Integrating Joint Feature Selection into Subspace Learning: A Formulation of 2DPCA for Outliers Robust Feature Selection

Imran Razzak, Raghib Abu Saris, Michael Blumenstein, Guandong Xu

***Abstract*—Since the principal component analysis and its variants are sensitive to outliers that affect their performance and applicability in real world, several variants have been proposed to improve the robustness. However, most of the existing methods are still sensitive to outliers and are unable to select useful features. To overcome the issue of sensitivity of PCA against outliers, in this paper, we introduce two-dimensional outliers-robust principal component analysis (ORPCA) by imposing the joint constraints on the objective function. ORPCA relaxes the orthogonal constraints and penalizes the regression coefficient, thus, it selects important features and ignores the same features that exist in other principal components. It is commonly known that square Frobenius norm is sensitive to outliers. To overcome this issue, we have devised an alternative way to derive objective function. Experimental results on four publicly available benchmark datasets show the effectiveness of joint feature selection and provide better performance as compared to state-of-the-art dimensionality-reduction methods.**

Keywords*—*PCA, 2DPCA, Outliers, Dimensionality Reduction, Principal Component Analysis

I. Introduction

With the recent advancement in data acquisition devices, acquiring data at faster rates and increased resolution has improved substantially over recent years. The data interpretation process, however, is facing several challenges due to high dimensionality. Not only for the classification, dimensionality reduction is also a serious challenge for several other domains such as data visualization, data compression, pattern recognition, and computer vision. The aim of dimensionality reduction is to transform the high-dimensional data into low-dimensional representation by preserving the quality of the data so that it could be classified efficiently. To deal with this issue, several vector-based methods are in use during the last two decades such as Principal Component Analysis (PCA) [22], Linear Discriminant Analysis (LDA) [1], [33], [15], LPP [4], SPP [13], SPPE [39], Isomap[35] and NPE [4]. Principal Component Analysis is one of the extensively used unsupervised dimensionality reduction method that projects high-dimensional representation into linear orthogonal space. However, one of the major drawbacks is that PCA is linear combination and loading are non-zero. This makes PCA data interpretation difficult, and it is still sensitive to outliers (as its covariance matrix is derived from ℓ_2-norm that affects its performance. Thus, it fails to deal with outliers that often appears in real-world data. Moreover, before applying PCA and LDA, there is need to convert the image into one-dimensional vector, thus it may not exploit image's spatial structural information very well [22], [30], [42], [3], [12], [34], [23], [4] which is very important for image representation. To overcome these issues, several variants of PCA have been proposed to improve the effectiveness of dimensionality reduction and robustness against outliers.

Matrix-based subspace learning methods have been widely applied for dimensionality reduction [31], [32], [7], [21], [6]. Results showed that 2DPCA [31], 2DLDA [32], multilinear PCA [8], and JGSPCA [5] are far more efficient as compared to one-dimensional subspace learning, due to its direct formulation based on two-dimensional images. Two-dimensional subspace learning methods directly calculate the class scatter metrics from images, hence can reveal the spatial structural information of image that is quite important for image classification task. To select important features, several efforts have been made such as robust 2DPCA, utilization of nuclear norm, ℓ_1, $\ell_{2,1}$, and Frobenius-norm that showed considerable improvement against outliers and able to select discriminant patterns.

Recently ℓ_1-norm-based subspace learning methods have shown great performance against outliers for tensor data classification [25], [24]. Ke and Kanade presented matrix factorization as an ℓ_1-norm minimization problem that is able to handle missing data straightforwardly. Wang et al. presented robust 2DPCA with non-greedy ℓ_1-norm maximization in which all projection directions are optimized simultaneously [27]. Luo et al. extended it by learning the optimization matrix by maximizing the sum of the projected difference between each pair of instances, rather than the difference between each instance and the mean of the data [9]. Although, ℓ_1-based methods provided great performance, these methods do not relate to covariance matrix which characterizes the geometric structure of the data, where as F-norm can exploit efficiently the spatial structure that is embedded in the data. Several efforts have been made to utilize F-norm as subspace learning such as 2DPCA [31], [32], 2D-PCA [21], F-norm 2DPCA [6], NM-2DPCA [2], [28], N-2DNPP [37]. However, either these methods still suffer from the effect of outliers or not able to select important features. Furthermore, sensitivity of F-norm is another challenge. Wang et al. presented non-squared F-norm minimization to overcome this challenge [28]. However, it affects the selection of important features.

To overcome the aforementioned issue of robust feature

Manuscript received xx; revised xxx.

selection and sensitivity of Frobenius norm, in this paper, we present a novel formulation for PCA that combines the subspace learning and feature selection together in order to exclude the effect of redundant patterns and joint feature selection. We employed Frobenius norm as distance metric learning and seek the projection matrix by joint minimization of regularizer and penalty terms. We relax the orthogonality constraints of transformation matrix and introduce another transformation that helps to jointly select important features and enhances the robustness against outliers. To overcome the sensitivity issue due to squared Frobenius norm, we devised an efficient way to compute F-Norm. As result, the proposed objective function not only weakens the effect of large distance but also has rotational invariance property. We can describe the theoretical and empirical **key contributions** of this work as follows:

- We present outliers robust two-dimensional principal component analysis by efficiently integrating the robustness of traditional 2DPCA and the regularization term $||Q||_F^2$ that relaxes the orthogonal constraint.
- The regularization term $||Q||_F^2$ reduces the constraints and enables the objective function to select features jointly. Furthermore, the regularization parameter $||Q||_F^2$ is convex and can be easily optimized.
- To overcome the sensitivity issue of F-Norm against outliers, we efficiently derived the objective function.
- Penalty term penalizes all regression coefficients corresponding to single feature as a whole to make PCA possible to select features jointly. Hence, ORPCA approximates high-dimensional representation in flexible manner. As such, ORPCA has more freedom to select low-dimensional features efficiently.
- The one major drawback of F-norm is its sensitivity against outliers as outlying measurement arbitrarily skew the solution from desired due to squared objective function. As a result, F-norm is not able to utilize the underlying geometric structure in a real sense. To cope the sensitivity due to squared F-norm, recently, non-square F-norm have been used.
- The latter method is evaluated empirically on four benchmark datasets. Experimental evaluation (discriminant features, computationally and convergence analysis) shows the considerable improvement in most cases, while time complexity remains very attractive.

The rest of the paper is organized as follows. In section II, we present basic notations and related work. In section III, we present the motivation followed by the proposed objective function and its optimization. In section V, we provide detailed experimental evaluations. Finally, conclusion is drawn in section VI.

II. Related Work

Recently, subspace-learning techniques have shown their great performance and have been widely applied for high-dimensional data representation and classification. In the recent few years, researchers proposed number of methods to reduce the effect of outliers, and several variants have been presented in literature. PCA is one of the most widely used dimension-reduction approach. Unlike traditional PCA, two-dimensional PCA is based on two-dimensional image matrices rather than one dimensional vectors. As as result, input image does not need to be converted into one-dimensional vector before feature extraction process. In the following discussion, we first give the basic notations, short description of 2DPCA, followed by review of several variants of 2DPCA.

