Generalized 2-D Principal Component Analysis by Lp-Norm for Image Analysis

Jing Wang

***Abstract*—This paper proposes a generalized 2-D principal component analysis (G2DPCA) by replacing the L2-norm in conventional 2-D principal component analysis (2DPCA) with Lp-norm, both in objective and constraint functions. It is a generalization of previously proposed robust or sparse 2DPCA algorithms. Under the framework of minorization–maximization, we design an iterative algorithm to solve the optimization problem of G2DPCA. A closed-form solution could be obtained in each iteration. Then a deflating scheme is employed to generate multiple projection vectors. Our algorithm guarantees to find a locally optimal solution for G2DPCA. The effectiveness of the proposed method is experimentally verified.**

***Index Terms*—Convex maximization, generalized 2-D principal component analysis (G2DPCA), image analysis, Lp-norm, minorization–maximization (MM).**

I. Introduction

PRINCIPAL component analysis (PCA) [1], [2] has been widely applied in dimensionality reduction, signal reconstruction, and pattern classification. However, its quadratic formulation renders it vulnerable to noises. This problem facilitates many robust PCA algorithms which utilize L1-norm on the objective function, e.g., L1-PCA [3], R1-PCA [4], and PCA-L1 [5]. Besides robustness, sparsity is also a desired property [6]. By applying L0- or L1-norm on the constraint function of PCA, sparsity could be introduced, resulting in a series of sparse PCA (SPCA) algorithms [7]–[10]. A newly proposed algorithm called robust SPCA (RSPCA) [11] applies L1-norm both in objective and constraint functions of PCA, inheriting the merits of robustness and sparsity.

Considering that L0-, L1-, and L2-norm are all special cases of Lp-norm, it is natural to replace the L2-norm in traditional PCA with arbitrary norm, as proposed in PCA-Lp [12] and generalized PCA (GPCA) [13]. In PCA-Lp, the Lp-norm is imposed on the objective function of PCA. A greedy solution based on a gradient ascent method or a Lagrangian multiplier method and a nongreedy solution based on a Lagrangian multiplier method are proposed to solve PCA-Lp [12]. In GPCA, the Lp-norm is imposed both in objective and constraint functions of PCA. The successive linearization technique (SLT) [14], [15] is employed to solve GPCA [13].

When applying the above robust and sparse PCA algorithms in image analysis, each image should be reshaped into a long vector in prior. In this way, the spatial information in images is destroyed and extensive computations are usually inevitable due to high dimensionality of reshaped images [16]. Image-as-matrix methods represented by 2-D PCA (2DPCA) [16] offer insights for improving the above robust PCA and SPCA algorithms. Two related improvements are L1-norm-based 2DPCA (2DPCA-L1) [17] and 2DPCA-L1 with sparsity (2DPCAL1-S) [18], corresponding to the 2-D cases of PCA-L1 and RSPCA, respectively.

This paper proposes a generalized 2DPCA (G2DPCA) by replacing the L2-norm of conventional 2DPCA with Lp-norm, on both objective and constraint functions, thus greatly extending previous 2DPCA-based algorithms. The proposed algorithm is greatly enlightened by GPCA [13]. Besides the image-as-matrix representation, G2DPCA differs from GPCA mainly by designing an elegant solution under the framework of minorization–maximization (MM) [19] rather than SLT. MM theoretically guarantees to find a locally optimal solution for an optimization problem while SLT intends to linearize a nonsmooth problem, thus MM is more stronger than SLT.

The remainder of this paper is organized as follows. In Section II, some robust and sparse 2DPCA algorithms are reviewed and the G2DPCA algorithm is proposed. Section III introduces the techniques that would be used to solve G2DPCA. Section IV provides the solution of G2DPCA. Section V reports experimental results. Section VI concludes this paper.

II. Robust and Sparse 2DPCA Algorithms

The notations in this paper are described as follows. Lowercase letters denote scalars, boldface lowercase letters denote vectors, boldface uppercase letters denote matrices; $\mathrm{sign}(\cdot)$ denotes the sign function; $|\cdot|$ denotes the absolute value; $\mathbf{w} \circ \mathbf{v}$ denotes the Hadamard product, i.e., the element-wise product between two vectors; $|\mathbf{w}|^p$ denotes the element-wise power of the absolute value of a vector; $\mathrm{diag}(\mathbf{w})$ denotes a square and diagonal matrix by putting the elements of $\mathbf{w}$ on the main diagonal; $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_p$, and $\|\cdot\|_F$ denote L1-, L2-, Lp-, and Frobenius-norm, respectively. Note that the sign function and the absolute value function could

Manuscript received July 13, 2014; revised November 21, 2014 and March 7, 2015; accepted March 21, 2015. Date of publication April 16, 2015; date of current version February 12, 2016. This paper was recommended by Associate Editor B. Zhang.

The author is with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing 210096, China (e-mail: wangjing0@seu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2015.2416274

be applied on a scalar or on a vector in the element-wise manner.

Traditionally, robust and sparse 2DPCA algorithms focus on finding a single projection vector each time, then a deflation scheme [20] is implemented to extract multiple projection vectors [17], [18]. This strategy is also adopted in this paper.

A. 2DPCA

Suppose there are n training image samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i \in \mathbb{R}^{h\times w}$, $i = 1, 2, \ldots, n$. h and w are the height and width of the images, respectively. The images are assumed to be mean-centered, i.e., $1/n \sum_{i=1}^{n} \mathbf{X}_i = \mathbf{0}$. 2DPCA [16] finds its first projection vector $\mathbf{w} \in \mathbb{R}^w$ by solving the following optimization problem:

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_2^2, \quad \text{s.t. } \|\mathbf{w}\|_2^2 = 1. \tag{1}$$

The projection vector $\mathbf{w}$ could be obtained by calculating the eigen decomposition of an image covariance matrix and selecting the eigenvector with the largest eigenvalue.

B. 2DPCA-L1

2DPCA-L1 [17] could be formulated by replacing the L2-norm in the objective function of 2DPCA with L1-norm. That is, 2DPCA-L1 finds its first projection vector by solving the problem

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_1, \quad \text{s.t. } \|\mathbf{w}\|_2^2 = 1. \tag{2}$$

The projection vector $\mathbf{w}$ could be calculated by an iterative algorithm. Let k be the iteration number, $\mathbf{w}^k$ be the projection vector at the kth step, then $\mathbf{w}$ could be updated by

$$\mathbf{v}^k = \sum_{i=1}^{n} \mathbf{X}_i^T \operatorname{sign}\left(\mathbf{X}_i\mathbf{w}^k\right) \tag{3}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{v}^k}{\|\mathbf{v}^k\|_2}. \tag{4}$$

C. 2DPCAL1-S

2DPCAL1-S [18] could be formulated by applying L1-norm both in objective and constraint functions of 2DPCA as follows:

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_1, \quad \text{s.t. } \|\mathbf{w}\|_1 \le c, \|\mathbf{w}\|_2^2 = 1 \tag{5}$$

where c is a positive constant. The projection vector $\mathbf{w}$ could be updated iteratively by

$$\mathbf{v}^k = \sum_{i=1}^{n} \mathbf{X}_i^T \operatorname{sign}\left(\mathbf{X}_i\mathbf{w}^k\right) \tag{6}$$

$$u_i^k = v_i^k \frac{\left|w_i^k\right|}{\lambda + \left|w_i^k\right|}, \ i = 1, 2, \ldots, w \tag{7}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_2} \tag{8}$$

where $\mathbf{u}^k \in \mathbb{R}^w$ is a vector; w_i^k, v_i^k, and u_i^k are the ith elements of $\mathbf{w}^k$, $\mathbf{v}^k$, and $\mathbf{u}^k$, respectively; λ is a positive scalar which serves as a tuning parameter in this algorithm. When λ is set to be zero, 2DPCAL1-S reduces to 2DPCA-L1.

Notice that w with a subscript is different from w without a subscript in this paper. The former one indicates an element in the projection vector $\mathbf{w}$ while the latter one indicates the image width.

D. G2DPCA

Inspired by the above robust and sparse 2DPCA algorithms, we propose the G2DPCA as follows:

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s, \quad \text{s.t. } \|\mathbf{w}\|_p^p = 1 \tag{9}$$

where $s \ge 1$ and $p > 0$. It is obvious that 2DPCA and 2DPCA-L1 are two special cases of G2DPCA. 2DPCAL1-S is unique, but it is closely related to G2DPCA. Intuitively, 2DPCAL1-S originates from G2DPCA with $s = 1$ and $p = 1$ which leads to a projection vector with only one nonzero element. Then the L2-norm constraint is employed to fix this problem, resulting in 2DPCAL1-S. On the other hand, G2DPCA with $s = 1$ and $1 < p < 2$ behaves like 2DPCAL1-S since the Lp-norm constraint in G2DPCA behaves like the mixed-norm constraint in 2DPCAL1-S.

Instead of trying different objective functions for G2DPCA as in [13], we limit our attention to the optimization problem in (9) because it is representative. Also, we want to check how the s value would affect the performance of G2DPCA in image reconstruction and classification.

After obtaining the first r projection vectors $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r]$ for 2DPCA, 2DPCA-L1, 2DPCAL1-S, or G2DPCA, $1 \le r < w$, the $(r+1)$th projection vector $\mathbf{w}_{r+1}$ could be calculated similarly on the deflated samples [20]

$$\mathbf{X}_i^{\text{deflated}} = \mathbf{X}_i\left(\mathbf{I} - \mathbf{W}\mathbf{W}^T\right), \ i = 1, 2, \ldots, n. \tag{10}$$

This deflation procedure is implemented repeatedly to extract multiple projection vectors.

