Pattern Recognition 61 (2017) 524–536

Contents lists available at ScienceDirect

Pattern Recognition

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

Joint sparse principal component analysis

Shuangyan Yi [a], Zhihui Lai [b], Zhenyu He [a,*], Yiu-ming Cheung [c,d], Yang Liu [c,d]

[a] School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, China
[b] The College of Computer Science and Software Engineering, Shenzhen University, China
[c] Department of Computer Science, Hong Kong Baptist University, Hong Kong
[d] The Institute of Research and Continuing Education, Hong Kong Baptist University, Hong Kong

ARTICLE INFO

Article history:
Received 16 January 2016
Received in revised form
21 August 2016
Accepted 22 August 2016
Available online 24 August 2016

Keywords:
Dimensionality reduction
Joint sparse
$\ell_{2,1}$-norm

ABSTRACT

Principal component analysis (PCA) is widely used in dimensionality reduction. A lot of variants of PCA have been proposed to improve the robustness of the algorithm. However, the existing methods either cannot select the useful features consistently or is still sensitive to outliers, which will depress their performance of classification accuracy. In this paper, a novel approach called joint sparse principal component analysis (JSPCA) is proposed to jointly select useful features and enhance robustness to outliers. In detail, JSPCA relaxes the orthogonal constraint of transformation matrix to make it have more freedom to jointly select useful features for low-dimensional representation. JSPCA imposes joint sparse constraints on its objective function, i.e., $\ell_{2,1}$-norm is imposed on both the loss term and the regularization term, to improve the algorithmic robustness. A simple yet effective optimization solution is presented and the theoretical analyses of JSPCA are provided. The experimental results on eight data sets demonstrate that the proposed approach is feasible and effective.

1. Introduction

Dimensionality reduction is an important issue in data classification. It aims to learn a transformation matrix to project the high-dimensional data into a low-dimensional subspace so that the data can be effectively classified in the low-dimensional subspace. There are many methods for dimensionality reduction [1–5] and the classical methods are principal component analysis (PCA) [6–9] and linear discriminant analysis (LDA) [10–12]. PCA is an unsupervised method, which projects data information into an orthogonal linear space. LDA is a supervised method, which extracts discriminative data information by maximizing the inter-class scatter matrix and at the same time minimizing the intra-class scatter matrix [13,14].

It is well known that PCA is an unsupervised method and the unsupervised methods are important in the practical applications [15] since labeled data are expensive to obtain [16]. However, the original PCA is sensitive to outliers since its covariance matrix is derived from ℓ_2-norm and ℓ_2-norm is sensitive to outliers [7,17,18]. Thus, PCA fails to deal with the outliers that often appear in data sets in real-world applications. In terms of this problem, many variants of PCA [18,19,16] have been proposed to reduce the effect of outliers. One of the main strategies is to impose ℓ_1-norm on loss term [20,18,21,22,19]. In detail, PCA based on ℓ_1-norm maximization [18] uses a greedy strategy to solve the optimization problem and easy to get stuck in a local solution. Robust principal component analysis with non-greedy ℓ_1-norm maximization (RPCA) [19] is proposed to obtain a much better solution than that in [18]. Recently, $\ell_{2,1}$-norm has caused wide research interests [16,23,24]. Rotational invariant ℓ_1-norm PCA [23] imposes $\ell_{2,1}$-norm on loss term [16]. Optimal mean robust principal component analysis (OMRPCA) [16] based $\ell_{2,1}$-norm is proposed to learn the optimal transformation matrix and optimal mean simultaneously, which imposes $\ell_{2,1}$-norm on loss term.

Although the variants of PCA method mentioned above are able to reduce the effect of outliers to some extent, one major disadvantage of them is that each new feature in low-dimensional subspace is the linear combination of all the original features in high-dimensional space. Therefore, it is usually not good for classification due to the redundant features. Besides, it is often difficult to interpret the new features. Actually, the interpretation of the new features is very important especially when they have physical meanings in many applications such as gene representation and face recognition. To facilitate interpretation, sparse principal component analysis (SPCA) [25] is proposed. However, SPCA has no ability to jointly select the useful features because the ℓ_1-norm is imposed on each transformation vector and ℓ_1-norm cannot select the consistent features. Moreover, SPCA still suffers from the effect of outliers because the ℓ_2-norm is imposed on loss

[*] Corresponding author.
E-mail addresses: lai_zhi_hui@163.com (Z. Lai), zyhe@hitsz.edu.cn (Z. He), ymc@comp.hkbu.edu.hk (Y.-m. Cheung), csygliu@comp.hkbu.edu.hk (Y. Liu).

http://dx.doi.org/10.1016/j.patcog.2016.08.025
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

term.

In this paper, we propose joint sparse principal component analysis (JSPCA), which integrates feature selection into subspace learning to exclude the redundant features. Specifically, JSPCA imposes joint $\ell_{2,1}$-norms on both loss term and regularization term. In this way, our method can discard the useless features on one hand and reduce the effect of outliers on the other hand. The main contributions are described as follows:

(1) JSPCA relaxes the orthogonal constraint of transformation matrix and introduces another transformation matrix to together recover the original data from the subspace spanned by the selected features, which makes JSPCA have more freedom to jointly select useful features for low-dimensional representation.

(2) Unlike PCA and its existing extensions, JSPCA uses joint sparse constraints on the objective function, i.e., $\ell_{2,1}$-norm is imposed on the loss term and the transformation matrix, to do feature selection and learn the optimal transformation matrix simultaneously.

(3) A simple yet effective optimal solution of JSPCA is provided. Furthermore, a series of theoretical analyses including convergence analysis, essence of JSPCA, and computational complexity are provided to validate the feasibility and effectiveness of JSPCA.

The remainder of this paper is organized as follows. In Section 2, we review some existing dimensionality reduction methods. In Section 3, we present the JSPCA model with an effective solution. In Section 4, we give the analyses of JSPCA in theory. In Section 5, we perform experiments and provide the observations. Finally, conclusion is drawn in Section 6.

2. Related work

In this section, we first give the basic notations and then review several variants of PCA. Suppose the given data matrix is $X=[x_1,\ldots,x_n]\in R^{m\times n}$, where m denotes the original image space dimensionality and n denotes the number of training samples. Without loss of generality, $\{x_j\}_{j=1}^n$ is assumed to have zero mean. The problem of linear dimensionality reduction is to project the data from the high-dimensional original space into a low-dimensional subspace. That is, we need to find a transformation matrix $A=[a_1,a_2,\ldots,a_d]\in R^{m\times d}$ with $d\ll m$, where each transformation vector a_k is with m loadings ($k=1,2,\ldots,d$). Then, the transformed data denoted by Y can be shown as follows:

$$Y=A^TX\in R^{d\times n}. \tag{1}$$

Notations: For the matrix A, we denote the (i,j)-th element by a_{ij}, the i-th row by A^i. In this paper, we denote $\|A\|_{2,1}=\sum_{i=1}^m\left\|A^i\right\|_2$, where $\left\|A^i\right\|_2$ means the ℓ_2-norm of vector A^i and $\left\|A^i\right\|_2=\sqrt{\left\|A^i\right\|^T\left\|A^i\right\|}$.

The traditional PCA [6] based on ℓ_2-norm aims to project the high-dimensional data onto the low-dimensional linear subspace spanned by the leading eigenvectors of the data covariance matrix. RPCA [19] based on ℓ_1-norm aims to be robust to outliers by imposing ℓ_1-norm on the projected data. OMRPCA [16] based on $\ell_{2,1}$-norm aims to remove optimal mean automatically and enhance the robustness to outliers by imposing $\ell_{2,1}$-norm on the loss term.

All of the above methods focus on operating different norms such as ℓ_2-norm, ℓ_1-norm, and $\ell_{2,1}$-norm on the loss term. Although the above methods can get a prominent performance in many cases, one common disadvantage of the above methods is that each new feature is the linear combination of all the original features. To this end, the regularization term imposed by different norms is proposed to solve this problem. For example, based on PCA, SPCA [25] is proposed to learn a sparse projection matrix, where each new feature is the linear combination of some original features. Based on spectral regression [26], sparse subspace learning (SSL) [27] is proposed for learning a sparse projection matrix, which first regress the low-dimensional projection data and then solve the projection matrix. However, both SPCA and SSL still cannot exclude the redundant features. Furthermore, based on graph embedding [28], joint feature selection and subspace learning (JFSSL) [2] is proposed to integrate the ability of feature selection into subspace learning. Although JFSSL has the ability of feature selection, it is sensitive to outliers.

3. Joint sparse principal component analysis

In this section, we first present the motivation of this work. Then, we give the objective function of the proposed method. Finally, an iterative optimal solution is given for the proposed objective function.

3.1. Motivation of JSPCA

As the previous statement, SPCA attempts to interpret the selection of features. Intuitively, we use the right subfigure in Fig. 1

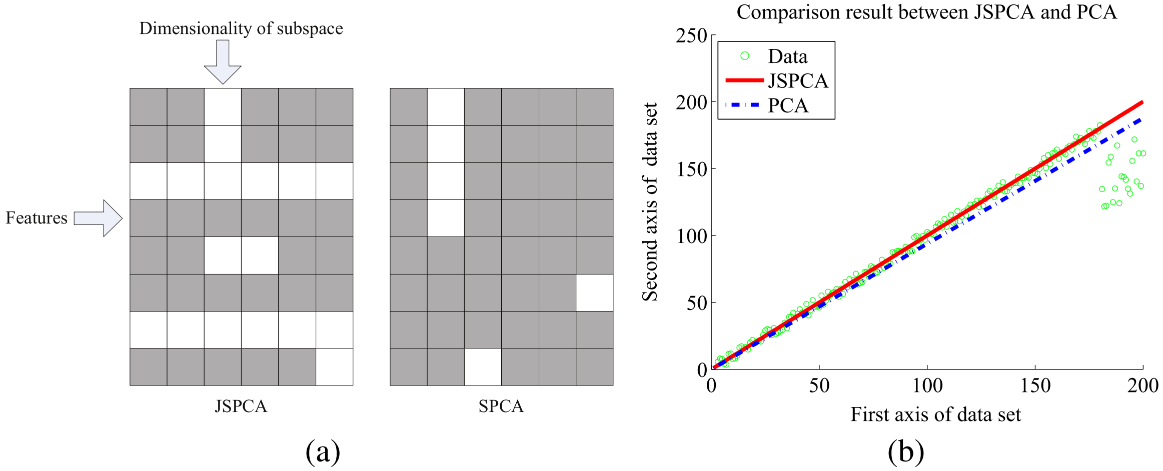


Fig. 1. Motivations of JSPCA. (a) Illustration of two transformation matrices got by JSPCA and SPCA, in which the white block means the zero loading and the gray block means the non-zero loading. JSPCA can tell us that the third and the seventh features are the useless features while SPCA cannot. (b) On a data set with some outliers, JSPCA shows more robustness to outliers than PCA.

