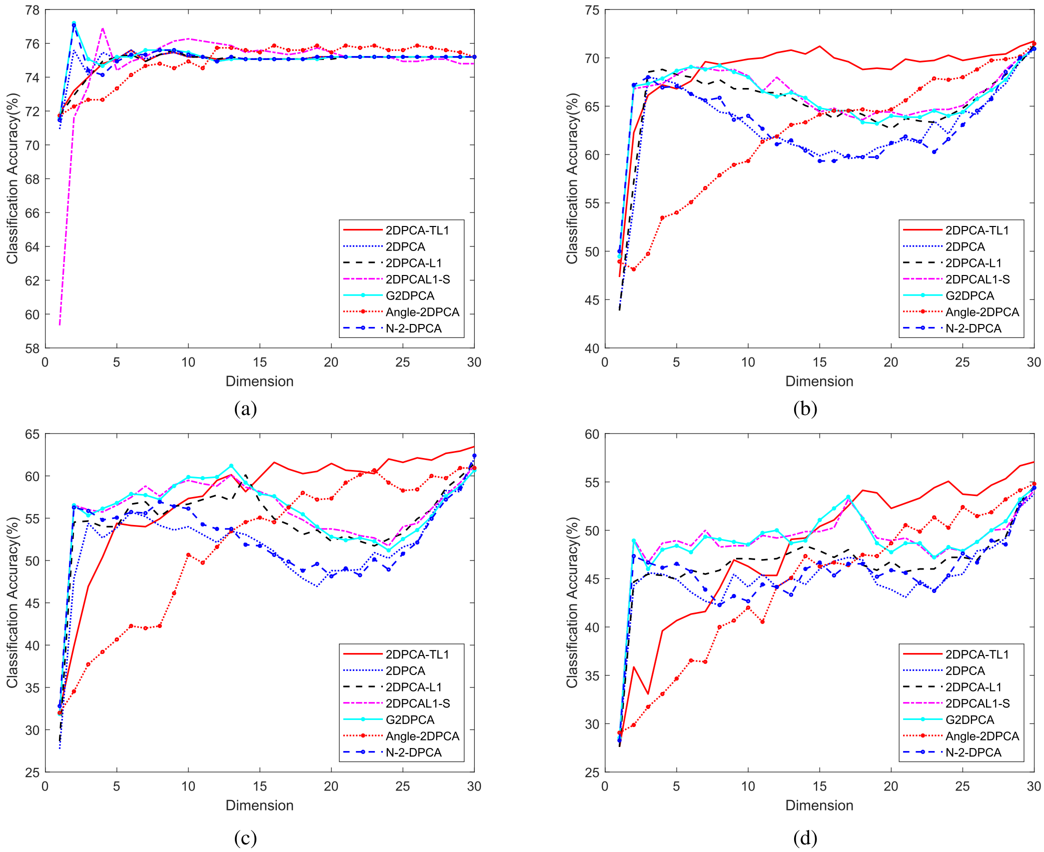


FIGURE 3. Average classification accuracy vs. dimension on Yale database with different noise intensities. (a) Original data. (b) 16 × 16 black and white noise. (c) 20 × 20 black and white noise. (d) 23 × 23 black and white noise.

TABLE 5. The average classification accuracies of Yale database under the optimal dimension.

Method	Accuracy(%)						
	2DPCA	2DPCAL1	Angle-2DPCA	N-2DPCA	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	75.6	75.6	75.86	77.06	76.93	**77.2**	75.6
With 16 × 16 block noise	70.93	71.33	71.46	70.93	70.93	70.93	**71.73**
With 20 × 20 block noise	62.13	61.46	60.93	62.4	61.2	61.2	**63.46**
With 23 × 23 block noise	53.73	54.53	54.8	54.4	54.13	54.4	**57.06**

TABLE 6. The average classification accuracies of ORL database under the optimal dimension.

Method	Accuracy(%)						
	2DPCA	2DPCAL1	Angle-2DPCA	N-2DPCA	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	96.62	96.62	96.62	96.62	96.62	96.62	**96.62**
With 16 × 16 block noise	85.31	85.06	85.43	85.43	85.43	85.56	**86.12**
With 20 × 20 block noise	72.31	72.06	70.81	72.37	71.81	72.12	**73.56**
With 23 × 23 block noise	53.87	55.87	50.06	54.5	55.37	55.87	**58.68**

TABLE 7. The average classification accuracies of Jaffe database under the optimal dimension.