Assume that $A_1, A_2..., A_N$ are a set of training images (mean centered) with size $m \times n$, where N is the number of images in the dataset. $V = [v_1, v_2, ...v_d] \in R^{n\times d}$ is the projection matrix, where v_1 is the first basis vector of two-dimensional PCA that maximizes the ℓ_1-norm-based dispersion of projected samples. In this paper, we denote $||X||_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q}|X_{i,j}|^2} = \sqrt{Tr(XX^H)}$ and $||X||_{2,1} = \sum_{j=1}^{q}||x_j||_2 = \sum_{j=}^{q}\sqrt{\sum_{i=1}^{p}|x_{i,j}|^2}$. The problem of linear dimensionality reduction is to project high-dimensional data to low dimensional space. The target is to find transformation matrix V with much lower dimensionality $(d << m)$.

$$V^* = \text{argmin}_{V^TV=I_d}\sum_{i=1}^{N}||A_i - A_iVV^T||_F^2$$

Where $\|\cdot\|_F$ denotes the Forbenius norm of matrix and is the sum of square of ℓ_2-norm of row/column vector of matrix. The above objective function is equivalent to the following objective function based on the fact $\sum_{n=1}^{N}||A_i - A_iVV^T||_F^2 + \sum_{n=1}^{N}||A_iV||_F^2 = \sum_{n=1}^{N}||A_i||_F^2$

$$V^* = \text{argmax}_{V^TV=I_d}\sum_{n=1}^{N}||A_iV||_F^2$$

where $tr(\cdot)$ is the trace function of matrix. As $V^* = \text{argmax}_{V^TV=I_d}\sum_{n=1}^{N}||A_iV||_F^2 = \text{argmax}_{V^TV=I_d}tr(\sum_{n=1}^{N}V^TA_i^TA_iV)$; if we let $S_t = \sum_{n=1}^{N}A_i^TA_i)$ denotes the covariance matrix, finding the orthogonal eigenvector of S_t corresponds to the first d largest eigenvalues. 2DPCA is sensitive to noise and outliers as optimal projection matrix of objective function mentioned above is not roubut in the sense that outlying measurement can skew the solution. To overcome this issue, 2DPCA-L1 was proposed which finds the basis vectors that maximizes the dispersion of the projected image in term of ℓ_1 norm.

$$V^* = \text{argmax}_{V^TV=I_d}\sum_{n=1}^{N}||A_iV||_{\ell_1}$$

subject to $\|V\|_{\ell_2} = 1$ where $\|\cdot\|\ell_1$ denotes the ℓ_1 norm and $\|\cdot\|\ell_2$ denotes the ℓ_2 norm of matrix. $\|D\|_{\ell_2} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|D(i,j)|^2}$. $D(i,j)$ denotes the i,j-th entry of matrix D, whereas $A_i(j,:)$ denotes the j-th row of A_i.

2DPCA based on $\ell_1 - norm$ is robust to outliers than 2DPCA. Computation of V is implemented by iterative method as: Basis vector $V(t+1)$ at the $(t+1)th$-step is updated based on the following

$$V(t+1)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{q}S_{ji}(t)a_{ji}}{\|\sum_{i=1}^{n}\sum_{j=1}^{q}S_{ji}(t)a_{ji}\|_{\ell_2}}$$

where S_{ji} is defined as sign$(V^T(t)a_{ji})$ and $A_i = \begin{bmatrix} a_{1,i}^T \\ a_{2,i}^T \\ \vdots \\ a_{n,i}^T \end{bmatrix}$

Since outlier does not have a precise mathematical meaning, the problem of robust PCA is still not well- defined. Several classical heuristics have been proposed to improve the robustness against outliers. Compared to the traditional PCA, ℓ_1 and $\ell_{2,1}$ based on matrix recovery based methods effectively improve the robustness of algorithms [29], [23], [38], [26]. Some work suggest that means, in the least squared sense, is not optimal of distance metrics such as ℓ_1, $\ell_{2,1}$ and nuclear norm [24], [27], [4], [36]. To improve their performance, simultaneously optimizing mean and projection matrix, the criterion function has been introduced [28]. Later, Song et al. presented robust PCA by simultaneously optimizing global mean and projection matrix [20]. Recently, a novel robust PCA (RPCA-AOM) is presented by by maximizing the sum of projected differences between each pair of data based on the ℓ_1-norm distance by avoiding the mean computation in solving the projection matrix [9]. However, RPCA-AOM does not well characterize the geometric structure of data and it is computationally expensive as well as difficult to solve the local optimal solution of RPCA-AOM.

Combination of nuclear norm with other ($\ell_1, \ell_{2,1}$) has shown great performance by providing sparse but also low-rank solution. Zhang combined nuclear norm and $\ell_{2,1}$-norm to extract neighborhood preserving features by minimizing reconstruction error due to Frobenius norm that is very sensitive to outliers [37], [36]. $\ell_{2,1}$ ensures the projection to be sparse in rows so that discriminative features are learned in the latent subspace whereas the nuclear-norm ensures the low-rank property by projecting data into their respective subspaces. The addition of nuclear norm with $\ell_{2,1}$-norm results not only sparse but also low-rank feature representation. Zhao et al. presented Local and global information (LLGDI) for effective semi-supervised dimensionality reduction [40]. LLGDI adopts a set of local classification functions in order to preserve local geometrical as well as discriminative information. Moreover, it also adopts global classification function that preserve the global discriminative information by solving the regression and dimensionality reduction simultaneously.

2DPCA and its variations cannot reveal the spatial structural information which is one of the core components in image representation [28], [16]. Moreover, features in low-dimensional subspace are linear combination of all features in high-dimensional space, thus, it usually consists of redundant features that affect the classification performance. However, it is quite difficult to interpret new feature set whereas it is quite important to extract new features especially when they have spatial meaning [14].

III. Motivation

As the aforementioned analysis in section I and section II, for the classification of high-dimensional noisy data, it is always important to find salient features that belong to specific part of image. Since the outlier does not have a precise mathematical meaning, thus the problem of RPCA problem is not well defined yet. Selection of important information by ignoring the redundant could help to improve the feature selection. However, most of the PCA-based methods are sensitive to outliers and select redundant features, thus are unable to select optimal feature set due to redundancy. Ignoring the features that already selected in other PCAs can help to encode further information that cannot be due to data redundancy. Furthermore, integrating feature selection into subspace learning could help to encode semantic information that helps to approximates high-dimensional data in a flexible way. Based on these above hypotheses, we have imposed the joint constraint on the objective and added a penalty term which helps to avoid redundant feature selection by avoiding selection of same features in different principal components, thus resulting partially sparse solution.

Sensitivity of F-norm is another challenge as the outlying measurement arbitrarily skew the solution from desired due to squared objective function. To overcome this issue, we have devised an alternative approach to derive objective function. Compared with traditional PCA-based on Frobenius norm, ORPCA not only selects featured jointly, but also weakens the effect of large distance and has rotational invariance properties.

IV. Outliers Robust 2DPCA

In this section, we present outliers robust dimensionality reduction approach (ORPCA) in detail. As described in earlier sections, the projection procedure consists of all the original features, thus, it may also have irrelevant and redundant features which could influence the performance of dimensionality reduction, in result affecting the classification performance. Furthermore, outliers strongly affect the feature selection which depresses the classification performance. In this work, we present a novel formulation for PCA that combines the subspace learning and feature selection together in order to exclude the effect of redundant patterns and joint feature selection. We employed Frobenius norm as distance metric learning and seeks the projection matrix by joint minimization of regularizer and penalty terms. We relax the orthogonal constraints of transformation matrix and introduce another transformation that helps to jointly select important features and discard the features that are already selected in other principal components. To overcome the sensitivity issue due to squared Frobenius norm, we devised an efficient way to compute F-Norm, as a result, ORPCA has more freedom to select robust features jointly for low dimensional representation that helps to minimizes the affect of outliers as well as redundancy. However, it does not guarantee fully sparse solution but it (joint feature selection and alternative derivation of objective function) make the objective function robust against outliers.

