Generalized Centered 2-D Principal Component Analysis

Gongyu Zhou, Student Member, IEEE, Guangquan Xu, Member, IEEE, Jianye Hao, Shizhan Chen, Member, IEEE, Junfeng Xu, and Xi Zheng, Member, IEEE

Abstract—Most existing robust principal component analysis (PCA) and 2-D PCA (2DPCA) methods involving the \(l_2\)-norm can mitigate the sensitivity to outliers in the domains of image analysis and pattern recognition. However, existing approaches neither preserve the structural information of data in the optimization objective nor have the robustness of generalized performance. To address the above problems, we propose two novel center-weight-based models, namely, centered PCA (C-PCA) and generalized centered 2DPCA with \(l_2,p\)-norm minimization (GC-2DPCA), which are developed for vector- and matrix-based data, respectively. The C-PCA can preserve the structural information of data by measuring the similarity between the data points and can also retain the PCA’s original desirable properties such as the rotational invariance. Furthermore, GC-2DPCA can learn efficient and robust projection matrices to suppress outliers by utilizing the variations between each row of the image matrix and employing power \(p\) of \(l_2\)-norm. We also propose an efficient algorithm to solve the C-PCA model and an iterative optimization algorithm to solve the GC-2DPCA model, and we theoretically analyze their convergence properties. Experiments on three public databases show that our models yield significant improvements over the state-of-the-art PCA and 2DPCA approaches.

Index Terms—2-D principal component analysis (2DPCA), center, dimensionality reduction, \(l_2,1\)-norm.

I. INTRODUCTION

DIMENSION reduction plays an important role as a data-preprocessing step in the domains of image analysis, machine learning, and data mining. Due to the curse of dimensionality, dimension reduction methods are typically used to extract the most characteristic features from the raw data, then the features are used for the subsequent visualization and classification tasks [1]–[3]. Among these methods, two representative linear dimensionality reduction methods are principal component analysis (PCA) [4] and linear discriminant analysis (LDA) [5], and the typical nonlinear dimensionality reduction methods are the locality preserving projections (LPPs) [6] and the kernel PCA (KPCA) [7]. The PCA and LDA keep the global geometric structures linear, while LPP can preserve the local neighborhood structure. The KPCA is the nonlinear form of PCA, which can better exploit the complicated spatial structures of high-dimensional data.

Standard PCA is often sensitive to outliers and noise at different levels of interference in high-dimensional datasets, since it is based on the minimum reconstruction error in a least-squares sense. To handle this issue, many robust PCA approaches have been developed. L1-PCA [8] was initially found to be better than the squared \(l_2\)-norm-based PCA, the function of the L1-PCA minimizes the \(l_1\)-norm-based reconstruction error. However, it usually has poor performance due to its large computation requirements, and it lacks the rotational invariance. In order to mitigate the problem, a robust PCA based on the rotational invariant \(R_1\)-norm, called the \(R_1\)-PCA [9], has been proposed to find the optimal basis vectors that will alleviate the impacts of outliers. However, the \(R_1\)-PCA model is solved using the subspace feature learning algorithm, which results in a very slow convergence rate.

To seek a more robust projection matrix, Kwak [10] uses the \(l_1\)-norm to measure the variance and proposes the PCA-L1 with a greedy algorithm (PCA-L1 greedy). In order to handle the convergence problem, Nie et al. proposed a non-greedy algorithm (PCA-L1 nongreedy) [11], which has a closed-form solution in every iteration when convergence exists.