III. Related Techniques

Before proceeding to the solution of G2DPCA problem, we will first introduce some related techniques that would be utilized.

A. MM Framework

Suppose $f(\mathbf{w})$ is the objective function to be maximized, under the MM framework [19], if there exists a surrogate function $g(\mathbf{w}|\mathbf{w}^k)$ which satisfies two key conditions that

$$f(\mathbf{w}^k) = g\left(\mathbf{w}^k \middle| \mathbf{w}^k\right) \tag{11}$$

$$f(\mathbf{w}) \ge g\left(\mathbf{w} \middle| \mathbf{w}^k\right) \text{ for all } \mathbf{w} \tag{12}$$

then $f(\mathbf{w})$ could be optimized by iteratively maximizing the surrogate function as follows:

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} g\left(\mathbf{w} \middle| \mathbf{w}^k\right). \tag{13}$$

One could see that

$$\begin{aligned} f(\mathbf{w}^{k+1}) &= f\left(\mathbf{w}^{k+1}\right) - g\left(\mathbf{w}^{k+1}\middle|\mathbf{w}^{k}\right) + g\left(\mathbf{w}^{k+1}\middle|\mathbf{w}^{k}\right) \\ &\geq f\left(\mathbf{w}^{k}\right) - g\left(\mathbf{w}^{k}\middle|\mathbf{w}^{k}\right) + g\left(\mathbf{w}^{k+1}\middle|\mathbf{w}^{k}\right) \\ &\geq f\left(\mathbf{w}^{k}\right) - g\left(\mathbf{w}^{k}\middle|\mathbf{w}^{k}\right) + g\left(\mathbf{w}^{k}\middle|\mathbf{w}^{k}\right) \\ &= f\left(\mathbf{w}^{k}\right) \end{aligned} \tag{14}$$

where the first inequality holds because $f(\mathbf{w}) - g(\mathbf{w}|\mathbf{w}^k)$ reaches its minimum at $\mathbf{w} = \mathbf{w}^k$ as a result of the two key conditions, while the second inequality holds because $g(\mathbf{w}|\mathbf{w}^k)$ reaches its maximum at $\mathbf{w} = \mathbf{w}^{k+1}$ as a result of the update rule. Therefore, the value of objective function monotonically increases during the iteration procedure and would converge to a local optimum.

The MM framework could turn a nonsmooth problem into a smooth one, thus could be used to solve the G2DPCA. A key point is to find a surrogate function that could be solved by purely analytic methods, using convenient inequalities. Some typical inequalities are listed in [19]. The MM framework is also referred to as "optimization transfer" [21], "auxiliary function method" [22], or "bound optimization" [23], etc.

B. First-Order Convexity Condition

Inequalities play a central role in designing MM algorithms. Below are some inequalities derived from the first-order convexity condition that will be utilized to solve G2DPCA. Given a convex and differentiable function $f(\mathbf{w})$ defined on a real vector space, the first-order condition of convex functions [24] states that

$$f(\mathbf{w}) \geq f(\mathbf{v}) + \nabla f(\mathbf{v})^T(\mathbf{w} - \mathbf{v}) \tag{15}$$

wherein the equality holds when $\mathbf{w} = \mathbf{v}$.

Lemma 1: Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, and $p \geq 1$, then

$$\|\mathbf{w}\|_p^p \geq p\Big[|\mathbf{v}|^{p-1} \circ \text{sign}(\mathbf{v})\Big]^T\mathbf{w} + (1-p)\|\mathbf{v}\|_p^p \tag{16}$$

holds and the inequality becomes equality when $\mathbf{w} = \mathbf{v}$.

Proof: If $\mathbf{v}$ has no zero element, i.e., all elements in $\mathbf{v}$ are not zeros, then $\|\mathbf{w}\|_p^p$ with $p \geq 1$ is convex and differentiable at $\mathbf{w} = \mathbf{v}$. The objective inequality (16) could be directly derived from the first-order convexity condition. If any element in $\mathbf{v}$ is zero, $\|\mathbf{w}\|_p^p$ would not be differentiable at $\mathbf{w} = \mathbf{v}$. Fortunately, the problem could be expanded into element form as

$$\sum_{i=1}^{d}|w_i|^p \geq p\sum_{i=1}^{d}|v_i|^{p-1}\text{sign}(v_i)w_i + (1-p)\sum_{i=1}^{d}|v_i|^p. \tag{17}$$

This inequality holds if

$$|w_i|^p \geq p|v_i|^{p-1}\text{sign}(v_i)w_i + (1-p)|v_i|^p \tag{18}$$

holds for all $i = 1, 2, \ldots, d$. It is easy to validate that (18) is true and the equality holds when $w_i = v_i$, no matter v_i is zero or not. This completes the proof. ■

Lemma 1 relaxes $\|\mathbf{w}\|_p^p$ with $p \geq 1$ to a linear function which would become much easier to handle. When $p = 1$, (16) reduces to

$$\|\mathbf{w}\|_1 \geq \text{sign}(\mathbf{v})^T\mathbf{w}. \tag{19}$$

This inequality is widely used to design solutions for robust PCA and 2DPCA algorithms [5], [11], [17], [18].

Lemma 2: Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{w} > \mathbf{0}$, $\mathbf{v} > \mathbf{0}$, and $0 < p < 1$. Specifically, $\mathbf{w} > \mathbf{0}$ and $\mathbf{v} > \mathbf{0}$ mean that all of the elements in $\mathbf{w}$ and $\mathbf{v}$ are larger than zero. Then

$$\|\mathbf{w}\|_p^p \leq p\Big[|\mathbf{v}|^{p-1} \circ \text{sign}(\mathbf{v})\Big]^T\mathbf{w} + (1-p)\|\mathbf{v}\|_p^p \tag{20}$$

holds and the inequality becomes equality when $\mathbf{w} = \mathbf{v}$.

Proof: Since $-\|\mathbf{w}\|_p^p$ is convex and differentiable at $\mathbf{w} = \mathbf{v}$ when $\mathbf{w} > \mathbf{0}$, $\mathbf{v} > \mathbf{0}$, and $0 < p < 1$, this lemma could be directly derived from the first-order convexity condition. ■

Lemma 3: Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v}$ has no zero element, let $0 < p < 2$, then

$$\|\mathbf{w}\|_p^p \leq \frac{p}{2}\mathbf{w}^T\text{diag}\Big(|\mathbf{v}|^{p-2}\Big)\mathbf{w} + \Big(1 - \frac{p}{2}\Big)\|\mathbf{v}\|_p^p \tag{21}$$

holds and the inequality becomes equality when $\mathbf{w} = \mathbf{v}$.

Proof: Assume $\mathbf{w}$ has no zero element, since $\mathbf{v}$ also has no zero element and $0 < p < 2$, we have $\mathbf{w} \circ \mathbf{w} > \mathbf{0}$, $\mathbf{v} \circ \mathbf{v} > \mathbf{0}$, and $0 < p/2 < 1$. According to Lemma 2, we have

$$\begin{aligned} \|\mathbf{w}\|_p^p &= \|\mathbf{w} \circ \mathbf{w}\|_{p/2}^{p/2} \\ &\leq \frac{p}{2}\Big[|\mathbf{v} \circ \mathbf{v}|^{p/2-1}\Big]^T(\mathbf{w} \circ \mathbf{w}) + \Big(1 - \frac{p}{2}\Big)\|\mathbf{v} \circ \mathbf{v}\|_{p/2}^{p/2} \\ &= \frac{p}{2}\mathbf{w}^T\text{diag}\Big(|\mathbf{v} \circ \mathbf{v}|^{p/2-1}\Big)\mathbf{w} + \Big(1 - \frac{p}{2}\Big)\|\mathbf{v} \circ \mathbf{v}\|_{p/2}^{p/2} \\ &= \frac{p}{2}\mathbf{w}^T\text{diag}\Big(|\mathbf{v}|^{p-2}\Big)\mathbf{w} + \Big(1 - \frac{p}{2}\Big)\|\mathbf{v}\|_p^p \end{aligned} \tag{22}$$

wherein the inequality becomes equality when $\mathbf{w} \circ \mathbf{w} = \mathbf{v} \circ \mathbf{v}$. This condition could be further guaranteed by $\mathbf{w} = \mathbf{v}$. Therefore, the inequalities in (21) and (22) would become equalities when $\mathbf{w} = \mathbf{v}$, satisfying our assumption that $\mathbf{w}$ has no zero element.

On the other hand, if any element in $\mathbf{w}$ is zero, we should expand (21) into element form in order to examine the zero points. That is

$$\sum_{i=1}^{d}|w_i|^p \leq \frac{p}{2}\sum_{i=1}^{d}w_i^2|v_i|^{p-2} + \Big(1 - \frac{p}{2}\Big)\sum_{i=1}^{d}|v_i|^p. \tag{23}$$

This inequality holds if

$$|w_i|^p \leq \frac{p}{2}w_i^2|v_i|^{p-2} + \Big(1 - \frac{p}{2}\Big)|v_i|^p \tag{24}$$

holds for all $i = 1, 2, \ldots, d$. For any $w_i \neq 0$, (24) holds and it becomes equality when $w_i = v_i$, as discussed above. For any $w_i = 0$, the corresponding inequality reduces to $(1 - p/2)|v_i|^p \geq 0$ which is always true, but would never become equality since $p < 2$ and $v_i \neq 0$. Therefore, (24) holds no matter w_i is zero or not, but the inequality would become equality only when $w_i = v_i$ in which case w_i is not zero.