Considering the appearance of outliers in the input data, we propose the following objective function

$$\min_{P,Q} J(P,Q) = \min_{P,Q} \sum_{j=1}^{N} \left\| A_j - A_j Q P^T \right\|_F^2 + \lambda \|Q\|_F^2 \quad (1)$$

where $P, Q \in R^{n\times d}$. Matrix Q is used to transform each sub-image into low-dimensional subspace and matrix P is used to recover the matrix A such that $A = [A_1, ..., A_N]$, where $A_j \in R^{m\times n}$. Furthermore, while we require the matrix P to be orthogonal ($P^T P = I_d$), we do not require the orthogonality of the matrix Q, thus ORPCA has more freedom to learn low-dimensional space. In addition, the regularization parameter $\|Q\|_F^2$ reduces the constraints and enables the ORPCA to select important features jointly. The penalty term penalizes the regression coefficient to make PCA possible to select features jointly and discard those features that have already been selected in other principal components. Moreover, regularization term $\|Q\|_F^2$ is convex that can be easily optimized. The parameter $\lambda \geq 0$ balances the loss and regularization terms. In short, we relaxed the orthogonal constraint of transformation matrix Q, introduce another transformation matrix P, and added an additional regularization term $\|Q\|_F^2$ to make the objective function robust and able to select features jointly.

A. Optimization

Squared F-norm is not robust in the sense that outlying measurements can arbitrarily skew the solution from the desired. We devised an efficient way to compute F-Norm to overcome its sensitivity challenge. Although the objective function is shown in Eq 1 is based on square F-norm, however, computation of P and Q are not squared. Compared with squared F-norm, the proposed derivation can weaken the effect of large distance but also has rotational invariance. ORPCA sees the projection matrix that makes the value of objective function small. The objective function has two main unknown terms P and Q. The following two theorems play a key role in determining the minimizers of the optimization problem 1.

***Theorem** 1:* The minimizers of the objective function given in the Equation 1 satisfy the following equation

$$Q = \left[\sum_{j=1}^{N} \left(\lambda I_n + A_j^T A_j \right) \right]^{-1} \left[\sum_{j=1}^{N} A_j^T A_j \right] P \quad (2)$$

Proof: According to the definition of Frobenius norm, the linearity and cyclic properties of trace function, and orthogonality of matrix P, the above objective function can be written in a more computationally traceable way as

$$J(P,Q) = \sum_{j=1}^{N} \left\| A_j - A_j Q P^T \right\|_F^2 + \lambda \|Q\|_F^2 \quad (3)$$

$$= \sum_{j=1}^{N} \text{tr} \left[\left(A_j^T - P Q^T A_j^T \right) \left(A_j - A_j Q P^T \right) \right] + \lambda \text{tr} \left(Q^T Q \right) \quad (4)$$

$$= \sum_{j=1}^{N} \text{tr} \left(A_j^T A_j - A_j^T A_j Q P^T - P Q^T A_j^T A_j + P Q^T A_j^T A_j Q P^T + \lambda \text{tr} \left(Q^T Q \right) \quad (5)$$

$$= \sum_{j=1}^{N} \text{tr} \left(A_j^T A_j - 2 A_j^T A_j Q P^T + A_j^T A_j Q Q^T \right) + \lambda \text{tr} \left(Q^T Q \right) \quad (6)$$

Now, differentiation Eq (6),

$$\frac{\partial J}{\partial Q} = \sum_{j=1}^{N} \left(-2 A_j^T A_j P + 2 A_j^T A_j Q \right) + 2\lambda Q. \quad (7)$$

Therefore,

$$\frac{\partial J}{\partial Q} = 0 \Rightarrow \sum_{j=1}^{N} \left(-2 A_j^T A_j P + 2 A_j^T A_j Q \right) + 2\lambda Q = 0 \quad (8)$$

Simplifying the above equation, we get

$$\sum_{j=1}^{N} (A_j^T A_j P) = \sum_{j=1}^{N} (A_j^T A_j Q) + \lambda Q \quad (9)$$

The above equation can be rewritten as

$$\left(\sum_{j=1}^{N} A_j^T A_j \right) P = \left(\sum_{j=1}^{N} (\lambda I_n + A_j^T A_j) \right) Q \quad (10)$$

Hence, we can write

$$P = \left(\sum_{j=1}^{N} (\lambda I_n + A_j^T A_j) \right) Q \left(\sum_{j=1}^{N} A_j^T A_j \right)^{-1} \quad (11)$$

■

Once matrix Q is known, we can optimize matrix P with respect to matirx Q.

***Theorem** 2:* If UDV^T is the singular value decomposition (SVD) of $\sum_{j=1}^{N} A_j^T A_j Q$, then

$$P = U I_{n\times d} V^T \quad (12)$$

is orthogonal and minimizes the Eq. (6) for a given matrix Q.

Proof: As we know that the matrices V and U are orthogonal matrices of sizes $d \times d$ and $n \times n$, respectively. As such,

$$P^T P = V I_{n\times d}^T U^T U I_{n\times d} V^T = I_d$$

■

The orthogonal constraint on matrix P reduces the feature redundancy and forces the objective function to be small. Below in table I, we describe an iterative algorithm of ORPCA for training samples $A_1, ..., A_n$ of size $m \times n$, and regularization parameter λ.

B. Convergence Analysis

First, we provide to the following lemma

***Lemma** 3:* For any nonzero matrix $P, Q \in R^{n \times d}$, the following results hold:

$$\|P\|_F - \frac{\|P\|_F^2}{2\|Q\|_F} \leq \|Q\|_F - \frac{\|Q\|_F^2}{2\|Q\|_F} \qquad (13)$$

Proof: We start with an obvious inequality $(\sqrt{S} - \sqrt{S_t})^2 \geq 0$, we have

$$\begin{aligned} &(\sqrt{S} - \sqrt{S_t})^2 \geq 0 \\ &\Rightarrow S - 2\sqrt{SS_t} + S_t \geq 0 \\ &\Rightarrow \sqrt{S} - \frac{S}{2\sqrt{S_t}} \leq \frac{1}{2} S_t \\ &\Rightarrow \sqrt{S} - \frac{S}{2\sqrt{S_t}} \leq \sqrt{S_t} - \frac{S_t}{2\sqrt{S_t}} \end{aligned}$$

Now substituting S and S_t by $||P||_F$ and $||Q||_F$ respectively, we arrive at Eq. 13. ■

Based on the above lemma 3, we provide the following convergence theorem.

***Theorem** 4:* Given all the variables in objective function equation 1, the iterative scheme of proposed ORPCA described in table 1 shows that objective function value is monotonically decreasing thus converges to local optima.

Proof: For given initial value of matrix P, say P_0, we can compute the matrix Q_0 by minimizing the objective function $J(P_0, Q)$. Consequently,

$$J(P_0, Q_0) \leq J(P_0, Q)$$

We can calculate matrix P_1 by minimizing the objective function $J(P, Q_0)$. Hence,

$$J(P_1, Q_0) \leq J(P_0, Q_0)$$

Since the matrix Q_1 minimizes the objective function $J(P_1, Q)$, we have

$$J(P_1, Q_1) \leq J(P_1, Q_0) \leq J(P_0, Q_0).$$

That is

$$\begin{aligned} & \sum_{j=1}^{N} \left\| A_j - A_j Q_1 P_1^T \right\|_F^2 + \lambda \|Q_1\|_F^2 \\ \leq \quad & \sum_{j=1}^{N} \left\| A_j - A_j Q_0 P_1^T \right\|_F^2 + \lambda \|Q_0\|_F^2 \\ \leq \quad & \sum_{j=1}^{N} \left\| A_j - A_j Q_0 P_0^T \right\|_F^2 + \lambda \|Q_0\|_F^2 \end{aligned}$$

Iteratively, we obtain

$$J(P_{t+1}, Q_{t+1}) \leq J(P_t, Q_t) \quad for\ t = 0, 1, 2, ...$$

Since the singular value decomposition (SVD) provides optimal P_t which decreases the value of objective function further. In other-words, the algorithms attains the optimal solution of the objective function in each iteration. Once, we compute the optimal value of matrix Q and P, in the following iteration, the matrix P_t converges to local optima. Moreover, the objective function is convex. The sequence $J(P_t, Q_t)$ is monotonically decreasing in each iteration. Thus, by the Monotonic Convergence Theorem, the objective function $J(P_t, Q_t)$ converges to a local optimal value.

$$\sum_{j=1}^{N} \text{tr} \left[\left(A_j^T - P_\infty Q_\infty^T A_j^T \right) \left(A_j - A_j Q_\infty P_\infty^T \right) \right] + \lambda \text{tr} \left(Q_\infty^T Q_\infty \right)$$

■

C. Numerical Algorithm

Below in table I, we describe an iterative algorithm of ORPCA for training samples $A_1, ..., A_n$ of size $m \times n$, and regularization parameter λ.