(a) to illustrate the learned transformation matrix by SPCA. Note that each row of the transformation matrix corresponds to an original feature while each column corresponds to a dimensionality of the subspace. For one fixed dimensionality of the subspace, the feature with zero loading is not selected. For example, the first four features are not selected on the second dimensionality but they are selected on the remaining subspace dimensionality. Besides, the eighth feature is not selected on the third dimensionality but it is selected on the remaining subspace dimensionality; the sixth feature is not selected on the sixth dimensionality but it is selected on the remaining subspace dimensionality. Since the feature loadings across all the subspace dimensionality cannot be ignorable, it still cannot tell us that which features are really useless as a whole. That is, the useless feature cannot be jointly excluded by SPCA. Inspired by SPCA, we aim to learn a transformation matrix with row-sparsity, which is shown in the left subfigure in Fig. 1(a). In this way, the learned transformation matrix can tell us that the third and seventh features are useless. This is the reason that why we add $\ell_{2,1}$-norm on the transformation matrix.

On the other hand, considering the largely appearing of outliers in real-world applications, we utilize $\ell_{2,1}$-norm on loss term to enhance the robustness to outliers. In order to test the robustness to outliers of JSPCA, 200 points near a straight line are generated with 20 outliers. Then, we apply PCA and JSPCA to this data set, respectively. From Fig. 1(b), we can see that PCA is significantly affected while JSPCA is affected much less. This is the reason that why we add $\ell_{2,1}$-norm on loss term.

3.2. Objective function of JSPCA

Considering the outliers appearing in data sets and the consistent selection of features, we propose the following optimization formulation:

$$\arg\min_{Q,P} J(Q,P) = \arg\min_{Q,P} \left\|X - PQ^TX\right\|_{2,1} + \lambda \|Q\|_{2,1}, \tag{2}$$

where transformation matrix $Q \in R^{m\times d}$ is first used to project the data matrix X onto a low-dimensional subspace and another transformation matrix $P \in R^{m\times d}$ is then used to recover the data matrix X. Here, we relax the orthogonal constraint of transformation matrix Q, introduce another transformation matrix P and add joint $\ell_{2,1}$-norms on both loss term and regularization term. In this way, JSPCA can have more freedom to learn a low-dimensional subspace that approximates to high-dimensional data in a flexible way. The loss term $\left\|X - PQ^TX\right\|_{2,1}$ is not squared and hence it enhances the robustness to outliers. The penalty term $\|Q\|_{2,1}$ penalizes all m regression coefficients corresponding to a single feature as a whole and hence our method is able to jointly select features. On the other hand, the regularization term $\|Q\|_{2,1}$ is convex and can be easily optimized. $\lambda \ge 0$, as a regularization parameter, is used to balance the loss term and regularization term.

Directly solving Eq. (2) is difficult as both loss term and regularization term are non-smooth [1]. Using some mathematical techniques for Eq. (2), we have,

$$\begin{aligned}
&\arg\min_{Q,P} \left\|X - PQ^TX\right\|_{2,1} + \lambda \|Q\|_{2,1}\\
&\quad= \arg\min_{Q,P} 2\mathrm{tr}((X - PQ^TX)^T D_1 (X - PQ^TX)) + 2\lambda \mathrm{tr}(Q^T D_2 Q)\\
&\quad= \arg\min_{Q,P} \mathrm{tr}((X - PQ^TX)^T \sqrt{D_1}^T \sqrt{D_1}(X - PQ^TX)) + \lambda \mathrm{tr}(Q^T \sqrt{D_2}^T \sqrt{D_2} Q)\\
&\quad= \arg\min_{Q,P} \mathrm{tr}((\sqrt{D_1}(X - PQ^TX))^T \sqrt{D_1}(X - PQ^TX)) + \lambda \mathrm{tr}((\sqrt{D_2} Q)^T \sqrt{D_2} Q)\\
&\quad= \arg\min_{Q,P} \left\|\sqrt{D_1}(X - PQ^TX)\right\|_F^2 + \lambda \left\|\sqrt{D_2} Q\right\|_F^2.
\end{aligned} \tag{3}$$

Hence, Eq. (2) becomes,

$$\arg\min_{Q,P} J(Q,P) = \arg\min_{Q,P} \left\|\sqrt{D_1}(X - PQ^TX)\right\|_F^2 + \lambda \left\|\sqrt{D_2} Q\right\|_F^2, \tag{4}$$

where

$$D_1 = \begin{bmatrix} \frac{1}{2\left\|(X - PQ^TX)^1\right\|_2} & & \\ & \frac{1}{2\left\|(X - PQ^TX)^2\right\|_2} & \\ & & \ddots \end{bmatrix}, \tag{5}$$

and

$$D_2 = \begin{bmatrix} \frac{1}{2\left\|Q^1\right\|_2} & & \\ & \frac{1}{2\left\|Q^2\right\|_2} & \\ & & \ddots \end{bmatrix}, \tag{6}$$

are two $m \times m$ diagonal matrices. Note that $(X - PQ^TX)^i$ $(i = 1, 2, \ldots, m)$ means the i-th row of matrix $X - PQ^TX$, and Q^i $(i = 1, 2, \ldots, m)$ means the i-th row of matrix Q. When $\left\|(X - PQ^TX)^i\right\|_2 = 0$, we let $D_1^{ii} = \frac{1}{2\left\|(X - PQ^TX)^i\right\|_2 + \zeta}$ (ζ is a very small constant). Similarly, when $\left\|Q^i\right\|_2 = 0$, we let $D_2^{ii} = \frac{1}{2\left\|Q^i\right\|_2 + \zeta}$. In this way, the smaller the D_2^{ii} is, the more important the i-th feature is. Moreover, we can see that if $\left\|(X - PQ^TX)^i\right\|_2$ and $\left\|Q^i\right\|_2$ are small, D_1 and D_2 are large and thus the minimization of $2\mathrm{tr}((X - PQ^TX)^T D_1 (X - PQ^TX)) + 2\lambda \mathrm{tr}(Q^T D_2 Q)$ in Eq. (3) tends to force $\left\|(X - PQ^TX)^i\right\|_2$ and $\left\|Q^i\right\|_2$ to be a very small value. After several iterations, some $\left\|(X - PQ^TX)^i\right\|_2$ and $\left\|Q^i\right\|_2 (i = 1, 2, \ldots, m)$ may be close to zero and thus we obtain a joint sparse Q and a small reconstruction loss.

Next, let $\sqrt{D_1}P = \bar{P}$, and $\sqrt{D_1}^{-1} Q = \bar{Q}$. Then, the formulation in Eq. (4) can be rewritten as,

$$\arg\min_{\bar{Q},\bar{P}} \left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T \sqrt{D_1}X\right\|_F^2 + \lambda \left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2. \tag{7}$$

In order to reduce the feature redundancy, we impose the orthogonal constraint $\bar{P}^T\bar{P} = I^{d\times d}$ for Eq. (7). Then, we have,

$$\begin{aligned}
&\arg\min_{\bar{Q},\bar{P}} J(\bar{Q},\bar{P}) = \arg\min_{\bar{Q},\bar{P}} \left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T \sqrt{D_1}X\right\|_F^2 + \lambda \left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2,\\
&\qquad\qquad \text{s.t.}\ \ \bar{P}^T\bar{P} = I^{d\times d},
\end{aligned} \tag{8}$$

where $\bar{Q} \in R^{m\times d}$ is first used to project the weighted data matrix $\sqrt{D_1}X$ and $\bar{P} \in R^{m\times d}$ is then used to recover it.

3.3. The optimal solution

The solution of Eq. (8) is divided into the below two steps:

Step 1: Given $\bar{P}$, there exists an optimal matrix $\bar{P}_\perp$ such that $[\bar{P}, \bar{P}_\perp]$ is $m \times m$ column orthogonal matrix. Then, optimization problem in Eq. (8) becomes,

$$\arg\min_{\bar{Q}} \left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T \sqrt{D_1}X\right\|_F^2 + \lambda \left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2. \tag{9}$$

The first part of Eq. (9) can be rewritten as,

$$\begin{aligned}\left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2 &= \left\|X^T\sqrt{D_1} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\right\|_F^2\\ &= \left\|X^T\sqrt{D_1}[\bar{P}, \bar{P}_\perp] - X^T\sqrt{D_1}\bar{Q}\bar{P}^T[\bar{P}, \bar{P}_\perp]\right\|_F^2\\ &= \left\|X^T\sqrt{D_1}\bar{P} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\bar{P}\right\|_F^2\\ &\quad + \left\|X^T\sqrt{D_1}\bar{P}_\perp - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\bar{P}_\perp\right\|_F^2\\ &= \left\|X^T\sqrt{D_1}\bar{P} - X^T\sqrt{D_1}\bar{Q}\right\|_F^2 + \left\|X^T\sqrt{D_1}\bar{P}_\perp\right\|_F^2. \end{aligned} \tag{10}$$

Since $\bar{P}$ is fixed, and $\left\|X^T\sqrt{D_1}\bar{P}_\perp\right\|_F^2$ is a constant, optimization problem in Eq. (8) becomes the following optimization problem:

$$\arg\min_{\bar{Q}} \left\|X^T\sqrt{D_1}\bar{P} - X^T\sqrt{D_1}\bar{Q}\right\|_F^2 + \lambda\left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2. \tag{11}$$

By the derivative of Eq. (11) with respect to $\bar{Q}$ to be 0, we get,

$$\bar{Q} = (\lambda\sqrt{D_1}D_2\sqrt{D_1} + \sqrt{D_1}XX^T\sqrt{D_1})^{-1}\sqrt{D_1}XX^T\sqrt{D_1}\bar{P}. \tag{12}$$

Hence,

$$Q = (\lambda D_2 + XX^T)^{-1}XX^T\sqrt{D_1}\bar{P}. \tag{13}$$

Step 2: Given $\bar{Q}$ to compute $\bar{P}$, optimization problem in Eq. (8) becomes,

$$\arg\min_{\bar{P}} \left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2, \quad \text{s.t. } \bar{P}^T\bar{P} = I^{I^{d\times d}}. \tag{14}$$

The first part of Eq. (14) can be rewritten as,

$$\begin{aligned}&\left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2\\ &\quad = \mathrm{tr}((\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X)^T(\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X))\\ &\quad = \mathrm{tr}((X^T\sqrt{D_1} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T)(\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X))\\ &\quad = \mathrm{tr}(X^TD_1X - X^T\sqrt{D_1}\bar{P}\bar{Q}^T\sqrt{D_1}X - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\sqrt{D_1}X\\ &\qquad + X^T\sqrt{D_1}\bar{Q}\bar{P}^T\bar{P}\bar{Q}^T\sqrt{D_1}X)\\ &\quad = \mathrm{tr}(X^TD_1X + X^T\sqrt{D_1}\bar{Q}\bar{Q}^T\sqrt{D_1}X) - 2\mathrm{tr}(\bar{Q}^T\sqrt{D_1}XX^T\sqrt{D_1}\bar{P}).\end{aligned} \tag{15}$$

Since $\bar{Q}$ is given, Eq. (14) becomes,

$$\arg\min_{\bar{P}} \mathrm{tr}(\bar{Q}^T\sqrt{D_1}XX^T\sqrt{D_1}\bar{P}), \quad \text{s.t. } \bar{P}^T\bar{P} = I^{d\times d}. \tag{16}$$

On the other hand, optimization problem in Eq. (14) is equal to,

$$\arg\min_{\bar{P}} \left\|X^T\sqrt{D_1} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\right\|_F^2, \quad \text{s.t. } \bar{P}^T\bar{P} = I^{d\times d}. \tag{17}$$

The update of $\bar{P}$ of minimizing Eq. (17) with the constraint of $\bar{P}^T\bar{P} = I^{d\times d}$ means that $\bar{P}$ is orthogonal in the columns. In order to compute $\bar{P}$, we introduce the following lemma [25].