Method	Accuracy(%)						
	2DPCA	2DPCAL1	Angle-2DPCA	N-2DPCA	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	99.37	99.53	99.06	99.37	99.37	99.37	**99.53**
With 16 × 16 block noise	91.4	92.34	91.56	92.03	92.81	92.81	**93.12**
With 20 × 20 block noise	74.84	74.84	73.28	74.53	75.15	74.68	**81.25**
With 23 × 23 block noise	48.75	49.37	49.06	47.5	49.37	49.06	**51.87**

the highest accuracy). In addition, it is easy to see that for all databases, the greater the noise intensity, the lower the classification accuracy, which is consistent with common sense.

From Figs. 3(a), 4(a), 5(a) and 6(a), for original databases (i.e. noise-free databases), the curves of all methods are relatively concentrated. The reason is that the extracted features of all methods tend to be similar when the data is noise-free,

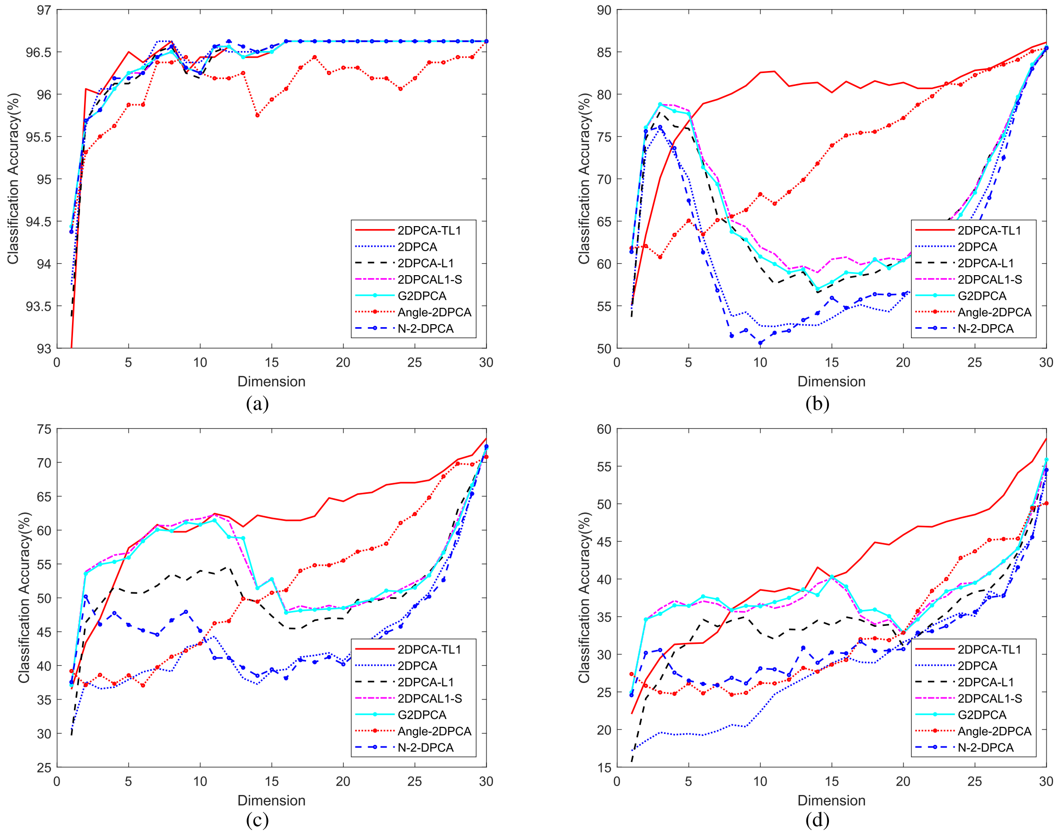


FIGURE 4. Average classification accuracy vs. dimension on ORL database with different noise intensities. (a) Original data. (b) 16×16 black and white noise. (c) 20×20 black and white noise. (d) 23×23 black and white noise.

TABLE 8. The average classification accuracies of COIL-20 database under the optimal dimension.

Method	Accuracy(%)						
	2DPCA	2DPCAL1	Angle-2DPCA	N-2DPCA	2DPCAL1-S	G2DPCA	2DPCA-Tℓ_1
Original data	99.82	99.82	99.61	99.82	99.63	**99.82**	99.67
With 36×36 block noise	89.57	90.03	89.34	89.46	89.9	89.92	**90.11**
With 48×48 block noise	73.01	74.01	72.44	72.17	73.82	73.61	**74.8**

leading to the relatively concentrated classification accuracy. The results of original data in Tables 5-8 also verify this view to some extent because the average classification accuracies of each method under the optimal dimension are also close. Although the performance of 2DPCA-Tℓ_1 is not always the best on original databases, but it is still relatively better than some methods.