TABLE I. ALGORITHMIC PROCEDURE OF ORPCA

```
Input: $A_j \in R^{m \times n}$ for $j = 1, ..., N$ where $A$ is centralized, and parameter $\lambda$.
Output: Matrix $P$ and Matrix $Q$

Step-I: Randomly initialize the matrix $P$
While do not converge do
    Step-II: Minimize the objective function with respect to matrix $Q$ by finding the matrix $Q$ using Eq.(2)
    Step-III: Compute the Singular Value Decomposition of $\sum_{j=1}^{N} A_j^T A_j Q$
    Step-IV: Update the matrix $P$ using Eq.(12) to minimize the objective function with respect to matrix $P$
end while
```

D. Connections to Other PCA algorithm

In the following discussion, we analyze the relations between our model and PCA based on $\ell_{2,1}$ norm to show its elegant properties (joint feature selection) over other methods. We further show that the traditional 2DPCA is a special case of ORPCA.

As discussed in earlier section, square Frobenius norm is not robust as outlying measurements can arbitrary skew the solution from desired solution. We devised an alternative approach to solve the objective function, thus outliers have less importance than the squared residual. Furthermore, objective function has a rotational invariance property while the ℓ_1-norm loss function does not have such desirable property. The challenges Frobenius norm has are no feature selection capability and sensitivity against outliers due to square and non-sparse output. We have solved these issues through joint feature selection and additional regularization term.

Below, theorem 5 validates our claim that proposed objective function provides robust and stable solution as compared to PCA-based methods based on $\ell_{2,1}$ norm.

***Theorem** 5:* If A is an $m \times n$ matrix, then $\|A\|_F \leq \|A\|_{2,1}$.

Proof: Recall that

$$\|A\|_F = \sqrt{\text{tr}\,(A^T A)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$$

$$\|A\|_{2,1} = \sum_{j=1}^{n} \|\mathbf{a}_j\|_2 = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} a_{i,j}^2}$$

where $\mathbf{a}_j$ is the jth column of A. With that in mind,

$$\|A\|_{2,1}^2 = \sum_{j=1}^{n} \|\mathbf{a}_j\|_2^2 + \underbrace{2\sum_{r=1}^{n}\sum_{s=1,s\neq r}^{n} \|\mathbf{a}_r\|_2\ \|\mathbf{a}_s\|_2}_{\text{nonegative term}}$$

$$\geq \sum_{j=1}^{n}\sum_{i=1}^{m} a_{i,j}^2 = \|A\|_F^2$$

■

From Theorem 5, we can deduce that

$$arg\min_{Q,P} \|X - PQ^T X\|_{2,1} + \lambda\|Q\|_{2,1} \geq arg\min_{Q,P} \|X - PQ^T X\|_F + \lambda\|Q\|_F \quad (14)$$

The above Eq. 14 shows that the objective function is robust and provide stable solution as compared to $\ell_{2,1}$. In other words, ℓ_1 and $\ell_{2,1}$ and square Frobenius norm penalizes the coefficients more than non squared Frobenius norm.

The additional penalty term, introduced in the objective function, excludes redundant features and provides robustness against outliers, i.e., the regularization parameter $\|Q\|_F^2$ reduces the constraints and enables our method to jointly select features. The following remark 1 shows that objective function penalizes all regression coefficients corresponding to single feature as a whole making it possible to discard the redundant features.

***Remark** 1:* Notice that, if regression coefficient $\lambda = 0$, then $Q = P$.

$$\left(\sum_{j=1}^{N} A_j^T A_j\right) P = \left(\sum_{j=1}^{N} A_j^T A_j\right) Q$$

$$Q = \left[\sum_{j=1}^{N} A_j^T A_j\right]^{-1} \left[\sum_{j=1}^{N} A_j^T A_j\right] P = P.$$

Moreover, the equation 11 simplifies to

$$J(Q,P) = \sum_{j=1}^{N} \left\|A_j - A_j P P^T\right\|_F^2$$

Hence, we can say that the proposed objection function degenerates to traditional 2DPCA. As such, the proposed objective function generalizes the 2DPCA. In this case, the optimal solution in Eq. 1 aim to find robust feature matrix.

V. Experimental Results

In order to evaluate the performance of the proposed ORPCA, in this section, we have discussed and compared the performance of proposed ORPCA on four commonly used image datasets including AR [11], Yale B [18] , ORL and CMU PIE. We have used k-nearest neighbour (where $k = 1$) for classification. The main contribution of this work is introducing joint feature selection in order to select useful features by effectively combining the robustness of traditional two-dimensional principal component analysis and the lasso regularization. Furthermore, we have introduced a penalty term in the objective function to exclude redundant features and provide robustness against outliers. Thus, to validate the our claims against outliers, we have corrupted each dataset with outliers to visualize the robustness of proposed approach in the presence of outliers. In addition, since 2D-RPCA is unsupervised method, we only compare its performance with unsupervised methods including PCA, 2DPCA, PCAℓ_1, 2DPCA-ℓ_1, OMF-2DPCA, and F-2DPCA on contaminated and non-contaminated benchmark datasets.

To validate the performance of of dimensionality reduction both persuasively and objectively, we have conducted several experiments on both original (non-contaminated) dataset and contaminated datasets. We have performed several of ORPCA at different λ value ($0 < \lambda < 1$ to find optimal λ. Once we have optimal value of λ, we have performed 10-fold validation.

A. Datasets

AR face dataset consists of 120 individual, 26 images per individual taken in two session, with total images 3120 [10]. The dataset was captured in two different session at different lightning condition and variable expressions. Face portion is cropped from their main images and then normalized to 32x32. Moreover, AR dataset consists of few images that are occluded with sunglasses, scarf or towels as shown in figure 1. In this experiment, we have considered face images with occlusion considered as noise images. Yale dataset consists of 64 images(except few 11-17,59-63), per subject with in total 2414 images under different lightning conditions from 38 individuals whereas half the dataset is corrupted by reflection or shadow. Figure 1 shows some reference of of Yale B dataset [41]. The database contains 5850 single light source images of 10 subjects (9 poses x 64 illumination conditions). For every subject in a particular pose, an image with ambient (background) illumination was also captured. ORL is face datset of 40 individuals with 10 images of each individual [17]. It consists of frontal views of faces with different expression and lightning conditions. CMU PIE dataset consists of 2856 frontal face images of 68 individual, 42 image per individual (with variation in lighting condition. We have selected 26 images randomly for training that consist of 7 noisy images [19].

