Generalized two dimensional principal component analysis by Lp-norm for image analysis

Jing Wang

***Abstract*—This paper proposes a generalized two dimensional principal component analysis (G2DPCA) by replacing the L2-norm in conventional two dimensional principal component analysis (2DPCA) with Lp-norm, both in the objective function and the constraint function. It is a generalization of previously proposed robust or sparse 2DPCA algorithms. Under the framework of minorization-maximization (MM), 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*—generalized 2DPCA (G2DPCA), Lp-norm, convex maximization, minorization-maximization (MM), image analysis.**

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. Sparse modeling seeks salient features from training data. It is an effective tool with applications in various areas such as signal processing, machine learning and pattern recognition, providing not only good interpretability but also excellent generalization performance [6]. However, traditional PCA and its robust improvements all generate dense results which are unsatisfactory. By applying L0-norm or L1-norm on the constraint function (also called penalty or regularization [7], [8]) of PCA, sparsity could be introduced to deal with this problem, resulting in a series of sparse PCA (SPCA) algorithms [9], [10], [11], [12], [13], [14], [15]. Problems with L0-norm are NP-hard and difficult to solve. Fortunately, developments in the theory of sparse representation and compressed sensing [16], [17], [18] have demonstrated that the solution of L0-norm problem is equivalent with the solution of corresponding L1-norm problem if certain conditions are satisfied. The L1-norm penalty alone is known as LASSO [19]. Combining L1-norm penalty and L2-norm penalty together will generate a mixed-norm penalty known as Elastic-Net [8]. Both LASSO and Elastic-Net have been extensively studied in the above SPCA algorithms.

J. Wang is with Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: wangjing0@seu.edu.cn, yuzhounh@gmail.com).

A newly proposed algorithm called robust sparse PCA (RSPCA) [20] utilizes L1-norm both on the objective function and the constraint function of traditional PCA, inheriting the merits of robustness and sparsity. Considering that L0-norm, L1-norm and L2-norm are all special cases of Lp-norm, it's natural to replace the L2-norm in traditional PCA with arbitrary norm, both on its objective function and constraint function, as proposed in generalized PCA (GPCA) [21].

Lp-norm based learning algorithms have received increasing attentions in recent years. Besides GPCA, several other typical examples include Lp-norm generalization of the least mean squared algorithm [22], [23], sparse logistic regression with Lp-norm penalty [24], Lp-norm multiple kernel learning [25], [26], Lp-norm sparse coding [27], [28], [29], [30], [31], [32], sparse support vector machine with Lp-norm penalty [33], Lp-norm multiple kernel Fisher discriminant analysis [34], generalized linear discriminant analysis with Lp-norm [35], PCA with Lp-norm (PCA-Lp) [36].

When Lp-norm is imposed on the objective function of PCA, $p \geq 1$ is required to guarantee the convexity of the optimization problem [21]. PCA-Lp [36] technically avoids the singular condition that may occur when $0 < p < 1$ by slightly moving the projection vector, but this trick could not guarantee to find a locally optimal solution theoretically. Therefore, for the Lp-norm that is imposed on the objective function, we limit our attention to the case of $p \geq 1$.

When Lp-norm is imposed on the constraint function, it could bridge the gap between the three traditional norms, i.e., L0-norm, L1-norm and L2-norm. On one hand, Lp-norm with $0 < p < 1$ is a tradeoff between L0-norm and L1-norm. It is nonconvex and it does not satisfy the definition of norm in mathematics. But it is very useful when the conditions that guarantee the equivalence between L0-norm problem and L1-norm problem are not satisfied. Studies in image restoration, feature selection, image classification, and compressed sensing [37], [38], [31], [32] have demonstrated that Lp-norm could achieve better solutions than L1-norm when $0 < p < 1$.

On the other hand, Lp-norm with $1 < p < 2$ is a tradeoff between L1-norm and L2-norm, thus behaving like the Elastic-Net. It's demonstrated that Lp-norm with $p > 1$ could facilitate structural sparsity [39] and group sparsity [28] in a model, consistent with the grouping effect found by Elastic-Net [8]. Also, experiments show that GPCA with $p > 1$ obtains the lowest error rates on three out of five UCI data sets [21]. Therefore, it's of great importance to explore Lp-norm constraint function with $p > 1$.

Among the Lp-norm based algorithms that are intended to improve traditional PCA, PCA-Lp and GPCA are two typical

representations. 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 [36]. In GPCA, the Lp-norm is imposed both on the objective function and the constraint function of PCA. The successive linearization technique (SLT) [40], [14] is employed to solve GPCA [21].

The above robust and sparse PCA algorithms are intrinsically image-as-vector methods. That is to say, when they are applied in image analysis, each image should be reshaped into a long vector in prior. By this way, the spatial information in images are destroyed and extensive computations are usually inevitable due to high dimensionality of reshaped images. Image-as-matrix methods represented by 2DPCA [41] offer insights for improving these robust and sparse PCA algorithms. Two related improvements are L1-Norm-Based 2DPCA (2DPCA-L1) [42] and 2DPCA-L1 with sparsity (2DPCAL1-S) [43]. 2DPCA-L1 is formulated by imposing L1-norm on the objective function of 2DPCA. 2DPCAL1-S is formulated by imposing L1-norm both on the objective function and the constraint function of 2DPCA. Iterative solutions for the two algorithms have been designed accordingly.

This paper proposes a generalized 2DPCA (G2DPCA) algorithm by replacing the L2-norm of conventional 2DPCA with Lp-norm, on both the objective function and the constraint function, thus greatly extending previous 2DPCA-based algorithms. The proposed G2DPCA is closely related to GPCA [21]. Besides the image-as-matrix representation, G2DPCA differs from GPCA by designing an elegant solution under the framework of minorization-maximization (MM) [44], [45] rather than SLT. MM theoretically guarantees to find a locally optimal solution for an optimization problem while SLT intends to linearize a nonsmooth function, thus MM is more reliable than SLT.

The proposed G2DPCA mainly has the following advantages. First, it's an image-as-matrix method, thus preserving spatial information in images and at the same time avoiding extensive computations required by corresponding image-as-vector algorithms. This also makes it possible to search for the optimal parameters thoroughly. Second, it closely relates to robust and sparse 2DPCA algorithms such as 2DPCA-L1 and 2DPCAL1-S, thus inheriting the merits of robustness and sparsity from these algorithms. Also, the Lp-norm incorporated greatly enriches the connotation of robust and sparse 2DPCA algorithms. Third, we design an iterative algorithm under the framework of MM to solve G2DPCA. A closed-form solution could be obtained in each iteration. Our algorithm guarantees to find a locally optimal solution for G2DPCA.

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. And 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; $\mathbf{X}^T$ denotes the transpose of matrix $\mathbf{X}$; $sign(\cdot)$ 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; $diag(\mathbf{w})$ 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-norm, L2-norm, Lp-norm and Frobenius norm respectively. Note that the sign function and the absolute value function could be applied on a vector in the element-wise manner, similar to the element-wise power.

For a vector $\mathbf{x} \in \mathbb{R}^n$ and a scalar $p > 0$, the Lp-norm of $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}. \tag{1}$$

Illustrations of Lp-norm in 1D and 2D space are shown in Fig. 1. The left panel shows that when $|x| > 1$, $|x|^2$ increases much faster with $|x|$ than $|x|$ does. This property guarantees that introducing L1-norm to the objective function could resist noises. From both subfigures, it could be observed that Lp-norm with $1 < p < 2$ is a tradeoff between L1-norm and L2-norm, thus behaving like the Elastic-Net. These are some of the intuitive reasons to improve 2DPCA-based algorithms by employing Lp-norm.

Traditionally, robust and sparse 2DPCA algorithms focus on finding a single projection vector each time, then a deflation scheme is implemented to extract multiple projection vectors [42], [43]. This strategy is also adopted in this paper.

A. 2DPCA

Suppose there are n training image samples $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$, where $\mathbf{X}_i \in \mathbb{R}^{h \times w}$, $i = 1, 2, \cdots, n$. h and w are the height and width of the images respectively. The images are assumed to be mean-centered, i.e., $\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i = \mathbf{0}$. 2DPCA [41] finds its first projection vector $\mathbf{w} \in \mathbb{R}^w$ by solving the following optimization problem

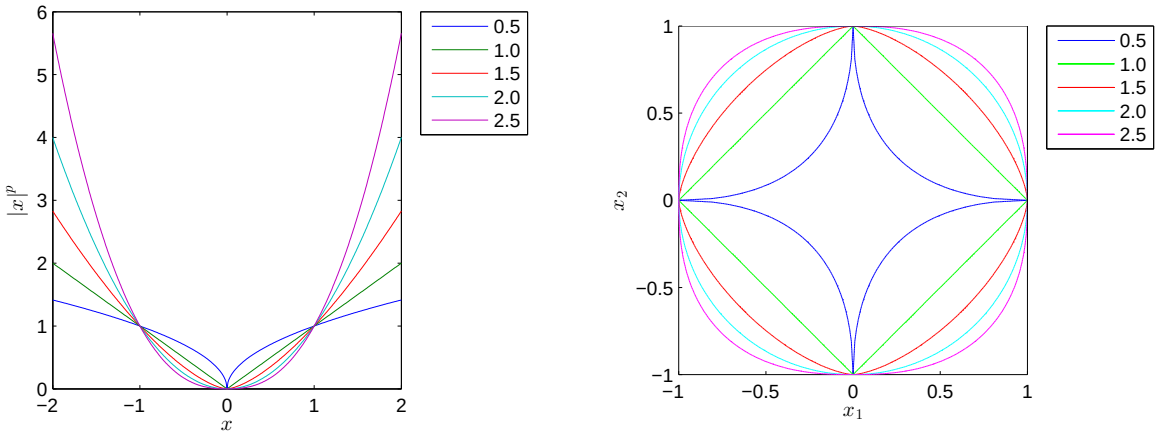
$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_2^2, \quad s.t.\ \|\mathbf{w}\|_2^2 = 1. \tag{2}$$

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. The solution of 2DPCA is neither robust nor sparse, thus facilitating its following alternatives.

B. 2DPCA-L1

2DPCA-L1 [42] 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 below

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_1, \quad s.t.\ \|\mathbf{w}\|_2^2 = 1. \tag{3}$$



(a) (b)

Fig. 1. (a) Diagrammatic representation of the function $|x|^p$. The p values correspond to the five curves are shown in the legend. (b) Diagrammatic representation of $\|\mathbf{x}\|_p^p = 1$ in 2D space, i.e., contour plot of $|x_1|^p + |x_2|^p = 1$. The p values correspond to the five contour lines are shown in the legend.

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 k-th step, then $\mathbf{w}^{k+1}$ could be updated by

$$\mathbf{v}^k = \sum_{i=1}^{n} \mathbf{X}_i^T sign(\mathbf{X}_i \mathbf{w}^k), \tag{4}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{v}^k}{\|\mathbf{v}^k\|_2}. \tag{5}$$

C. 2DPCAL1-S

2DPCAL1-S [43] could be formulated by applying L1-norm both on the objective function and the constraint function of 2DPCA as follows

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i \mathbf{w}\|_1, \quad s.t.\ \|\mathbf{w}\|_1 \leq c, \|\mathbf{w}\|_2^2 = 1, \tag{6}$$

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

$$\mathbf{v}^k = \sum_{i=1}^{n} \mathbf{X}_i^T sign(\mathbf{X}_i \mathbf{w}^k), \tag{7}$$

$$u_i^k = v_i^k \frac{|w_i^k|}{\lambda + |w_i^k|}, \ i = 1, 2, \cdots, w, \tag{8}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_2}, \tag{9}$$

where $\mathbf{u}^k \in \mathbb{R}^w$ is a vector; w_i^k, v_i^k and u_i^k are the i-th 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.

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. Without further declarations, this kind of notation is adopted likewise throughout this paper.

D. G2DPCA

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

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i \mathbf{w}\|_s^s, \quad s.t.\ \|\mathbf{w}\|_p^p = 1, \tag{10}$$

where $s \geq 1$, $p > 0$.

It's obvious that 2DPCA and 2DPCA-L1 are two special cases of G2DPCA. 2DPCAL1-S is a bit different since it combines L1-norm constraint and L2-norm constraint together. If the L1-norm constraint is eliminated from 2DPCAL1-S, then it reduces to 2DPCA-L1. If the L2-norm constraint is eliminated from 2DPCAL1-S and $c = 1$, then it reduces to G2DPCA with $s = 1$ and $p = 1$. It will be shown in Section IV that the projection vector $\mathbf{w}$ in G2DPCA with $s = 1$ and $p = 1$ has no more than one nonzero element. This is inappropriate from the respect of feature extraction. Intuitively, the L2-norm constraint is employed to fix the problem encountered by G2DPCA with $s = 1$ and $p = 1$, resulting in 2DPCAL1-S. From another viewpoint, 2DPCAL1-S might be approximated by G2DPCA with $s = 1$ and $1 \leq p \leq 2$ since the mixed-norm constraint could be approximated by Lp-norm constraint, as shown in Fig. 1. There still remains a difference that in 2DPCAL1-S the equality constraint is imposed on the L2-norm rather than the mixed-norm while in G2DPCA the equality constraint is imposed on the Lp-norm. For a conclusion, 2DPCAL1-S is closely related to G2DPCA, but it is still unique.

The objective function in G2DPCA could be replaced by other types of convex functions as in the work of GPCA [21]. Here we limit our attention to the optimization problem in (10) for three reasons. First, from the theoretical viewpoint, the basic ideas would be similar when solving different problems formulated by changing the objective function. The problem in (10) is representative, relating closely to other classical 2DPCA-based algorithms as shown before. Therefore, it suffices to show the superiority of G2DPCA. It is also impossible

to validate all potential convex objective functions. Second, from the experimental viewpoint, the results of GPCA in image classification tasks in [21] show that the objective function in (10) outperforms another objective function where only two objective functions are compared. Therefore, it is reasonable to focus on the objective function in (10). Third, we also want to check how the s value would affect the performance of G2DPCA in image reconstruction and classification. The formulations of 2DPCA and 2DPCA-L1 differ only on the s value. It's likely that the s value would make certain influences on our results. For these reasons, we will focus on the problem in (10) in the following paper.