To summarize, (21) holds when $\mathbf{v}$ has no zero element and $0 < p < 2$. When $\mathbf{w} = \mathbf{v}$, the inequality becomes equality. This completes the proof. ■

Lemma 3 relaxes $\|\mathbf{w}\|_p^p$ with $0 < p < 2$ to a quadratic function which also becomes much easier to handle. When $p = 1$, (21) reduces to

$$\|\mathbf{w}\|_1 \leq \frac{1}{2}\mathbf{w}^T\text{diag}\Big(|\mathbf{v}|^{-1}\Big)\mathbf{w} + \frac{1}{2}\|\mathbf{v}\|_1. \tag{25}$$

This inequality is used to solve 2DPCAL1-S [18].

C. Linear Optimization Problem With Lp-Norm Constraint

Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, let $p,\ q \in [1, \infty]$ be two scalars with $1/p + 1/q = 1$, Hölder's inequality [25] states that

$$\sum_{i=1}^{d} |v_i w_i| \leq \|\mathbf{v}\|_q \|\mathbf{w}\|_p \tag{26}$$

wherein the equality holds if and only if there exists a positive real scalar c satisfying $|w_i|^p = c|v_i|^q,\ i = 1, 2, \ldots, d$. The following lemma could be given based on Hölder's inequality.

Lemma 4: Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v} \neq \mathbf{0}$, let $p,\ q \in [1, \infty]$ be two scalars satisfying $1/p + 1/q = 1$, then the optimization problem

$$\max_{\mathbf{w}} \mathbf{v}^T \mathbf{w}, \quad \text{s.t. } \|\mathbf{w}\|_p^p = 1 \tag{27}$$

has a closed-form solution

$$\mathbf{w} = \frac{|\mathbf{v}|^{q-1} \circ \text{sign}(\mathbf{v})}{\|\mathbf{v}\|_q^{q-1}}. \tag{28}$$

Proof: According to Hölder's inequality in (26), we have

$$\mathbf{v}^T \mathbf{w} \leq \sum_{i=1}^{d} |v_i w_i| \leq \|\mathbf{v}\|_q \|\mathbf{w}\|_p = \|\mathbf{v}\|_q. \tag{29}$$

Therefore, the maximum of the objective function is obtained when both inequalities become equalities. The first equality holds when

$$\text{sign}(w_i) = \text{sign}(v_i),\ \ i = 1, 2, \ldots, d. \tag{30}$$

The second equality holds when

$$|w_i|^p = c|v_i|^q,\ \ i = 1, 2, \ldots, d. \tag{31}$$

Since $\mathbf{v} \neq \mathbf{0}$, the constant c could then be calculated by

$$c = \frac{\sum_{i=1}^{d} |w_i|^p}{\sum_{i=1}^{d} |v_i|^q} = \frac{\|\mathbf{w}\|_p^p}{\|\mathbf{v}\|_q^q} = \frac{1}{\|\mathbf{v}\|_q^q}. \tag{32}$$

Substituting (32) into (31), we have

$$|w_i| = \left(c|v_i|^q\right)^{1/p} = \left(\frac{|v_i|^q}{\|\mathbf{v}\|_q^q}\right)^{1/p} = \frac{|v_i|^{q-1}}{\|\mathbf{v}\|_q^{q-1}},\ \ i = 1, 2, \ldots, d. \tag{33}$$

Considering (30), we have

$$w_i = \frac{|v_i|^{q-1}}{\|\mathbf{v}\|_q^{q-1}} \text{sign}(v_i),\ \ i = 1, 2, \ldots, d. \tag{34}$$

Rewriting the equations into vector form will complete this proof. ■

IV. Solution of Generalized 2DPCA

With the above techniques, the solution for the G2DPCA problem in (9) is provided as follows. Considering that the constraint set could be either convex or nonconvex depending on the p value, we divide the G2DPCA problem into two cases, the same as in GPCA [13].

A. Case 1

In case 1, $p \geq 1$ and the constraint set is convex. Then the optimization problem of G2DPCA states

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i \mathbf{w}\|_s^s, \quad \text{s.t. } \|\mathbf{w}\|_p^p = 1 \tag{35}$$

where $s \geq 1$, $p \geq 1$, and $\mathbf{w} \in \mathbb{R}^w$. This problem could be turned into iteratively maximizing a surrogate function under the MM framework as follows. Let $\mathbf{w}^k$ be the projection vector at the kth step in the iteration procedure. It could be regarded as a constant vector that is irrelevant with respect to $\mathbf{w}$. According to Lemma 1, The objective function could be linearized as

$$\sum_{i=1}^{n} \|\mathbf{X}_i \mathbf{w}\|_s^s \geq s\left(\mathbf{v}^k\right)^T \mathbf{w} + (1 - s) \sum_{i=1}^{n} \|\mathbf{X}_i \mathbf{w}^k\|_s^s \tag{36}$$

wherein

$$\mathbf{v}^k = \sum_{i=1}^{n} \mathbf{X}_i^T \left[\left| \mathbf{X}_i \mathbf{w}^k \right|^{s-1} \circ \text{sign}\left(\mathbf{X}_i \mathbf{w}^k\right) \right] \tag{37}$$

and the inequality (36) becomes equality when $\mathbf{w} = \mathbf{w}^k$. Denote the objective function as $f(\mathbf{w})$, denote the linearized function as $g(\mathbf{w}|\mathbf{w}^k)$, we have $f(\mathbf{w}^k) = g(\mathbf{w}^k|\mathbf{w}^k)$ and $f(\mathbf{w}) \geq g(\mathbf{w}|\mathbf{w}^k)$ for all $\mathbf{w}$, satisfying the two key conditions of the MM framework. Therefore, $g(\mathbf{w}|\mathbf{w}^k)$ is a feasible surrogate function of $f(\mathbf{w})$. According to the MM framework, the G2DPCA problem could be turned into iteratively maximizing the surrogate function as follows:

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} g\left(\mathbf{w} \middle| \mathbf{w}^k\right), \quad \text{s.t. } \|\mathbf{w}\|_p^p = 1. \tag{38}$$

By dropping the term irrelevant to $\mathbf{w}$ in the surrogate function, maximizing the surrogate function leads to a linear optimization problem with Lp-norm constraint

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} \left(\mathbf{v}^k\right)^T \mathbf{w}, \quad \text{s.t. } \|\mathbf{w}\|_p^p = 1. \tag{39}$$

Since $p \geq 1$, according to Lemma 4, its solution is

$$\mathbf{w}^{k+1} = \frac{\left|\mathbf{v}^k\right|^{q-1} \circ \text{sign}\left(\mathbf{v}^k\right)}{\left\|\mathbf{v}^k\right\|_q^{q-1}} \tag{40}$$

where q satisfies $1/p + 1/q = 1$. The solution could be rewritten in a two-step procedure as

$$\mathbf{u}^k = \left|\mathbf{v}^k\right|^{q-1} \circ \text{sign}\left(\mathbf{v}^k\right) \tag{41}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\left\|\mathbf{u}^k\right\|_p}. \tag{42}$$

This completes the solution in case 1.

Two extreme conditions of case 1, i.e., $p = 1$ and $p = \infty$ are discussed as follows. When $p = 1$, since $1/p + 1/q = 1$, we will have $q = \infty$. Let $j = \arg\max_{i \in [1,w]} |v_i^k|$, i.e., $|v_j^k|$ is the largest value in $|\mathbf{v}^k|$. By taking the limit of (40) when p approaches 1, we have

$$w_i^{k+1} = \begin{cases} \text{sign}\left(v_j^k\right), & i = j \\ 0, & i \neq j \end{cases} \tag{43}$$

for $i = 1, 2, \ldots, w$. It shows that there is only one nonzero element, 1 or -1, in the final result of $\mathbf{w}$ when $p = 1$. That is why 2DPCAL1-S should be formulated by combining L1- and L2-norm constraints together rather than using L1-norm constraint alone. Similarly, when p approaches infinity, the limit of (40) is

$$\mathbf{w}^{k+1} = \text{sign}\Big(\mathbf{v}^k\Big). \tag{44}$$

All elements in the final result of $\mathbf{w}$ should be either 1 or -1. In practice, when p is large enough, all the elements in the projection vector tend to have very close absolute values.

When $s = 1$ and $p = 2$, G2DPCA degenerates to 2DPCA-L1. By substituting $s = 1$ and $p = 2$ into (40) we could obtain the same solution as in [17]. It tells that the solution in [17] could be explained from the MM viewpoint. The solution of 2DPCAL1-S in [18] could also be explained from the MM viewpoint though 2DPCAL1-S is not exactly a special case of G2DPCA.