Lemma 1. *Let $Z^{n\times m}$ and $V^{n\times d}$ be two matrices. Consider the constrained minimization problem,*

$$\arg\min_{P} \left\|Z - VP^T\right\|^2, \quad s.t. \ P^TP = I^{d\times d}. \tag{18}$$

Suppose the SVD of Z^TV is EDU^T, then the optimal solution is $P = EU^T$.

According to Lemma 1, we have $Z^TV = \sqrt{D_1}\ XX^T\sqrt{D_1}\bar{Q}$. Let the SVD of $\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = EDU^T$, we have,

$$\bar{P} = EU^T. \tag{19}$$

Thus,

$$P = \sqrt{D_1}^{-1}EU^T. \tag{20}$$

In fact, before we compute $\bar{Q}$ in Eq. (12), we need to compute the input of matrix $\bar{P}$, D_1 and D_2, which cannot be obtained directly. Therefore, we need to compute them in the designed iterative algorithm. Once $\bar{P}$, $\bar{Q}$, D_1 and D_2 are obtained, we can obtain P and Q according to Eqs. (20) and (13). According to the obtained P and Q, we get the new D_1 and D_2. Iterating the above procedures will give the local optimal solutions of the algorithm. Algorithm 1 gives the details.

4. Discussion and analysis

In this section, we will further give the theoretical analysis of the proposed method, which includes convergence analysis, essence of the optimization algorithm, connection to weighted PCA and computational complexity analysis.

4.1. Convergence analysis

Before giving the proof of convergence of the proposed optimal algorithm, we need to give the following lemma [29].

Lemma 2. *For any nonzero vectors $p, q \in R^c$, the following result holds:*

$$\|p\|_2 - \frac{\|p\|_2^2}{2\|q\|_2} \le \|q\|_2 - \frac{\|q\|_2^2}{2\|q\|_2}. \tag{21}$$

Based on Lemma 2, we give the following proposed theorem.

Theorem 1. *Given all the variables in Eq. (2) except for P, Q, the optimal problem in Eq. (2) will monotonically decrease the objective function value in each iteration and converge to the local optimum solution.*

Proof. For simplicity, we denote the objective function in Eq. (2) as $J(Q, P) = J(Q, P, D_1, D_2)$, suppose for the $t-1$-th iteration, we obtain $P^{(t-1)}$, $Q^{(t-1)}$, $D_1^{(t-1)}$ and $D_2^{(t-1)}$. From Eq. (13), we can find that,

$$J(Q^{(t)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}) \le J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}). \tag{22}$$

Since the SVD gives the optimal $P^{(t)}$ that further decreases the objective value, we have,

$$J(Q^{(t)}, P^{(t)}, D_1^{(t-1)}, D_2^{(t-1)}) \le J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}). \tag{23}$$

Once the optimal $P^{(t)}$ and $Q^{(t)}$ are obtained, we have,

$$\begin{aligned}&\mathrm{tr}((X - P^{(t)}Q^{(t)T}X)^TD_1^{(t-1)}(X - P^{(t)}Q^{(t)T}X)) + \lambda\mathrm{tr}(Q^{(t)T}D_2^{(t-1)}Q^{(t)})\\ &\quad \le \mathrm{tr}((X - P^{(t-1)}Q^{(t-1)T}X)^TD_1^{(t-1)}(X - P^{(t-1)}Q^{(t-1)T}X))\\ &\qquad + \lambda\mathrm{tr}(Q^{(t-1)T}D_2^{(t-1)}Q^{(t-1)}).\end{aligned} \tag{24}$$

That is,

$$\begin{aligned}&\mathrm{tr}\left(\sum_{i=1}^{m}\frac{\left\|X - P_i^{(t)}Q_i^{(t)T}X\right\|_2^2}{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2}\right) + \lambda\mathrm{tr}\left(\sum_{i=1}^{m}\frac{\left\|Q_i^{(t)}\right\|_2^2}{\left\|Q_i^{(t-1)}\right\|_2}\right)\\ &\quad \le \mathrm{tr}\left(\sum_{i=1}^{m}\frac{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2^2}{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2}\right) + \lambda\mathrm{tr}\left(\sum_{i=1}^{m}\frac{\left\|Q_i^{(t-1)}\right\|_2^2}{\left\|Q_i^{(t-1)}\right\|_2}\right).\end{aligned} \tag{25}$$

On one hand, according to Lemma 2, we have,

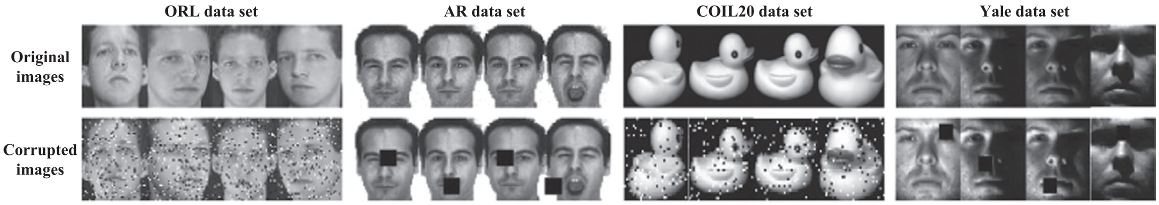


Fig. 2. Visualization of some original data sets and their corresponding corruptions, where the original images from four data sets are shown in the first row and the corresponding corrupted images are shown in the second row. Specifically, each image of AR and Yale data sets is corrupted by 10 × 10 block occlusions while ORL and COIL20 data sets are corrupted by 10% salt & pepper noises.

$$\left\|X - P_i^{(t)}Q_i^{(t)T}X\right\|_2 - \frac{\left\|X - P_i^{(t)}Q_i^{(t)T}X\right\|_2^2}{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2} \le \left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2 - \frac{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2^2}{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2}. \tag{26}$$

Using the matrix calculus for Eq. (26), we have the formulation as follows:

$$\sum_{i=1}^{m}\left(\left\|X - P_i^{(t)}Q_i^{(t)T}X\right\|_2 - \frac{\left\|X - P_i^{(t)}Q_i^{(t)T}X\right\|_2^2}{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2}\right) \le \sum_{i=1}^{m}\left(\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2 - \frac{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2^2}{\left\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\right\|_2}\right). \tag{27}$$

On the other hand, according to Lemma 2, we have,

$$\left\|Q_i^{(t)}\right\|_2 - \frac{\left\|Q_i^{(t)}\right\|_2^2}{\left\|Q_i^{(t-1)}\right\|_2} \le \left\|Q_i^{(t-1)}\right\|_2 - \frac{\left\|Q_i^{(t-1)}\right\|_2^2}{\left\|Q_i^{(t-1)}\right\|_2}, \tag{28}$$

Similarly, using the matrix calculus for Eq. (28), we have the formulation as follows:

$$\sum_{i=1}^{m}\left(\left\|Q_i^{(t)}\right\|_2 - \frac{\left\|Q_i^{(t)}\right\|_2^2}{\left\|Q_i^{(t-1)}\right\|_2}\right) \le \sum_{i=1}^{m}\left(\left\|Q_i^{(t-1)}\right\|_2 - \frac{\left\|Q_i^{(t-1)}\right\|_2^2}{\left\|Q_i^{(t-1)}\right\|_2}\right). \tag{29}$$

By combining Eqs. (25) and (27) with Eq. (29), we have,

$$\begin{aligned} J(Q^{(t)}, P^{(t)}, D_1^{(t-1)}, D_2^{(t-1)}) &= \left\|X - P^{(t)}Q^{(t)T}X\right\|_{21} + \lambda \left\|Q^{(t)}\right\|_{21} \\ &\le \left\|X - P^{(t-1)}Q^{(t-1)T}X\right\|_{21} + \lambda \left\|Q^{(t-1)}\right\|_{21} \\ &= J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}). \end{aligned} \tag{30}$$

That is,

$$\begin{aligned} J(Q^{(t)}, P^{(t)}) &= J(Q^{(t)}, P^{(t)}, D_1^{(t)}, D_2^{(t)}) \le J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}) \\ &= J(Q^{(t-1)}, P^{(t-1)}). \end{aligned} \tag{31}$$

Therefore, the optimization problem in Eq. (2) can converge to the local optimum solution. □

Algorithm 1. JSPCA algorithm

```
Input: Training sample set $X$, parameter $\lambda$, and dimensionality $d$.
  1: Initialize $D_1 = I^{m\times m}$, $D_2 = I^{m\times m}$ and random $\bar{P}^{m\times d}$.
  2: while not converge do
    2.1: Compute $\bar{Q}$ according to Eq. (12)
    2.2: Compute $Q$ according to Eq. (13)
    2.3: Compute $\bar{P}$ according to Eq. (19)
    2.4: Compute $P$ according to Eq. (20)
    2.5: Compute $D_1$ according to Eq. (5)
    2.6: Compute $D_2$ according to Eq. (6)
  end while
  3: Normalize each column vectors of $Q$ to be identity vectors.
Output: Transformation matrix $Q$.
```

4.2. *Essence of JSPCA*

4.2.1. *The intrinsic relationship between P and Q*

In order to explore the essence of P and Q in Eq. (2), we need to explore the essence of $\bar{P}$ and $\bar{Q}$ in Eq. (8).