The optimization problem of G2DPCA is intrinsically a convex maximization problem which is hard to solve in general cases [7]. A major property of convex maximization, known as the maximum principle [46] states that the maximum of a convex function on a compact set is only obtained on the boundary. It means that the following optimization problem

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s, \quad s.t.\ \|\mathbf{w}\|_p^p \leq 1 \tag{11}$$

is equivalent with (10). Therefore, with a slight abuse of notation, we also refer $\|\mathbf{w}\|_p^p = 1$ with $0 < p < 1$ as a nonconvex set, and refer $\|\mathbf{w}\|_p^p = 1$ with $p \geq 1$ as a convex set.

Whether the global optimum of a convex maximization problem could be guaranteed theoretically still remains to be an open problem. To obtain good locally optimal solutions, we might borrow ideas from a wider field termed global optimization [47]. Some general methods for global optimization include deterministic methods, e.g., branch and bound, cutting plane; stochastic methods, e.g., simulated annealing, Monte-Carlo sampling; heuristic methods, e.g., evolutionary algorithms, swarm-based optimization algorithms; and response surface methodology based methods, e.g., Bayesian optimization. For recent reviews, see [48], [49].

The topic of global optimization goes beyond the scope of this paper. Nevertheless, a special case of the stochastic methods, i.e., the multistart method [50], [51] is widely suggested to be an efficient method for finding a good locally optimal solution in PCA-based algorithms [5], [20], [21], [36]. 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 G2DPCA by the corresponding projection vectors of 2DPCA. This method makes the most of the relationship between 2DPCA and G2DPCA, therefore it is expected to find a good locally optimal solution. The disadvantage is that we have to calculate 2DPCA in prior which is tolerable in practice.

Based on the above discussions, we target a locally optimal solution for the optimization problem of G2DPCA in (10) throughout this paper. An iterative algorithm under the MM framework is designed for this purpose. Only a single projection vector is calculated each time. The solution would be given in Section IV.

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

$$\mathbf{X}_i^{deflated} = \mathbf{X}_i(\mathbf{I} - \mathbf{W}\mathbf{W}^T),\ i = 1, 2, \cdots, n. \tag{12}$$

This deflation procedure is implemented repeatedly to extract multiple projection vectors. For a comparison of different deflation schemes, see [52]. Though the deflating scheme offers a possible way for solving 2DPCA, it is certainly more preferable to be solved by eigen decomposition alone. For the other three 2DPCA-based algorithms, however, there seems to be no better choices yet.

III. Related techniques

Before proceeding to the solution of G2DPCA problem, we will first introduce some related techniques that would be utilized, including the MM framework, the first-order convexity condition, and a linear optimization problem with Lp-norm constraint. The MM framework is employed to turn the nonsmooth optimization problem of G2DPCA into smooth ones. The inequalities derived from the first-order convexity condition would play a central role in designing algorithms under the MM framework. When $p \geq 1$, the nonlinear optimization problem of G2DPCA could be turned into iteratively optimizing a linear problem with Lp-norm constraint. The analytic solution of the new problem would be given based on the Hölder's inequality.

A. The MM framework

Suppose $f(\mathbf{w})$ is the objective function to be maximized. Under the MM framework [44], if there exists a surrogate function $g(\mathbf{w}|\mathbf{w}^k)$ which satisfies two key conditions that

$$f(\mathbf{w}^k) = g(\mathbf{w}^k|\mathbf{w}^k), \tag{13}$$

$$f(\mathbf{w}) \geq g(\mathbf{w}|\mathbf{w}^k) \text{ for all } \mathbf{w}, \tag{14}$$

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

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} g(\mathbf{w}|\mathbf{w}^k). \tag{15}$$

One could see that

$$\begin{aligned} f(\mathbf{w}^{k+1}) &= f(\mathbf{w}^{k+1}) - g(\mathbf{w}^{k+1}|\mathbf{w}^k) + g(\mathbf{w}^{k+1}|\mathbf{w}^k) \\ &\geq f(\mathbf{w}^k) - g(\mathbf{w}^k|\mathbf{w}^k) + g(\mathbf{w}^{k+1}|\mathbf{w}^k) \\ &\geq f(\mathbf{w}^k) - g(\mathbf{w}^k|\mathbf{w}^k) + g(\mathbf{w}^k|\mathbf{w}^k) \\ &= f(\mathbf{w}^k) \end{aligned} \tag{16}$$

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 $\mathbf{w} = \mathbf{w}^{k+1}$. 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 problem, 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 [44]. The MM framework is also referred to as "optimization transfer" [45], "auxiliary function method" [53] or "bound optimization" [54] in other literatures.

B. The first-order convexity condition

Inequalities play a central role in designing MM algorithms. Below are some inequalities derived from the first-order condition of convex functions [7] which will be utilized to solve G2DPCA. Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, given a convex and differentiable function $f(\mathbf{w})$ defined on a real vector space, we have

$$f(\mathbf{w}) \geq f(\mathbf{v}) + \nabla f(\mathbf{v})^T(\mathbf{w}-\mathbf{v}), \tag{17}$$

wherein the equality holds when $\mathbf{w} = \mathbf{v}$. The equality defines a supporting hyperplane to the feasible set of $\mathbf{w}$ at the point $\mathbf{v}$. Intuitively, Any linear function tangent to the graph of a convex function lies below the function and is a minorizer at the point of tangency [44].

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

$$\|\mathbf{w}\|_p^p \geq p\left[|\mathbf{v}|^{p-1} \circ sign(\mathbf{v})\right]^T \mathbf{w} + (1-p)\|\mathbf{v}\|_p^p \tag{18}$$

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

Proof. If all elements in $\mathbf{v}$ are not zeros, then $\|\mathbf{w}\|_p^p$ is differentiable at $\mathbf{w} = \mathbf{v}$, we have

$$\left.\frac{\partial \|\mathbf{w}\|_p^p}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{v}} = p\,|\mathbf{v}|^{p-1} \circ sign(\mathbf{v}). \tag{19}$$

Additionally, $\|\mathbf{w}\|_p^p$ with $p \geq 1$ is convex. Then the objective inequality (18) could be easily derived from the first-order convexity condition and the inequality becomes equality when $\mathbf{w} = \mathbf{v}$.

If any element in $\mathbf{v}$ is zero, $\|\mathbf{w}\|_p^p$ would not be differentiable at $\mathbf{w} = \mathbf{v}$ since the absolute value function is not differentiable at the zero point, therefore the objective inequality could not be obtained directly. Fortunately, it could be expanded into element form as

$$\sum_{i=1}^{d} |w_i|^p \geq p\sum_{i=1}^{d} |v_i|^{p-1} sign(v_i)w_i + (1-p)\sum_{i=1}^{d} |v_i|^p. \tag{20}$$

This inequality holds if

$$|w_i|^p \geq p\,|v_i|^{p-1} sign(v_i)w_i + (1-p)\,|v_i|^p,\ i = 1, 2, \cdots, d. \tag{21}$$

For any $v_i \neq 0$, the inequality in (21) holds since $|w_i|^p$ is convex and differentiable at $w_i = v_i$. At the same time, the inequality becomes equality when $w_i = v_i$. For any $v_i = 0$, the inequality in (21) reduces to $|w_i|^p \geq 0$ since $sign(0) = 0$ and $0^0 = 1$. The reduced inequality is always true. At the same time, the inequality becomes equality when $w_i = 0$, thus $w_i = v_i$. To summarize, (21) holds and the inequalities becomes equalities when $w_i = v_i$ no matter v_i is zero or not, for all $i = 1, 2, \cdots, d$. Consequently, (18) holds and the inequality becomes equality when $\mathbf{w} = \mathbf{v}$ no matter $\mathbf{v}$ has zero elements or not. This completes the proof. Lemma 1 relaxes $\|\mathbf{w}\|_p^p$ with $p \geq 1$ to a linear function which becomes much easier to handle.

When $p = 1$, (18) reduces to

$$\|\mathbf{w}\|_1 \geq sign(\mathbf{v})^T\mathbf{w} \tag{22}$$

wherein the inequality becomes equality when $\mathbf{w} = \mathbf{v}$. This inequality is widely used in algorithms where L1-norm is imposed on the objective function of PCA or 2DPCA [5], [20], [42], [43].

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

$$\|\mathbf{w}\|_p^p \leq p\left[|\mathbf{v}|^{p-1} \circ sign(\mathbf{v})\right]^T \mathbf{w} + (1-p)\|\mathbf{v}\|_p^p \tag{23}$$

holds wherein 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, in other words, all of the elements in $\mathbf{v}$ are not zeros, let $0 < p < 2$, then

$$\|\mathbf{w}\|_p^p \leq \frac{p}{2}\mathbf{w}^T diag\left(|\mathbf{v}|^{p-2}\right)\mathbf{w} + (1-\frac{p}{2})\|\mathbf{v}\|_p^p \tag{24}$$

holds wherein 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 $0 < p/2 < 1$. By treating $\mathbf{w} \circ \mathbf{w}$, $\mathbf{v} \circ \mathbf{v}$ and $p/2$ as a whole respectively, we could derive from Lemma 2 that

$$\begin{aligned}
\|\mathbf{w}\|_p^p =& \|\mathbf{w} \circ \mathbf{w}\|_{p/2}^{p/2} \\
\leq& \frac{p}{2}\left[|\mathbf{v} \circ \mathbf{v}|^{p/2-1}\right]^T (\mathbf{w} \circ \mathbf{w}) + (1-\frac{p}{2})\|\mathbf{v} \circ \mathbf{v}\|_{p/2}^{p/2} \\
=& \frac{p}{2}\mathbf{w}^T diag\left(|\mathbf{v} \circ \mathbf{v}|^{p/2-1}\right)\mathbf{w} + (1-\frac{p}{2})\|\mathbf{v} \circ \mathbf{v}\|_{p/2}^{p/2} \\
=& \frac{p}{2}\mathbf{w}^T diag\left(|\mathbf{v}|^{p-2}\right)\mathbf{w} + (1-\frac{p}{2})\|\mathbf{v}\|_p^p,
\end{aligned} \tag{25}$$

wherein the inequality becomes equality when $\mathbf{w} \circ \mathbf{w} = \mathbf{v} \circ \mathbf{v}$. The equality could be further guaranteed by $\mathbf{w} = \mathbf{v}$. Therefore, the inequalities in (24) and (25) 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 (24) into element form in order to examine the zero points. That is

$$\sum_{i=1}^{d} |w_i|^p \leq \frac{p}{2}\sum_{i=1}^{d} w_i^2\,|v_i|^{p-2} + (1-\frac{p}{2})\sum_{i=1}^{d} |v_i|^p. \tag{26}$$

The expanded inequality holds if the following inequalities hold

$$|w_i|^p \leq \frac{p}{2}w_i^2\,|v_i|^{p-2} + (1-\frac{p}{2})\,|v_i|^p,\ i = 1, 2, \cdots, d. \tag{27}$$

For any $w_i \neq 0$, the corresponding inequality 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, the inequalities in (27) hold no matter w_i is zero or not. But the inequalities would become equalities only when $w_i = v_i$ in which case w_i is not zero, $i = 1, 2, \cdots, d$.

To summarize, (24) 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 would be much easier to handle.

When $p = 1$, (24) reduces to

$$\|\mathbf{w}\|_1 \leq \frac{1}{2}\mathbf{w}^T diag\left(|\mathbf{v}|^{-1}\right)\mathbf{w} + \frac{1}{2}\|\mathbf{v}\|_1 \tag{28}$$

wherein $\mathbf{v}$ has no zero element. The equality holds when $\mathbf{w} = \mathbf{v}$. This inequality is used to design the solution for 2DPCAL1-S in [43].

C. A 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 $1/p + 1/q = 1$, then the Hölder's inequality [55] states that

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

The equality holds if and only if there exists a positive real scalar c satisfying $|w_i|^p = c\,|v_i|^q$, $i = 1, 2, \cdots, d$. Based on the Hölder's inequality, we will give a closed-form solution to a linear optimization problem with Lp-norm constraint as below.

Lemma 4. Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v} \neq \mathbf{0}$, and let $p,\ q \in [1, \infty]$ be two scalars satisfying $1/p + 1/q = 1$. Specifically, $\mathbf{v} \neq \mathbf{0}$ means that $\mathbf{v}$ is not a vector with all zeros. Then the optimization problem

$$\max_{\mathbf{w}} \mathbf{v}^T\mathbf{w}, \quad s.t.\ \|\mathbf{w}\|_p^p = 1 \tag{30}$$

has a closed-form solution

$$\mathbf{w} = \frac{|\mathbf{v}|^{q-1} \circ sign(\mathbf{v})}{\|\mathbf{v}\|_q^{q-1}}. \tag{31}$$

Proof. According to the Hölder's inequality in (29), we have

$$\mathbf{v}^T\mathbf{w} \leq \sum_{i=1}^{d} |v_i w_i| \leq \|\mathbf{v}\|_q \|\mathbf{w}\|_p = \|\mathbf{v}\|_q. \tag{32}$$

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

$$sign(w_i) = sign(v_i),\ i = 1, 2, \cdots, d. \tag{33}$$

The second equality holds when

$$|w_i|^p = c\,|v_i|^q,\ i = 1, 2, \cdots, d. \tag{34}$$

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

$$c = \frac{\sum_{i=1}^{d} |w_i|^p}{\sum_{i=1}^{d} |v_i|^q} = \frac{\|\mathbf{w}\|_p^p}{\|\mathbf{v}\|_q^q} = \frac{1}{\|\mathbf{v}\|_q^q}. \tag{35}$$

Substituting (35) into (34), we have

$$|w_i| = \left(c\,|v_i|^q\right)^{1/p} = \left(\frac{|v_i|^q}{\|\mathbf{v}\|_q^q}\right)^{1/p} = \frac{|v_i|^{q-1}}{\|\mathbf{v}\|_q^{q-1}},\ i = 1, 2, \cdots, d. \tag{36}$$

Considering (33), we have

$$w_i = \frac{|v_i|^{q-1}}{\|\mathbf{v}\|_q^{q-1}} sign(v_i),\ i = 1, 2, \cdots, d. \tag{37}$$

Rewriting the equations into vector form will complete this proof.