B. Case 2

In case 2, $0 < p < 1$ and the constraint set is nonconvex. By applying the method of Lagrange multipliers, maximizing the optimization problem of G2DPCA equals maximizing the Lagrangian as follows:

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s - \lambda\Big(\|\mathbf{w}\|_p^p - 1\Big) \tag{45}$$

where $s \geq 1$, $0 < p < 1$, $\lambda > 0$, and $\mathbf{w} \in \mathbb{R}^w$. According to Lemmas 1 and 3, the Lagrangian could be relaxed as

$$\begin{aligned}
&\sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s - \lambda\Big(\|\mathbf{w}\|_p^p - 1\Big) \\
&\quad \geq s\Big(\mathbf{v}^k\Big)^T\mathbf{w} + (1-s)\sum_{i=1}^{n}\Big\|\mathbf{X}_i\mathbf{w}^k\Big\|_s^s \\
&\qquad -\lambda\frac{p}{2}\mathbf{w}^T\text{diag}\bigg(\Big|\mathbf{w}^k\Big|^{p-2}\bigg)\mathbf{w} - \lambda\Big(1-\frac{p}{2}\Big)\Big\|\mathbf{w}^k\Big\|_p^p + \lambda
\end{aligned} \tag{46}$$

wherein $\mathbf{v}^k$ is defined in (37) and the inequality becomes equality when $\mathbf{w} = \mathbf{w}^k$. The $\mathbf{w}^k$ in the relaxed function is required to have no zero element which could be guaranteed by replacing $\mathbf{w}^k$ with $\mathbf{w}^k + \varepsilon$, where ε is a random scalar that is sufficiently close to zero. Let the Lagrangian be denoted as $f(\mathbf{w})$ and let the relaxed function be denoted as $g(\mathbf{w}|\mathbf{w}^k)$. Again, we have $f(\mathbf{w}^k) = g(\mathbf{w}^k|\mathbf{w}^k)$ and $f(\mathbf{w}) \geq g(\mathbf{w}|\mathbf{w}^k)$ for all $\mathbf{w}$, satisfying the two key conditions of the MM framework. Therefore, $g(\mathbf{w}|\mathbf{w}^k)$ is a feasible surrogate function of the Lagrangian. According to the MM framework, maximizing the Lagrangian could be turned into iteratively maximizing the surrogate function as follows:

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} g\Big(\mathbf{w}|\mathbf{w}^k\Big). \tag{47}$$

After dropping the terms irrelevant to $\mathbf{w}$, we will reach the following quadratic optimization problem:

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} s(\mathbf{v}^k)^T\mathbf{w} - \lambda\frac{p}{2}\mathbf{w}^T\text{diag}\bigg(\Big|\mathbf{w}^k\Big|^{p-2}\bigg)\mathbf{w}. \tag{48}$$

TABLE I
ALGORITHM PROCEDURE OF G2DPCA

Input: $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$, $s \in [1, \infty)$, $p \in (0, \infty]$, r.
Output: $\mathbf{W}$.

Initialize $\mathbf{W} = [\,]$, $\mathbf{X}_i^0 = \mathbf{X}_i$, $i = 1, 2, \cdots, n$.
for $t = 1, 2, \cdots, r$ **do**
  Initialize $k = 0$, $\delta = 1$, $\mathbf{w}^0$.
  $\mathbf{w}^0 = \frac{\mathbf{w}^0}{\|\mathbf{w}^0\|_p}$.
  $f^0 = \sum_{i=1}^{n}\|\mathbf{X}_i\mathbf{w}^0\|_s^s$.
  while $\delta > 10^{-4}$ **do**
    $\mathbf{v}^k = \sum_{i=1}^{n}\mathbf{X}_i^T\Big[\Big|\mathbf{X}_i\mathbf{w}^k\Big|^{s-1} \circ sign(\mathbf{X}_i\mathbf{w}^k)\Big]$.
    if $0 < p < 1$
      $\mathbf{u}^k = \Big|\mathbf{w}^k\Big|^{2-p} \circ \mathbf{v}^k$,
      $\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_p}$.
    elseif $p = 1$
      $j = \arg\max_{i\in[1,w]}\big|v_i^k\big|$,
      $w_i^{k+1} = \begin{cases} sign(v_j^k), & i = j, \\ 0, & i \neq j. \end{cases}$
    elseif $p < \infty$
      $q = p/(p-1)$,
      $\mathbf{u}^k = \Big|\mathbf{v}^k\Big|^{q-1} \circ sign(\mathbf{v}^k)$,
      $\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_p}$.
    elseif $p = \infty$
      $\mathbf{w}^{k+1} = sign(\mathbf{v}^k)$.
    end if
    $f^{k+1} = \sum_{i=1}^{n}\|\mathbf{X}_i\mathbf{w}^{k+1}\|_s^s$.
    $\delta = \big|f^{k+1} - f^k\big|/f^k$.
    $k \leftarrow k + 1$.
  end while
  $\mathbf{w}_t = \mathbf{w}^k$.
  $\mathbf{W} \leftarrow [\mathbf{W}, \mathbf{w}_t]$.
  $\mathbf{X}_i = \mathbf{X}_i^0(\mathbf{I} - \mathbf{W}\mathbf{W}^T)$, $i = 1, 2, \cdots, n$.
end for

Its solution is

$$\mathbf{w}^{k+1} = \frac{s}{\lambda p}\Big|\mathbf{w}^k\Big|^{2-p} \circ \mathbf{v}^k. \tag{49}$$

Since $2 - p > 0$, this solution indicates that $\mathbf{w}^k$ is no longer required to have no zero element. Since ε is sufficiently close to zero, we could treat this solution as the solution of the problem in (48) when $\mathbf{w}^k$ has zero elements. Considering the constraint $\|\mathbf{w}\|_p^p = 1$ and $\lambda > 0$, we have

$$\lambda = \frac{s}{p}\bigg\|\Big|\mathbf{w}^k\Big|^{2-p} \circ \mathbf{v}^k\bigg\|_p. \tag{50}$$

Then the update rule is

$$\mathbf{w}^{k+1} = \frac{\Big|\mathbf{w}^k\Big|^{2-p} \circ \mathbf{v}^k}{\bigg\|\Big|\mathbf{w}^k\Big|^{2-p} \circ \mathbf{v}^k\bigg\|_p}. \tag{51}$$

The above solution equals the two-step procedure below

$$\mathbf{u}^k = \Big|\mathbf{w}^k\Big|^{2-p} \circ \mathbf{v}^k \tag{52}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\Big\|\mathbf{u}^k\Big\|_p}. \tag{53}$$

This completes the solution in case 2.

Notice that the solution in case 2 is also feasible when $1 \leq p < 2$ since the inequality in (46) holds when p is in the range of $(0, 2)$. Therefore, we have two different solutions when $1 \leq p < 2$. In practice, we find that the solution in

case 1 converges much faster than the solution in case 2 when $1 \leq p < 2$ thus being more preferred.

The above completes the solution of G2DPCA problem. From the results in (40) and (51), it could be observed that a closed-form solution is obtained in each iteration for either case. The solution successfully avoids zero-finding problems [26], learning rates [12], or extra tuning parameters [11], [18] which are usually encountered in solving Lp-norm related algorithms. The algorithm procedure of G2DPCA is listed in Table I. Notice that 0 on the superscript denotes the initialization, k or $k+1$ on the superscript denotes iteration number, and $|\mathbf{w}|^p$ or the like denotes the element-wise power of the absolute value of a vector.

V. Experiments

Two benchmark face databases, the Olivetti Research Laboratory (ORL) database [27] and the Face Recognition Technology (FERET) database [28] are used in our experiments. In order to evaluate the proposed G2DPCA algorithm, we compare it with three state-of-the-art algorithms, i.e., 2DPCAL1-S [18], GPCA [13], and RSPCA [11] in the tasks of image reconstruction and classification on the two databases. GPCA and RSPCA are modified to be the 1-D counterparts of G2DPCA and 2DPCAL1-S, respectively in order to make a fair comparison. For G2DPCA and GPCA, we search the optimal parameter set from $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$. A wider range has also been tried, but no better reconstruction or classification results could be found. For 2DPCAL1-S and RSPCA, a parameter ρ relates to the λ in (7) via $\lambda = 10^{-\rho}$ is tuned, consistent with [18]. When ρ is small enough, 2DPCAL1-S approximates to G2DPCA with $s = 1$ and $p = 1$, RSPCA approximates to GPCA with $s = 1$ and $p = 1$; when ρ is large enough, 2DPCAL1-S approximates to 2DPCA-L1, RSPCA approximates to PCA-L1. These extreme conditions guarantee that the optimal ρ value in 2DPCAL1-S or RSPCA could be located in a finite range. Therefore, we search for the optimal ρ value from $[-3.0{:}0.1{:}3.0]$ in our experiments.

As for initialization, the multistart method [29] is widely suggested to be an efficient method for finding a good locally optimal solution in PCA-based algorithms [5], [11]–[13]. That is, random initializations are tried multiple times and the initialization with the maximal objective function value is finally chosen. In this paper, however, we directly initialize the projection vectors of the 1-D algorithms by the corresponding components of PCA, and initialize the projection vectors of the 2-D algorithms by the corresponding components of 2DPCA. This method makes the most of the relationships between these algorithms, therefore it is expected to find a good locally optimal solution.

Major source codes have been made publicly available at https://github.com/yuzhounh/G2DPCA.

A. ORL Face Database

The ORL face database [27] contains 400 images from 40 subjects, ten images per subject. The images are taken with tolerances for different facial expressions, different rotation angles, and different scaling ratios. The image size is 112×92. We further resize the images to 56×46 to reduce the computational time. Fig. 1 shows some sample images from this database.