Substituting Eq. (12) into Eq. (16), we get the following optimization problem:

$$\arg\min_{\bar{P}} \ \mathrm{tr}(\bar{P}^T\sqrt{D_1}XX^T\sqrt{D_1}(\lambda\sqrt{D_1}D_2\sqrt{D_1} + \sqrt{D_1}XX^T\sqrt{D_1})^{-1}\sqrt{D_1} XX^T\sqrt{D_1}\bar{P}), \ \text{s.t.} \quad \bar{P}^T\bar{P} = I^{d\times d}. \tag{32}$$

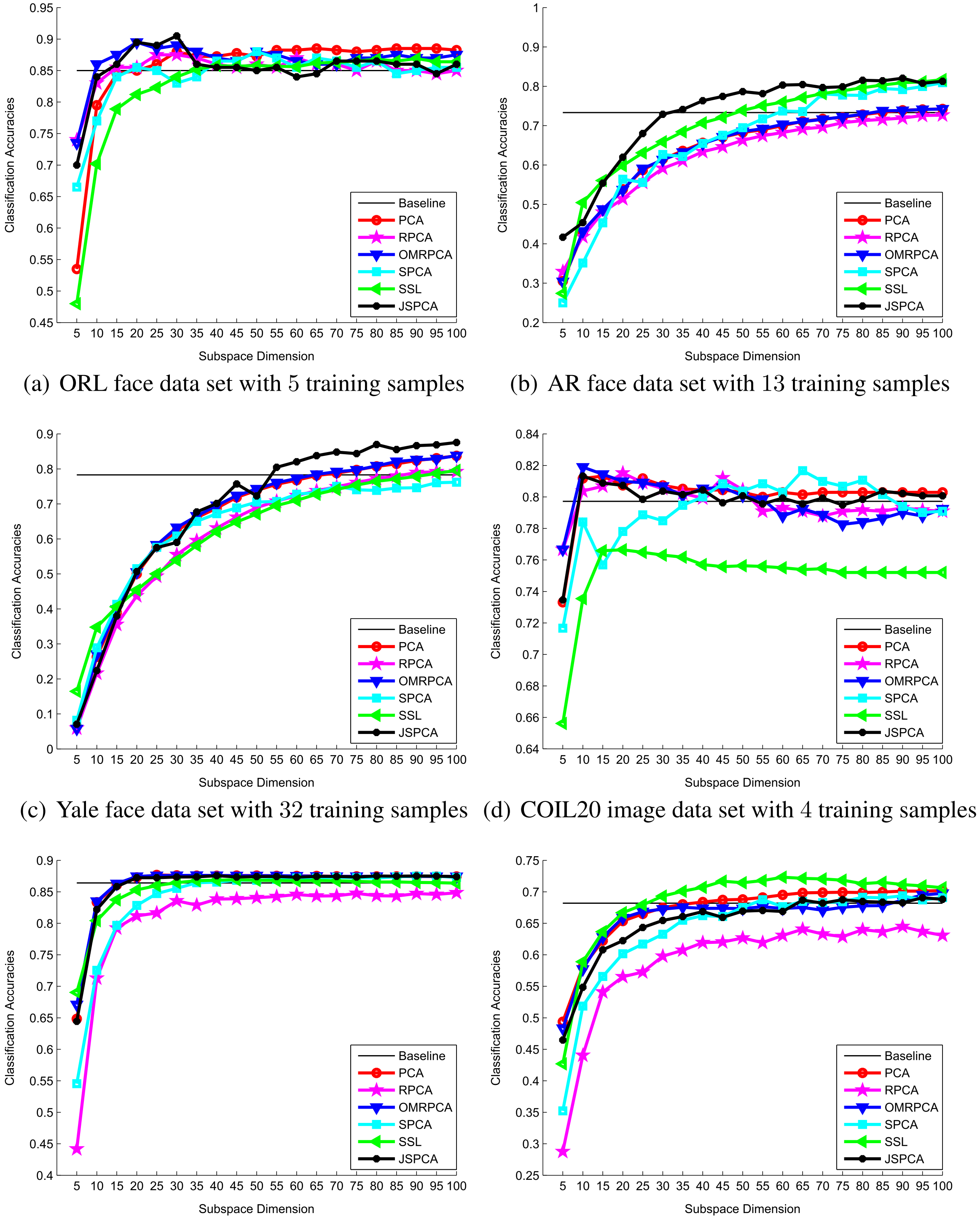
It is clear that the optimal solution $\bar{P}$ is the standard eigendecomposition of the following eigen equation:

$$\sqrt{D_1}XX^T\sqrt{D_1}(\lambda\sqrt{D_1}D_2\sqrt{D_1} + \sqrt{D_1}XX^T\sqrt{D_1})^{-1}\sqrt{D_1}XX^T\sqrt{D_1}\bar{P} = \bar{P}\Sigma, \tag{33}$$

where Σ is the eigenvalue matrix. Therefore, $\bar{P}$ contains the eigenvectors corresponding to the larger eigenvalues of Eq. (33). If $\bar{P}$ contains all the eigenvectors of Eq. (33), since Eqs. (12) and (33) can be rewritten as $\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = \bar{P}\Sigma$. Furthermore,

Table 1
Classification performance (average classification accuracy with standard deviation) on the facial data sets with different training samples.

Data sets	Baseline	PCA	RPCA	OMRPCA	SPCA	SSL	JSPCA
ORL/3	0.7450 ± 0.0314	0.7432 ± 0.0288	0.8016 ± 0.0201	0.8070 ± 0.0148	0.7641 ± 0.0289	0.7300 ± 0.0268	**0.8095 ± 0.0176**
ORL/5	0.8500 ± 0.0281	0.8630 ± 0.0265	0.8635 ± 0.0321	0.8710 ± 0.0224	0.8640 ± 0.0227	0.8690 ± 0.0201	**0.9050 ± 0.0236**
ORL/7	0.9233 ± 0.0292	0.9042 ± 0.0297	0.9356 ± 0.0245	0.9350 ± 0.0257	0.9198 ± 0.0211	0.9217 ± 0.0173	**0.9392 ± 0.0241**
AR/8	0.7218 ± 0.0105	0.7017 ± 0.0101	0.7262 ± 0.0135	0.7308 ± 0.0063	0.7608 ± 0.0152	0.7636 ± 0.0215	**0.7700 ± 0.0122**
AR/13	0.7333 ± 0.0088	0.7423 ± 0.0197	0.7396 ± 0.0153	0.7423 ± 0.0138	0.8090 ± 0.0267	0.8169 ± 0.0075	**0.8206 ± 0.0152**
AR/18	0.8733 ± 0.0112	0.8507 ± 0.0117	0.8508 ± 0.0091	0.8612 ± 0.0056	0.8682 ± 0.0200	0.8745 ± 0.0089	**0.8780 ± 0.0172**
Yale/22	0.6984 ± 0.0118	0.7188 ± 0.0026	0.6873 ± 0.0050	0.7172 ± 0.0078	0.6560 ± 0.0313	0.7518 ± 0.0187	**0.7700 ± 0.0086**
Yale/27	0.7457 ± 0.0048	0.7615 ± 0.0102	0.7342 ± 0.0086	0.7652 ± 0.0114	0.7081 ± 0.0222	0.7736 ± 0.0103	**0.8135 ± 0.0085**
Yale/32	0.7831 ± 0.0149	0.8266 ± 0.1068	0.7520 ± 0.0164	0.7780 ± 0.0130	0.7360 ± 0.0702	0.7958 ± 0.0130	**0.8455 ± 0.1085**



(a) ORL face data set with 5 training samples

(b) AR face data set with 13 training samples

(c) Yale face data set with 32 training samples

(d) COIL20 image data set with 4 training samples

(e) USPS data set with 20 training samples

(f) Isolate data set with 4 training samples

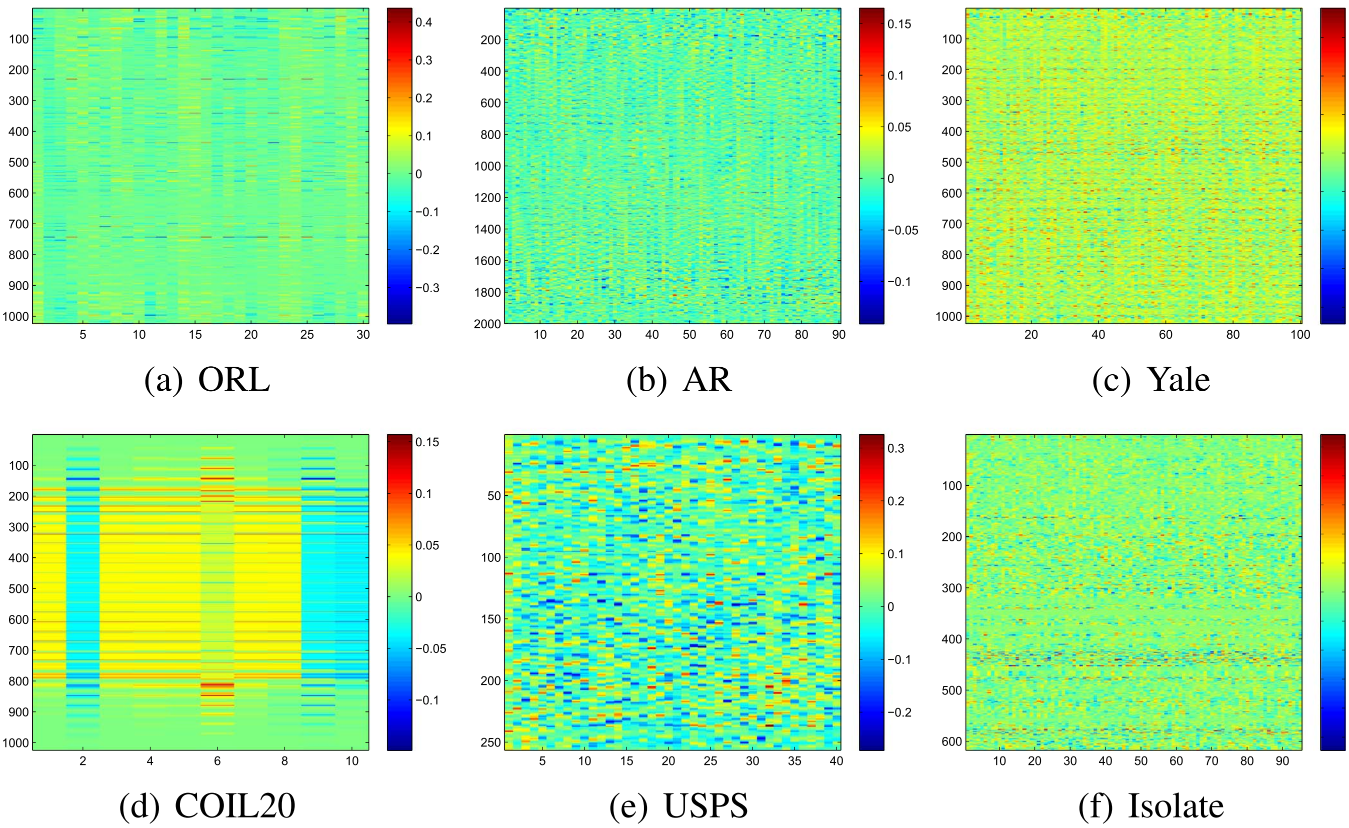
Fig. 3. Classification accuracies of some subspace learning methods versus the dimensions on six data sets.

$\bar{P}^T\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = \Sigma$.