IV. The solution of generalized 2DPCA

With the above techniques, the solution for G2DPCA in (10) 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 [21].

A. Case 1

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

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s, \quad s.t.\ \|\mathbf{w}\|_p^p = 1, \tag{38}$$

where $s \geq 1$, $p \geq 1$, $\mathbf{w} \in \mathbb{R}^w$. This problem could be turned into iteratively maximizing a surrogate function under the MM framework, as shown below. Let $\mathbf{w}^k$ be the projection vector at the k-th 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 convex objective function could be linearized as

$$\begin{aligned}\sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s \geq & s\sum_{i=1}^{n} \left[\left|\mathbf{X}_i\mathbf{w}^k\right|^{s-1} \circ sign(\mathbf{X}_i\mathbf{w}^k)\right]^T \mathbf{X}_i\mathbf{w} \\ & + (1-s)\sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}^k\|_s^s,\end{aligned} \tag{39}$$

wherein the inequality becomes equality when $\mathbf{w} = \mathbf{w}^k$. Let the objective function be denoted as $f(\mathbf{w})$, and let the linearized function be denoted as $g(\mathbf{w}|\mathbf{w}^k)$. That is to say

$$f(\mathbf{w}) = \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s, \tag{40}$$

and

$$\begin{aligned}g(\mathbf{w}|\mathbf{w}^k) = & s\sum_{i=1}^{n} \left[\left|\mathbf{X}_i\mathbf{w}^k\right|^{s-1} \circ sign(\mathbf{X}_i\mathbf{w}^k)\right]^T \mathbf{X}_i\mathbf{w} \\ & + (1-s)\sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}^k\|_s^s.\end{aligned} \tag{41}$$

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 framwork. Therefore, $g(\mathbf{w}|\mathbf{w}^k)$ is a feasible surrogate function of $f(\mathbf{w})$. According to the MM framework, the optimization problem in (38) could be turned into iteratively maximizing the surrogate function as follows

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} g(\mathbf{w}|\mathbf{w}^k), \quad s.t.\ \|\mathbf{w}\|_p^p = 1. \tag{42}$$

Define

$$\mathbf{v}^k = \sum_{i=1}^{n} \mathbf{X}_i^T \left[\left|\mathbf{X}_i\mathbf{w}^k\right|^{s-1} \circ sign(\mathbf{X}_i\mathbf{w}^k)\right]. \tag{43}$$

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

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} (\mathbf{v}^k)^T\mathbf{w}, \quad s.t.\ \|\mathbf{w}\|_p^p = 1. \tag{44}$$

According to Lemma 4, the solution of this problem is

$$\mathbf{w}^{k+1} = \frac{\left|\mathbf{v}^k\right|^{q-1} \circ sign(\mathbf{v}^k)}{\|\mathbf{v}^k\|_q^{q-1}}, \tag{45}$$

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

$$\mathbf{u}^k = \left|\mathbf{v}^k\right|^{q-1} \circ sign(\mathbf{v}^k), \tag{46}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_p}. \tag{47}$$

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 $j = \arg\max_{i\in[1,w]} \left|v_i^k\right|$, i.e., $\left|v_j^k\right|$ is the largest value in $\left|\mathbf{v}^k\right|$. By taking the limit of (45) when p approaches 1, we have

$$w_i^{k+1} = \begin{cases} sign(v_j^k), & i = j, \\ 0, & i \neq j, \end{cases} \tag{48}$$

for $i = 1, 2, \cdots, w$. This result shows that there is at most one nonzero element, 1 or -1, in the final result of $\mathbf{w}$ when $p = 1$. That's why 2DPCAL1-S should be formulated by combining L1-norm constraint and L2-norm constraint together rather than using L1-norm constraint alone. Similarly, when p approaches infinity, the limit of (45) is

$$\mathbf{w}^{k+1} = sign(\mathbf{v}^k). \tag{49}$$

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 (45) we could obtain the same solution as in [42]. It tells that the solution in [42] could be explained from the MM viewpoint. The solution of 2DPCAL1-S in [43] 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 [56], maximizing the optimization problem of G2DPCA equals maximizing the Lagrangian as follows

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s - \lambda\left(\|\mathbf{w}\|_p^p - 1\right), \tag{50}$$

where $s \geq 1$, $0 < p < 1$, $\lambda > 0$, $\mathbf{w} \in \mathbb{R}^w$. This problem could be turned into iteratively maximizing a surrogate function under the MM framework, as shown below. Let $\mathbf{w}^k$ be the projection vector at the k-th step in the iteration procedure. If any element in $\mathbf{w}^k$ is zero, then replace it with $\mathbf{w}^k + \varepsilon$ to make sure that it has no zero element, where ε is a random scalar that is sufficiently close to zero. According to Lemma 1 and Lemma 3, we have

$$\begin{aligned} &\sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s - \lambda\left(\|\mathbf{w}\|_p^p - 1\right) \\ \geq& s(\mathbf{v}^k)^T\mathbf{w} + (1-s)\sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}^k\|_s^s \\ &- \lambda\frac{p}{2}\mathbf{w}^T diag\left(\left|\mathbf{w}^k\right|^{p-2}\right)\mathbf{w} - \lambda(1-\frac{p}{2})\|\mathbf{w}^k\|_p^p + \lambda, \end{aligned} \tag{51}$$

wherein $\mathbf{v}^k$ is defined in (43) and the inequality becomes equality when $\mathbf{w} = \mathbf{w}^k$. Let the Lagrangian be denoted as $f(\mathbf{w})$ and let the relaxed function be denoted as $g(\mathbf{w}|\mathbf{w}^k)$. That is to say

$$f(\mathbf{w}) = \sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}\|_s^s - \lambda\left(\|\mathbf{w}\|_p^p - 1\right), \tag{52}$$

and

$$\begin{aligned} g(\mathbf{w}|\mathbf{w}^k) &= s(\mathbf{v}^k)^T\mathbf{w} + (1-s)\sum_{i=1}^{n} \|\mathbf{X}_i\mathbf{w}^k\|_s^s \\ &- \lambda\frac{p}{2}\mathbf{w}^T diag\left(\left|\mathbf{w}^k\right|^{p-2}\right)\mathbf{w} - \lambda(1-\frac{p}{2})\|\mathbf{w}^k\|_p^p + \lambda. \end{aligned} \tag{53}$$

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 framwork. 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

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} g(\mathbf{w}|\mathbf{w}^k). \tag{54}$$

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

$$\mathbf{w}^{k+1} = \arg\max_{\mathbf{w}} s(\mathbf{v}^k)^T\mathbf{w} - \lambda\frac{p}{2}\mathbf{w}^T diag\left(\left|\mathbf{w}^k\right|^{p-2}\right)\mathbf{w}. \tag{55}$$

Its solution is

$$\mathbf{w}^{k+1} = \frac{s}{\lambda p}\left|\mathbf{w}^k\right|^{2-p} \circ \mathbf{v}^k. \tag{56}$$

Since $2 - p > 0$, this solution indicates that $\mathbf{w}^k$ is no longer required to have no zero element. Therefore, we could treat this solution as the solution of the problem in (55) when $\mathbf{w}^k$ has zero elements. Considering the constraint $\|\mathbf{w}\|_p^p = 1$ and $\lambda > 0$, we have

$$\lambda = \frac{s}{p}\|\left|\mathbf{w}^k\right|^{2-p} \circ \mathbf{v}^k\|_p. \tag{57}$$

Then the update rule is

$$\mathbf{w}^{k+1} = \frac{\left|\mathbf{w}^k\right|^{2-p} \circ \mathbf{v}^k}{\|\left|\mathbf{w}^k\right|^{2-p} \circ \mathbf{v}^k\|_p}. \tag{58}$$

The above solution equals the two-step procedure below

$$\mathbf{u}^k = \left|\mathbf{w}^k\right|^{2-p} \circ \mathbf{v}^k, \tag{59}$$

$$\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_p}. \tag{60}$$

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 (51) 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 (45) and (58) it could be observed that a closed-form solution is obtained in each iteration for both cases. The solution successfully avoids the zero-finding problems [28], [32], learning rates [25], [36] or other extra tuning parameters [20], [43] which are usually encountered in solving related algorithms. The algorithm procedure of G2DPCA is listed in Table I.

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**
    $\mathbf{v}^k = \sum_{i=1}^n \mathbf{X}_i^T \left[\left| \mathbf{X}_i \mathbf{w}^k \right|^{s-1} \circ sign(\mathbf{X}_i \mathbf{w}^k) \right]$.
    if $0 < p < 1$
      $\mathbf{u}^k = \left| \mathbf{w}^k \right|^{2-p} \circ \mathbf{v}^k$,
      $\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_p}$.
    elseif $p = 1$
      $j = \arg\max_{i \in [1,w]} \left| v_i^k \right|$,
      $w_i^{k+1} = \begin{cases} sign(v_j^k), & i = j, \\ 0, & i \neq j. \end{cases}$
    elseif $p < \infty$
      $q = p/(p-1)$,
      $\mathbf{u}^k = \left| \mathbf{v}^k \right|^{q-1} \circ sign(\mathbf{v}^k)$,
      $\mathbf{w}^{k+1} = \frac{\mathbf{u}^k}{\|\mathbf{u}^k\|_p}$.
    elseif $p = \infty$
      $\mathbf{w}^{k+1} = sign(\mathbf{v}^k)$.
    end if
    $f^{k+1} = \sum_{i=1}^n \|\mathbf{X}_i \mathbf{w}^{k+1}\|_s^s$.
    $\delta = \left| f^{k+1} - f^k \right| / f^k$.
    $k \leftarrow k + 1$.
  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

V. EXPERIMENTS

Two benchmark face databases ORL [57] and FERET [58] are used in our experiments. In order to evaluate the proposed G2DPCA algorithm, we compare it with three state-of-art algorithms, i.e., 2DPCAL1-S [43], GPCA [21] and RSPCA [20] in the tasks of image reconstruction and classification. The three algorithms implemented in this paper are targeting the same optimization problems as their original algorithms. However, small differences exist in the respects of iterative solutions and tuning parameters in order to make a fair comparison.

The formulation and solution of 2DPCAL1-S used in this paper are described in Section II. GPCA and RSPCA in this paper are the one dimensional counterparts of G2DPCA and 2DPCAL1-S respectively, with their definitions described below. Let the n training image samples $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be transformed into corresponding vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, where $\mathbf{y}_i \in \mathbb{R}^d$, $i = 1, 2, \cdots, n$, and $d = h \times w$. Define $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d \times n}$. GPCA could be formulated by replacing the L2-norm both in the objective function and the constraint function of traditional PCA with Lp-norm. That is, GPCA finds its first projection vector $\mathbf{w} \in \mathbb{R}^d$ by solving the following optimization problem

$$\max_{\mathbf{w}} \|\mathbf{Y}^T \mathbf{w}\|_s^s, \quad s.t.\ \|\mathbf{w}\|_p^p = 1, \tag{61}$$

where $s \geq 1$, $p > 0$. It could be observed that PCA and PCA-L1 are two special cases of GPCA. RSPCA could be formulated by applying L1-norm both on the objective function and the constraint function of traditional PCA. That is, RSPCA finds its first projection vector $\mathbf{w} \in \mathbb{R}^d$ by solving the optimization problem

$$\max_{\mathbf{w}} \|\mathbf{Y}^T \mathbf{w}\|_1, \quad s.t.\ \|\mathbf{w}\|_1 \leq c, \|\mathbf{w}\|_2^2 = 1, \tag{62}$$

where c is a positive constant. After obtaining the first few projection vectors for GPCA and RSPCA, a similar deflation procedure as in Section II is implemented repeatedly to extract multiple projection vectors. It's quite direct to obtain iterative solutions for GPCA and RSPCA under the MM framework based on the solutions of G2DPCA and 2DPCAL1-S. And the new solutions are guaranteed to be locally optimal, the same as their original solutions.

The differences of the tuning parameters between the algorithms in this paper and their original versions are listed below. For 2DPCAL1-S, the tuning parameter λ in (8) relates to the tuning parameter ρ in the original 2DPCAL1-S [43] via $\lambda = 10^{-\rho}$. The latter one, i.e., ρ is adopted in our experiments for simplicity. For GPCA, the objective function is fixed to be the convex function (i) in [21] in order to be consistent with the proposed G2DPCA. Therefore, GPCA has two tuning parameters s and p. For RSPCA, a tuning parameter ρ corresponding to the one in 2DPCAL1-S is used to replace the sparsity controlling parameter k in [20]. In the following paper, we will search the optimal parameters for each of the four algorithms in a fine granularity under different image analysis tasks. Their optimal performances are then compared.