We have resized the images in each dataset to 32×32 pixel. For training and evaluation purpose on non-contaminated datasets, we have divided 70%/30% and 80%/20% into training/testing. In order to validate the robustness of proposed

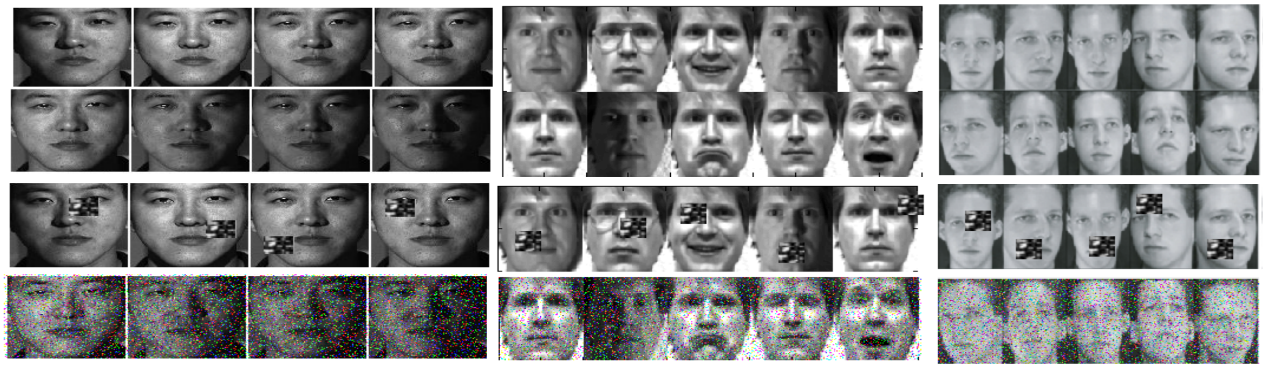

Fig. 1. Sample images of CMU PIE, Yale and ORL. First two rows real dataset, Row 3 contaminated with block and Row 4 is contaminated with salt and paper noise 15%

TABLE II. DATASET DESCRIPTION

Database	Image Size	Subject	Image per Subject	Number of Images	Detail of Face Images
[0.5ex] AR	576 × 768	126	26	3276	All frontal views of: neutral expression, smile, anger, scream, left light on, right light on, all sides lights on, wearing sun glasses, wearing sun glassses and left light on, wearing sun glasses and right light on, wearing scarf, wearing scarf and left light on, wearing scarf and right light on; second sessions repeated same conditions.
CMU PIE	640 × 486	337		750,000	15 view points and 19 illumination conditions while displaying a range of facial expressions.
Yale	640 × 480	28	576	16128	9 poses and 64 illumination conditions.
ORL	92 × 112	40	10	400	All frontal and slight tilt of the head

method against outliers, 20% images have been selected randomly and various types of noise (i.e. block occlusions, salt and peeper etc). We have added random noise (salt and peeper) with intensity of 10%, 15% on randomly selected images in each dataset as shown in figure 1. Similarly, we have added block occlusion of variable sizes at random locations with variable size $(5 \times 5,\ 10 \times 10,\ 10 \times 15)$ as shown in figure 1. In order to evaluate performance of the proposed ORPCA on corrupted datasets, we have randomly selected 60% and 70% and 80% samples for each subject form each dataset as training set.

B. Parameter Selection

The objective function in equation 1 has only one parameter λ required to be optimal. λ controls the regression coefficient. The greater value of λ could result in heavy penalty on regression coefficient that could affect the structural information, similarly smaller value of λ leads to 2DPCA. In order to find optimal range of regression coefficient λ, we have performed several experiments on each dataset. Initially, we have selected λ value $0 \leq \lambda \leq 4$ and then narrow down its range after few experiments based on its convergence and better accuracy. We have noticed that λ provided better performance between 0.15 to 0.25 for original datasets and between 0.1 to 0.3 for corrupted datasets. ORPCA achieved better performance over reasonable range of λ. The value of λ marginally varies for different datasets, however, it provided best accuracy on interval [0.1,0.3], ideally when λ is close to 0.2. We have also noticed that accuracy was reduced when $\lambda = 0$ or $\lambda \rightarrow 0$ which validates our claim made in earlier section, 2DPCA is a special case of ORPCA. Results showed that accuracy of ORPCA is exactly the same as 2DPCA when $\lambda = 0$. Thus, we can conclude that the optimal value of λ is very crucial to achieve better robustness. Table 2 and Table 3 show that ORPCA achieved better accuracy over reasonable range of λ and robust to different setting of λ as long as it is in the range mentioned above. After selection of range of optimal λ generically, we performed experiment for each dataset to find optimal λ explicitly for that dataset.

The objective function in equation 1 has only one parameter λ required to be optimal. λ controls the regression coefficient. The greater value of λ could result in heavy penalty on regression coefficient that could affect the structural information, similarly smaller value of λ leads to 2DPCA. In order to find optimal λ, we have performed several experiments with different λ value with $0 \leq \lambda \leq 4$ and narrow down its range after few experiments based on its convergence and better accuracy. Firstly, we evaluated on difference of 0.5 to find optimal interval where it provided better result followed by several experiments in selected interval. We have noticed that λ provided good accuracy between 0.15 to 0.25 for original datasets whereas it provided good accuracy between 0.1 to 0.3 for corrupted datasets. ORPCA achieved better performance over reasonable range of λ. The value of λ marginally varies for different datasets, however, it provided best accuracy on interval [0.1,0.3], ideally when λ is close to 0.2. We have also noticed that accuracy was reduced when λ=0 or $\lambda \rightarrow 0$. Furthermore, as claimed in earlier section, ORPCA is a special case of 2DPCA, accuracy of ORPCA is same as 2DPCA when $\lambda = 0$ which validates the claim ”ORPCA is a special

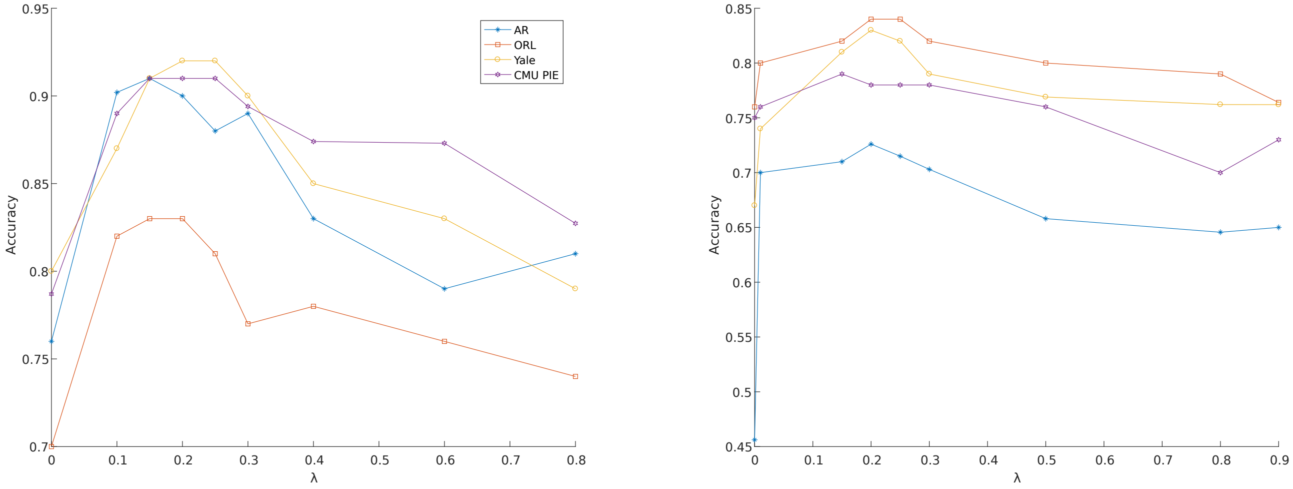

Fig. 2. Classification performance at different value of λ for real (left) and contaminated (right) datasets

TABLE III. AVERAGE CLASSIFICATION ACCURACY (ACCURACY ±CORRESPONDING STANDARD DEVIATION) ON REAL DATASET AT OPTIMAL RESULT OF ORPCA

Dataset	PCA	RPCA	2DPCA	PCA2Dℓ_1	OMF-2DPCA	F-2DPCA	ORPCA [Proposed]
AR	0.6832 ± 0.005	0.6459 ± 0.008	0.7589 ± 0.0071	0.8477 ± 0.0023	0.8577 ± 0.0011	0.8782 ± 0.021	0.8932 ± 0.003
ORL	0.7891 ± 0.0028	0.8009 ± 0.0091	0.8843 ± 0.0411	0.8637 ± 0.0071	0.8623± 0.019	0.8754± 0.023	0.9254 ± 0.0091
Yale B	0.6886 ± 0.0031	0.5976 ± 0.0061	0.7911 ± 0.0091	0.7305 ± 0.0071	0.6743 ± 0.021	0.6643 ± 0.019	0.6934 ± 0.0131
CMU PIE	0.7445 ± 0.0091	0.7666 ± 0.0027	0.8987 ± 0.0026	0.8607 ± 0.0015	0.8608 ± 0.018	0.8522 ± 0.025	0.8947 ± 0.0041