From the above analysis, we can obtain the following interesting conclusions: If $\lambda > 0$, $\bar{Q}$ is sparse in row, Eq. (12) indicates that the optimization solution for the objective function in Eq. (2) is to find a row-sparse matrix $\bar{Q}$ ($Q = \sqrt{D_1}\bar{Q}$) and an orthogonal matrix $\bar{P}$ $\left(P = \sqrt{D_1}^{-1}\bar{P}\right)$ to diagonalize $\sqrt{D_1}XX^T\sqrt{D_1}(\lambda\sqrt{D_1}D_2\sqrt{D_1} + \sqrt{D_1}XX^T\sqrt{D_1})^{-1}$ $\sqrt{D_1}XX^T\sqrt{D_1}$. When $\sqrt{D_1}XX^T\sqrt{D_1}$ is full rank, if $\lambda = 0$ or $\lambda \to 0$, $\bar{Q}$ is not sparse and $\bar{Q} = \bar{P}(Q = D_1P)$ or $\bar{Q} \to \bar{P}$ $(Q \to D_1P)$. At this moment, the optimal solution in Eq. (2) aims to find the optimal non-sparse column orthogonal matrix $\bar{Q}$ ($Q = \sqrt{D_1}\bar{Q}$) to diagonalize scatter matrix $\sqrt{D_1}XX^T\sqrt{D_1}^T$, i.e., $\bar{Q}^T\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = \Sigma$ or $\bar{Q}^T\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} \to \Sigma$. This is the degenerated weighted PCA, which is similar to the traditional PCA but differ from it.

4.2.2. *Connection to weighted PCA*

When $\lambda = 0$, Eq. (8) becomes,



(a) ORL (b) AR (c) Yale

(d) COIL20 (e) USPS (f) Isolate

Fig. 4. Projection matrix got by JSPCA on six data sets.

$$\arg\min_{\bar{Q},\bar{P}} J(\bar{Q},\bar{P}) = \arg\min_{\bar{Q},\bar{P}} \left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2, \quad \text{s.t.}\ \ \bar{P}^T\bar{P} = I^{d\times d}. \tag{34}$$

In fact, when $\lambda = 0$, we get $\bar{P} = \bar{Q}$ which have been discussed in Section 4.2.1. At this moment, we have,

$$\left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2 = \left\|\sqrt{D_1}X - \bar{Q}\bar{Q}^T\sqrt{D_1}X\right\|_F^2. \tag{35}$$

Obviously, the optimal $\bar{Q}$ under this case is exactly the first d transformation vectors of weighted scatter matrix $(\sqrt{D_1}X)(\sqrt{D_1}X)^T$. Therefore, the proposed method can degenerate into weighted PCA whose weight matrix $\sqrt{D_1}$ is adaptive and can be induced by the penalty term $\|Q\|_{2,1}$ itself. It is just the weight matrix $\sqrt{D_1}$ that makes our method robust to outliers. In other words, the essence of JSPCA is to add the sparsity to weighted PCA.

4.2.3. The learned subspace by JSPCA

According to the statement in Section 4.2.1, when $\lambda = 0$, the proposed JSPCA can degenerate into weighted PCA whose generalized eigen equation is shown as follows:

$$(\sqrt{D_1}X)(\sqrt{D_1}X)^T\xi = \alpha\xi. \tag{36}$$

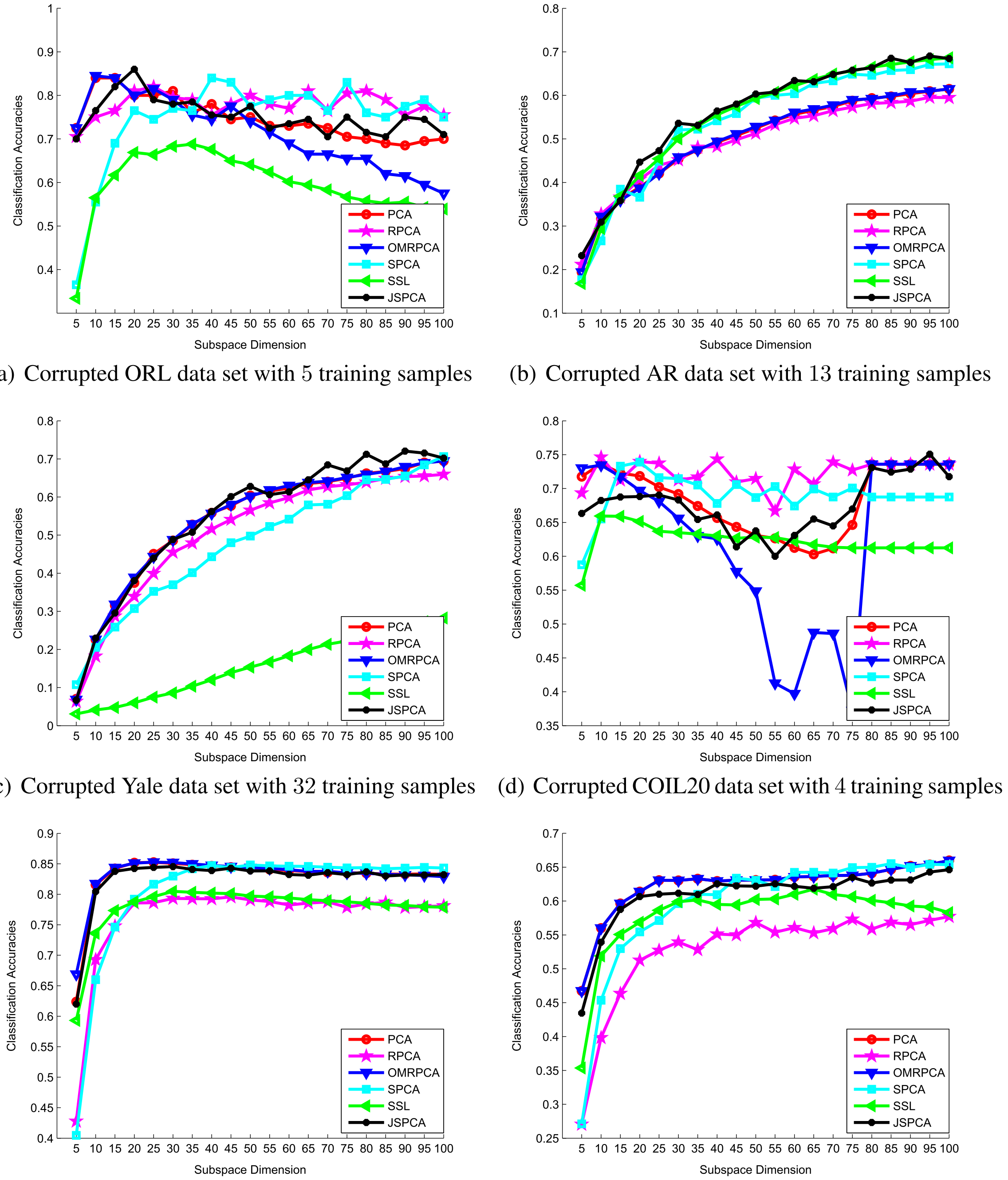
Obviously, $(\sqrt{D_1}X)(\sqrt{D_1}X)^T$ is a symmetric matrix. Then, we have $SVD((\sqrt{D_1}X)(\sqrt{D_1}X)^T) = E\Lambda E^T$. Equivalently, we have $(\sqrt{D_1}X)(\sqrt{D_1}X)^T E = E\Lambda$. Therefore, the first d eigenvectors corresponding to

Table 2
Classification performance (average classification accuracy with standard deviation) on the non-facial data sets with different training samples.

Data sets	Baseline	PCA	RPCA	OMRPCA	SPCA	SSL	JSPCA
COIL20/4	0.7972 ± 0.0274	**0.8132 ± 0.0187**	0.7769 ± 0.0275	0.7841 ± 0.0293	0.8056 ± 0.0233	0.7665 ± 0.0216	0.8013 ± 0.0234
COIL20/5	0.8182 ± 0.0238	0.8279 ± 0.0230	0.8101 ± 0.0204	0.8103 ± 0.0287	**0.8349 ± 0.0172**	0.7799 ± 0.0145	0.8117 ± 0.0231
COIL20/6	0.8442 ± 0.0149	0.8527 ± 0.0225	0.8441 ± 0.0156	0.8383 ± 0.0147	**0.8679 ± 0.0122**	0.8114 ± 0.0156	0.8153 ± 0.0090
USPS/10	0.8149 ± 0.0114	0.8191 ± 0.0140	0.7774 ± 0.0208	0.8195 ± 0.0142	0.8154 ± 0.0137	0.8158 ± 0.0131	**0.8199 ± 0.0131**
USPS/15	0.8420 ± 0.0112	**0.8460 ± 0.0102**	0.8126 ± 0.0105	0.8449 ± 0.0112	0.8435 ± 0.0114	0.8337 ± 0.0088	0.8446 ± 0.0120
USPS/20	0.8743 ± 0.0086	**0.8767 ± 0.0097**	0.8485 ± 0.0090	0.8761 ± 0.0082	0.8748 ± 0.0085	0.8690 ± 0.0071	0.8756 ± 0.0083
Isolate/4	0.6820 ± 0.0162	0.7021 ± 0.0159	0.6448 ± 0.0216	0.6975 ± 0.0203	0.6937 ± 0.0161	**0.7231 ± 0.0212**	0.6905 ± 0.0052
Isolate/5	0.7119 ± 0.0110	0.7187 ± 0.0198	0.6677 ± 0.0261	0.7049 ± 0.0229	0.7175 ± 0.0069	**0.7411 ± 0.0071**	0.6965 ± 0.0248
Isolate/6	0.7234 ± 0.0222	0.7221 ± 0.0128	0.6714 ± 0.0172	0.7181 ± 0.0142	0.7252 ± 0.0142	**0.7507 ± 0.0140**	0.7077 ± 0.0188
MNIST/20	0.7822 ± 0.0134	0.7932 ± 0.0059	0.7653 ± 0.0145	0.7969 ± 0.0062	**0.7990 ± 0.0102**	0.7368 ± 0.0154	0.7944 ± 0.0041
MNIST/30	0.8140 ± 0.0055	0.8297 ± 0.0097	0.7992 ± 0.0143	**0.8312 ± 0.0080**	0.8264 ± 0.0072	0.7665 ± 0.0042	0.8190 ± 0.0075
MNIST/40	0.8389 ± 0.0074	**0.8547 ± 0.0050**	0.8307 ± 0.0065	0.8542 ± 0.0050	0.8525 ± 0.0075	0.7849 ± 0.0091	0.8374 ± 0.0103
COIL100/10	0.7899 ± 0.0055	0.8165 ± 0.0054	0.8117 ± 0.0075	0.8192 ± 0.0058	0.8051 ± 0.0041	0.7799 ± 0.0081	**0.8201 ± 0.0031**
COIL100/20	0.8773 ± 0.0051	0.9022 ± 0.0059	0.8958 ± 0.0096	**0.9113 ± 0.0066**	0.8967 ± 0.0080	0.8669 ± 0.0065	0.9018 ± 0.0045
COIL100/30	0.9203 ± 0.0034	0.9401 ± 0.0036	0.9310 ± 0.0091	**0.9453 ± 0.0044**	0.9395 ± 0.0047	0.9148 ± 0.0020	0.9394 ± 0.0067

Table 3
Classification performance (average classification accuracy with standard deviation) on eight corrupted data sets with different training samples.