It's worthwhile to mention that 2DPCAL1-S and RSPCA could reduce to existing algorithms in some extreme conditions. When the ρ value in 2DPCAL1-S is small enough, the L1-norm constraint in the corresponding Lagrangian of 2DPCAL1-S would have a much larger weight than the L2-norm constraint, and 2DPCAL1-S would approximate to G2DPCA with $s = 1$ and $p = 1$. When the ρ value in 2DPCAL1-S is large enough, 2DPCAL1-S would approximate to 2DPCA-L1, as discussed in [43]. Similarly, RSPCA with a small ρ value would approximate to GPCA with $s = 1$ and $p = 1$, and RSPCA with a large ρ value would approximate to PCA-L1. These extreme conditions guarantee that the optimal ρ value in 2DPCAL1-S or RSPCA could be located in a finite

range. In practice, we mainly search ρ in the range of $[-3, 3]$ for the image reconstruction task and search ρ in the range of $[-4, 4]$ for the image classification task.

As for the initialization of the above algorithms, there are generally two kinds of suggestions [5], [36], [20], [21], [43]. The first one is to initialize the 1D algorithms by the corresponding components of PCA, and initialize the 2D algorithms by the corresponding components of 2DPCA. This is expected to find a better local optimum with higher probability in fewer iterations than random initializations. A variant in this category is to initialize 1D algorithms by the sample with the largest L2-norm in order to avoid the calculation of PCA. This method also turns out be satisfactory in practice [5], [36]. The second one is the multistart method [50], [51]. It is to run an algorithm multiple times with random initializations and then choose the initialization with optimized objective function value. The first kind of suggestion is adopted throughout this paper for consistency.

Previous studies have reached many consensuses about the performances of existing algorithms in image analysis. Some typical conclusions are, 2D algorithms could generally achieve much better results in image reconstruction and classification tasks than 1D algorithms [41], [42], [43]; a 2D algorithm takes much less time to train a projection vector than its corresponding 1D algorithm [41]; applying L1-norm in the objective function of traditional PCA or 2DPCA could resist noises thus improving its image reconstruction performance [5], [42], [20], [21], [43], [36]; applying L1-norm in the constraint of traditional PCA or 2DPCA could lead to sparse projection vectors and improved classification performances [20], [21], [43]. These results will be repeated in the following experiments. However, more emphasis will be placed on the performance of G2DPCA in image analysis, especially about the effect of Lp-norm.

Our scripts are written in Matlab(R). The experiments are run on a Dell(R) workstation with dual quad-core 2.27 GHz Intel(R) Xeon(R) processors and 23.6 GB memory.

A. ORL face database

We first conduct experiments on the ORL face database [57]. The ORL face database contains 400 images from 40 subjects, 10 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. 2 shows some sample images from this database.

Fig. 2. Sample images of the ORL face database.

In order to compare the computational time of the two solutions of G2DPCA when $1 < p < 2$, we train the first projection vector on the whole image set of the ORL database by setting $s = 2$ and $p = [1.1 : 0.1 : 1.9]$. That is, p ranges from 1.1 to 1.9 with a step of 0.1. The training time is shown in Fig. 3. It shows that the solution in case 1 generally converges faster than the solution in case 2. Similar results could be obtained when different s values are tried, when different number of projection vectors are extracted and when only a subset of the ORL database is trained. Therefore, when $1 < p < 2$, the solution in case 1 is more preferred.

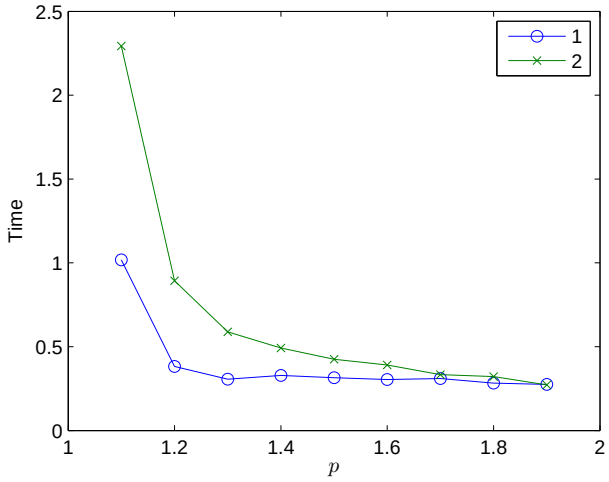


Fig. 3. Training time (seconds) of the first projection vector by G2DPCA with $s = 2$ and $p = [1.1 : 0.1 : 1.9]$ on the ORL database. The blue curve corresponds to the solution in case 1 and the green curve corresponds to the solution in case 2.

To study the sparsity of the projection vector of G2DPCA with different s values and p values, we train the first projection vector on the whole image set of the ORL database. 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. Fig. 4 shows the sparse rates of the first projection vector of G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 2.0]$. When $p \leq 1$, there is only one nonzero element appeared in the projection vector. When $1 < p < 1.5$, the sparse rate decreases with increasing p value. When $p \geq 1.5$, the projection vector would be dense. Therefore, the sparse rate of a projection vector is closely related to the p value. Different s values, i.e., $s = [1.0 : 0.1 : 3.0]$ have also been tried and we get almost the same results. The only difference observed is that with specific s values, the projection vector might become a little sparse when $p = 1.5$ or 1.6. Therefore, the sparse rate is stable with different s values. For a conclusion, a small p value is preferred if one wants to extract sparse projection vectors.

To evaluate the reconstruction performance of G2DPCA, we conduct an experiment on a polluted ORL database. 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 $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_m$ be m (m = 320)

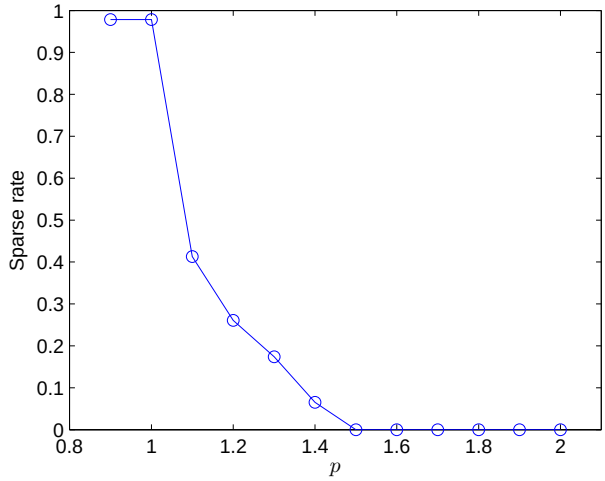


Fig. 4. Sparse rates of the first projection vector by G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 2.0]$ on the ORL database.

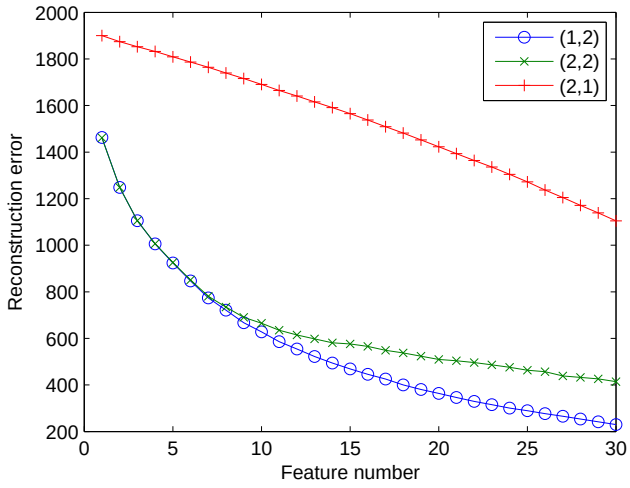


Fig. 5. Reconstruction errors of G2DPCA in three special cases on the ORL database. The (s, p) pairs for the three cases are shown in the legend.

clean images which are mean-centered, then the average reconstruction error of G2DPCA is defined as

$$\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{Z}_i(\mathbf{I}-\mathbf{W}\mathbf{W}^T)\|_F. \tag{63}$$

Fig. 5 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 results, both the 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 [42], [43]. The figure shows that applying L1-norm on the objective function of 2DPCA would improve its reconstruction performance and applying L1-norm on the constraint of 2DPCA would deteriorate its reconstruction performance. As an illustration, the reconstructed images of the three special cases on two sample images are shown in Fig. 6.

Fig. 6. The reconstructed images of G2DPCA in three special cases on two sample images from the polluted ORL database. The first column are the images to be reconstructed. Some images have random noises while others don't. 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 investigate the reconstruction performance of G2DPCA with different s values, we set $s = [1.0 : 0.1 : 3.0]$ and $p = 2$ in G2DPCA. By averaging the reconstruction errors with different feature numbers which are in the range of $[1, 30]$, we obtain the results in Fig. 7. From the figure, the reconstruction error increases with s value. And the lowest reconstruction error is obtained when $s = 1$ in which case G2DPCA reduces to 2DPCA-L1. It demonstrates that G2DPCA with a small s value could resist noises in image reconstruction.

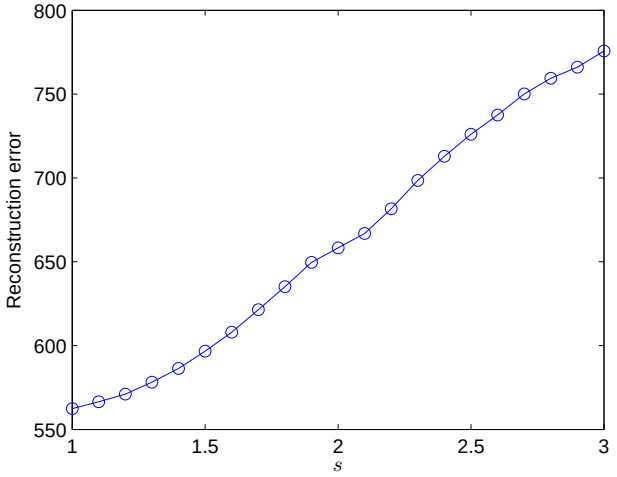


Fig. 7. Reconstruction errors of G2DPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = 2$ on the ORL database.

To examine the reconstruction performance of G2DPCA with different p values, we set $s = 2$ and $p = [0.9 : 0.1 : 3.0]$ in G2DPCA. By averaging the reconstruction errors across feature numbers as above, we obtain the results in Fig. 8. The figure shows that when $p > 2.1$, the reconstruction error increases sharply. The detailed results when $p = [0.9 : 0.1 : 2.1]$ are listed in Table II. From the results, the lowest

reconstruction error is achieved when $p = 2$ in which case G2DPCA reduces to 2DPCA.

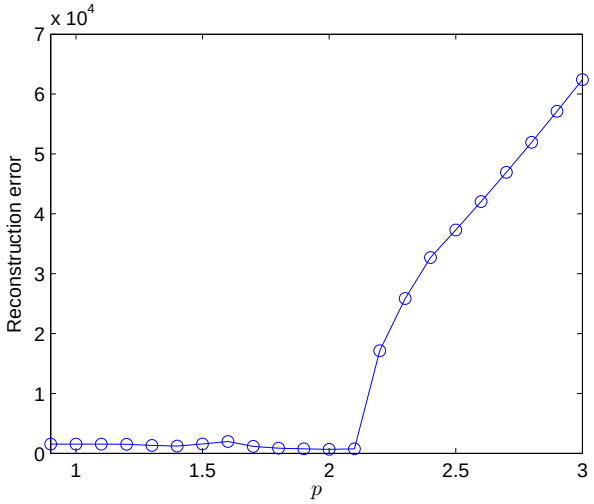


Fig. 8. Reconstruction errors of G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 3.0]$ on the ORL database.

TABLE II
RECONSTRUCTION ERRORS OF G2DPCA WITH $s = 2$ AND $p = [0.9 : 0.1 : 2.1]$ ON THE ORL DATABASE.

p	0.9	1.0	1.1	1.2	1.3	1.4	1.5
$\times 10^3$	1.53	1.53	1.53	1.53	1.32	1.22	1.56
p	1.6	1.7	1.8	1.9	2.0	2.1	
$\times 10^3$	1.99	1.16	0.86	0.75	**0.66**	0.75	

Based on the above results, it's reasonable to assume that the lowest reconstruction error of G2DPCA is obtained when $1.0 \leq s \leq 3.0$ and $0.9 \leq p \leq 3.0$, thus could be located by simultaneously varying s and p values in this range. With this consideration, 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 likewise. For each algorithm, we search its optimal tuning parameters in a wide range. Specifically, we set $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$ in G2DPCA, and the same parameter set is applied on GPCA considering the similarity between G2DPCA and GPCA. For 2DPCAL1-S and RSPCA, the tuning parameter ρ is set to be $[-3.0 : 0.1 : 3.0]$, similar but in a much finer granularity compared with the setting in [43].

The reconstruction errors of the four algorithms are shown in Fig. 9. From this figure, the reconstruction error of G2DPCA is relatively stable with different s values, but it is greatly affected by various p values. The same conclusion could be obtained from the results of GPCA. For 2DPCAL1-S and RSPCA, the reconstruction error is stable when $\rho < -1$ or $\rho > 2$, but it is greatly affected by ρ value when $-1 < \rho < 2$. 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 III. 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. Fig. 10 shows the reconstruction errors of the G2DPCA, 2DPCAL1-S, GPCA and RSPCA with different feature numbers when respective optimal parameters are applied. With tolerances for random errors, we could make conclusions from these results 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 III
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

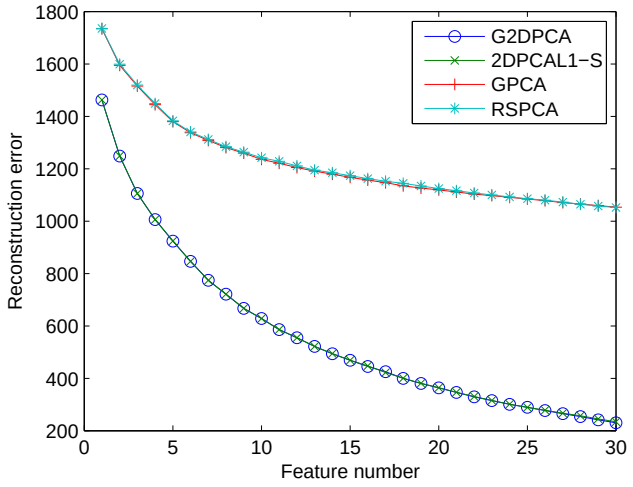


Fig. 10. Reconstruction errors of G2DPCA, 2DPCAL1-S, GPCA and RSPCA with different feature numbers on the ORL database when respective optimal parameters are applied.