Then, we investigate the robustness of our 2DPCA-ℓ_1 to noise. To see this, we compare the average classification accuracies with the different dimensions of reduced space on Yale database, ORL database, Jaffe database and COIL-200 database, with $i \times i$ black and white noise, as plotted in Figs. 3(b)-(d), 4(b)-(d), 5(b)-(d), and 6(b)-(c). Here $i = 16, 20, 23$ for the Yale, ORL, and Jaffe database, and $i = 36, 48$ for the COIL-20 database. From Figs. 6(b)-(c), on COIL-20 database, 2DPCA-ℓ_1 is only slightly better than 2DPCA-L1 which is the best one among the other six methods. The reason may be that the features of different objects in this database are quite different, the classification accuracies have no significant difference based on the extracted features of different methods. However, from Figs. 3(b)-(d), 4(b)-(d), and 5(b)-(d), 2DPCA-ℓ_1 is significantly better than other methods on Yale, ORL and, Jaffe database. Overall, our 2DPCA-ℓ_1 is superior to the other six methods on all noised databases, especially for Yale, ORL, and Jaffe database. This may be because that Tℓ_1 criterion is more robust due to its boundedness and Lipschitz-continuity. Combined with Tables 5-8, we can see that, under the optimal dimension, our 2DPCA-Tℓ_1 also outperforms the other six methods on all the noised databases. In most cases, the accuracy of 2DPCA-Tℓ_1 is at most 5% higher than that of classical 2DPCA. Compared with 2DPCA-L1, 2DPCAL1-S, N-2-DPCA, G2DPCA, and Angle-2DPCA, the accuracy of our method is 1% to 3% higher than theirs. At the same time, it is easy to see that the greater the noise intensity, the more obvious the advantage of 2DPCA-ℓ_1.