Data sets	PCA	RPCA	OMRPCA	SPCA	SSL	JSPCA
ORL/5	0.8400 ± 0.0200	0.8200 ± 0.0199	0.8450 ± 0.0128	0.8400 ± 0.0209	0.6880 ± 0.0261	**0.8600 ± 0.0145**
AR/13	0.6147 ± 0.0063	0.5955 ± 0.0195	0.6147 ± 0.0088	0.6724 ± 0.0067	0.6863 ± 0.0173	**0.6904 ± 0.0192**
Yale/32	0.6945 ± 0.0288	0.6594 ± 0.0291	0.6945 ± 0.0168	0.7062 ± 0.0200	0.2825 ± 0.0268	**0.7204 ± 0.0112**
COIL20/4	0.7368 ± 0.0116	0.7463 ± 0.0191	0.7360 ± 0.0198	0.7390 ± 0.0109	0.6594 ± 0.0168	**0.7507 ± 0.0092**
USPS/20	0.8525 ± 0.0104	0.7956 ± 0.0197	**0.8528 ± 0.0197**	0.8482 ± 0.0117	0.8049 ± 0.0103	0.8454 ± 0.0147
Isolate/4	**0.6599 ± 0.0134**	0.5771 ± 0.0238	0.6599 ± 0.0182	0.6552 ± 0.0309	0.6176 ± 0.0185	0.6463 ± 0.0179
MNIST/40	0.8401 ± 0.0074	0.7793 ± 0.0114	**0.8416 ± 0.0091**	0.8407 ± 0.0077	0.7398 ± 0.0181	0.8084 ± 0.0075
COIL100/10	0.6652 ± 0.0102	0.6558 ± 0.0055	0.6766 ± 0.0093	0.6885 ± 0.0087	0.5274 ± 0.0125	**0.7976 ± 0.0093**



(a) Corrupted ORL data set with 5 training samples

(b) Corrupted AR data set with 13 training samples

(c) Corrupted Yale data set with 32 training samples

(d) Corrupted COIL20 data set with 4 training samples

(e) Corrupted USPS data set with 20 training samples

(f) Corrupted Isolate data set with 4 training samples

Fig. 5. Classification accuracies of some subspace learning methods versus the dimensions on six corrupted data sets.

the increasing ordered eigenvalues of the eigenfunction of Eq. (36) are exactly the first d columns of E. $\Phi = span\{E\}$ is the subspace spanned by eigenvectors of the generalized eigenfunction (36) of weighted PCA.

In the following, we discuss the relationship between Q got by JSPCA and the eigenvectors of Eq. (36).

Substituting Eq. (19) into Eq. (13), we have,

$$Q = (\lambda D_2 + XX^T)^{-1} XX^T \sqrt{D_1} E U^T. \tag{37}$$

Note that when $\lambda \to 0$ and XX^T is full rank, we have $Q \to \sqrt{D_1} EU^T$. Denote our sparse subspace as Ω. Then, $\Omega = span\{Q\} \to span\{\sqrt{D_1} E\} = span\{E\} = \Phi$. When $D_1 = I$, Eq. (36) becomes $XX^T \xi = \alpha\xi$, which is the eigen equation of traditional PCA. The D_1 in JSPCA is not usually I.

4.3. Computational complexity analysis

The main computational complexity of JSPCA have two steps in each iteration, the first step is to compute $Q = (\lambda D_2 + XX^T)^{-1} XX^T \sqrt{D_1} \bar{P}$ with $O(m^3)$. The second step is to compute the SVD of $\sqrt{D_1} XX^T \sqrt{D_1} \bar{Q} = EDU^T$, whose computational complexity is also $O(m^3)$ at most. Therefore, the computational complexity of one iteration will be up to $O(m^3)$. If the algorithm needs t iteration steps, then the total computational complexity is in the order of $O(tm^3)$.

5. Experiments

To evaluate JSPCA, we compare it with the traditional PCA and its variants including PCA [6], RPCA [19], OMRPCA [16], SPCA [25], and SSL [27]. In order to compare the dimensionality reduction performance of different methods objectively and persuasively, we test these methods on each data set using the nearest neighbor (NN) classifier to obtain the classification accuracy.

Additionally, in order to test the robustness to outliers of JSPCA, we simulate the following two levels of corruptions:

(1) *Block occlusions*: The block occlusions are randomly added to different locations in each image with block size of 10 × 10.

(2) *Random pixel corruptions*: The pixels are randomly chosen from each image and corrupted by salt & pepper noises. The rate of corrupted pixels is 10%.

Our experiments are divided into two groups: One is the experiments on the original data sets; the other is the experiments on the corrupted data sets. The so-called corrupted data sets are constructed in this way: we add the block occlusions on AR and Yale (Extended Yale B) data sets; we add random pixel corruptions on ORL, COIL20, USPS, Isolate, MNIST, and COIL100 data sets. Fig. 2 shows some original images in the first row and the corresponding corrupted images in the second row.

5.1. Data sets

The ORL face data set, including frontal views of faces with different facial expressions and lighting conditions, contains 40 individuals and each individual contains 10 face images. Here, we resize each image to 56 × 46 pixels.

The AR face data set [30,31], including frontal views of faces with different facial expressions, lighting conditions and occlusions (glasses and scarf), contains 120 individuals in which each individual contains 26 images. Here, we resize each image to 50 × 40 pixels.

The Yale face data set [32,33] contains 2414 frontal face images of 38 individuals [32] under different lighting conditions. Each individual contains about 64 images and half of the images are corrupted by shadows or reflection. Here, each image is cropped and resized to 50 × 40 pixels.

The COIL20 image data set [34] contains 20 individuals in which each individual contains 72 images and each image is taken at pose intervals of 5°. Here, each image is converted to a gray-scale image of 32 × 32 pixels.

The USPS data set [35] contains totally 9298 digit images from 0 to 9, each of which is of size 16 × 16 pixels, with 256 gray levels per pixel.

The used Isolate data set [36] contains totally 150 speakers who spoke the name of each letter of the alphabet twice. The speakers are grouped into sets of 30 speakers each and are referred to as Isolate1 through isolate5. Here, we refer Isolate1 as the used Isolate data set where the dimensionality is 617 and size is 1560.

The used MNIST data set (http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html) contains totally 4000 digit images from 0 to 9, each of which is of size 28 × 28 pixels, with 784 gray levels per pixel.

The COIL100 data set (http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html) contains 100 individuals in which each individual contains 72 images and each image is taken at pose intervals of 5°. Here, each image is converted to a gray-scale image of 32 × 32 pixels.

5.2. Experiments on the original data sets

5.2.1. Experiments on the facial data sets

On the ORL data set, we randomly select 3, 5, and 7 samples

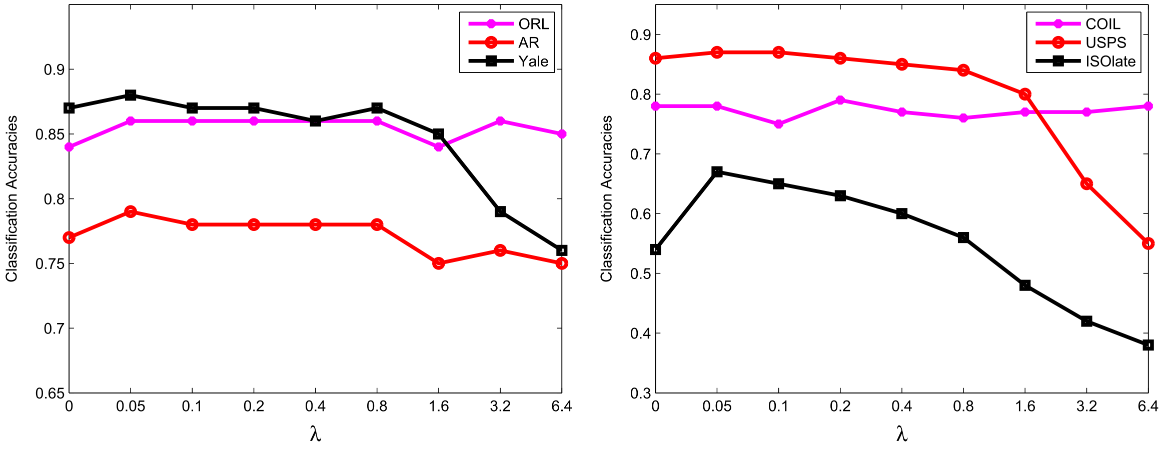


Fig. 6. Parameter selection on six data sets.