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

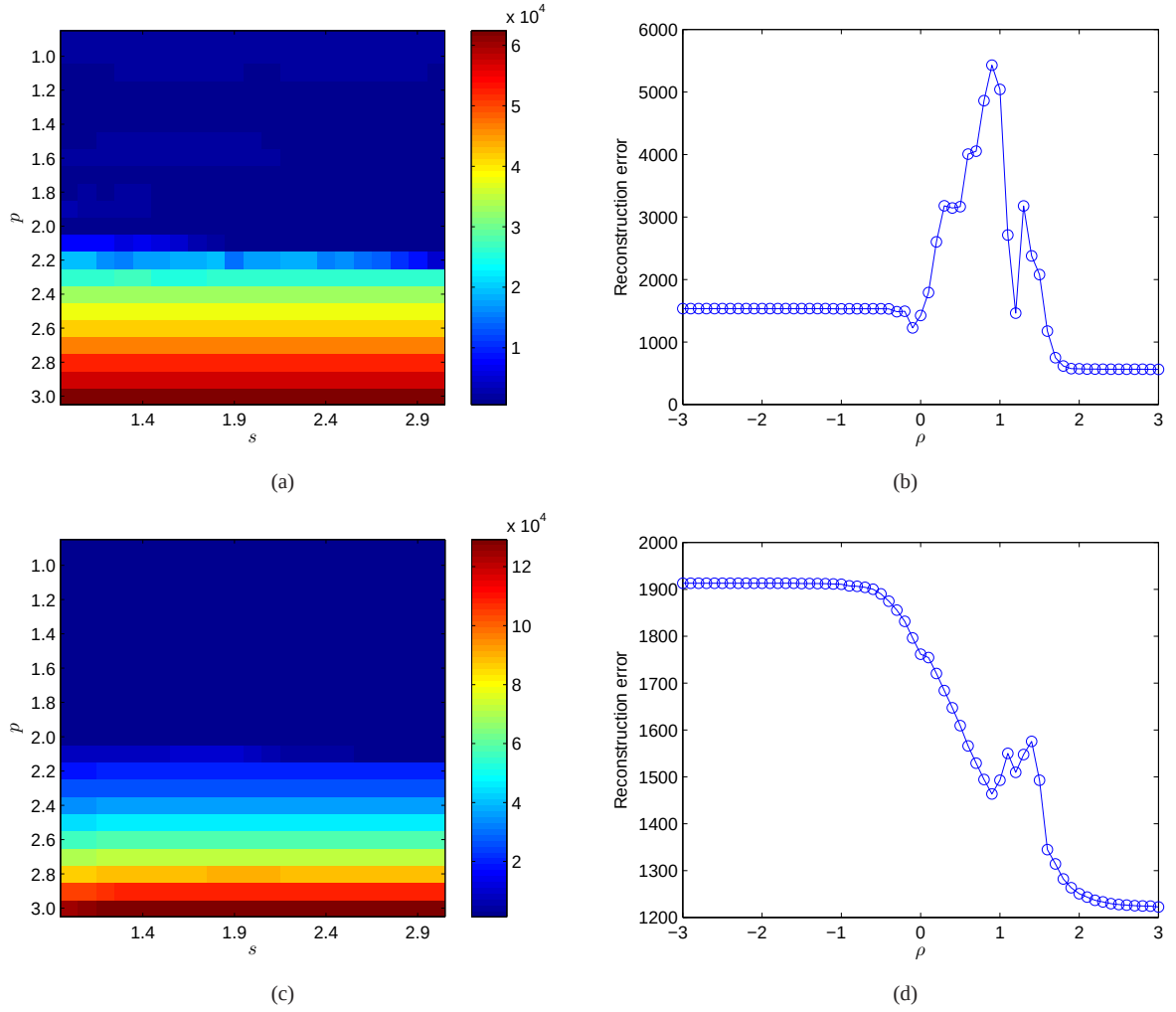


Fig. 9. Reconstruction errors of four algorithms with different tuning parameters on the ORL database. (a) G2DPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (b) 2DPCAL1-S with $\rho = [-3.0 : 0.1 : 3.0]$. (c) GPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (d) RSPCA with $\rho = [-3.0 : 0.1 : 3.0]$.

Then we proceed to investigate the classification performance of G2DPCA on the ORL database. The proposed algorithm is 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 reported.

Fig. 11 shows the classification accuracies of G2DPCA in three special cases with different feature numbers. From the results, the two curves corresponding to 2DPCA and 2DPCA-L1 are very close which means that they obtain nearly the same classification performance. Both the results of 2DPCA and 2DPCA-L1 outperform the result of G2DPCA with $s = 2$ and $p = 1$. The figure also indicates that the classification performance of G2DPCA is sensitive to p value, but not sensitive to s value.

To investigate the classification performance of G2DPCA with different s values, we set $s = [1.0 : 0.1 : 5.0]$ and $p = 2$ in G2DPCA. By averaging the classification accuracies with different feature numbers, we obtain the results in Fig. 12.

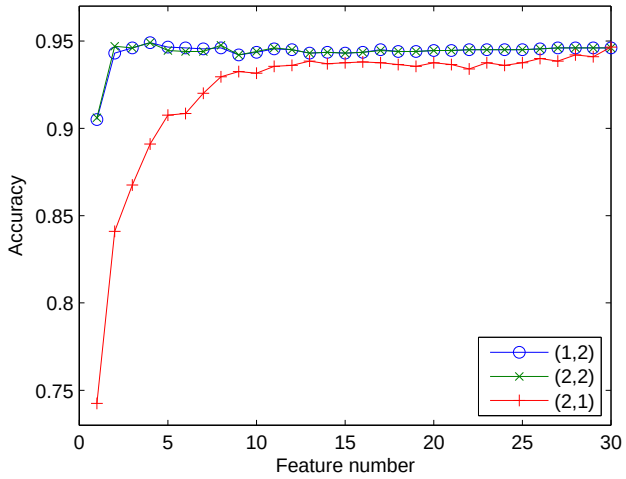


Fig. 11. Classification accuracies of G2DPCA in three special cases on the ORL database. The (s, p) pairs for the three cases are shown in the legend.

When $s < 3.5$, the classification accuracies are in the range of $[0.9435, 0.9438]$. When $s > 3.5$, the results become worse but differences are subtle. This figure demonstrates that changing s value in G2DPCA would not greatly affect its classification performance.

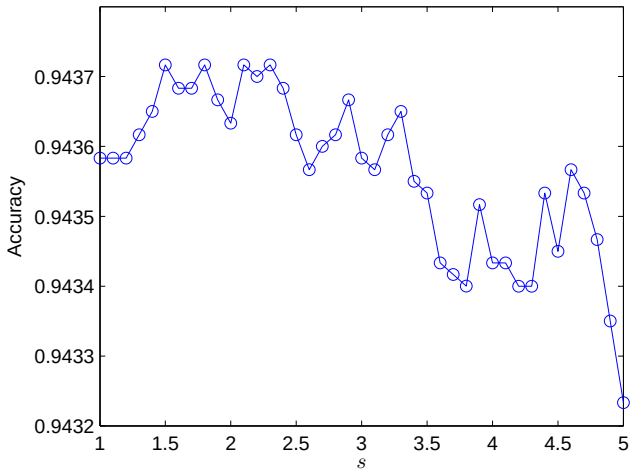


Fig. 12. Classification accuracies of G2DPCA with $s = [1.0 : 0.1 : 5.0]$ and $p = 2$ on the ORL database.

To investigate the classification performance of G2DPCA with different p values, we set $s = 2$ and $p = [0.9 : 0.1 : 5.0]$ in G2DPCA. The classification accuracies with different p values are shown in Fig. 13. The highest classification accuracy is 0.9468 when $p = 1.5$. When $p > 2.5$, the classification accuracy increases slowly with p value and finally converges to 0.9220. When $p = 2$, G2DPCA reduces to 2DPCA and the corresponding accuracy is 0.9436. This result shows that the classification performance of 2DPCA could be improved by applying Lp-norm on its constraint.

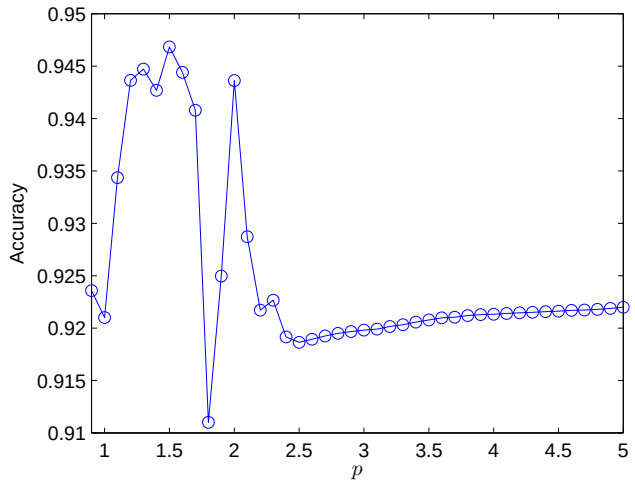


Fig. 13. Classification accuracies of G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 5.0]$ on the ORL database.

Based on the above results, we assume that the optimal classification performance of G2DPCA is obtained when $1.0 \leq s \leq 3.0$ and $0.9 \leq p \leq 3.0$. The optimal parameters could be located by simultaneously varying s and p values. With this consideration, we proceed to compare the classification performance of G2DPCA with other three algorithms, i.e., 2DPCAL1-S, GPCA and RSPCA. The parameters of G2DPCA and GPCA are set to be $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. The parameters of 2DPCAL1-S and RSPCA are set to be $\rho = [-4.0 : 0.1 : 4.0]$.

The classification accuracies of the four algorithms are shown in Fig. 14. From Fig. 14(a), the accuracy of G2DPCA is insensitive with respect to s value except when $p = 1.8$ and $p = 2.2$. At the same time, the accuracy of G2DPCA is greatly affected by p value. Fig. 14(b) shows that the accuracy of 2DPCAL1-S peaks at $\rho = 1.7$. Fig. 14(c) shows that the accuracy of GPCA is also greatly affected by p value but not by s value, as G2DPCA. Fig. 14(d) shows the accuracy of RSPCA generally increases with ρ value. A wider range for the ρ value in 2DPCAL1-S and RSPCA, i.e., $\rho = [-6.0 : 0.1 : 6.0]$ is also tried and we find the accuracy tends to be stable when $\rho < -3$ or $\rho > 3$. The reason is that when $\rho < -3$, 2DPCAL1-S approximates to G2DPCA with $s = 1$ and $p = 1$, RSPCA approximates to GPCA with $s = 1$ and $p = 1$; when $\rho > 3$, 2DPCAL1-S approximates to 2DPCA-L1, RSPCA approximates to PCA-L1. Therefore, when ρ is small or large enough, the accuracies of 2DPCAL1-S and RSPCA converge to the accuracies in their respective extreme conditions.

The highest classification accuracies and corresponding parameters of the four algorithms are listed in Table IV. 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. Fig. 15 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. The optimal parameters are different between G2DPCA and GPCA, or between 2DPCAL1-S and RSPCA. The result also demonstrates that applying Lp-norm on the objective function and the constraint function of traditional 2DPCA could slightly improve its classification performance on the ORL database.

TABLE IV
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

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

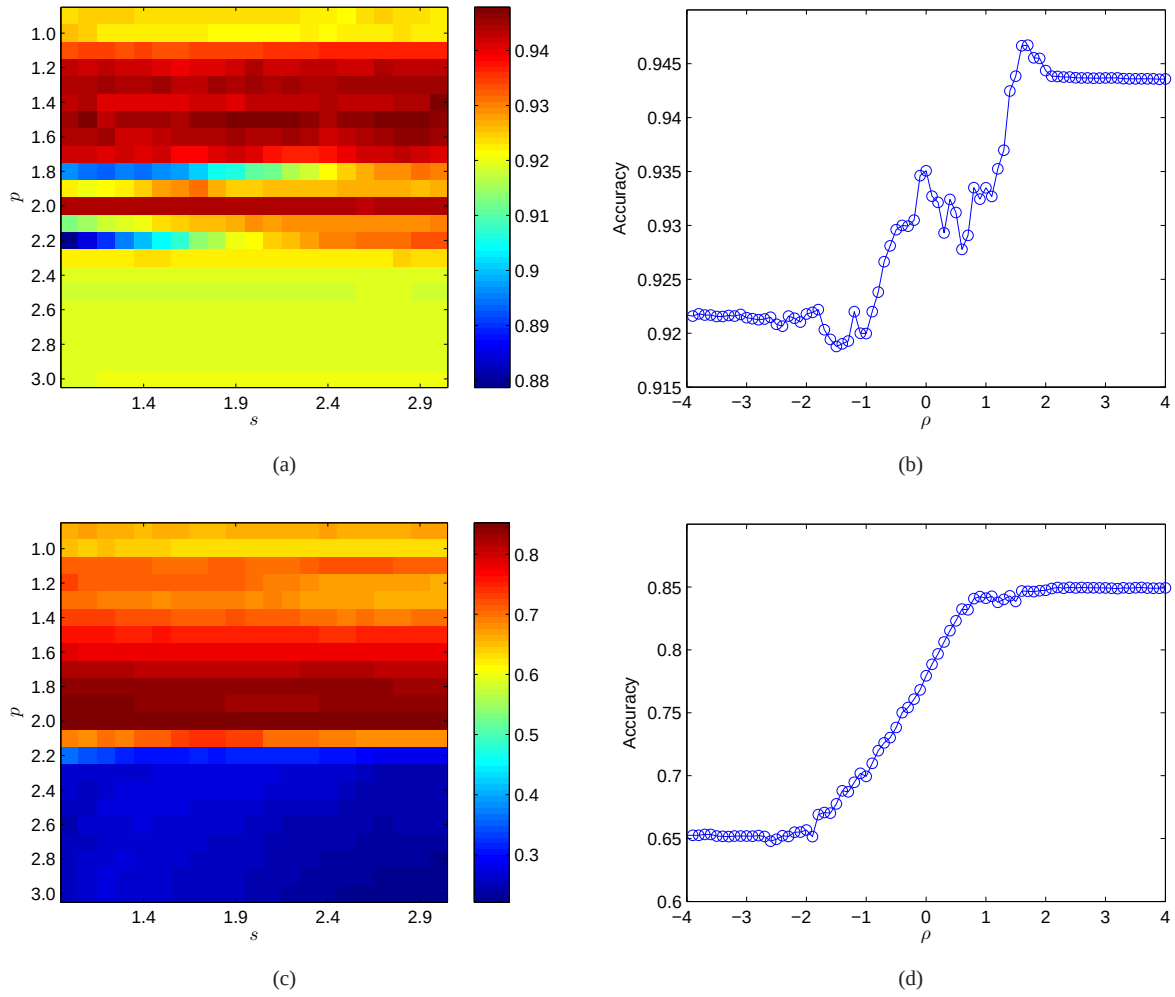


Fig. 14. Classification accuracies of the four algorithms with different tuning parameters on the ORL database. (a) G2DPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (b) 2DPCAL1-S with $\rho = [-4.0 : 0.1 : 4.0]$. (c) GPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (d) RSPCA with $\rho = [-4.0 : 0.1 : 4.0]$.