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

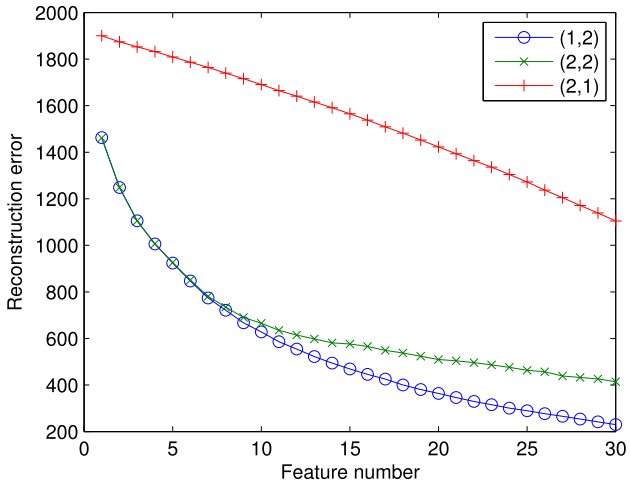


Fig. 2. Reconstruction errors of G2DPCA in three special cases on the ORL database. The (s, p) pairs for the three cases are shown in the legend.

To evaluate the reconstruction performance of G2DPCA, we conduct an experiment on a polluted ORL database, similar as in [5], [17], and [18]. Specifically, 20% of the total 400 images are randomly selected and occluded with a rectangular noise whose size is at least 20×20, locating at a random position. The noise consists of random black and white dots. Let $\mathbf{W}$ be the projection matrix trained on the whole polluted ORL database, let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_m$ be m ($m = 320$) clean images which are mean-centered, then the average reconstruction error of G2DPCA is defined as

$$\frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{Z}_i\left(\mathbf{I}-\mathbf{W}\mathbf{W}^T\right)\right\|_F. \tag{54}$$

Fig. 2 shows the reconstruction errors of G2DPCA in three special cases with different number of extracted features. Among the three cases, G2DPCA with $s = 2$ and $p = 2$ corresponds to traditional 2DPCA, G2DPCA with $s = 1$ and $p = 2$ corresponds to 2DPCA-L1. From the figure, both reconstruction results of 2DPCA and 2DPCA-L1 are much better than that of G2DPCA with $s = 2$ and $p = 1$. When the feature number is larger than seven, the reconstruction error of 2DPCA-L1 is lower than that of 2DPCA, consistent with the results in [17] and [18]. The figure shows that applying L1-norm on the objective function of 2DPCA would improve its reconstruction performance. As an illustration, the reconstructed images of the three special cases on two sample images are shown in Fig. 3.

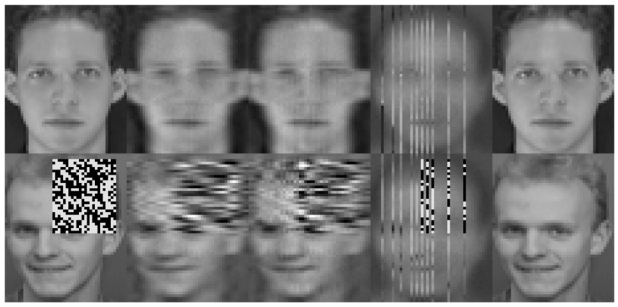


Fig. 3. Reconstructed images of G2DPCA in three special cases on two sample images from the polluted ORL database. The first column is the images to be reconstructed. One image has random noises while the other does not. The following three columns are the reconstructed images by using the first ten projection vectors of G2DPCA wherein the (s, p) pairs are set to be $(1, 2)$, $(2, 2)$, and $(2, 1)$ in order. The last column shows the original images for comparison.

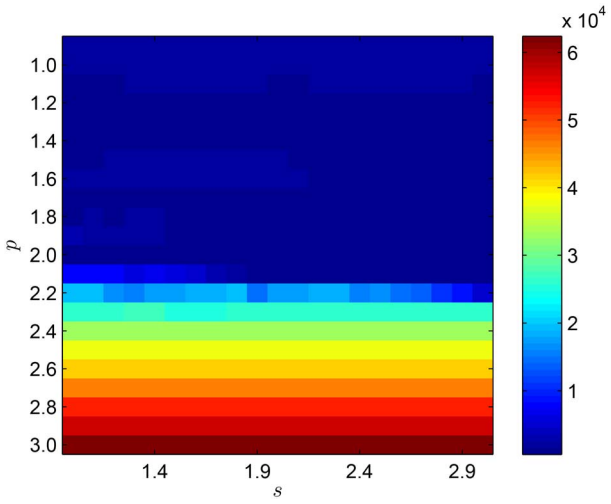


Fig. 4. Reconstruction errors of G2DPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the ORL database.

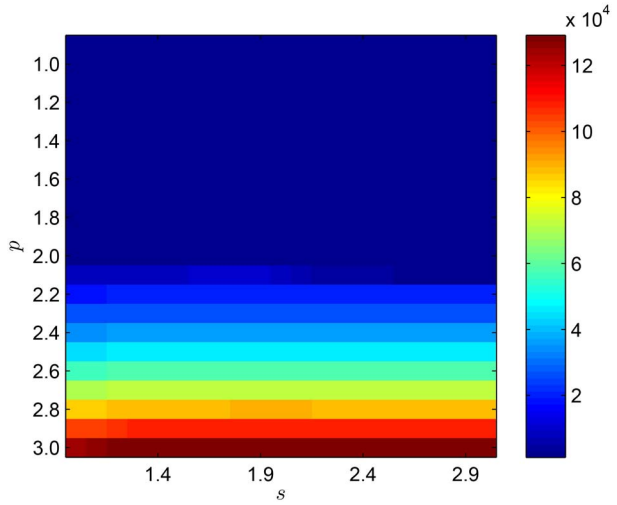


Fig. 5. Reconstruction errors of GPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the ORL database.

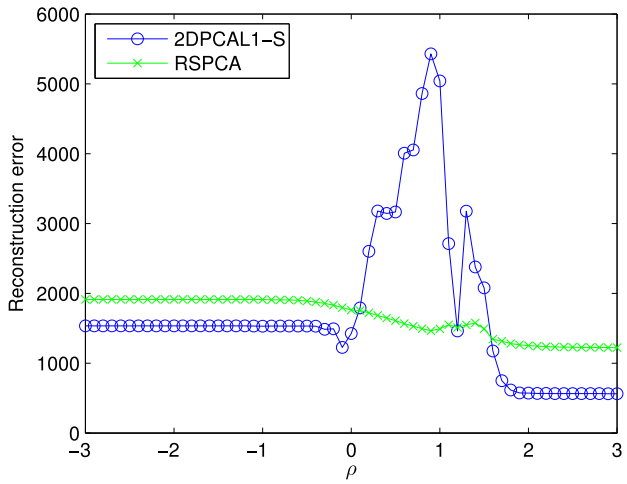


Fig. 6. Reconstruction errors of 2DPCAL1-S and RSPCA with $\rho = [-3.0{:}0.1{:}3.0]$ on the ORL database.

Then, we proceed to compare the reconstruction performance of G2DPCA with other three algorithms, i.e., 2DPCAL1-S, GPCA, and RSPCA. The average reconstruction errors of the three algorithms could be defined similarly. By averaging the reconstruction errors with different feature numbers which are in the range of $[1, 30]$, we obtain the results in Figs. 4–6. From the results, the reconstruction errors of G2DPCA and GPCA are relatively stable with different s values, but are greatly affected by various p values. Both results of 2DPCAL1-S and RSPCA show that the lowest reconstruction error is obtained when ρ is set to be a large value in which case 2DPCAL1-S approximates to 2DPCA-L1 and RSPCA approximates to PCA-L1.

The lowest reconstruction errors and corresponding parameters of the four algorithms are listed in Table II. As special cases of G2DPCA and GPCA, the results of 2DPCA-L1, 2DPCA, PCA-L1, and PCA are also listed in the table for comparison. With tolerances for random errors, we could conclude that the best reconstruction performances of G2DPCA and 2DPCAL1-S are achieved when they reduce to 2DPCA-L1, and the best reconstruction performances of GPCA and RSPCA are achieved when they reduce to PCA-L1.

TABLE II
RECONSTRUCTION ERRORS OF EIGHT ALGORITHMS ON THE ORL DATABASE

Algorithms	Optimal parameters	Reconstruction error ($\times 10^3$)
G2DPCA	$s = 1.0$, $p = 2.0$	**0.5624**
2DPCAL1-S	$\rho = 2.9$	0.5630
2DPCA-L1	-	**0.5624**
2DPCA	-	0.6583
GPCA	$s = 1.1$, $p = 2.0$	1.2184
RSPCA	$\rho = 3.0$	1.2222
PCA-L1	-	1.2220
PCA	-	1.3036

Apparently, the reconstruction performances of 2-D algorithms are much better than those of 1-D algorithms when the same number of features are extracted. However, the meaning of this comparison is limited concerning the differences between the two categories of algorithms. Among these differences, a major one is that the maximal number of features

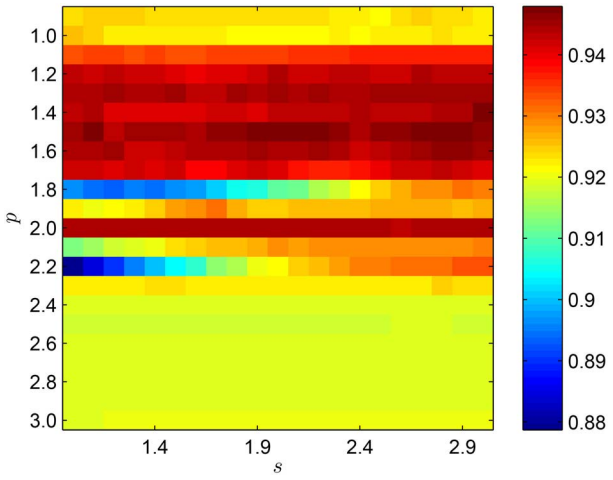


Fig. 7. Classification accuracies of G2DPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the ORL database.