The aforementioned methods for image classification generally convert the input image data into 1-D vectors. However, this process gives rise to the loss of inherent spatial (structural) information in an image. In order to leverage more structural information and improve performance, Yang et al. proposed the 2DPCA [12], which discards the transformation between the 1-D vector and images and avoids destroying the topology structures of image pixels. Motivated by the 2DPCA, many image-as-matrix methods have been developed, such as the 2DLDA [13] and the multilinear PCA (MPCA) [14]. All of these 2-D-based methods have been empirically shown to perform better compared to their corresponding methods [15], [16]. Inspired by the \(l_1\)-norm PCA, the 2DPCA-L1 [17] and the 2DPCA-L1-S [18] imposed the sparse

Digital Object Identifier 10.1109/TCYB.2019.2931957
constraint in the 2DPCA-L1 and successively appear. More important, the robust 2DPCA-L1 with a non-greedy algorithm (2DPCA-L1-nongreedy) is found [19].

Compared with the traditional PCA/2DPCA, the $l_1$-norm-based PCA/2DPCA technique is robust, but it lacks rotational invariance. Most existing robust PCA/2DPCAs only consider one of the two factors: the covariance matrix and the reconstruction error. Therefore, Wang and Gao proposed the AnglePCA [20] and F-2DPCA [21]. These methods alleviate the sensitivity to outliers in some ways by synthesizing the above two factors. An $F$-norm-based model can obtain robust projection matrices instead of using the squared $F$-norm and has become an active topic in dimensionality reduction and data mining [21]–[24]. In addition, a $l_2$-norm-based model can retain the PCA's desirable properties (rotational invariance and the weighted matrix related to the covariance matrix and the reconstruction error). However, AnglePCA and F-2DPCA fail to extract internal information and more complicated structures from the image matrix. These methods also lack robust generalizable performance [25].

In this article, we first propose two novel center-weighted-based models, called the centered PCA (C-PCA) and the generalized centered 2DPCA with $l_{2,p}$-norm minimization (GC-2DPCA). In these two models, the first point of innovation is that we consider the data similarity in order to better retain the data structures. It is similar to the linear KPCA (L-KPCA) [26], but L-KPCA explicitly neglects the reconstruction error. By leveraging this insight, our proposed models can learn the connected projection matrices for recognition and reconstruction tasks. Even though these ideas are simple, applying them together is powerful for recognition tasks. The C-PCA can retain the structural information of data by measuring the similarity between the data points, and it can also retain the PCA's original desirable properties such as rotational invariance. The C-PCA only applies the centered-weighted matrix and a more robust reconstruction weighted matrix than other 2DPCAs.

**II. RELATED WORKS**

Assume that we have a set of $n$ sample images $X = [x_1, x_2, x_3, \ldots, x_n]$, where $x_i \in \mathbb{R}^d$ ($i = 1, 2, \ldots, n$) denotes the $i$th training image. Suppose that the data are centered, that is, the mean from the data is zero. $d$ is the dimensionality of the sample space. The traditional PCA method maximizes the variance in the data in the project subspace in order to obtain the projection matrix $W = [w_1, w_2, \ldots, w_k] \in \mathbb{R}^{dk}$, and to solve the following optimization function:

$$
\max_{W^TW = I} \text{Tr}(W^T S W) = \max_{W^TW = I} \sum_{i=1}^{n} \|W^T x_i\|_2^2
$$

where $\text{Tr}(\cdot)$ is the trace operator of a matrix, $I$ is the identity matrix, and $S_i = \sum_{i=1}^{n} x_i x_i^T$ is the covariance matrix. We denote the $l_1$-norm and the $l_2$-norm of a vector by $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively. In fact, the function (1) can be reformulated as the objective function (2) [4]

$$
\min_{W^TW = I} \sum_{i=1}^{n} \|x_i - WW^T x_i\|_2^2.
$$

The objective functions (1) and (2) employing the squared $l_2$-norm, are more sensitive to outliers, and they magnify the effect of outliers. It is well known that the $l_1$-norm is robust to outliers. Therefore, many robust methods that directly substitute the $l_1$-norm for the squared $l_2$-norm have been proposed [10], [11]

$$
\max_{W^TW = I} \sum_{i=1}^{n} \|W^T x_i\|_1.
$$

The solution of the objective function (3) is not equivalent to function (2), since function (2) is the true goal of PCA. Therefore, Wang et al. proposed the AnglePCA, which considers the relation between the reconstruction error and variance of projected data [20]. It seeks the projection matrix $W$ using the objective function

$$
\max_{W^TW = I} \sum_{i=1}^{n} \frac{\|W^T x_i\|_2}{\|x_i - W W^T x_i\|_2}.
$$