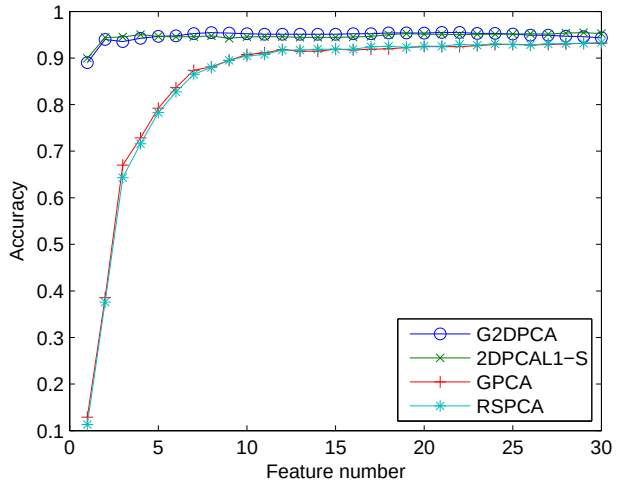


Fig. 15. Classification accuracies of G2DPCA, 2DPCAL1-S, GPCA and RSPCA with different feature numbers on the ORL database when respective optimal parameters are applied.

Then the sparse rates of the projection vectors of G2DPCA with $s = 2.9$ and $p = 1.5$ are calculated, as shown in Fig. 16. The first thirty projection vectors in ten repetitions are included for comparison. Since the training samples in the ten repetitions are randomly chosen, their results are different. From the results, the projection vectors are slightly sparse which is consistent with our previous conclusion that the projection vectors of G2DPCA with $p = 1.5$ behave at the fine edge between density and sparsity on the ORL database.

B. FERET face database

To further corroborate the above conclusions about G2DPCA in image reconstruction and classification, we conduct experiments on a subset of the FERET face database [58]. The experimental scheme is similar as what we have done on the ORL database.

The subset of the FERET database used in our experiments contains 1400 images from 200 subjects, 7 images per subject. The images are taken with different facial expressions and view angles. The image size is 80 × 80. We further resize the

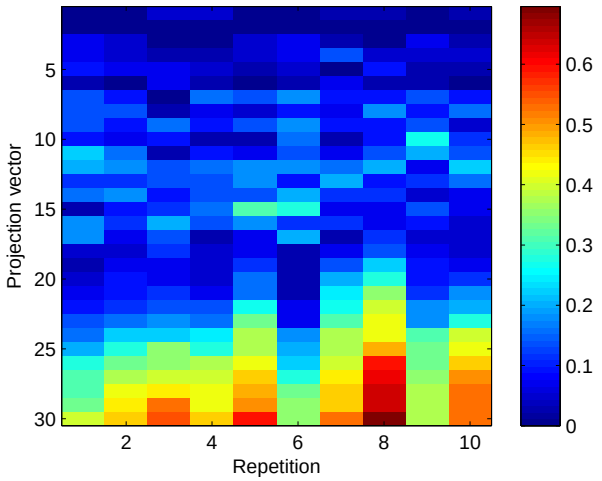


Fig. 16. The sparse rates of the projection vectors of G2DPCA with $s = 2.9$ and $p = 1.5$ on the ORL database.

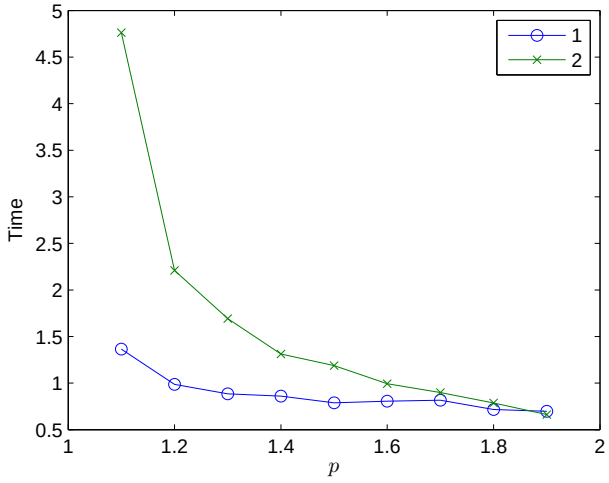


Fig. 18. Training time (seconds) of the first projection vector by G2DPCA with $s = 2$ and $p = [1.1 : 0.1 : 1.9]$ on the FERET database. The blue curve corresponds to the solution in case 1 and the green curve corresponds to the solution in case 2.

images to 40 × 40 to reduce the computational time. Fig. 17 shows some sample images from this database.

Fig. 17. Sample images of the FERET face database.

To compare the computational time of the two solutions when $1 < p < 2$, we train the first projection vector of G2DPCA on the whole database by setting $s = 2$ and $p = [1.1 : 0.1 : 1.9]$. The training time is shown in Fig. 18. Again, the solution in case 1 generally converges faster than the solution in case 2, thus being more preferred.

Fig. 19 shows the sparse rates of the first projection vector of G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 2.0]$. When $p \leq 1$, there is only one nonzero element appeared in each projection vector. When $1 < p < 1.3$, the sparse rate decreases with increasing p value. When $p \geq 1.3$, the projection vector would be dense. Additional experiments indicate that changing s value would hardly affect the sparse rate. Therefore, to extract sparse projection vectors, it's necessary to set p as a small value.

Then we conduct an experiment on a polluted FERET database to evaluate the reconstruction performance of G2DPCA. 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.

Fig. 20 shows the reconstructed images of G2DPCA in three special cases on two sample images. Fig. 21 shows the reconstruction errors of G2DPCA in the three special cases with different feature numbers. From the result, the reconstruction error of 2DPCA-L1 is lower than that of 2DPCA when the

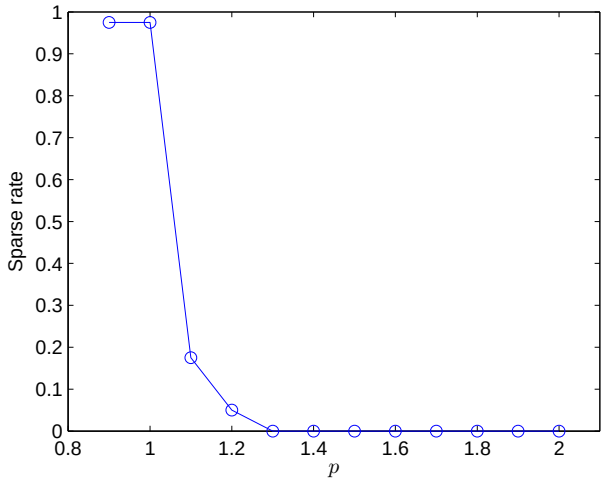


Fig. 19. Sparse rates of the first projection vector by G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 2.0]$ on the FERET database.

feature number is larger than five. Also, both the reconstruction errors of 2DPCA-L1 and 2DPCA are much lower than that of G2DPCA with $s = 2$ and $p = 1$.

By setting $s = [1.0 : 0.1 : 3.0]$ and $p = 2$ in G2DPCA, we get the reconstruction errors with different s values, as shown in Fig. 22. Again, the reconstruction error increases with s value and the lowest reconstruction error is obtained when $s = 1$ in which case G2DPCA reduces to 2DPCA-L1.

By setting $s = 2$ and $p = [0.9 : 0.1 : 3.0]$ in G2DPCA, we get the reconstruction errors with different p values, as shown in Fig. 23. From the figure, the reconstruction error increases very fast when $p > 2.1$. Table V lists the detailed results when $p = [0.9 : 0.1 : 2.1]$. It shows that the lowest reconstruction error is obtained when $p = 2$ in which case G2DPCA reduces to 2DPCA.

Through the above results, one might guess that the optimal

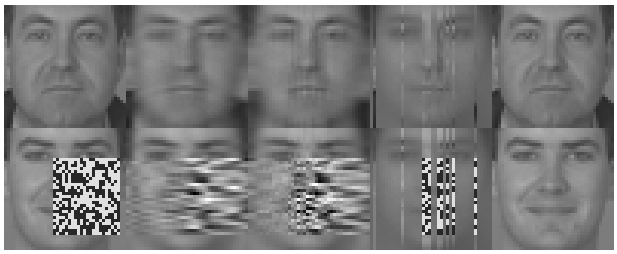


Fig. 20. The reconstructed images of G2DPCA in three special cases on two sample images from the polluted FERET database. The first column are the images to be reconstructed. Some images have random noises while others don't. The following three columns are the reconstructed images by using the first 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.

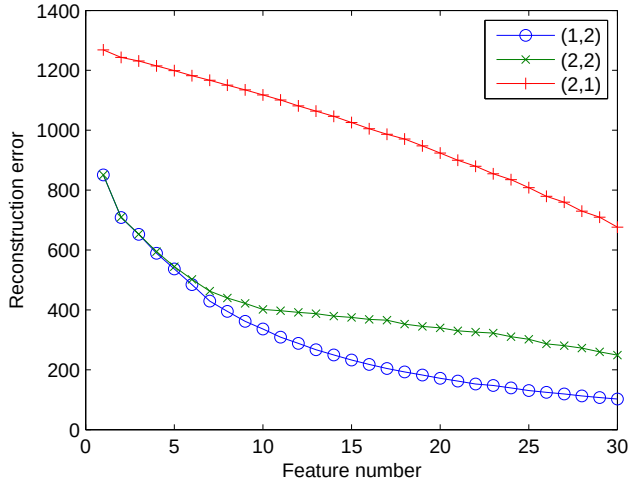


Fig. 21. Reconstruction errors of G2DPCA in three special cases on the FERET database. The (s, p) pairs for the three cases are shown in the legend.

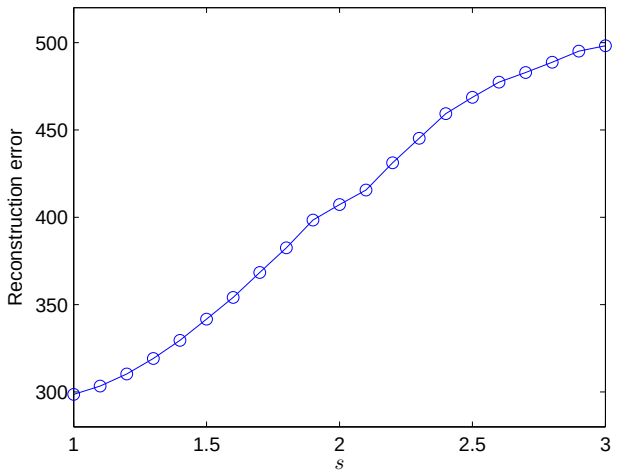


Fig. 22. Reconstruction errors of G2DPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = 2$ on the FERET database.

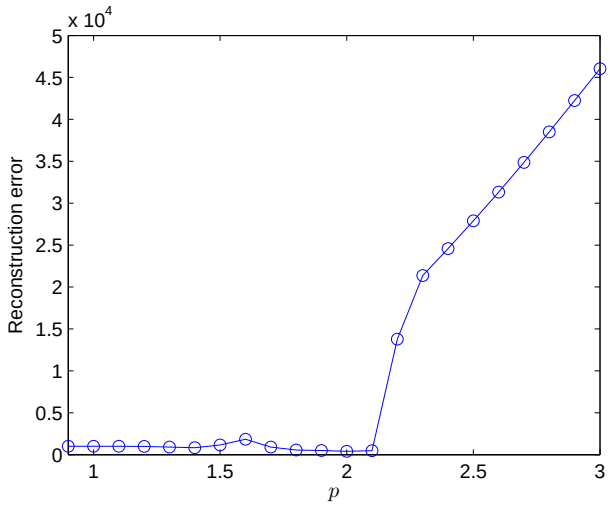


Fig. 23. Reconstruction errors of G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 3.0]$ on the FERET database.