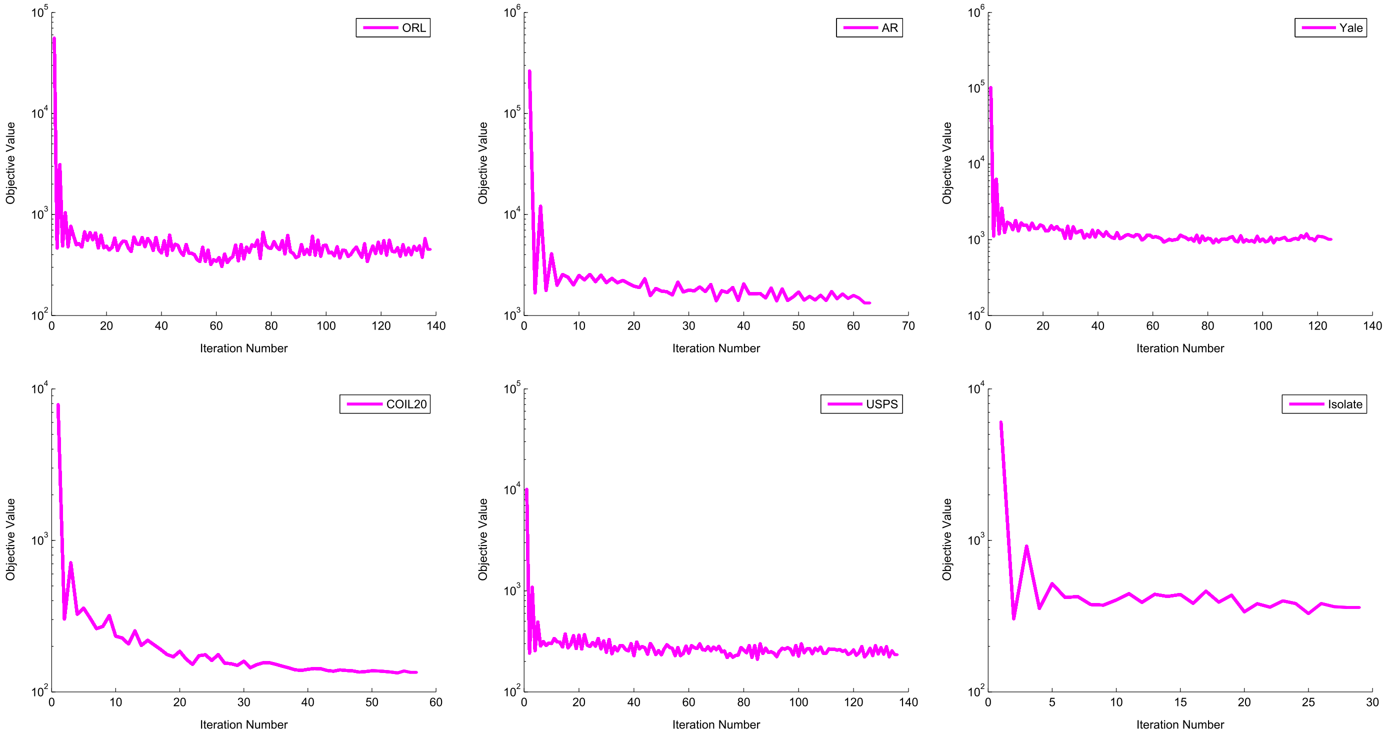
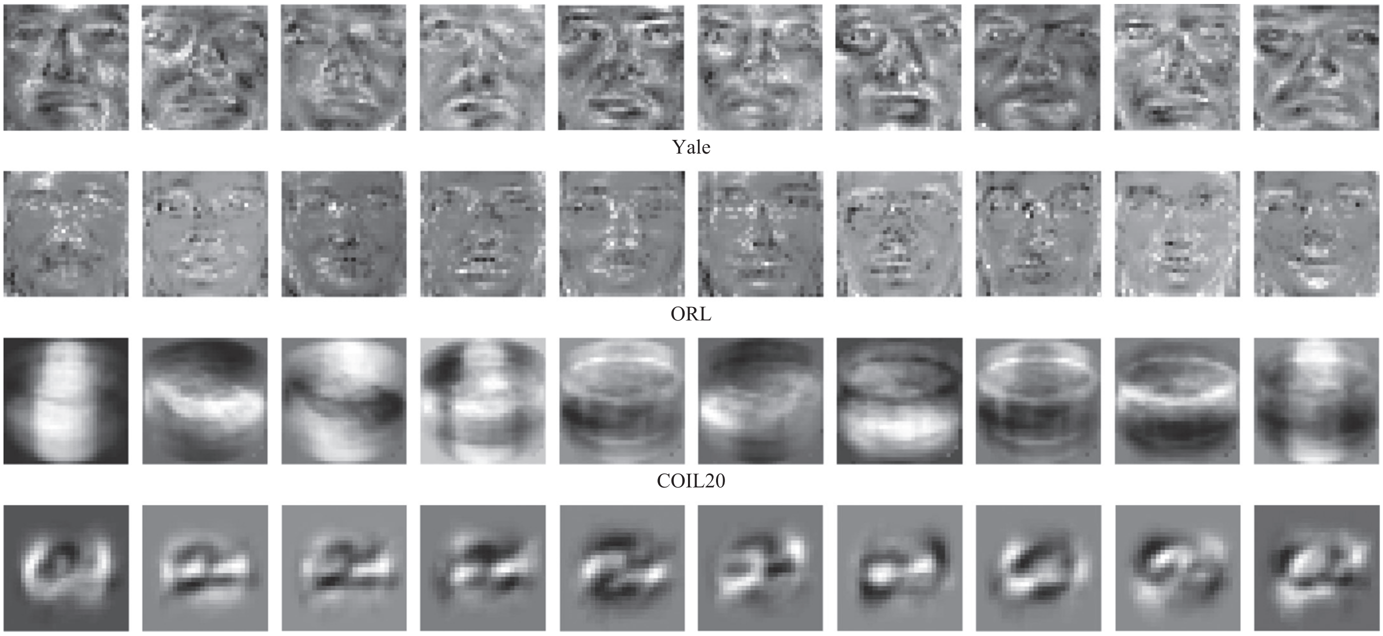


Fig. 7. Objective value versus iteration number on six data sets.



Yale

ORL

COIL20

MNIST

Fig. 8. Visualization of the first ten bases images of JSPCA on different data sets. Each basis image corresponds to a basis vector from projection matrix Q.

from per individual as training set and use the remaining as testing set. On the AR data set, we randomly select 8, 13, and 18 samples from per individual as training set and use the remaining as testing set. On the Yale data set, we randomly select 22, 27, and 32 samples from per individual as training set and use the remaining as testing set. When the optimal subspace dimensionality is fixed, the optimal transformation matrix with row-sparsity is learned and used for feature extraction. On each data set, each experiment is independently performed 20 times and the average classification accuracy with standard deviation are calculated and reported in Table 1. Figs. 3(a)–(c) show the variations of the classification accuracy with different subspace dimensionality. From Fig. 3(a), the experimental results show that JSPCA obtains the best classification accuracy on the 30th dimensionality. From Fig. 3(b), we can see that the classification accuracy of each method varies with different subspace dimensionality, in which 13 samples per individual are randomly selected as the training set and the remaining is used as testing set. The experimental results show that JSPCA can obtain the best classification accuracy compared with the other methods. From Fig. 3(c), the experimental results show that JSPCA can obtain the best classification accuracy compared with the other methods.

Table 4
Reconstruction error comparisons of six PCA methods on the training samples of COIL20 data set using different dimensions.

Methods	10	20	30	40	50	60
PCA	1.8442×10^5	1.4265×10^5	1.2485×10^5	1.0937×10^5	9.9127×10^4	9.1058×10^4
RPCA	2.0528×10^5	1.5905×10^5	1.3409×10^5	1.2248×10^5	1.0911×10^5	9.6751×10^4
OMRPCA	1.9450×10^5	1.4503×10^5	1.2750×10^5	1.1498×10^5	1.0348×10^5	9.6341×10^4
SPCA	4.5913×10^5	4.4378×10^5	4.3226×10^5	4.1953×10^5	4.1361×10^5	4.1428×10^5
SSL	4.5851×10^5	4.5392×10^5	4.5264×10^5	4.5236×10^5	4.5375×10^5	4.5386×10^5
JSPCA	2.4480×10^5	2.0410×10^5	1.8658×10^5	1.8138×10^5	1.7298×10^5	1.6691×10^5

Table 5
Reconstruction error comparisons of six PCA methods on the training samples of AR data set using different dimensions.

Methods	50	60	70	80	90	100
PCA	2.3924×10^6	2.2343×10^6	2.0741×10^6	1.9710×10^5	1.9196×10^6	1.8540×10^6
RPCA	2.5253×10^6	2.4010×10^6	2.2146×10^6	2.1242×10^6	2.1115×10^6	1.9867×10^6
OMRPCA	2.3622×10^6	2.2078×10^6	2.0536×10^6	1.9511×10^6	1.8988×10^6	1.8348×10^6
SPCA	8.6334×10^6	8.3803×10^6	8.1822×10^6	7.9305×10^6	7.7310×10^6	7.4437×10^6
SSL	9.8424×10^6	9.8421×10^6	9.8418×10^6	9.8413×10^6	9.8412×10^6	9.8410×10^6
JSPCA	3.5027×10^6	3.4986×10^6	3.5156×10^6	3.5730×10^6	3.5860×10^6	3.6118×10^6

In fact, it is difficult to get high classification accuracy on these facial data sets due to the different variations such as occlusions in AR, illuminations in Yale and pose variations in ORL. However, the experimental results from Table 1 and Figs. 3(a)–(c) show that JSPCA can obtain the better classification accuracy than other compared methods. This is because JSPCA uses the $\ell_{2,1}$-norm to constrain the projection matrix. In this way, the projection matrix with row-sparsity can indicate the importance degree of the features. On the other hand, JSPCA uses the $\ell_{2,1}$-norm to constrain the loss term, and the loss term can be gradually trending to the smaller value by D_1 (see Eq. (5)). Therefore, JSPCA is robust to the negative influence of the data set with complicated variations to some extent. The obtained projection matrix is intuitively displayed in the first row of Fig. 4.

5.2.2. Experiments on the non-facial data sets

On the COIL20 data set, we randomly select 4, 5, and 6 samples from per subject as training set and use the remaining as testing set. On the USPS data set, we randomly select 10, 15, and 20 samples from per subject as training set and use the remaining as testing set. On the Isolate data set, we randomly select 4, 5, and 6 samples from per subject as training set and use the remaining as testing set. When the optimal subspace dimensionality is fixed, the optimal transformation matrix with row-sparsity is learned and used for feature extraction. Each experiment is independently performed 20 times and the average classification accuracy with standard deviation are calculated and reported in Table 2. Figs. 3(d)–(f) show the variations of the classification accuracy with different subspace dimensionality. From Fig. 3(d), the experimental results show that JSPCA obtains the approximated classification accuracy with the other compared methods. From Fig. 3(e), the experimental results show that JSPCA obtains the approximated classification accuracy with the other compared methods. From Fig. 3(f), the experimental results show that SSL obtains the best classification accuracy and JSPCA gets an approximated result to SSL. To sum up, from Table 2 and Figs. 3(d)–(f), we can see that the experimental results obtained by JSPCA approximate to that of other compared methods. The obtained projection matrix on some non-facial data sets are intuitively displayed in the second row of Fig. 4.

5.3. *Experiments on the corrupted data sets*

Table 3 lists the experimental results with the optimal subspace dimensionality on the eight corrupted data sets where ORL face data set uses 5 samples per individual as the training set, AR face data set uses 13 samples per individual as the training set, Yale face data set uses 32 samples per individual as the training set, COIL20 data set uses 4 samples per subject as the training set, USPS data set uses 20 samples per subject as the training set, and Isolate data set uses 4 samples per subject as the training set. Fig. 5 shows the variation of the classification accuracy of each method versus the different dimensionality on the corrupted data set. Compared with Fig. 3, we can see that JSPCA not only outperforms the other compared methods on the three original facial data sets, but also outperforms them on the corrupted non-facial data sets. Moreover, JSPCA approximates the OMRPCA and SPCA on the corrupted non-facial data sets.

Although JSPCA performs well on the corrupted data sets compared to the other methods, JSPCA suffers from the influence of random corruptions inevitably due to its ability of feature selection. Therefore, JSPCA is more robust to the slight complicated variations in the original data sets rather than the added random corruptions on the original data sets.