TABLE IV. COMPARATIVE EVALUATION BASED ON AVERAGE CLASSIFICATION ACCURACY ((ACCURACY ±CORRESPONDING STANDARD DEVIATION)) ON CONTAMINATED DATASETS AT OPTIMAL RESULT OF ORPCA

Dataset	PCA	RPCA	2DPCA	PCA2Dℓ_1	OMF-2DPCA	F-2DPCA	ORPCA
AR	0.5741 ± 0.0023	0.5387 ± 0.0022	0.6576 ± 0.0049	0.6277 ± 0.0053	0.781 ± 0.019	0.773 ± 0.021	0.8121 ± 0.014
ORL	0.6385 ± 0.0012	0.7411 ± 0.00321	0.8161 ± 0.0094	0.838 ± 0.0021	0.832 ± 0.016	0.856 ± 0.019	0.8892 ± 0.013
Yale B	0.5153 ± 0.0034	0.4865 ± 0.0083	0.5983 ± 0.0043	0.621 ± 0.0091	0.8109± 0.0031	0.80 ± 0.0017	0.82892 ± 0.0071
CMU PIE	0.577 ± 0.0032	0.5981 ± 0.0007	0.7181 ± 0.0091	0.6886 ± 0.0083	0.836± 0.021	0.8221± 0.012	0.8513 ± 0.008

case of 2DPCA, it degenerates to 2DPCA when $\lambda = 0$". Moreover, it indicates that λ is very important to achieve better robustness. Table 2 and Table 3 show that ORPCA achieved better accuracy over reasonable range of λ and robust to different setting of λ as long as it is in the range mentioned above. After selection of range of optimal λ generically, we performed experiment for each dataset to find optimal λ explicitly for that datasets.

C. Evaluation on Original Dataset

In order to compare the performance of proposed objective function both persuasively and objectively, the classification is performed based on nearest neighbour. We have performed 10 fold validation on each dataset. We performed several experiments with variable sample size per individual i.e 60% and 70% and 80% samples for each individual subject and rest of samples are used for validation. The classification performance with different subspace dimensionality at optimal value of $\lambda = 0.18$ is shown in Table 2. Notice that, due to the dataset complexity (variations, pose, illumination and occlusions), getting high accuracy is quite challenging. Table 2 shows that proposed ORPCA achieved better classification in comparison to state-of-the-art methods as shown in table III, IV. Furthermore, we have notice that ORPCA selected important features that plays important role in classification.

D. Evaluation on Corrupted Dataset

In order to validate the robustness of proposed ORPCA against outliers and joint selection of features, we corrupted the dataset with outliers. In this experiment, we have randomly selected 70% of images for corrupted datasets as a training set and consider rest of the images as a validation datasets. We have performed several experiments with different subspace dimensionality. Experimental results showed that the proposed ORPCA achieved much better performance as compared to state-of-the-art methods in the presence of outliers that validate the robustness of proposed approach against outliers. Notice that ORPCA performed well for corrupted data however, it partially sufer from random corruption due to its joint feature selection ability.

E. Computational Complexity

Computation complexity of ORPCA has 3 steps in each iteration. First step is to compute Q using equation

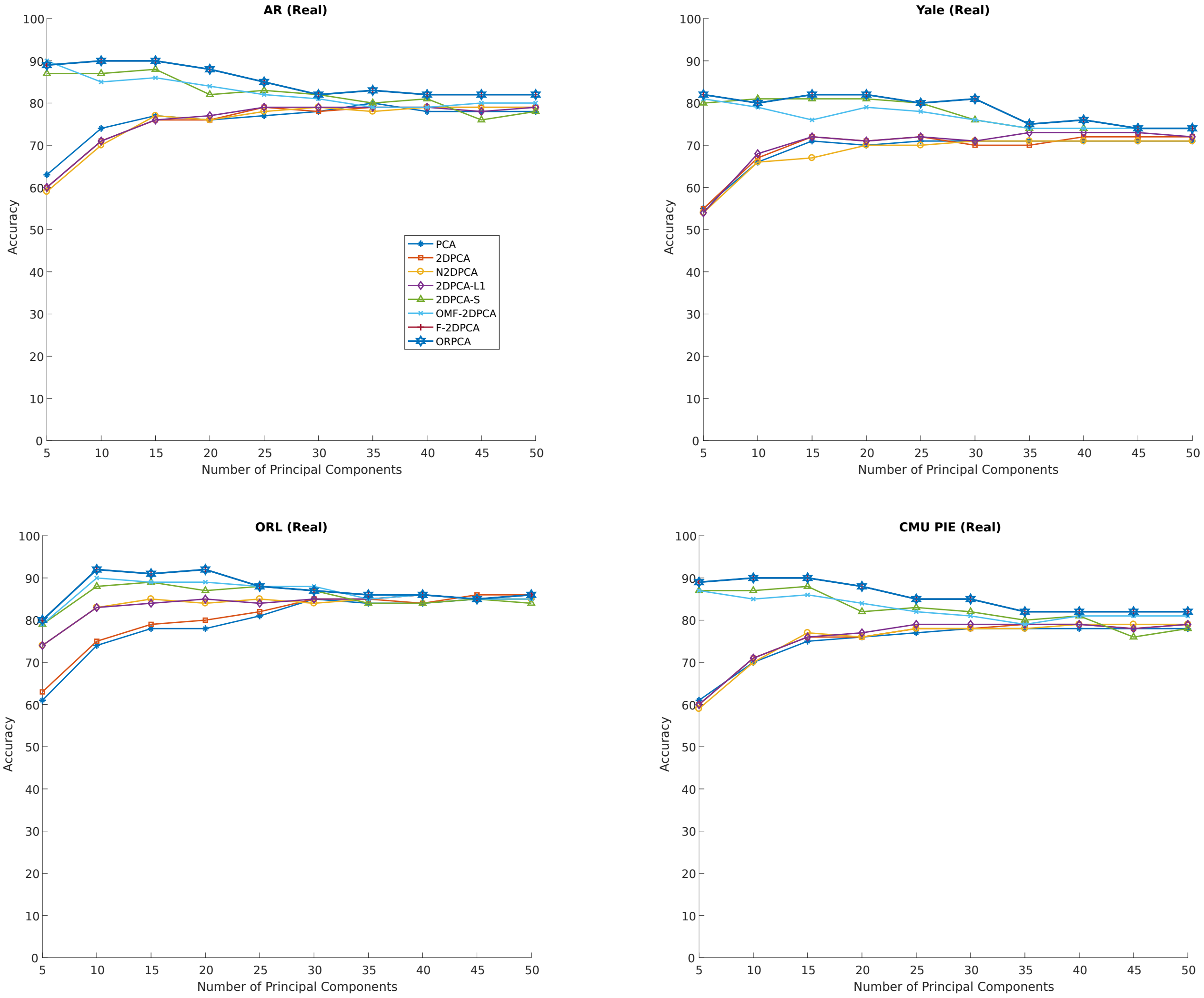

Fig. 3. Comparative evaluation on real dataset (AR, Yale, ORL, and CMUIPIE

$Q = \left[\sum_{j=1}^{N}\left(\lambda I_n + A_j^T A_j\right)\right]^{-1}\left[\sum_{j=1}^{N} A_j^T A_j\right] P$. Computational complexity of Q is $O(n^3)$ as $A_j^T A_j$ is the core step in computation of matrix Q. The second step is to compute the SVD of $\sum_{j=1}^{N} A_j^T A_j Q$; its computational complexity is also $O(n^3)$. Third step is to computation $P = U I_{n \times d} V^T$. Computation complexity of P is also $O(n^3)$. Thus, computational complexity of one iteration is $O(n^3)$. If the algorithm need t iteration to converge, it computation complexity will be $O(tn^3)$.