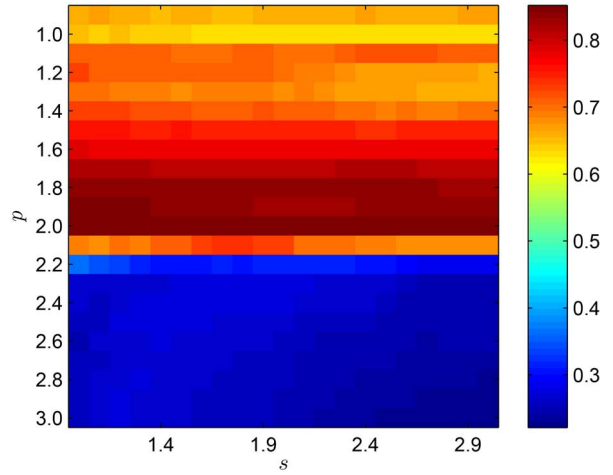


Fig. 8. Classification accuracies of GPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the ORL database.

that could be extracted by 1-D algorithms is much larger than that could be extracted by 2-D algorithms. As a result, the same number of 1-D components account for much less variance than 2-D components. Therefore, it is unsurprising that the reconstruction errors of 2-D algorithms are much lower than those of 1-D algorithms.

Then, we proceed to compare the classification performance of G2DPCA with other three algorithms on the ORL database. These algorithms are employed to extract features, then the nearest neighbor classifier is applied to do classification. In the ORL database, we randomly choose five images from each subject for testing and use the remaining images for training. The procedure is repeated ten times and the average classification accuracy is recorded. The classification accuracies with feature numbers in the range of $[1, 30]$ are further averaged, then the final results are reported, as shown in Figs. 7–9. From the results, the accuracies of G2DPCA and GPCA are insensitive with respect to s values, but are greatly affected by p values. The accuracy of 2DPCAL1-S peaks at $\rho = 1.7$. The accuracy of RSPCA generally increases with ρ value and becomes stable when ρ is large enough.

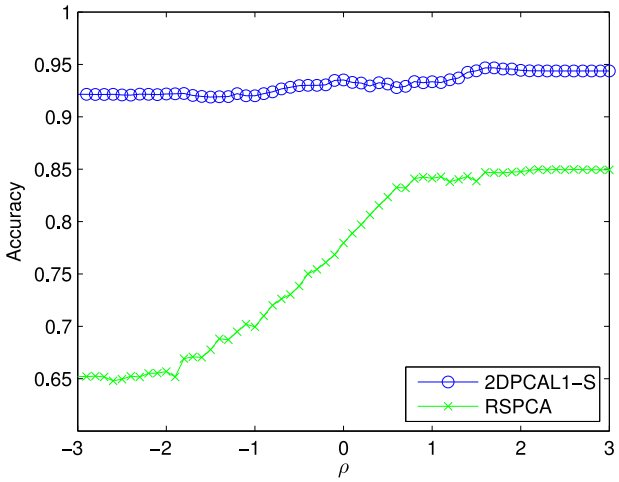


Fig. 9. Classification accuracies of 2DPCAL1-S and RSPCA with $\rho = [-3.0{:}0.1{:}3.0]$ on the ORL database.

TABLE III
CLASSIFICATION ACCURACIES OF EIGHT ALGORITHMS ON THE ORL DATABASE

Algorithms	Optimal parameters	Accuracy
G2DPCA	$s = 2.9$, $p = 1.5$	**0.9479**
2DPCAL1-S	$\rho = 1.7$	0.9467
2DPCA-L1	-	0.9436
2DPCA	-	0.9436
GPCA	$s = 2.3$, $p = 2.0$	0.8521
RSPCA	$\rho = 2.4$	0.8498
PCA-L1	-	0.8493
PCA	-	0.8515

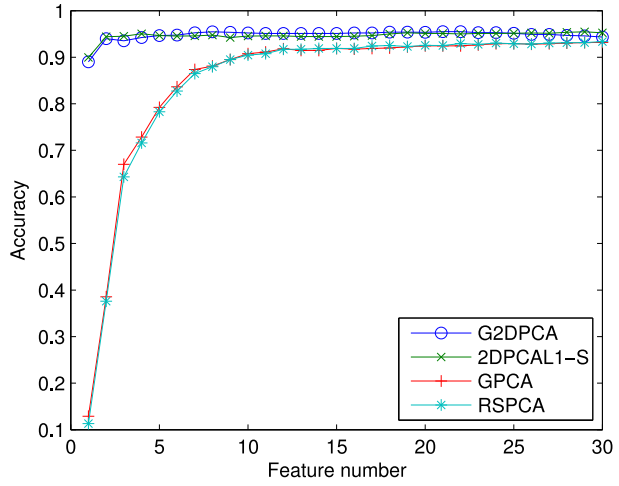


Fig. 10. Classification accuracies of G2DPCA, 2DPCAL1-S, GPCA, and RSPCA with different feature numbers on the ORL database when respective optimal parameters are applied.

The highest classification accuracies and corresponding parameters of the four algorithms are listed in Table III, including the results of 2DPCA-L1, 2DPCA, PCA-L1, and PCA for comparison. Fig. 10 shows the detailed accuracy results with different feature numbers when the optimal parameters are applied in the four algorithms. From the results, the accuracy of G2DPCA is slightly higher than that of 2DPCAL1-S, and the accuracy of GPCA is slightly higher than that of RSPCA.

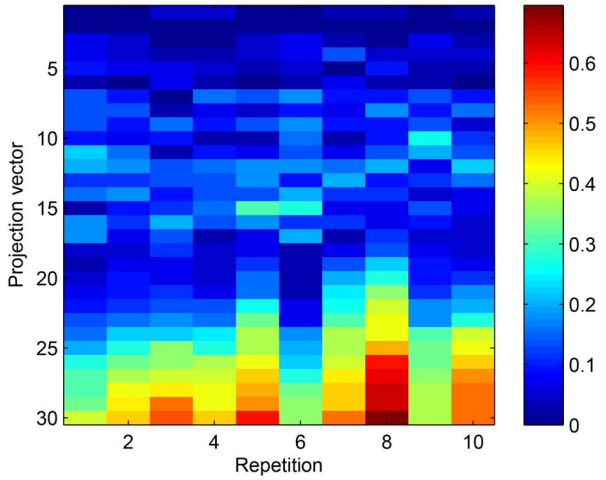


Fig. 11. Sparse rates of the first 30 projection vectors of G2DPCA with $s = 2.9$ and $p = 1.5$ in ten subsets of the ORL database which are randomly generated.

It also demonstrates that applying Lp-norm on the objective and constraint functions of traditional 2DPCA could improve its classification performance.

The optimal classification performances of the 2-D algorithms are much better than those of the 1-D algorithms since the variance explained by each 2-D component is much larger than that explained by corresponding 1-D component, as discussed before.

Then the sparsity of the projection vectors of G2DPCA with optimal parameters, i.e., $s = 2.9$ and $p = 1.5$ are studied. Define sparse rate of a vector as the ratio of zero elements in that vector. In practice, elements with absolute values smaller than 10^{-4} are treated as zeros. The sparse rates of the first 30 projection vectors in the above image classification task are calculated, as shown in Fig. 11. There are ten random repetitions. From the results, the projection vectors are slightly sparse which indicates that some irrelevant features are eliminated by G2DPCA and sparsity is helpful in image classification.

B. FERET Face Database

To further examine the performance of G2DPCA in image reconstruction and classification, we conduct experiments on a subset of the FERET face database [28], similar to the experiments on the ORL database. The FERET database contains 1400 images from 200 subjects, seven images per subject. The images are taken with different facial expressions and view angles. The image size is 80×80. We further resize the images to 40×40 to reduce the computational time. Fig. 12 shows some sample images from this database.

Fig. 13 shows the reconstruction errors of G2DPCA in the three special cases with different feature numbers. Fig. 14 shows the reconstructed images of G2DPCA in three special cases on two sample images. From the result, the reconstruction error of 2DPCA-L1 is lower than that of 2DPCA when the feature number is larger than five. Also, both reconstruction errors of 2DPCA-L1 and 2DPCA are much lower than that of G2DPCA with $s = 2$ and $p = 1$.

Fig. 12. Sample images of the FERET face database.

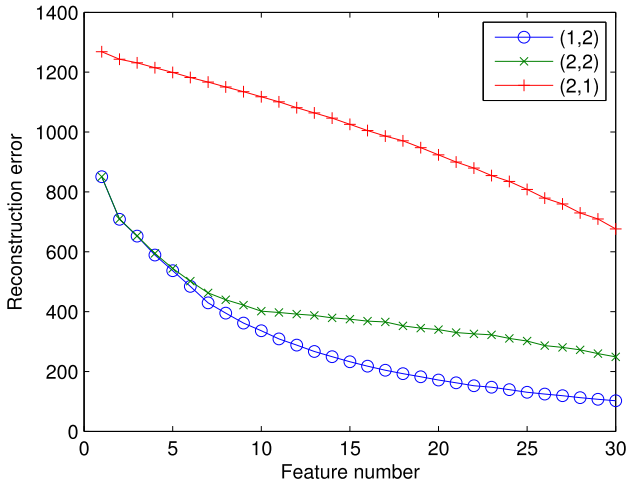


Fig. 13. Reconstruction errors of G2DPCA in three special cases on the FERET database. The (s, p) pairs for the three cases are shown in the legend.