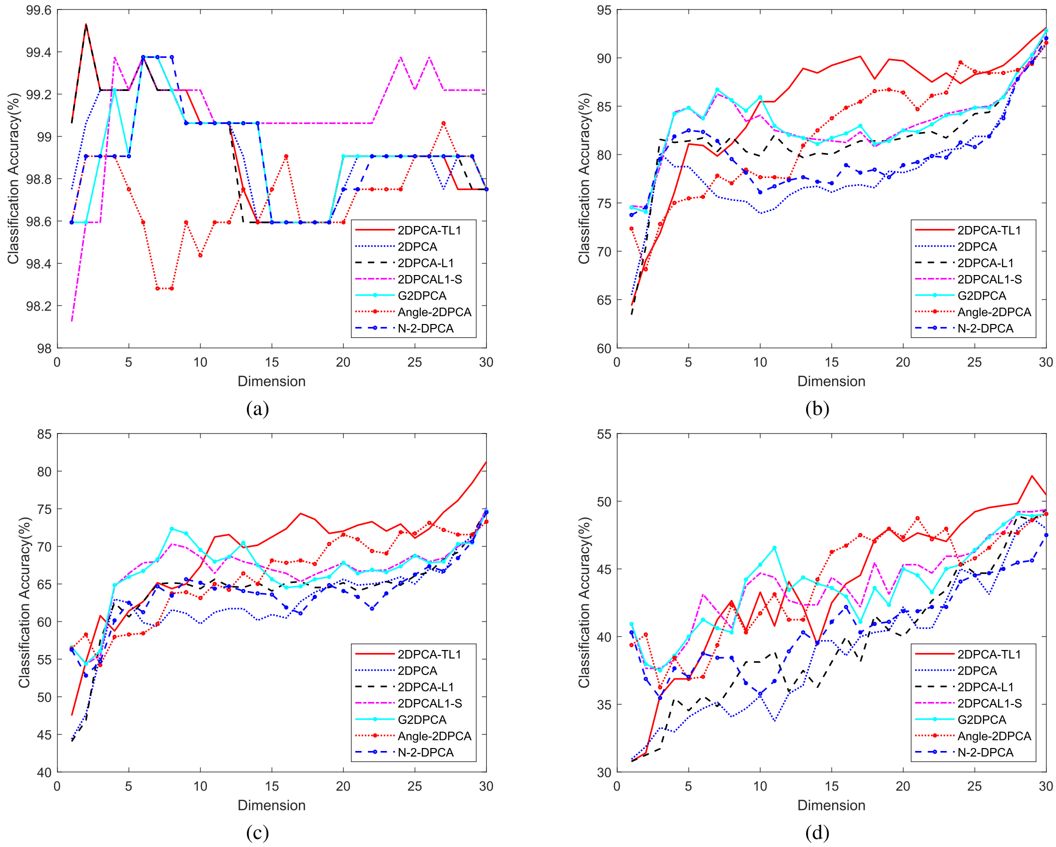


FIGURE 5. Average classification accuracy vs. dimension on Jaffe database with different noise intensities. (a) Original data. (b) 16×16 black and white noise. (c) 20×20 black and white noise. (d) 23×23 black and white noise.

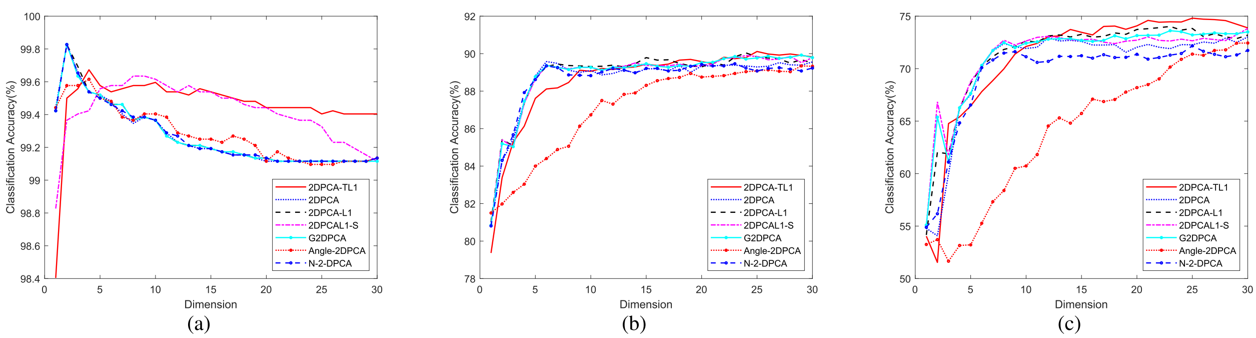
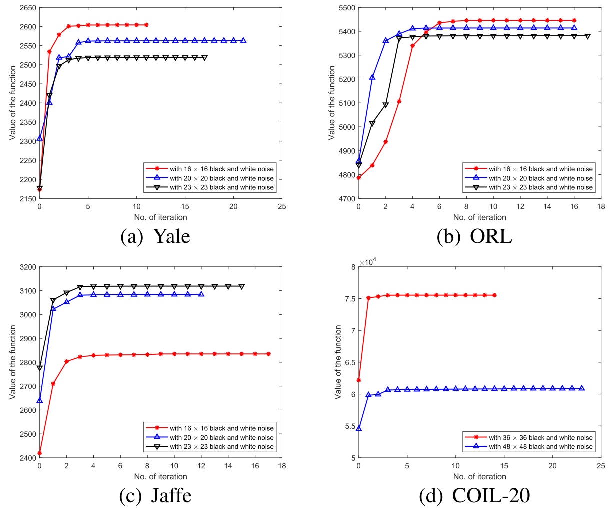


FIGURE 6. Average classification accuracy vs. dimension on COIL-20 database with different noise intensities. (a) Original data. (b) 36×36 black and white noise. (c) 48×48 black and white noise.

It is also worth mentioning that classical 2DPCA, 2DPCA-L1, 2DPCAL1-S, N-2-DPCA, and G2DPCA are vulnerable to the variation of dimensions, and the classification accuracy may descend as dimensions increase. For example, Figs. 3(b)-(c), Figs. 4(b)-(c), and Fig. 6(a) are all in this situation. We speculate the reason for this phenomenon is that when the reduced dimension is higher than a certian dimension, some useless or disturbing information may also be contained, causing negative effects. However, our 2DPCA-Tℓ_1 is stable to the variation of dimensions with a basic uptrend along with the dimensions.

C. CONVERGENCE EXPERIMENTS

At last, to observe the convergence of 2DPCA-Tℓ_1, we test the variations of the objective function (8) under different noise intensities on Yale, ORL, Jaffe, and COIL-20 database. Fig. 7 shows the convergence of the objective functions along with the number of iteration. It is easy to see that these objective functions are non-decreasing functions of iterations. And the objective function of 2DPCA-Tℓ_1 can converge quickly, generally within about 25 steps. This shows the stability of our 2DPCA-Tℓ_1.