While the KPCA allows us to generalize the traditional PCA to a nonlinear dimensionality reduction, its reconstruction error for data is not equivalent to the objective function (2). Since the aforementioned robust PCA methods are related to the learning linear subspace projection matrix, we only consider the simple linear kernel $k(x, y) = (x^T y)$. The essence of the kernel is to compare the similarity between two objects and, thus, the L-KPCA fulfills the following objective function using a kernel trick [26]:

$$
\max_{W^TW = I} \sum_{i=1}^{n} \|W^T H x_i\|_2^2
$$

GC-2DPCA generally obtains a more efficient recognition weighted matrix and a more robust reconstruction weighted matrix than other 2DPCAs.

Our main contributions are as follows.

1) We propose the C-PCA model to effectively preserve the structural information of the data. In order to assess the scalability of our model, we extend from the vector-based AnglePCA to the vector-based C-AnglePCA.

2) We propose the matrix-based GC-2DPCA model. This model not only can effectively preserve the structural information for data but also obtains a more robust projection matrix for data.

3) The convergence analysis of the objective function for the GC-2DPCA is given. We demonstrate that the

- **2DPCA-L1-nongreedy** is found [19].

- **IEEE TRANSACTIONS ON CYBERNETICS**

- **Authorized licensed use limited to: Macquarie University. Downloaded on November 07,2020 at 07:19:52 UTC from IEEE Xplore. Restrictions apply.**
where $H = I - (1/n)11^T$ is the centering matrix. To extract more structural information, we propose a centered projection matrix method. Meanwhile, the weighted matrix-based reconstruction error to the function (2) is preserved.

For the 2DPCA model, we assume that the data are denoted by $X = [X_1, X_2, \ldots, X_n] \in \mathbb{R}^{n \times m}$ which is centralized, that is, $\sum_{i=1}^{n} X_i = 0$. 2DPCA seeks the projection matrix $W = [w_1, w_2, \ldots, w_k] \in \mathbb{R}^{m \times k}$ by [12]

$$\max \sum_{i=1}^{n} W_i^T X_i W = \max \sum_{i=1}^{n} \|X_i W\|^2_F$$

(6)

where $\| \cdot \|_F$ denotes the Frobenius norm.

The F-2DPCA adopts an intuitive and reasonable use of the F-norm instead of the squared F-norm, and it is equivalent to function (6). The objective function (7) is called the F-2DPCA [21], and it is denoted as follows:

$$\min \sum_{i=1}^{n} \|X_i - X_i W W^T\|_F^2$$

$$\rightarrow \min \sum_{i=1}^{n} \|X_i - X_i W W^T\|_F^2.$$  

(7)

III. OUR SOLUTIONS: C-PCA AND GC-2DPCA

We present two novel dimensionality reduction models based on the PCA and the 2DPCA, namely, the C-PCA and the GC-2DPCA, for data classification and representation.

A. Centered PCA

In this section, we consider combining the PCA and the L-KPCA in order to maximize our robustness and efficiency. We use the centering matrix and propose the C-PCA. The C-PCA can well characterize the geometric structure. In the previous research, there is no difference between recognition accuracy weight and reconstruction error weight as they are regarded (without distinction) as the same weight. We first propose the projection matrices $W_{re}$ and $W_{ra}$ to calculate them separately. We seek the projection matrices $W_{re}$ and $W_{ra}$ by solving the functions (8) and (9) as follows:

$$\max \sum_{i=1}^{n} \|W_{re}^T x_i\|^2_2$$

(8)

$$\max \sum_{i=1}^{n} \|H^{-1}W_{ra} W_{ra}^T x_i\|^2_2$$

$$= \max \sum_{i=1}^{n} \text{Tr}((H W_{ra})^T x_i x_i^T H W_{ra})$$

(9)

$$= \max \sum_{i=1}^{n} \|W_{ra}^T H x_i\|^2_2$$

where $W_{ra}$ represents the recognition accuracy weight, $W_{re}$ represents the reconstruction error weight, and $H^{-1}$ represents the pseudoinverse matrix of $H$.