F. Convergence Verification

To verify the convergence of algorithm I, we tested different variations of parameters on all four datasets. The convergence of proposed ORPCA is shown in figure 5. It shows the convergence of objective function 1 along with each iteration. It can be found that objective function is non-decreasing functions of iterations. As theorem 4 proves that ORPCA converges to local optima so does the case in figure 5 that shows that algorithm converges to local optima.

VI. Discussion

We notice that methods based on matrix perform better as compared to vector-based methods. Results show that proposed ORPCA finds the representative features from high-dimensional space that are used for classification. Unlike 2DPCA based on ℓ_1-norm, ORPCA has rotational invariance property and has the freedom to jointly select the important and contributive features such as nose, eyes, lips in case of face image, while contours of different objects in non-facial datasets. Traditional methods are not able to interpret new features whereas it is quite important to interpret new features especially when they have spatial meaning. Results showed that ORPCA outperforms other PCA-based methods especially

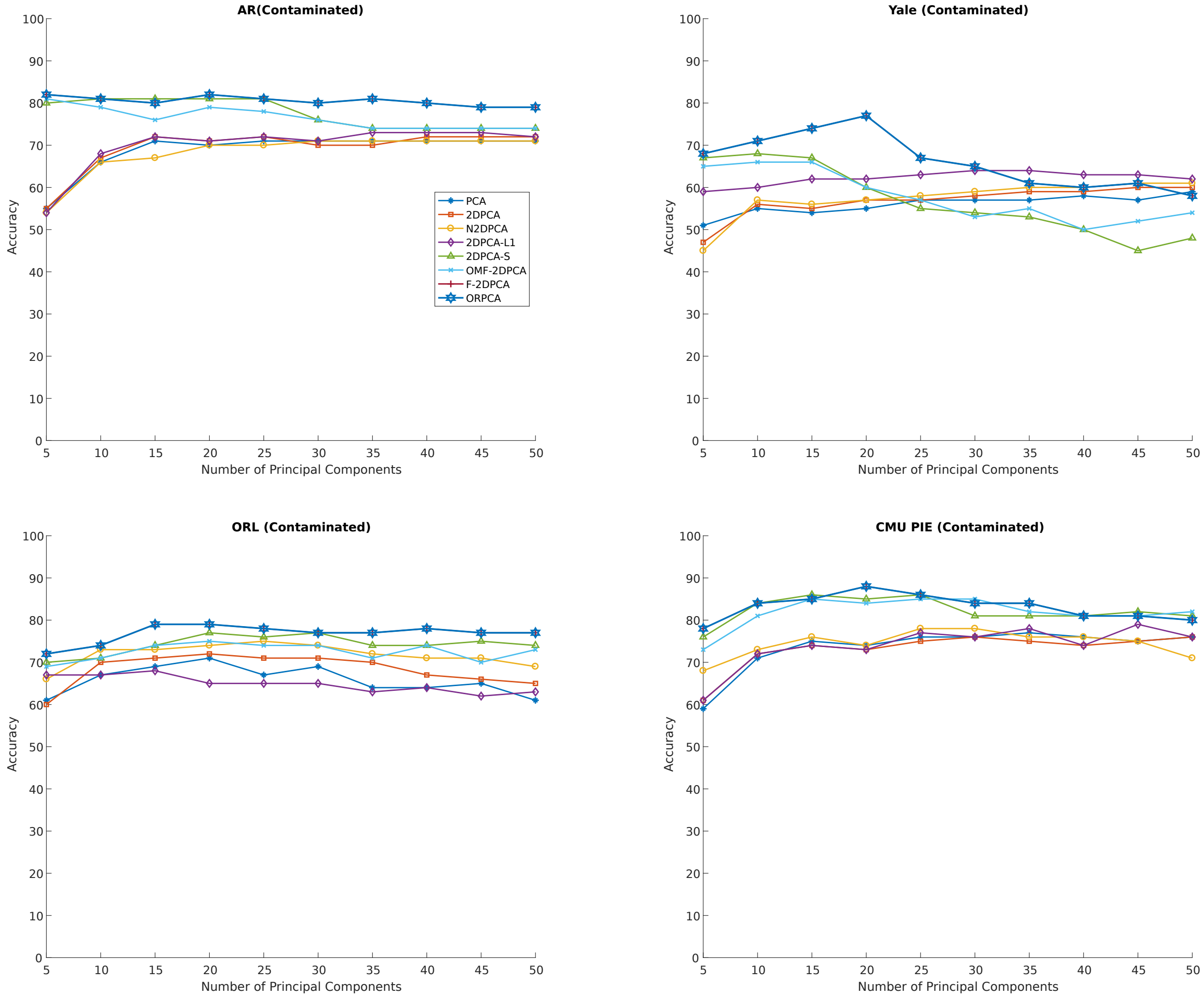

Fig. 4. Comparative evaluation on corrupted dataset (AR, Yale, ORL, and CMUIPIE)

in the presence of outliers. This shows that proposed approach suppress the role of outliers. The proposed approach reveals the geometric structure due to the fact that it select the features by maintaining the spatial structural information of the image. It is due to the fact, that the solution of ORPCA relates to the weighted image covariance matrix which characterizes the spatial structure. We notice that the performance drop significantly with the increase in projection vectors. Table VI shows the comparative analysis of reconstruction error on four detest. Notice that ORPCA has marginally poor reconstruction error as compared to others. This is due to the joint feature selection and ignoring the features that exist in other principal components.

Comparing with aforementioned experimental evaluation, we have the following interesting observations.

(I) The Objective function of the ORPCA degenerates into 2DPCA in case of P is equal to Q and $\lambda = 0$. Thus, optimal Q in this case is the transformation matrix to accommodate the robustness against outliers in 2DPCA.

(II) Penalty term introduced in the objective function excludes redundant features and provides robustness against outliers, i.e., the regularization term $\|Q\|_F^2$ reduces the constraints and enables our method to jointly select features. In other-words, penalty term penalizes all regression coefficients corresponding to single feature as a whole to make PCA possible to select discriminant features jointly.

(III) Theoretical analysis shown in theorem 4 indicates that ORPCA is convergent to local optima as shown in figure 5.

(IV) We have noticed that discrimianant features selected by ORPCA are those important and contributive features such as nose, eyes, lips in case of face

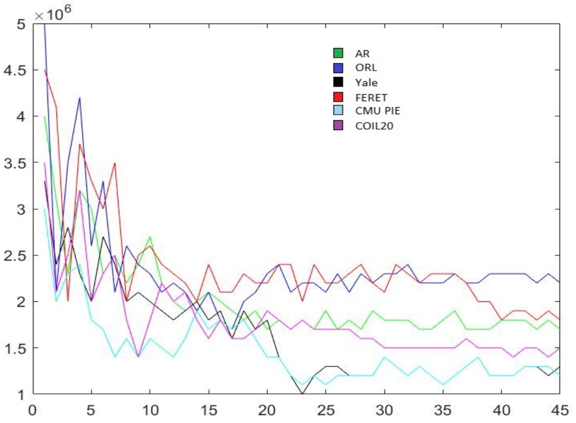

Fig. 5. Convergence curve of ORPCA on four datasets

image, while contours of different objects in non-facial datasets.

VII. Conclusion

In this paper, we presented a robust dimensionality reduction method that by relaxing the orthogonal constraints of the transformation matrix and imposing a penalty function on regularization term. In contrast to previous work on robustness in PCA, we jointly select the important features. Introduction of penalty function results in the robustness against outliers by reducing their impact in projection matrix. Compared with state-of-the-art methods, our evaluation results show the improvement in effectiveness of ORPCA for image reconstruction and classification. In conclusion, the numerical results suggest that our method is superior to previous approaches. However, this calls for further analysis and variations of the ORPCA. For example, having more than one P and one Q, offers more flexibility in accommodating the discriminant features.