(a) Yale (b) ORL (c) Jaffe (d) COIL-20

FIGURE 7. Variation of objective function value along the number of iteration for 2DPCA-Tℓ_1 on Yale, ORL, Jaffe, and COIL-20 database with different noise intensities.

V. CONCLUSION

A novel two-dimensional principal component, 2DPCA-Tℓ_1, is proposed. Compared with the existing two-dimensional PCAs, our method employs Tℓ_1 criterion as the distance metric. The main difference between Tℓ_1 criterion and ℓ_2-norm, ℓ_1-norm, ℓ_p-norm is that Tℓ_1 criterion is bounded and Lipschitz-continuous. The above two properties imply that Tℓ_1 criterion is more robust, resulting in making our 2DPCA-Tℓ_1 less affected by noise remarkably. To solve the optimization problem required by our 2DPCA-Tℓ_1, an modified gradient ascent algorithm is provided. Experimental results on several real databases have shown the effectiveness and advantages of our method

REFERENCES

[1] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Edu. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[2] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.

[3] A. Baccini, P. Besse, and A. D. Falguerolles, "A L_1-norm PCA and a heuristic approach," in *Ordinal and Symbolic Data Analysis*, E. Diday, Y. Lechevalier, and O. Opitz, Eds. New York, NY, USA: Springer, 1996, pp. 359–368.

[4] C. Ding, D. Zhou, X.-F. He, and H.-Y. Zha, "R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Internal Conf. Mach. Learn.*, 2006, pp. 281–288.

[5] N. Kwak, "Principal component analysis based on ℓ_1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[6] N. Kwak, "Principal component analysis by L_p-norm maximization," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 594–609, May 2014.

[7] C. Kim and D. Klabjan, "A simple and fast algorithm for L1-norm kernel PCA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1842–1855, Aug. 2020.

[8] J. Fan and T. W. S. Chow, "Exactly robust kernel principal component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 749–761, Mar. 2020.

[9] E. J. Candes, X. D. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, May 2011.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.

[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York, NY, USA: Academic, 1991.

[12] X. F. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, Dec. 2003, pp. 153–160.

[13] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[14] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional pca: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[15] X. Li, Y. Pang, and Y. Yuan, "L1-norm-based 2DPCA," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1170–1175, Aug. 2010.

[16] R. Wang, F. P. Nie, X. J. Yang, F. F. Gao, and M. L. Yao, "Robust 2DPCA with non-greedy ℓ_1-norm maximization for image analysis," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1108–1112, May 2015.

[17] H. Wang and J. Wang, "2DPCA with L1-norm for simultaneously robust and sparse modelling," *Neural Netw.*, vol. 46, pp. 190–198, Oct. 2013.

[18] J. Wang, "Generalized 2-D principal component analysis by Lp-norm for image analysis," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 792–803, Mar. 2016.

[19] Q. Wang and Q. Gao, "Robust 2DPCA and its application," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1152–1158.

[20] Q. Gao, L. Ma, Y. Liu, X. Gao, and F. Nie, "Angle 2DPCA: A new formulation for 2DPCA," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1672–1678, May 2018.

[21] T. Li, M. Li, Q. Gao, and D. Xie, "F-norm distance metric based robust 2DPCA and face recognition," *Neural Netw.*, vol. 94, pp. 204–211, Oct. 2017.

[22] Q. Wang, Q. Gao, X. Gao, and F. Nie, "Optimal mean two-dimensional principal component analysis with F-norm minimization," *Pattern Recognit.*, vol. 68, pp. 286–294, Aug. 2017.

[23] Q. X. Gao, S. Xu, F. Chen, C. Ding, X. B. Gao, and Y. S. Li, "R-1-2-DPCA and face recognition," *IEEE Trans. Cybern.*, vol. 49, no. 4, pp. 1212–1223, Apr. 2019.

[24] F. Zhang, J. Yang, J. Qian, and Y. Xu, "Nuclear norm-based 2-DPCA for extracting features from images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2247–2260, Oct. 2015.

[25] H. Du, Q. Hu, M. Jiang, and F. Zhang, "Two-dimensional principal component analysis based on Schatten p-norm for image feature extraction," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 55–62, Oct. 2015.

[26] G. Zhou, G. Xu, J. Hao, S. Chen, J. Xu, and X. Zheng, "Generalized centered 2-D principal component analysis," *IEEE Trans. Cybern.*, early access, Aug. 19, 2020, doi: 10.1109/TCYB.2019.2931957.