$H$ is an idempotent matrix. We have the following properties: 1) its eigenvalues can only be 0 or 1 and 2) trace$(H) = \text{rank}(H) = m - 1$, which means that $H^{-1}$ does not exist; and $H = H^{-1}$ and $H = H^T$.

Algorithm 1 lists the pseudocode for solving the functions (8).

The KPCA with the linear kernel is exactly equivalent to the standard PCA [7]. That is, after the kernel Gram matrix is reduced to the form $X^T X$, the kernel Gram matrix is equivalent to the standard Gram matrix as the principal components will not change. After establishing the connection between (2) (PCA) and (5) (L-KPCA), we can also derive the conclusion that (8) is equivalent to (9). From the above analysis, we can obtain $W_{re} = H^{-1} W_{ra}$ and the solution of (9), that is, $W_{ra} = HW_{ra}$.

The established equivalence of the two equations can help to derive the new model. In fact, based on the connection, we can improve the efficiency of the proposed method. The detailed analysis is as follows. The PCA is better than the L-KPCA in terms of the data reconstruction loss, but the L-KPCA is better than the PCA when considering the similarity between data. Therefore, we combine the two equations in order to generate a new model. The projection matrix that is learned by the new model will naturally have the advantages of the PCA and the L-KPCA models, which are improved by classification and robustness.

First, (8) is equivalent to (1), which is the traditional PCA. Second, we consider Theorem 1 and, thus, (9) is equivalent to (5), which is the L-KPCA. The projection matrix of the former represents the reconstruction error weight, and the latter represents the recognition accuracy weight.

If there is no special description below, in this article, we use $W$ to represent $W_{re}$.

Algorithm 1 lists the pseudocode for solving the function (8).

Therefore, the C-PCA can obtain two different but connected projection weighted matrices. Specifically, the recognition accuracy weight is equivalent to the matrix in the L-KPCA [26], and the reconstruction error weight is the same as the weight in the traditional PCA. The C-PCA has one additional simple matrix multiplication operation than PCA; this operation is not as complicated as the decomposition matrix, so the C-PCA is slightly higher than PCA in average time consumption. In order to analyze the connection of our method with the L-KPCA, we first need the following theorem. Denote $D^x \in \mathbb{R}^{n \times n}$ as a distance matrix, where the $(i, j)$th element is $d^x_{ij} = \|x_i - x_j\|^2_2$.

Theorem 1: $W^T H D^x H W = -2W^T H X X^T H W$.

Proof: Since $d^x_{ij} = \|x_i - x_j\|^2_2 = x_i^T x_i + x_j^T x_j - 2x_i^T x_j$, we have $D^x = \text{diag}(XX^T)11^T + 11^T \text{diag}(XX^T) - 2XX^T$, where
diag(XX^T) is a diagonal matrix with the diagonal elements of XX^T. Note that H_1 = 1^T H = 0 and according to the definition of H, we have W^T H D^T H W = -2 W^T H X X^T H W.