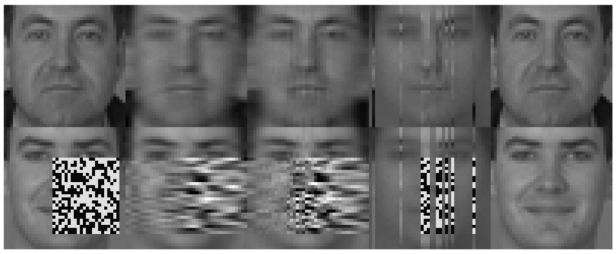


Fig. 14. Reconstructed images of G2DPCA in three special cases on two sample images from the polluted FERET database. The first column is the images to be reconstructed. One image has random noises while the other does not. The following three columns are the reconstructed images by using the first ten projection vectors of G2DPCA wherein the (s, p) pairs are set to be $(1, 2)$, $(2, 2)$, and $(2, 1)$ in order. The last column shows the original images for comparison.

To compare the reconstruction performance of G2DPCA with other three algorithms, we conduct an experiment on a polluted FERET database. Similarly, 20% of the total 1400 images are randomly selected and occluded with a rectangular noise whose size is at least 20×20, locating at a random position. The noise consists of random black and white dots. The reconstruction errors of the four algorithms are shown in Figs. 15–17. From the results, the reconstruction errors of G2DPCA and GPCA are relatively stable with different s values, but are greatly affected by various p values. Both results of 2DPCAL1-S and RSPCA show that the lowest reconstruction error is obtained when ρ is large enough in which case 2DPCAL1-S approximates to 2DPCA-L1 and RSPCA approximates to PCA-L1.

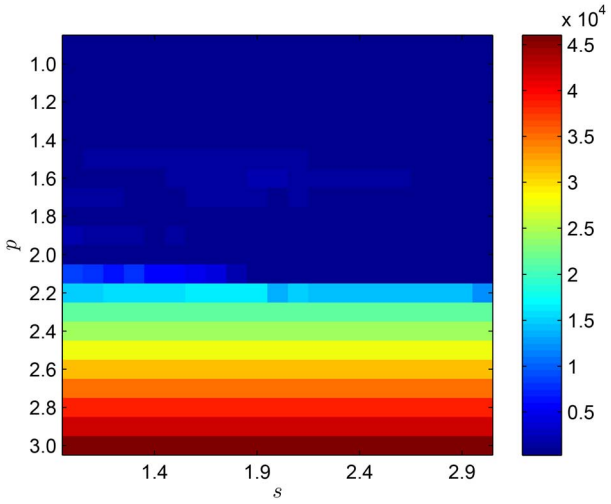


Fig. 15. Reconstruction errors of G2DPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the FERET database.

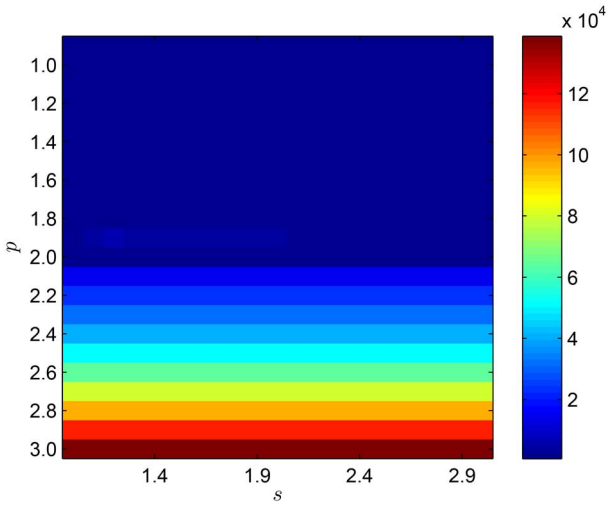


Fig. 16. Reconstruction errors of GPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the FERET database.

The lowest reconstruction errors and corresponding parameters of the four algorithms are listed in Table IV, including the results of 2DPCA-L1, 2DPCA, PCA-L1, and PCA for comparison. With tolerances for random errors, the results demonstrate that the best reconstruction performances of G2DPCA and 2DPCAL1-S are achieved when they reduce to 2DPCA-L1, and the best reconstruction performances of GPCA and RSPCA are achieved when they reduce to PCA-L1, the same as what we have concluded from the experimental results on the ORL database.

To compare the classification performance of G2DPCA with other three algorithms on the FERET database, we randomly choose four images from each subject for testing and use the remaining images for training. The procedure is repeated ten times. The average classification accuracies are shown in Figs. 18–20. From the figures, the accuracies of G2DPCA and GPCA are generally sensitive to p values but insensitive to s values. And the best classification performances of 2DPCAL1-S and RSPCA are achieved when $0 < \rho < 1$.

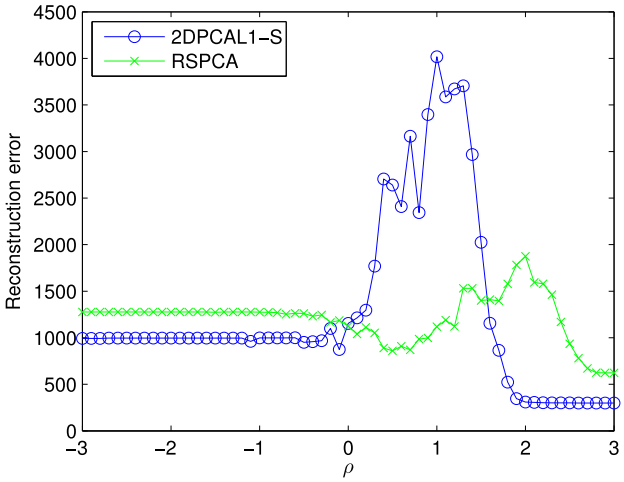


Fig. 17. Reconstruction errors of 2DPCAL1-S and RSPCA with $\rho = [-3.0{:}0.1{:}3.0]$ on the FERET database.

TABLE IV
RECONSTRUCTION ERRORS OF EIGHT ALGORITHMS ON THE FERET DATABASE

Algorithms	Optimal parameters	Reconstruction error ($\times 10^3$)
G2DPCA	$s = 1.0$, $p = 2.0$	**0.2985**
2DPCAL1-S	$\rho = 2.9$	0.2987
2DPCA-L1	-	**0.2985**
2DPCA	-	0.4072
GPCA	$s = 1.1$, $p = 2.0$	0.6217
RSPCA	$\rho = 3.0$	0.6223
PCA-L1	-	0.6219
PCA	-	0.6538

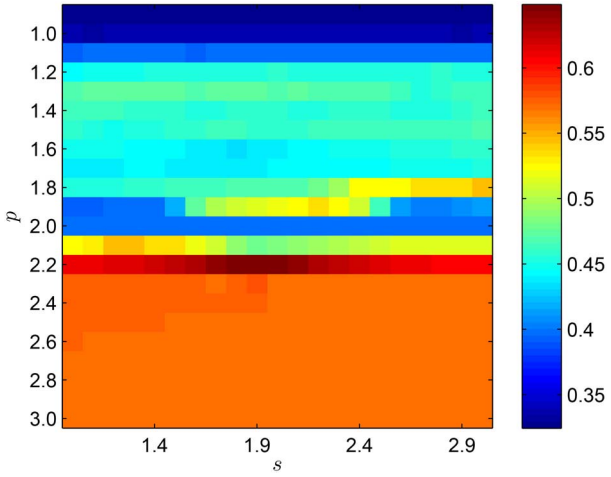


Fig. 18. Classification accuracies of G2DPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the FERET database.

The highest classification accuracies and corresponding parameters of the four algorithms are listed in Table V, including the results of 2DPCA-L1, 2DPCA, PCA-L1, and PCA for comparison. Fig. 21 shows the classification accuracies of the four algorithms with different feature numbers when respective optimal parameters are applied. From the results, the classification performance of G2DPCA is much better than that of 2DPCAL1-S, and the classification performance of GPCA is worse than that of RSPCA. The result also demonstrates that

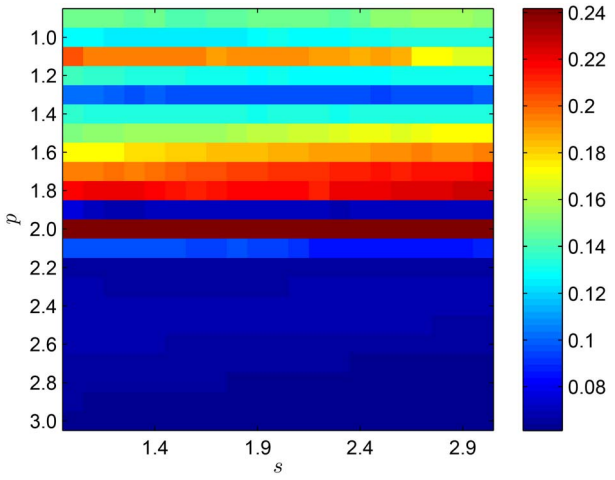


Fig. 19. Classification accuracies of GPCA with $s = [1.0{:}0.1{:}3.0]$ and $p = [0.9{:}0.1{:}3.0]$ on the FERET database.