TABLE V
RECONSTRUCTION ERRORS OF G2DPCA WITH $s = 2$ AND $p = [0.9 : 0.1 : 2.1]$ ON THE FERET DATABASE.

p	0.9	1.0	1.1	1.2	1.3	1.4	1.5
$\times 10^3$	1.00	1.00	1.00	0.99	0.90	0.84	1.16
p	1.6	1.7	1.8	1.9	2.0	2.1	
$\times 10^3$	1.85	0.90	0.56	0.49	**0.41**	0.47	

reconstruction performance of G2DPCA would be achieved when it reduces to 2DPCA-L1. To verify this assumption, we should still search the optimal (s, p) pair in a wide range as before. Then we proceed to compare the reconstruction performance of G2DPCA with other three algorithms, i.e., 2DPCAL1-S, GPCA and RSPCA. The parameters of G2DPCA and GPCA are set to be $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. The parameters of 2DPCAL1-S and RSPCA are set to be $\rho = [-3.0 : 0.1 : 3.0]$.

The reconstruction errors of the four algorithms are shown in Fig. 24. From this figure, the reconstruction error of G2DPCA is relatively stable with different s values, but is greatly affected by various p values. This conclusion also applies to GPCA. The reconstruction error of 2DPCAL1-S is stable when $\rho < -1$ or $\rho > 2$. The reconstruction error of RSPCA is stable when $\rho < -1$ or $\rho > 3$. 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. This is consistent with the reconstruction result on the FERET database in [43] where the lowest reconstruction error of 2DPCAL1-S is obtained when 2DPCAL1-S reduces to 2DPCA-L1.

The lowest reconstruction errors and corresponding parameters of the four algorithms are listed in Table VI. 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. Fig. 25 shows the reconstruction errors of the four algorithms with different feature numbers when their respective optimal parameters are applied. With tolerances

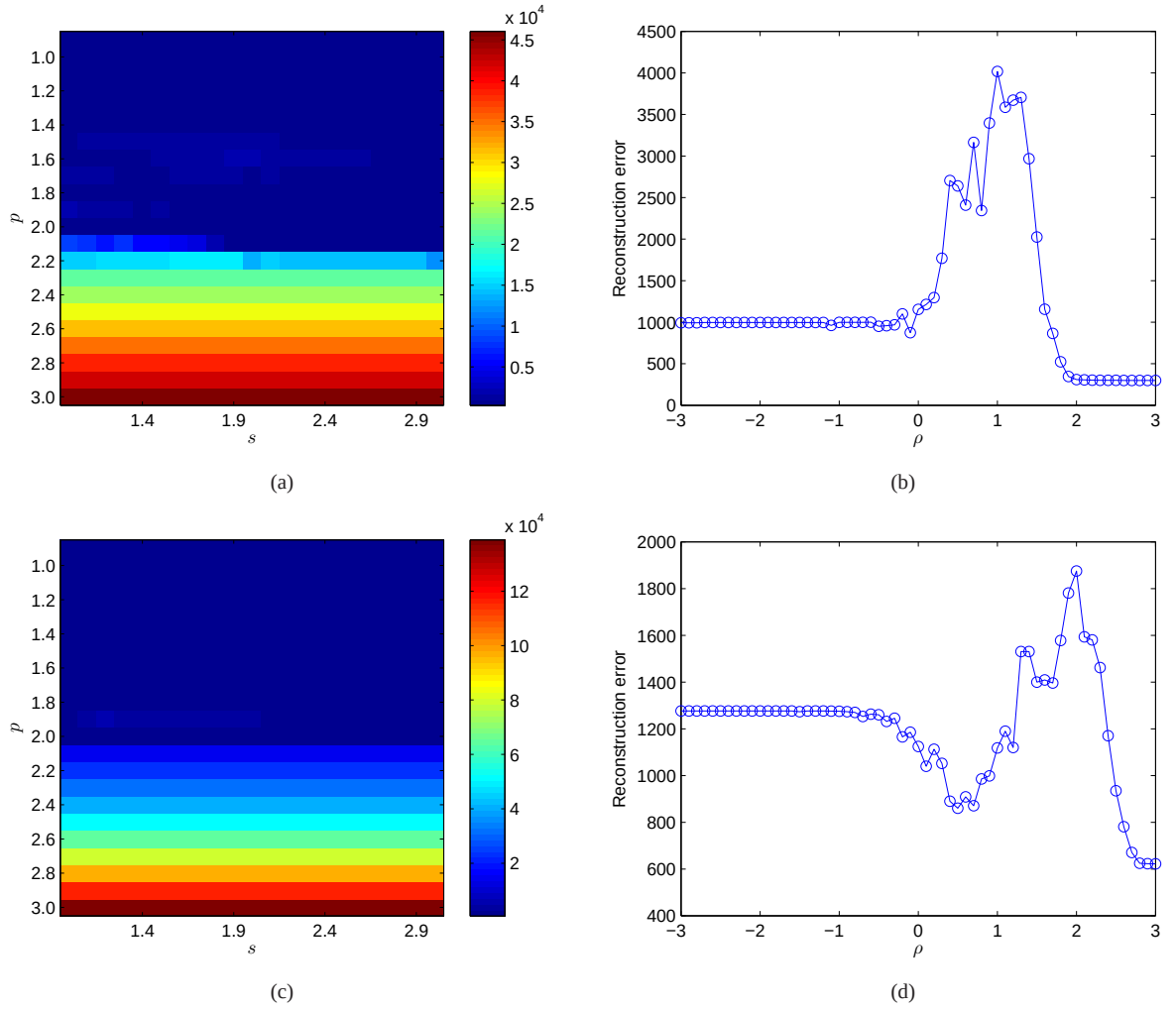


Fig. 24. Reconstruction errors of the four algorithms with different tuning parameters on the FERET database. (a) G2DPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (b) 2DPCAL1-S with $\rho = [-3.0 : 0.1 : 3.0]$. (c) GPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (d) RSPCA with $\rho = [-3.0 : 0.1 : 3.0]$.

for random errors, the results again 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.

TABLE VI
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

To investigate the classification performance of G2DPCA 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 and then the

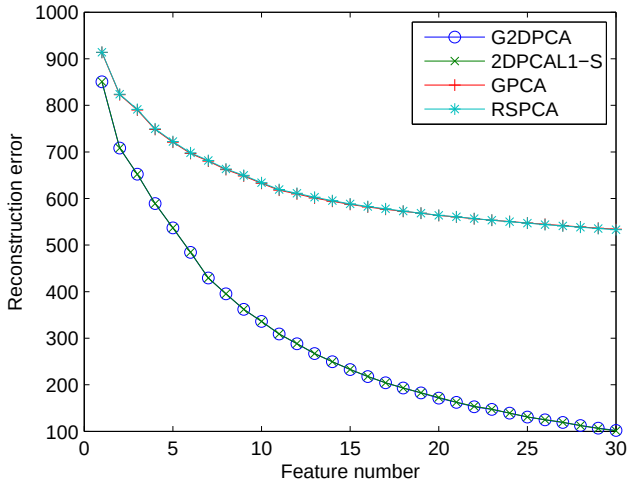


Fig. 25. Reconstruction errors of G2DPCA, 2DPCAL1-S, GPCA and RSPCA with different feature numbers on the FERET database when respective optimal parameters are applied.

average classification accuracy is reported.

Fig. 26 shows the classification accuracies of G2DPCA in three special cases with different feature numbers. From the figure, the results corresponding to 2DPCA and 2DPCA-L1 are nearly the same. Both results are much better than that of G2DPCA with $s = 2$ and $p = 1$ when the feature number is smaller than 20. When the feature number is larger than 20, the results of the three cases are very close. This figure indicates that the classification performance of G2DPCA is sensitive to p value, but not sensitive to s value.

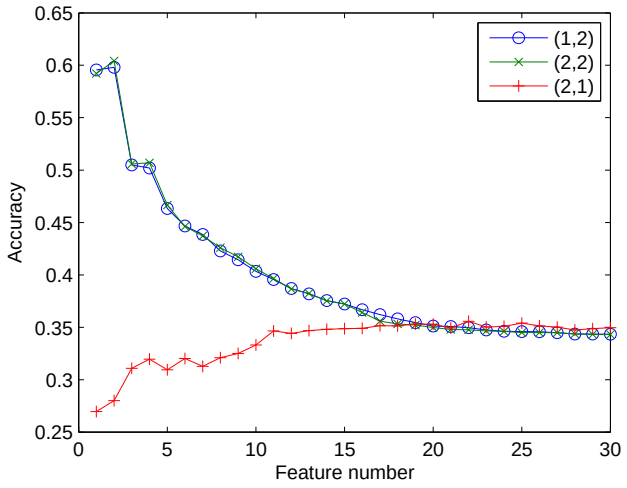


Fig. 26. Classification accuracies of G2DPCA in three special cases on the FERET database. The (s, p) pairs for the three cases are shown in the legend.

Fig. 27 shows the classification accuracies of G2DPCA with $s = [1 : 1 : 50]$ and $p = 2$. From the figure, the classification accuracy of 2DPCA, i.e. G2DPCA with $s = 2$ and $p = 2$ is the lowest among the results with different s values. When $s \leq 15$, the accuracy generally increases with s value. When $s > 15$, the accuracy becomes stable with different s values. The highest classification accuracy is 0.4169, obtained when $s = 15$. Therefore, applying Lp-norm on the objective function of 2DPCA could improve its classification performance on the FERET database.

Fig. 28 shows the classification accuracies of G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 3.5]$. When $p < 2.3$, the classification accuracy is greatly affected by p value. When $p \geq 2.3$, the classification accuracy becomes stable with increasing p value. The highest classification accuracy is 0.6467 when $p = 2.2$, much higher than 0.3983 when $p = 2$ in which case G2DPCA reduces to 2DPCA. This indicates a substantial improvement on the traditional 2DPCA. In other words, applying Lp-norm on the constraint of 2DPCA could also improve its classification performance.

According to the above results, we calculate the classification accuracies of G2DPCA with $s = [1 : 1 : 20]$ and $p = [0.9 : 0.1 : 2.5]$ to search for the optimal parameters, as shown in Fig. 29. From the result, the highest classification accuracy is 0.6467 when $s = 2$ and $p = 2.2$.

It's better to search the optimal (s, p) pair in a finer granularity. The above result shows that the optimal parameters

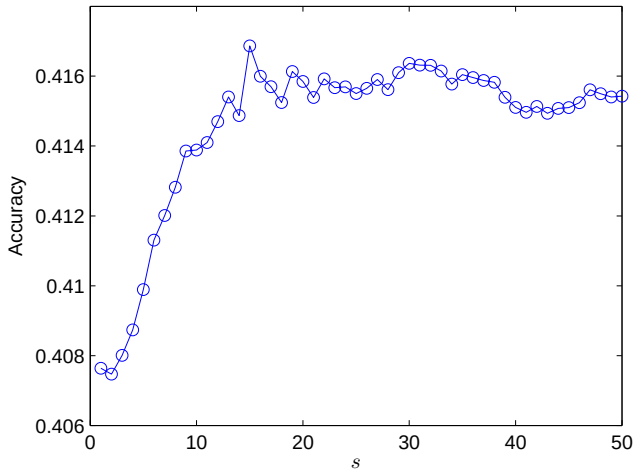


Fig. 27. Classification accuracies of G2DPCA with $s = [1 : 1 : 50]$ and $p = 2$ on the FERET database.

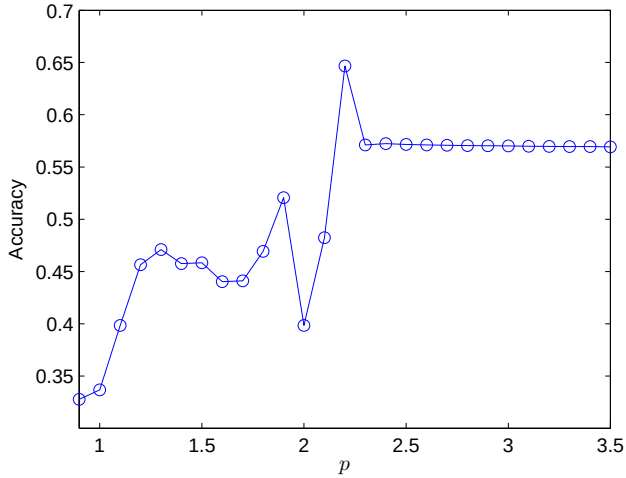


Fig. 28. Classification accuracies of G2DPCA with $s = 2$ and $p = [0.9 : 0.1 : 3.5]$ on the FERET database.

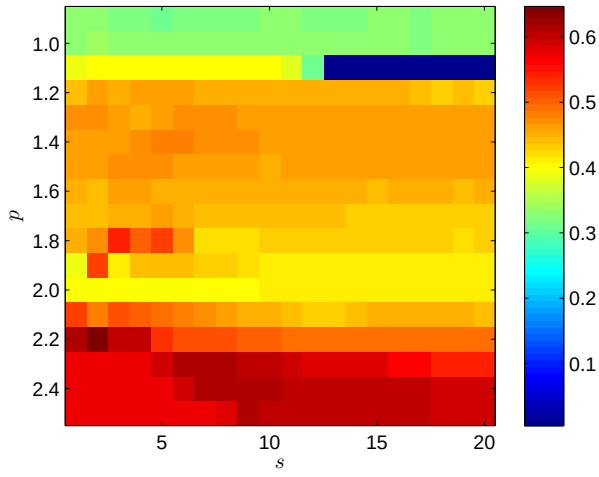


Fig. 29. Classification accuracies of G2DPCA with $s = [1 : 1 : 20]$ and $p = [0.9 : 0.1 : 2.5]$ on the FERET database.

again fail in the range of $1 \leq s \leq 3$ and $0.9 \leq p \leq 3$. To be consistent with previous experiments, we calculate the 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 results are compared with other three algorithms, i.e., 2DPCAL1-S, GPCA and RSPCA, as shown in Fig. 30. The parameter set for each algorithm is chosen the same as the experiments on the ORL database.