According to Theorem 1, we can know that the L-KPCA essentially seeks the recognition weight by minimizing the distance between similar data points. According to Theorem 2, we know that the objective function (1) and the following objective function are equivalent

\[
\max_{W^T W = I} \sum_{i,j} \| W^T (x_i - x_j) \|^2_2. \tag{10}
\]

According to Theorem 2, we know that the objective function of the conventional PCA (10) and the L-KPCA (5) is not consistent to some extent. If H^{-1} exists, then (1) is absolutely the same as (5). Equation (1) is more inclined to capture the fine reconstruction error weight than (5). Meanwhile, the recognition accuracy weight that is obtained by (5) is more outstanding than that of (1). However, the C-PCA can take into account the advantages of both and obtain two different but connected projection weighted matrices. For example, we randomly produce some outliers and two classified data points using MATLAB. Then, we plot the optimal projection directions of the L-KPCA, PCA, and C-PCA using the artificial dataset in Fig. 1.

Fig. 1 shows the following information. After the projection of the sample points on the L-KPCA and C-PCA planes, the L-KPCA and C-PCA have the same distinguishable effect in this figure. However, the overall distance from the sample point to the C-PCA hyperplane is closer than that of the L-KPCA. Meanwhile, the overall distance from the sample point to the C-PCA and PCA hyperplanes is the same. After the points are projected onto the respective C-PCA and PCA planes, it is shown that the C-PCA is better than the PCA in terms of its distinguishability. In conclusion, Fig. 1 demonstrates that the C-PCA acquires advantages of the L-KPCA and PCA. Specifically, the C-PCA has good performance in classification and robustness.

As mentioned in Section I, most PCA methods will lose the structural information of the data. The idea of adopting the PCA to the C-PCA is to use to solve the problem that exists in most PCA methods. We solve the reconstruction weight using the original 1-D method, and then we can obtain the recognition weight using the transformation formula. According to the previous analysis of Theorem 1, the transformation formula essentially seeks the recognition weight by minimizing the distance between similar data points. For example, the idea of our C-PCA model is applied to AnglePCA, which forms a C-AnglePCA to improve scalability.

B. Generalized Centered 2DPCA With the l_{2,p}-Norm Minimization

Recall that the proposed C-PCA only considers optimizing the recognition matrix. In this section, we propose the GC-2DPCA. We extend the proposed C-PCA to its 2-D robust version and develop a novel robust 2DPCA with special l_{2,p}-norm minimization.

By simple algebra, we have

\[
\min_{W^T W = I} \sum_{i=1}^n \| X_i - X_i W W^T \|_F^2
\]

\[
= \min_{W^T W = I} \sum_{i=1}^n \sum_{j=1}^r \| X_i(j,:) - X_i(j,:) W W^T \|_2^2 \tag{11}
\]

where \(X_i(j,:)\) denotes the \(j\)th row of \(X_i\).

We propose a generalized 2DPCA formulation

\[
\min_{W^T W = I} \sum_{i=1}^n \sum_{j=1}^r \| X_i(j,:) - X_i(j,:) W W^T \|_2^2
\]

\[
\rightarrow \min_{W^T W = I} \sum_{i=1}^n \sum_{j=1}^r \| X_i(j,:) - X_i(j,:) W W^T \|_p^2 \tag{12}
\]

where \(0 < p \leq 2\).

Equation (12) uses the special \(l_{2,p}\)-norm to replace the squared \(l_2\)-norm in (11) and achieves robustness; meanwhile, (12) adopts a similar idea as (4) and (7). The objective function (12) is called GC-2DPCA. Compared with the F-2DPCA, our proposed method can make further use of the variations between each row of the image matrix, and the GC-2DPCA is more robust to outliers. Meanwhile, we also keep a predominant property (i.e., rotational invariance), and given an arbitrary rotation matrix \(\Gamma (\Gamma \Gamma^T = I)\), we have \(\| \Gamma X_i(j,:) W \|_2^p = \| X_i(j,:) W \|_2^p\).