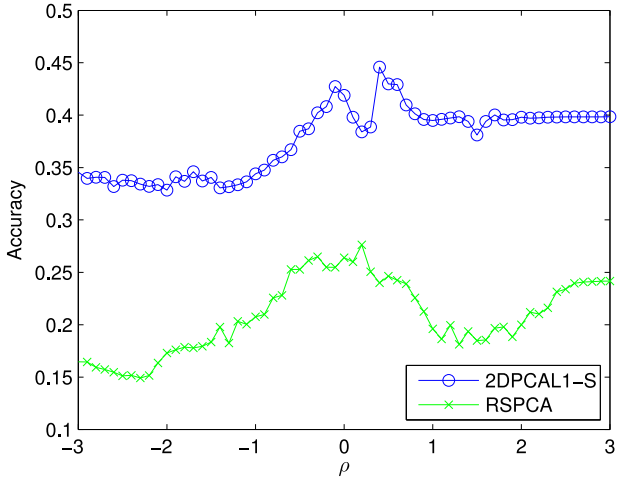


Fig. 20. Classification accuracies of 2DPCAL1-S and RSPCA with $\rho = [-3.0{:}0.1{:}3.0]$ on the FERET database.

TABLE V
CLASSIFICATION ACCURACIES OF EIGHT ALGORITHMS ON THE FERET DATABASE

Algorithms	Optimal parameters	Accuracy
G2DPCA	$s = 1.9$, $p = 2.2$	**0.6484**
2DPCAL1-S	$\rho = 0.4$	0.4458
2DPCA-L1	-	0.3985
2DPCA	-	0.3983
GPCA	$s = 1.3$, $p = 2.0$	0.2417
RSPCA	$\rho = 0.2$	0.2763
PCA-L1	-	0.2415
PCA	-	0.2406

applying Lp-norm both in objective and constraint functions of 2DPCA could greatly improve its classification performance. However, the same operation on PCA just slightly improves its classification performance. The best classification performance among the 1-D algorithms on the FERET database is achieved by RSPCA.

As for the sparsity of the projection vectors of G2DPCA with optimal parameters, all of the results turn out to be dense since the optimal p value is larger than 2.0.

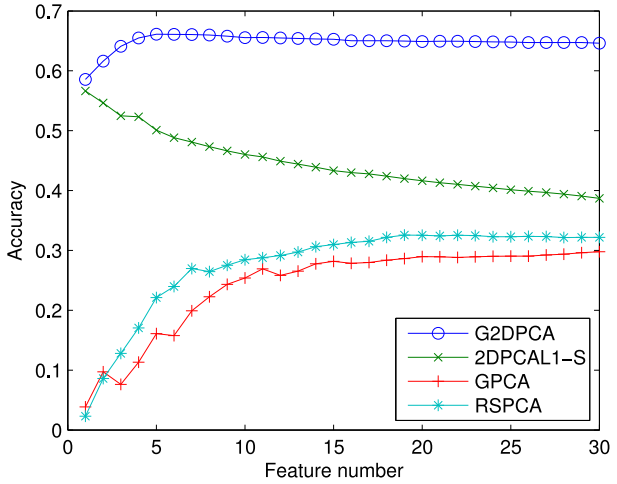


Fig. 21. Classification accuracies of G2DPCA, 2DPCAL1-S, GPCA, and RSPCA with different feature numbers on the FERET database when respective optimal tuning parameters are applied.

VI. CONCLUSION

A general 2DPCA algorithm based on Lp-norm, called G2DPCA is proposed for image analysis in this paper. It applies Lp-norm both in objective and constraint functions of conventional 2DPCA. An iterative algorithm is designed to solve the optimization problem of G2DPCA under the MM framework, and a closed-form solution is obtained in each iteration. Then a deflating scheme is employed to extract multiple projection vectors. The solution of G2DPCA is guaranteed to be locally optimal.

Two face databases, i.e., ORL and FERET are employed in the experiments regarding image reconstruction and classification. Generally speaking, all results from the experiments show that G2DPCA is insensitive to s value but sensitive to p value. In task of image reconstruction, the optimal reconstruction performance of G2DPCA is achieved when it reduces to 2DPCA-L1. In task of image classification, the optimal (s, p) pair differs on different databases, $(2.9, 1.5)$ for the ORL database, and $(1.9, 2.2)$ for the FERET database, respectively. Our results demonstrate the superiority of G2DPCA in image classification over seven existing algorithms, i.e., 2DPCAL1-S, 2DPCA-L1, 2DPCA, GPCA, RSPCA, PCA-L1, and PCA. However, how to determine the optimal (s, p) pair theoretically remains to be an unsolved problem. Another unsolved problem is to find the locally optimal solution for G2DPCA with $0 < s < 1$ and $p > 0$ if it exists. Additionally, it is worthwhile to extend some L1-norm based subspace learning algorithms such as linear discriminant analysis with L1-norm (LDA-L1) [30], [31] and discriminant locality preserving projections with L1-norm (DLPP-L1) [32] to corresponding Lp-norm cases.

Some questions that remain unclear concerning the experimental results are listed below. First, the accuracy by some 2-D algorithms is decreasing with feature number on the FERET database, as shown in Fig. 21. This is strange. Second, on the FERET database, the optimal classification performance of G2DPCA is better than that of 2DPCAL1-S, but the optimal classification performance of GPCA is worse than

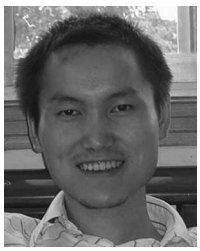
that of RSPCA. Considering that GPCA and RSPCA are the 1-D counterparts of G2DPCA and 2DPCAL1-S, respectively, this result is difficult to explain. Third, among the four 2-D algorithms, i.e., G2DPCA, 2DPCAL1-S, 2DPCA-L1, and 2DPCA, the best reconstruction performance is obtained by 2DPCA-L1 or by G2DPCA and 2DPCAL1-S when they reduce to 2DPCA-L1; among the four 1-D algorithms, i.e., GPCA, RSPCA, PCA-L1, and PCA, the best reconstruction performance is obtained by PCA-L1 or by GPCA and RSPCA when they reduce to PCA-L1. It is difficult to explain why G2DPCA, 2DPCAL1-S, GPCA, and RSPCA could not achieve better reconstruction performances considering the flexibility of their tuning parameters. These questions might be discussed in the future work when the performances of these algorithms on more databases are examined and when we know more about the intrinsic properties of these algorithms.

ACKNOWLEDGMENT

The author would like to thank Prof. H. Wang and Prof. G. Xue for their supervision and the reviewers for their valuable insights which help to enrich and consolidate this paper.

REFERENCES

[1] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer, 2004.
[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[3] Q. Ke and T. Kanade, "Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. San Diego, CA, USA, 2005, pp. 739–746.
[4] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2006, pp. 281–288.
[5] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.
[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[7] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.
[8] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, 2007.
[9] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *J. Multivar. Anal.*, vol. 99, no. 6, pp. 1015–1034, 2008.
[10] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.
[11] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse PCA by L_1-norm maximization," *Pattern Recognit.*, vol. 45, no. 1, pp. 487–497, 2012.
[12] N. Kwak, "Principal component analysis by L_p-norm maximization," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 594–609, May 2014.
[13] Z. Liang, S. Xia, Y. Zhou, L. Zhang, and Y. Li, "Feature extraction based on Lp-norm generalized principal component analysis," *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1037–1045, 2013.
[14] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. Int. Conf. Mach. Learn.*, vol. 98. Madison, WI, USA, 1998, pp. 82–90.
[15] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, Mar. 2010.
[16] 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.
[17] 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.
[18] H. Wang and J. Wang, "2DPCA with L1-norm for simultaneously robust and sparse modelling," *Neural Netw.*, vol. 46, pp. 190–198, Oct. 2013.
[19] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.
[20] L. Mackey, "Deflation methods for sparse PCA," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21. Whistler, BC, Canada, 2008, pp. 1017–1024.
[21] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *J. Comput. Graph. Stat.*, vol. 9, no. 1, pp. 1–20, 2000.
[22] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2001, pp. 556–562.
[23] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.
[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[25] W. H. Yang, "On generalized Hölder inequality," *Nonlin. Anal. Theory Methods Appl.*, vol. 16, no. 5, pp. 489–498, 1991.
[26] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 217–224.
[27] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, Sarasota, FL, USA, 1994, pp. 138–142.
[28] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[29] R. Martí, "Multi-start methods," in *Handbook of Metaheuristics*. New York, NY, USA: Springer, 2003, pp. 355–368.
[30] F. Zhong and J. Zhang, "Linear discriminant analysis based on L1-norm maximization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.
[31] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L1-norm," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.
[32] F. Zhong, J. Zhang, and D. Li, "Discriminant locality preserving projections based on L1-Norm maximization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2065–2074, Nov. 2014.



Jing Wang received the B.S. degree from the School of Biological Science and Medical Engineering, Southeast University, Nanjing, China, in 2010, where he is currently pursuing the Ph.D. degree with the Research Center for Learning Science.

He was a Visiting Student with the Center for Brain and Learning Sciences, Beijing Normal University, Beijing, China, in 2012. His current research interests include neuroimaging and machine learning.