5.4. *Parameter settings*

For the proposed optimization problem in Eq. (9), there are one parameter, i.e., λ. We first use the grid search method to search the optimal parameter λ in the scope of $[10^{-6}, 10^6]$, and then narrow the search scope to be $[0, 6.4]$. As can be seen, Fig. 6 shows the influence of different settings of λ on six data sets. On ORL data set, the impact is small and the more robust parameter range is $[0.05, 0.8]$. On AR data set, the best classification performance can be got when $\lambda = 0.05$ and the more robust parameter range is $[0, 0.8]$. On Yale data set, the best classification performance is got when $\lambda = 0.05$ and the more robust parameter range is $[0, 0.8]$. On COIL20 data set, the best classification performance corresponds to $\lambda = 0.2$ and the more robust parameter range is $[0, 0.4]$. On USPS and Isolate data sets, the best classification performance corresponds to $\lambda = 0.05$. We can see that JSPCA can achieve better classification performance over a reasonable range of λ, and is robust to the different settings of λ as long as the values are in the reasonable range. Overall, the better classification performance is usually achieved when λ is close to 0.05. However, when $\lambda \to 0$, the classification accuracy will decrease. This indicates that the regularization parameter λ is also important for JSPCA to achieve its best performance.

5.5. *Observations*

Based on the experimental results shown in the above subsections, we have the following observations and analyses:

(1) *Sparsity of JSPCA:* The regularization term of JSPCA is imposed by $\ell_{2,1}$-norm, which is defined to encourage the rows of the

projection matrix to be zero. Hence, the projection matrix Q can be used to indicate the significance of the features, which are intuitively displayed in Fig. 4. This shows that the regularization term of JSPCA imposed by $\ell_{2,1}$-norm can exclude the redundant features and improve the classification performance.

(2) *Convergence of JSPCA:* Theoretical analysis in Section 4.1 indicates that JSPCA is convergent. Fig. 7 shows the convergence curves of JSPCA on six data sets where the max iteration number is 140.

(3) *Bases images of JSPCA:* To further observe JSPCA, we give the bases images of JSPCA on four data sets (see Fig. 8). Specifically, for the facial data sets, the selected features are those important features such as eyes, nose, mouth, and facial contour. For those non-facial data sets, the selected features are the different contours of different subjects.

5.6. Discussion

In this paper, JSPCA is proposed to find representative features from the original high-dimensional space. The found representative features have been used for classification tasks. Although JSPCA outperforms the other PCA methods in most of classification experiments, a series of PCA methods including JSPCA achieve the low classification accuracy overall. This is because these PCA methods do not use class labels to extract discriminative features. Any dimensionality reduction method without using class labels does not always extract effective features for classification. In the future, our method would be extended to the supervised method to solve the skewed/imbalanced classification problem.

In order to further rich the proposed method, the found representative features are also used for reconstruction experiments as shown in Tables 4 and 5.

From these two tables, we can see that JSPCA achieves a better reconstruction than SPCA and SSL. This is because JSPCA is able to select the effective features for reconstruction while SPCA and SSL are not. Moreover, JSPCA has a worse reconstruction than PCA, RPCA, and OMRPCA. This is a matter of course because JSPCA inevitably suffers from the loss of some information.

6. Conclusion

In this paper, JSPCA is designed by relaxing the orthogonal constraint of transformation matrix Q, introducing another transformation matrix P and imposing joint $\ell_{2,1}$-norms on both loss term and regularization term. The proposed method has more freedom to jointly select the useful features for a low-dimensional representation and is robust to outliers. A simple yet effective algorithm is designed for the optimization problem. A series of theoretical analyses are discussed which reveal some intrinsic qualities of the proposed method. In essence, JSPCA is the weighted PCA with sparsity. Experiments on eight benchmark data sets show the feasibility and effectiveness of JSPCA compared to the original PCA and its variants.

Acknowledgments

This study was supported by the Shenzhen Research Council (Grant nos. JSGG20150331152017052, JCYJ20140819154343378, JCYJ20160406161948211), partially by the Faculty Research Grant of Hong Kong Baptist University (HKBU) (Project codes: FRG2/14-15/075 and FRG2/15-16/049), by the Knowledge Transfer Office of HKBU (Grant no. MPCF-005-2014/2015), by the National Natural Science Foundation of China (Grant nos. 61272366, 61272203, 61503317, 61672444, 61672183) and by the National Science and Technology Research and Development Program (Grant no. 2015BAK36B00).

References

[1] F. Nie, H. Huang, X. Cai, C.H. Ding, Efficient and robust feature selection via joint l2,1-norms minimization, in: Advances in Neural Information Processing Systems, 2010, pp. 1813–1821.
[2] Q. Gu, Z. Li, J. Han, Joint feature selection and subspace learning, in: International Joint Conference on Artificial Intelligence, 2011, pp. 1294–1299.
[3] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, A survey of multilinear subspace learning for tensor data, Pattern Recognit. 44 (7) (2011) 1540–1551.
[4] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, Z. Zhu, Robust face recognition via occlusion dictionary learning, Pattern Recognit. 47 (4) (2014) 1559–1572.
[5] J. Huang, X. You, Y. Yuan, F. Yang, L. Lin, Rotation invariant iris feature extraction using Gaussian Markov random fields with non-separable wavelet, Neurocomputing 73 (4) (2010) 883–894.
[6] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–591.
[7] F. De La Torre, M.J. Black, A framework for robust subspace learning, Int. J. Comput. Vis. 54 (1–3) (2003) 117–142.
[8] D. Skočaj, A. Leonardis, H. Bischof, Weighted and robust learning of subspace representations, Pattern Recognit. 40 (5) (2007) 1556–1569.
[9] D. Skocaj, A. Leonardis, Weighted and robust incremental method for subspace learning, in: IEEE International Conference on Computer Vision, 2003, pp. 1494–1501.
[10] P.N. Belhumeur, J.P. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[11] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[12] Z. Lai, W.K. Wong, Z. Jin, J. Yang, Y. Xu, Sparse approximation to the eigensubspace for discrimination, IEEE Trans. Neural Netw. Learn. Syst. 23 (12) (2012) 1948–1960.
[13] Y. Yang, D. Xu, F. Nie, S. Yan, Y. Zhuang, Image clustering using local discriminant models and global integration, IEEE Trans. Image Process. 19 (10) (2010) 2761–2773.
[14] Z. Fan, Y. Xu, D. Zhang, Local linear discriminant analysis framework using sample neighbors, IEEE Trans. Neural Netw. 22 (7) (2011) 1119–1132.
[15] A.M. Martínez, A.C. Kak, Pca versus lda, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233.
[16] F. Nie, J. Yuan, H. Huang, Optimal mean robust principal component analysis, in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 1062–1070.
[17] F. De la Torre, M.J. Black, Robust principal component analysis for computer vision, in: Eighth IEEE International Conference on Computer Vision, 2001, pp. 362–369.
[18] N. Kwak, Principal component analysis based on l1-norm maximization, IEEE Trans. Pattern Anal. Mach. Intell. 30 (9) (2008) 1672–1680.
[19] F. Nie, H. Huang, C. Ding, D. Luo, H. Wang, Robust principal component analysis with non-greedy l1-norm maximization, in: International Joint Conference on Artificial Intelligence, 2011, pp. 1433–1438.
[20] J.P. Brooks, J. Dulá, E.L. Boone, A pure l1-norm principal component analysis, Comput. Stat. Data Anal. 61 (2013) 83–98.
[21] V. Choulakian, L1-norm projection pursuit principal component analysis, Comput. Stat. Data Anal. 50 (6) (2006) 1441–1451.
[22] J. Gao, Robust l1 principal component analysis and its Bayesian variational inference, Neural Comput. 20 (2) (2008) 555–572.
[23] C. Ding, D. Zhou, X. He, H. Zha, R1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization, in: Proceedings of the 23rd International Conference on Machine learning, 2006, pp. 281–288.
[24] D. Kong, C. Ding, H. Huang, Robust nonnegative matrix factorization using l21-norm, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 673–682.
[25] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.
[26] X. Niyogi, Locality preserving projections, in: Neural Information Processing Systems, 2004, pp. 153–160.
[27] D. Cai, X. He, J. Han, Spectral regression: A unified approach for sparse subspace learning, in: Seventh IEEE International Conference on Data Mining, 2007, pp. 73–82.
[28] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[29] Z. Zheng, Sparse locality preserving embedding, in: 2nd International Congress on Image and Signal Processing, 2009, pp. 1–5.
[30] Z. Jiang, Z. Lin, L.S. Davis, Label consistent k-svd: learning a discriminative dictionary for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2651–2664.

[31] A.M. Martinez, The AR face database, CVC Technical Report 24.

[32] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, N. Yu, Non-negative low rank and sparse graph for semi-supervised learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2328–2335.

[33] K.-C. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.

[34] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process. 20 (5) (2011) 1327–1336.

[35] J.J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (5) (1994) 550–554.

[36] C. Hou, F. Nie, D. Yi, Y. Wu, Feature selection via joint embedding learning and sparse regression, in: International Joint Conference on Artificial Intelligence, 2011, pp. 1324–1329.

Shuangyan Yi received the M.S. degree in Mathematics Department from Harbin Institute of Technology Shenzhen Graduate School, China. She is currently pursing the Ph.D. degree in Computer Science and Technology at Harbin Institute of Technology Shenzhen Graduate School, China. Her current research interests include object tracking, pattern recognition and machine learning.

Zhihui Lai received the B.S degree in Mathematics from South China Normal University, M.S. degree from Jinan University, and the Ph.D. degree in Pattern Recognition and Intelligence System from Nanjing University of Science and Technology (NUST), China, in 2002, 2007 and 2011, respectively. He has been a Research Associate, Postdoctoral Fellow and Research Fellow at The Hong Kong Polytechnic University since 2010. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research. He serves as an Associate Editor on International Journal of Machine Learning and Cybernetics. For more information, the readers are referred to the website http://www.scholat.com/laizhihui.

Zhenyu He received his Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2007. He is currently an Associated Professor with the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. His research interests include sparse representation and its applications, deep learning and its applications, pattern recognition, image processing, and computer vision.

Yiu-ming Cheung is a Full Professor at the Department of Computer Science in Hong Kong Baptist University. He received Ph.D. degree at Department of Computer Science and Engineering from the Chinese University of Hong Kong. His current research interests focus on artificial intelligence, visual computing, and optimization. Prof. Cheung is the Founding and Past Chairman of Computational Intelligence Chapter of IEEE Hong Kong Section. Also, he is now serving as an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems, Knowledge and Information Systems, and International Journal of Pattern Recognition and Artificial Intelligence, among others. He is a Senior Member of IEEE and ACM. More details can be found at http://www.comp.hkbu.edu.hk/~ymc/.

Yang Liu received the B.S. and M.S. degrees in Automation from National University of Defense Technology, in 2004 and 2007, respectively. He received the Ph. D. degree in Computing from The Hong Kong Polytechnic University in 2011. Between 2011 and 2012, he was a Postdoctoral Research Associate in the Department of Statistics at Yale University. Dr. Liu is currently a Research Assistant Professor in the Department of Computer Science at Hong Kong Baptist University. His research interests include cognitive science, machine learning, applied mathematics, as well as their applications in brain modeling, high-dimensional data mining, visual content analysis, and music therapy.