In fact, the objective function (12) can be further written in the following form:

\[
\min_{W^T W = I} \sum_{i=1}^n \sum_{j=1}^r \| X_i(j,:) - X_i(j,:) W W^T \|_2^p
\]

\[
= \min_{W^T W = I} \sum_{i,j} \text{Tr}(\Pi) E_{ij} \tag{13}
\]

where \(\Pi = X_i(j,:)^T X_i(j,:) - W^T X_i(j,:)^T X_i(j,:) W\)

\[
E_{ij} = \| X_i(j,:) - X_i(j,:) W W^T \|_2^{p-2} \tag{14}
\]
is a reconstruction error weight containing image row information.

Thus, the above (13) objective function can be written as
\[
\max_{W^T W = I} \sum_{i=1}^{n} \sum_{j=1}^{r} \text{Tr}(W^T X_i(j,:)T X_i(j,:) W)E_{ij}. \tag{15}
\]

We are considering a new optimization approach to solve the two unknown variables \(W\) and \(E_{ij}\) in the maximization function (15). By means of decoupling the functional relation between \(W\) and \(E_{ij}\), we can optimize the two variables in an alternative way by optimizing one variable while keeping the other fixed.

Specifically, when \(E_{ij}^{(t)}\) is known, we can learn \(W\) by maximizing (15). In the \((t+1)\)th iteration, with \(E\) as a constant matrix, its \(i\)th row and \(j\)th column element is \(E_{ij}\), and the function (15) finally becomes
\[
\max_{W^T W = I} \text{Tr}(W^T C W) = \max_{W^T W = I} \text{Tr}(W^T Z) \tag{16}
\]
where \(C = \sum_{i=1}^{n} \sum_{j=1}^{r} X_i(j,:)T E_{ij} X_i(j,:), Z = CW\).

We solve the optimization function (15) based on Theorem 3.

**Theorem 3:** Suppose the SVD of \(Z\) is \(Z = PAQ^T\), where \(P \in \mathbb{R}^{c \times c}\), \(A \in \mathbb{R}^{c \times k}\), and \(Q \in \mathbb{R}^{k \times k}\). The solution of the optimization function (15) is derived as \(W = P L_{c \times k} Q^T\).

**Proof:** Based on the SVD of \(Z\), we have
\[
\text{Tr}(W^T Z) = \text{Tr}(W^T PAQ^T) = \text{Tr}(AQ^T P^T) = \text{Tr}(\Theta) = \sum_{j} \lambda_{ij}\theta_{ij}
\]
where \(\Theta = Q^T W^T P\); and \(\lambda_{ij}\) and \(\theta_{ij}\) represent the \((j,\ j)\)th element of matrices \(A\) and \(\Theta\), respectively. Recall the constraint of \(W^T W = I\), where we can have \(\Theta \Theta^T = I\) and \(\theta_{ij} \leq 1\).

Therefore, the maximum with \(\Theta = I\). Combining \(\Theta = I\) and \(\Theta = Q^T W^T P\) yields \(W = P L_{c \times k} Q^T\).

Now, we can extend the objective function (16) to the formulation of the centered 2DPCA. Algorithm 2 lists the pseudocode of solving the following problem:
\[
\max_{W_{ra}^T W_{re}=I} \sum_{i=1}^{n} \text{Tr}\left((H^{-1}_{ra} W_{ra})^T C H^{-1}_{re} W_{re}\right) = \max_{W_{ra}^T W_{re}=I} \sum_{i=1}^{n} \left\|W_{ra}^T C W_{re}\right\|_2^2 = \max_{W_{ra}^T W_{re}=I} \sum_{i=1}^{n} \left\|W_{ra}^T C W_{re}\right\|_2^2 \tag{17}
\]
where \(W_{ra} = HW_{re}\) is the recognition weighted matrix.

**C. Convergence Analysis**

A similar convergence conclusion to PCA can be established for the C-PCA in Algorithm 1. For brevity, we skip the convergence analysis in this article.