[27] M. Nikolova, "Local strong homogeneity of a regularized estimator," *SIAM J. Appl. Math.*, vol. 61, no. 2, pp. 633–658, Jan. 2000.

[28] J. Lv and Y. Fan, "A unified approach to model selection and sparse recovery using regularized least squares," *Ann. Statist.*, vol. 37, no. 6A, pp. 3498–3528, Dec. 2009.

[29] R. Ma, J. Miao, L. Niu, and P. Zhang, "Transformed ℓ_1 regularization for learning sparse deep neural networks," *Neural Netw.*, vol. 119, pp. 286–298, Nov. 2019.

[30] S. Zhang and J. Xin, "Minimization of transformed L_1 penalty: Theory, difference of convex function algorithm, and robust application in compressed sensing," *Math. Program.*, vol. 169, no. 1, pp. 307–336, May 2018.

[31] S. Zhang and J. Xin, "Minimization of transformed L_1 penalty: Closed form representation and iterative thresholding algorithms," *Commun. Math. Sci.*, vol. 15, no. 2, pp. 511–537, 2017.

[32] S. Zhang, P. Yin, and J. Xin, "Transformed Schatten-1 iterative thresholding algorithms for low rank matrix completion," *Commun. Math. Sci.*, vol. 15, no. 3, pp. 839–862, 2017.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[34] P. Jain and P. Kar, "Non-convex optimization for machine learning," *Found. Trends Mach. Learn.*, vol. 10, nos. 3–4, pp. 142–363, Dec. 2017.

[35] M. Yu, L. Shao, X. Zhen, and X. He, "Local feature discriminant projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1908–1914, Sep. 2016.

[36] F. Zhong and J. Zhang, "Linear discriminant analysis based on L1-norm mazimization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.

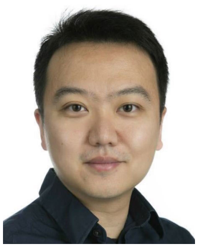
[37] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[38] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, Dec. 1994, pp. 138–142.

[39] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[40] S. Nene, S. Nayar, and H. Murase, "Columbia object image library," Tech. Rep. CUCS-006-96, 1996.

XIANGFEI YANG received the M.S. degree from the Tianjin University of Commerce, Tianjin, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Statistics, Capital University of Economics and Business, Beijing, China. His research interests include statistical machine learning and non-convex optimization.



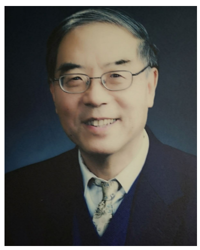
WENSI WANG (Member, IEEE) received the B.S. degree from the Beijing University of Technology, China, in 2005, and the M.Sc. and Ph.D. degrees in microelectronics from the University College Cork, Ireland, in 2007 and 2012, respectively. From 2012 to 2015, he worked as a Postdoctoral Researcher and a Research Engineer with the Tyndall National Institute, Ireland. His research interests include data analysis in the Internet of Things (IoT) and medical electronics.

LIMING LIU received the Ph.D. degree from the Department of Mathematics, China Agricultural University, Beijing, China, in 1999. She is currently a Professor with the School of Statistics, Capital University of Economics and Business. Her research interests include mathematical statistics and optimization methods.

YUANHAI SHAO received the B.S. degree in information and computing science from the College of Mathematics, Jilin University, Changchun, China, in 2006, and the master's degree in applied mathematics and the Ph.D. degree in operations research and management from the College of Science, China Agricultural University, Beijing, China, in 2008 and 2011, respectively.

He is currently a Professor with the School of Management, Hainan University, Haikou, China. His current research interests include support vector machines, nonparallel support vector machines, data mining, machine learning, and optimization methods. He has authored or coauthored over 100 refereed articles on these areas.

LITING ZHANG received the B.S. degree in electronics and information engineering from Chengdu University, Chengdu, China, in 2019. She is currently pursuing the M.S. degree with the Faculty of Information Technology, College of Electronic Science and Technology, Beijing University of Technology. Her research interests include data analysis in environmental protection field and machine learning.



NAIYANG DENG received the M.Sc. degree from the Department of Mathematics, Peking University, China, in 1967. He was a Professor with the College of Science, China Agricultural University. He has published over 100 articles. His research interests include operational research, optimization, machine learning, and data mining. He was an Honorary Director of the China Operations Research Society, a Managing Editor of the *Journal of Operational Research*, and an Abstract Editor of the *International Journal of Operations Research*.