References

[1] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Technical report, Yale University New Haven United States, 1997.

[2] Yudong Chen, Zhihui Lai, Jiajun Wen, and Can Gao. Nuclear norm based two-dimensional sparse principal component analysis. *International Journal of Wavelets, Multiresolution and Information Processing*, 16(02):1840002, 2018.

[3] Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust pca via stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 404–412, 2013.

[4] Xiaofei He and Partha Niyogi. Locality preserving projections. In *Advances in neural information processing systems*, pages 153–160, 2004.

[5] Zohaib Khan, Faisal Shafait, and Ajmal Mian. Joint group sparse pca for compressed hyperspectral imaging. *IEEE Transactions on Image Processing*, 24(12):4934–4942, 2015.

[6] Tao Li, Mengyuan Li, Quanxue Gao, and Deyan Xie. F-norm distance metric based robust 2dpca and face recognition. *Neural Networks*, 94:204–211, 2017.

[7] Xuelong Li, Yanwei Pang, and Yuan Yuan. L1-norm-based 2dpca. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):1170–1175, 2010.

[8] Haiping Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. Mpca: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, 19(1):18–39, 2008.

[9] Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander G Hauptmann, and Qinghua Zheng. Avoiding optimal mean 2, 1-norm maximization-based robust pca for reconstruction. *Neural computation*, 29(4):1124–1150, 2017.

[10] Aleix M Martinez. The ar face database. *CVC technical report*, 1998.

[11] Aleix M Martínez and Avinash C Kak. Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233, 2001.

[12] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[13] Lishan Qiao, Songcan Chen, and Xiaoyang Tan. Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 43(1):331–341, 2010.

[14] Imran Razzak, Michael Blumenstein, and Guandong Xu. Multiclass support matrix machines by maximizing the inter-class margin for single trial eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2019.

[15] Muhammad Imran Razzak, Muhammad Khurram Khan, Khaled Alghathbar, and Rubiyah Yousaf. Face recognition using layered linear discriminant analysis and small subspace. In *2010 10th IEEE International Conference on Computer and Information Technology*, pages 1407–1412. IEEE, 2010.

[16] Muhammad Imran Razzak, Raghib Abu Saris, Michael Blumenstein, and Guandong Xu. Robust 2d joint sparse principal component analysis with f-norm minimization for sparse modelling: 2d-rjspca. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[17] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142. IEEE, 1994.

[18] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 53–58. IEEE, 2002.

[19] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 53–58. IEEE, 2002.

[20] Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise l 1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 688–701. ACM, 2017.

[21] Chunwei Tian, Qi Zhang, Jian Zhang, Guanglu Sun, and Yuan Sun. 2d-pca representation and sparse representation for image recognition. *Journal of Computational and Theoretical Nanoscience*, 14(1):829–834, 2017.

[22] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[23] Namrata Vaswani, Thierry Bouwmans, Sajid Javed, and Praneeth Narayanamurthy. Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery. *IEEE Signal Processing Magazine*, 35(4):32–55, 2018.

[24] Haixian Wang, Qin Tang, and Wenming Zheng. L1-norm-based common spatial patterns. *IEEE Transactions on Biomedical Engineering*, 59(3):653–662, 2012.

[25] Haixian Wang and Jing Wang. 2dpca with l1-norm for simultaneously robust and sparse modelling. *Neural Networks*, 46:190–198, 2013.

TABLE V. AVERAGE RECONSTRUCTION ERROR ($\times 10^3$) AND CORRESPONDING STANDARD DEVIATION OF EACH APPROACH ON THE EXTENDED YALE B,AR, AND CMU PIE DATABASES

Methods/Dataset	2DPCA	N-2DPCA	2DPCA-L1	2DPCAL1-S	OMF-2DPCA	F-Norm 2DPCA	ORPCA
AR	118.03 ± 2.98	119.21 ± 2.43	118.37 ± 3.47	118.02 ± 2.99	116.88 ± 2.16	117.97 ± 3.0	118.15 ± 2.11
Yale B	177.53 ± 1.6	176.59 ± 2.31	178.03 ± 2.5	177.11 ± 1.7	177.25 ± 1.78	176.93 ± 1.89	177.43 ± 1.66
CMU PIE	107.41 ± 2.1	106.41 ± 1.09	107.19 ± 1.46	107.37 ± 1.91	107.21 ± 1.33	106.87 ± 2.21	108.32 ± 1.45
CMU PIE	74.12 ± 0.80	80.47 ± 0.52	74.12 ± 0.80	80.15 ± 0.59	73.78 ± 1.21	73.80 ± 0.85	75.41 ± 1.43

[26] Lei Wang, Bangjun Wang, Zhao Zhang, Qiaolin Ye, Liyong Fu, Guangcan Liu, and Meng Wang. Robust auto-weighted projective low-rank and sparse recovery for visual representation. *Neural Networks*, 117:201–215, 2019.

[27] Qianqian Wang and Quanxue Gao. Robust 2dpca and its application. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 79–85, 2016.

[28] Qianqian Wang, Quanxue Gao, Xinbo Gao, and Feiping Nie. Optimal mean two-dimensional principal component analysis with f-norm minimization. *Pattern Recognition*, 68:286–294, 2017.

[29] Fei Wen, Rendong Ying, Peilin Liu, and Robert C Qiu. Robust pca using generalized nonconvex regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[30] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

[31] Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 26(1):131–137, 2004.

[32] Jian Yang, David Zhang, Xu Yong, and Jing-yu Yang. Two-dimensional discriminant transform for face recognition. *Pattern recognition*, 38(7):1125–1129, 2005.

[33] Qiaolin Ye, Liyong Fu, Zhao Zhang, Henghao Zhao, and Meem Naiem. Lp-and ls-norm distance based robust linear discriminant analysis. *Neural Networks*, 105:393–404, 2018.

[34] Shuangyan Yi, Zhihui Lai, Zhenyu He, Yiu-ming Cheung, and Yang Liu. Joint sparse principal component analysis. *Pattern Recognition*, 61:524–536, 2017.

[35] Yan Zhang, Zhao Zhang, Jie Qin, Li Zhang, Bing Li, and Fanzhang Li. Semi-supervised local multi-manifold isomap by linear embedding for feature extraction. *Pattern Recognition*, 76:662–678, 2018.

[36] Zhao Zhang, Fanzhang Li, Mingbo Zhao, Li Zhang, and Shuicheng Yan. Joint low-rank and sparse principal feature coding for enhanced robust representation and visual classification. *IEEE Transactions on Image Processing*, 25(6):2429–2443, 2016.

[37] Zhao Zhang, Fanzhang Li, Mingbo Zhao, Li Zhang, and Shuicheng Yan. Robust neighborhood preserving projection by nuclear/l2, 1-norm regularization for image feature extraction. *IEEE Transactions on Image Processing*, 26(4):1607–1622, 2017.

[38] Zhao Zhang, Jiahuan Ren, Weiming Jiang, Zheng Zhang, Richang Hong, Shuicheng Yan, and Meng Wang. Joint subspace recovery and enhanced locality driven robust flexible discriminative dictionary learning. *arXiv preprint arXiv:1906.04598*, 2019.

[39] Zhao Zhang, Shuicheng Yan, and Mingbo Zhao. Pairwise sparsity preserving embedding for unsupervised subspace learning and classification. *IEEE Transactions on Image Processing*, 22(12):4640–4651, 2013.

[40] Mingbo Zhao, Tommy WS Chow, Zhou Wu, Zhao Zhang, and Bing Li. Learning from normalized local and global discriminative information for semi-supervised regression and dimensionality reduction. *Information Sciences*, 324:286–309, 2015.

[41] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

[42] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.