**Algorithm 2 : GC-2DPCA**

**Input:** data set \(\{X_i \in \mathbb{R}^{c \times c} : i = 1, 2, \ldots, n\}\), \(k, p\), where \(X_i\) is normalized. Initialize \(W_{re}^{(1)} \in \mathbb{R}^{c \times k}\) which satisfies \((W_{re}^{(1)})^T (W_{re}^{(1)}) = I, t = 1\).

1: while not converge do
2: For all training samples, calculate \(E_{ij}^{(t)}(i = 1, \ldots, N; j = 1, \ldots, r)\) by Eq. (14);
3: Calculate \(Z^{(t)}\) according to Eq. (16), i.e., \(Z^{(t)} = \sum_{i=1}^{n} X_i(j,:)T E_{ij} X_i(j,:) W_{re}^{(t)}\);
4: Calculate the SVD of \(Z\) as \(Z = PAQ^T\), then \(W_{re}^{(t+1)} = PQ^T\);
5: \(t = t + 1\);
6: end while

**Output:** The reconstruction weight \(W_{re}\) and the recognition weight \(W_{ra}\).

**Theorem 4:** Algorithm 2 will converge to a local optimal solution of the objective function (13).

**Proof:** The Lagrangian function of the function (13) is
\[
L(W) = \sum_{i,j} \left\|X_i(j,:) - X_i(j,:) W W^T \right\|_2^p - \text{Tr}(\Lambda^T (W^T W - I))\tag{18}
\]
where the Lagrangian multiplies \(\Lambda = (\Lambda_{pq})\) for enforcing the orthonormal constrains \(W^T W = I\). The KKT condition for the optimal solution specifies that the gradient of \(L\) must be zero, that is
\[
\frac{\partial L(W)}{\partial W} = p \sum_{i,j} \left(\left|X_i(j,:) - X_i(j,:) W W^T \right|_2^{p-2}\right) X_i(j,:) W - \Lambda^T = 0. \tag{19}
\]

By simple algebra, we have
\[
\sum_{i,j} \left|X_i(j,:) - X_i(j,:) W W^T \right|_2^{p-2} X_i(j,:) W = W(\Lambda/p)^T. \tag{20}
\]

According to the aforementioned analysis, the optimal solution of the objective function (16) can be obtained in step 4 in Algorithm 2. Thus, the converged solution of Algorithm 2 satisfies the KKT condition of the objective function (16). The Lagrangian auxiliary function (16) is
\[
L_2(W) = 2 \text{Tr}(W^T Z) - \text{Tr}(\Lambda^T (W^T W - I)). \tag{21}
\]

Taking the derivation of (21) with respect to \(W\) and setting it to zero, we have
\[
Z = W \Lambda_1. \tag{22}
\]

Equation (22) is formally similar to (20). The main difference between (20) and (22) is that \(W\) of \(Z\) is known in each iteration in Algorithm 2. Suppose we obtain the optimal solution \(W^*\) in the \((t+1)\)th iteration; thus, we have \(W^* = W^{(t)} = W^{(t+1)}\). According to the definition of \(E_{ij}\), we
can find that (22) and (20) are the same. It means that the converged solution of Algorithm 2 satisfies the KKT condition of (13), that is

\[ \frac{\partial L}{\partial \mathbf{W}} \bigg|_{\mathbf{W} = \mathbf{W}^*} = 0. \]  

(23)

Thus, the converged solution of Algorithm 2 is a local solution of (13).

IV. RELATIONSHIP TO THE RELEVANT 2DPCA METHODS

In this section, we show the relationship of GC-2DPCA with the other related robust 2DPCA algorithms.

A. Connection to Standard 2DPCA

When \( p = 2 \), the weighted \( E_i \) becomes a constant and, thus, the objective function (12) becomes

\[ \mathbf{W}^* = \arg \min_{\mathbf{W}^T \mathbf{W} = 1} \sum_{i,j} \left\| \mathbf{X}_i(j,:) - \mathbf{X}_i(j,:) \mathbf{W} \mathbf{W}^T \right\|_2^2 \]  

(24)

in which case the GC-2DPCA reduces to the standard 2DPCA [12]. The standard 2DPCA applies the squared \( F \)-norm, which is very sensitive to noise and outliers in given data, in order to model the reconstruction error. This may result in higher reconstruction errors in practice since most real data contain various noise, errors, or even corruptions.