From Fig. 30, the results of G2DPCA and GPCA are generally sensitive to p value but insensitive to s value. Some exceptions include G2DPCA with $p = 1.8$, 1.9, 2.1 or 2.2 and GPCA with $p = 1.1$ or 1.6. The highest classification accuracy of G2DPCA is obtained when $p = 2.2$, and the highest classification accuracy of GPCA is obtained when $p = 2.0$. Both for 2DPCAL1-S and RSPCA, the highest classification accuracy is obtained when $0 < \rho < 1$. A wider range of ρ value, i.e., $\rho = [-6.0 : 0.1 : 6.0]$ is also tried for the two algorithms. We find that the classification accuracy of 2DPCAL1-S is stable when $\rho < -3$ or $\rho > 2$, and the classification accuracy of RSPCA is stable when $\rho < -4$ or $\rho > 3$. The reason is that when the ρ value is small or large enough, 2DPCAL1-S and RSPCA would approximate to their respective extreme conditions which have stable classification accuracies, as discussed before. Additionally, if the ρ value is chosen from $[-3.0 : 1.0 : 3.0]$, the highest classification accuracy of 2DPCAL1-S would be obtained when $\rho = 0$. This is consistent with the classification result of 2DPCAL1-S on the FERET database in [43].

The highest classification accuracies and corresponding parameters of the four algorithms are listed in Table VII. 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. Fig. 31 shows the classification accuracies of the four algorithms with different feature numbers when respective optimal parameters are applied. From these 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. This demonstrates that applying Lp-norm both on the objective function and constraint function of 2DPCA could greatly improve its classification performance on the FERET database. However, the same operation on PCA just slightly improves its classification performance, from 0.2406 to 0.2417. The best classification performance among the 1D algorithms on the FERET database is achieved by RSPCA.

TABLE VII
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

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 much larger than 1.3.

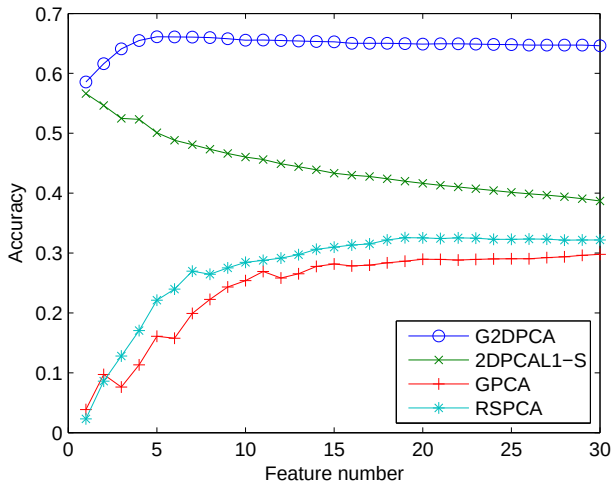


Fig. 31. Classification accuracies of G2DPCA, 2DPCAL1-S, GPCA and RSPCA with different feature numbers on the FERET database when respective optimal tuning parameters are applied.

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 on the objective function and the constraint function 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.

In the experiments on two face databases, i.e., ORL and FERET, we find that the sparse rates of the projection vectors in G2DPCA are closely related to p value, and a small p value is required to generate sparse projection vectors. 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.

Some questions that remain unclear concerning the experimental results are listed below. First, the accuracy by some 2D algorithms is decreasing with feature number on the FERET database, as shown in Fig. 26 and Fig. 31. 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 one dimensional counterparts of G2DPCA and 2DPCAL1-S

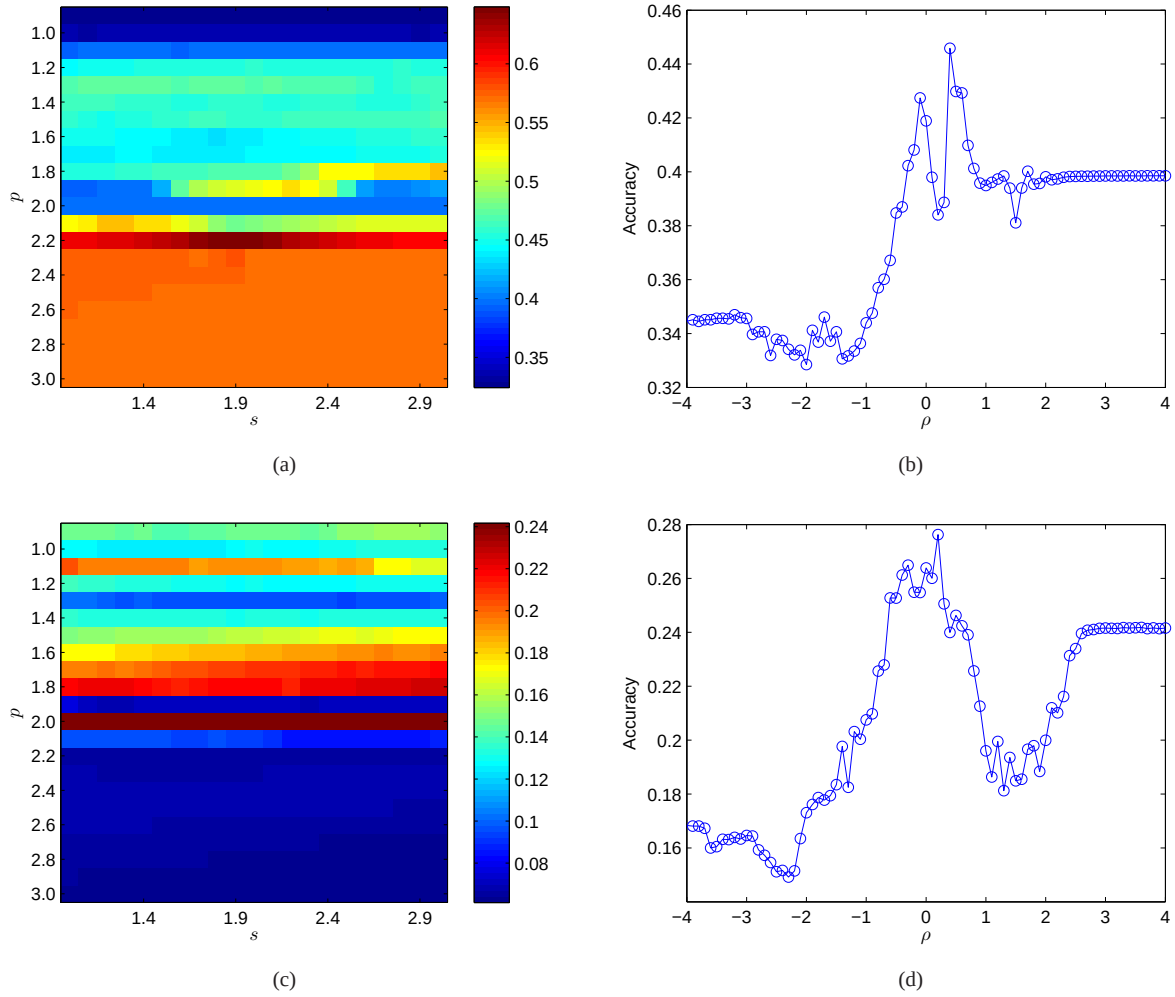


Fig. 30. Classification accuracies of the four algorithms with different tuning parameters on the FERET database. (a) G2DPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (b) 2DPCAL1-S with $\rho = [-4.0 : 0.1 : 4.0]$. (c) GPCA with $s = [1.0 : 0.1 : 3.0]$ and $p = [0.9 : 0.1 : 3.0]$. (d) RSPCA with $\rho = [-4.0 : 0.1 : 4.0]$.

respectively, this result is difficult to explain. Third, among the four 2D 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 1D 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's 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. Finally, it is also an interesting problem to find the locally optimal solution of G2DPCA with $0 < s < 1$ and $p > 0$ if it exists.

ACKNOWLEDGMENT

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

REFERENCES

[1] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.
[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, 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 Conference on Computer Vision and Pattern Recognition*, vol. 1, 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 International Conference on Machine learning*, 2006, pp. 281–288.
[5] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1672–1680, 2008.
[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
[7] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2009.
[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[9] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[10] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM review*, vol. 49, no. 3, pp. 434–448, 2007.

[11] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of multivariate analysis*, vol. 99, no. 6, pp. 1015–1034, 2008.

[12] A. d'Aspremont, F. Bach, and L. E. Ghaoui, "Optimal solutions for sparse principal component analysis," *The Journal of Machine Learning Research*, vol. 9, pp. 1269–1294, 2008.

[13] 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.

[14] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *The Journal of Machine Learning Research*, vol. 11, pp. 517–553, 2010.

[15] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Machine Learning*, vol. 85, no. 1–2, pp. 3–39, 2011.

[16] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[17] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[18] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[20] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse PCA by L_1-norm maximization," *Pattern Recognition*, vol. 45, no. 1, pp. 487–497, 2012.

[21] Z. Liang, S. Xia, Y. Zhou, L. Zhang, and Y. Li, "Feature extraction based on Lp-norm generalized principal component analysis," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1037–1045, 2013.

[22] C. Gentile, "The robustness of the p-norm algorithms," *Machine Learning*, vol. 53, no. 3, pp. 265–299, 2003.

[23] J. Kivinen, M. K. Warmuth, and B. Hassibi, "The p-norm generalization of the LMS algorithm for adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1782–1793, 2006.

[24] Z. Liu, F. Jiang, G. Tian, S. Wang, F. Sato, S. J. Meltzer, and M. Tan, "Sparse logistic regression with Lp penalty for biomarker identification," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, 2007.

[25] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien, "Efficient and accurate ℓ_p-norm multiple kernel learning." in *Proc. Advances in Neural Information Processing Systems*, vol. 22, no. 22, 2009, pp. 997–1005.

[26] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "ℓ_p-norm multiple kernel learning," *The Journal of Machine Learning Research*, vol. 12, pp. 953–997, 2011.

[27] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 395–407, 2009.

[28] J. Liu and J. Ye, "Efficient ℓ_1/ℓ_q norm regularization," *arXiv:1009.4766*, 2010.

[29] M.-J. Lai and J. Wang, "An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems," *SIAM Journal on Optimization*, vol. 21, no. 1, pp. 82–101, 2011.

[30] J. Yan and W.-S. Lu, "New algorithms for sparse representation of discrete signals based on $\ell_p - \ell_2$ optimization," in *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2011, pp. 73–78.

[31] Q. Lyu, Z. Lin, Y. She, and C. Zhang, "A comparison of typical ℓ_p minimization algorithms," *Neurocomputing*, vol. 119, no. 0, pp. 413–424, 2013.

[32] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. International Conference on Computer Vision*, 2013, pp. 217–224.

[33] Z. Liu, S. Lin, and M. T. Tan, "Sparse support vector machines with L_p Penalty for Biomarker Identification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 1, pp. 100–107, 2010.

[34] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler, "ℓ_p norm multiple kernel Fisher discriminant analysis for object and image categorisation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3626–3632.

[35] J. H. Oh and N. Kwak, "Generalization of linear discriminant analysis using L_p-norm," *Pattern Recognition Letters*, vol. 34, no. 6, pp. 679–685, 2013.

[36] N. Kwak, "Principal component analysis by L_p-norm maximization," *IEEE Transactions on Cybernetics*, vol. 44, no. 5, pp. 594–609, 2014.

[37] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[38] R. Mazumder, J. H. Friedman, and T. Hastie, "SparseNet: Coordinate descent with nonconvex penalties," *Journal of the American Statistical Association*, vol. 106, no. 495, pp. 1125–1138, 2011.

[39] J. Duchi and Y. Singer, "Boosting with structural sparsity," in *Proc. International Conference on Machine Learning*, 2009, pp. 297–304.

[40] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. International Conference on Machine Learning*, vol. 98, 1998, pp. 82–90.

[41] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[42] X. Li, Y. Pang, and Y. Yuan, "L1-norm-based 2DPCA," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 4, pp. 1170–1175, 2010.

[43] H. Wang and J. Wang, "2DPCA with L1-norm for simultaneously robust and sparse modelling," *Neural Networks*, vol. 46, no. 0, pp. 190–198, 2013.

[44] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[45] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of computational and graphical statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[46] B. Gidas, W.-M. Ni, and L. Nirenberg, "Symmetry and related properties via the maximum principle," *Communications in Mathematical Physics*, vol. 68, no. 3, pp. 209–243, 1979.

[47] R. Horst and H. E. Romeijn, *Handbook of global optimization*. Springer, 2002, vol. 2.

[48] L. Liberti, "Introduction to global optimization," *Lecture of Ecole Polytechnique, Palaiseau F*, vol. 91128, p. 12, 2008.

[49] C. A. Floudas and C. E. Gounaris, "A review of recent advances in global optimization," *Journal of Global Optimization*, vol. 45, no. 1, pp. 3–38, 2009.

[50] R. Martí, "Multi-start methods," in *Handbook of metaheuristics*. Springer, 2003, pp. 355–368.

[51] C. G. E. Boender and A. R. Kan, "Bayesian stopping rules for multistart global optimization methods," *Mathematical Programming*, vol. 37, no. 1, pp. 59–80, 1987.

[52] L. Mackey, "Deflation Methods for Sparse PCA," in *Proc. Advances in Neural Information Processing Systems*, vol. 21, 2008, pp. 1017–1024.

[53] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Advances in neural information processing systems*, 2001, pp. 556–562.

[54] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[55] W. H. Yang, "On generalized Hölder inequality," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 16, no. 5, pp. 489–498, 1991.

[56] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[57] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. The Second IEEE Workshop on Applications of Computer Vision*, 1994, pp. 138–142.

[58] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.