B. Connection to \( R_1 \)-2DPCA and OMF-2DPCA

If \( p = 1 \), then the objective function (12) becomes

\[ \mathbf{W}^* = \arg \min_{\mathbf{W}^T \mathbf{W} = 1} \sum_{i,j} \left\| \mathbf{X}_i(j,:) - \mathbf{X}_i(j,:) \mathbf{W} \mathbf{W}^T \right\|_2. \]  

(25)

To some extent, this is equivalent to the objective function and that is called the \( R_1 \)-2DPCA [33].

If we optimize the mean \( \bar{\mathbf{X}} \) in the objective function, then we can obtain

\[ \mathbf{W}^* = \arg \min_{\mathbf{W}^T \mathbf{W} = 1} \sum_{i,j} \left\| \mathbf{X}_i(j,:) - \bar{\mathbf{X}} - (\mathbf{X}_i(j,:) - \bar{\mathbf{X}}) \mathbf{W} \mathbf{W}^T \right\|_2. \]  

(26)

To some extent, this is equivalent to the objective function that is called the OMF-2DPCA [23]. In the objective function (25) or (26), the reconstruction error of each data point is measured using the \( l_{2,1} \)-norm rather than the squared \( F \)-norm as in the standard 2DPCA. Thus, compared with the standard 2DPCA, the \( R_1 \)-2DPCA and OMF-2DPCA are robust to outliers. However, the large reconstruction error still plays an important role in the objective function (25) or (26).

By referring to the related papers [2, 21, 23, 33], we can find that the experimental performance of these methods is almost the same as that of the F-2DPC and Angle-2DPCA. Therefore, in Section V, only the F-2DPCA and Angle-2DPCA are assessed.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conduct several experiments to evaluate the proposed C-PCA and GC-2DPCA models on three public face databases (CMU-PIE, AR, and Extended Yale B) and compare them to our proposed vector- and matrix-based methods, respectively. For the 1-D benchmark models, we choose the following representative methods: the traditional PCA [4], the PCA-L1 greedy [10], the PCA-L1 nongreedy [11], the HQ-PCA [34], and the AnglePCA [20]. For the 2-D benchmark models, we choose the 2DPCA [12], the 2DPCA-L1 [17], the 2DPCAL1-S [18], the N-2DPCA [35], and the F-2DPCA [21].

To compare our approach with the benchmarks, we used the same setting with [20] and [21] except for the range of noise ratio. The parameter used in the cited papers is intervenient with the range 0.05–0.15, and our parameter is intervenient with range 0.05–0.35 as we want to verify the noise immunity and recognition rate of our proposed method to a greater extent. In addition, we tune \( p \) from \( [0.5, 1, 1.5] \) for the GC-2DPCA model.

In our experiments, we use the 1-nearest neighbor (1NN) and set \( p = 1.5 \) for the classification. We set the number of projection vectors as 300 and 25 in the 1-D and 2-D methods, respectively. Furthermore, we use the following reconstruction error to measure the quality of 2-D methods:

\[ \text{error} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i^{\text{clean}} - \mathbf{x}_i^{\text{clean}} \mathbf{w}_r \mathbf{w}_r^T \right\|_F \]  

(27)

where \( N \) is the number of training data, \( \mathbf{w}_r \) is the learned reconstruction weight, and \( \mathbf{x}_i \) is the \( i \)th clean training sample.

The CMU-PIE face database [36], from which 1632 frontal-face images of 68 individuals are chosen, is used to assess the intraclass variation, such as illumination and expression. In our experiment, we choose 21 images of each person, among which we use 10 images for training and 11 images for testing. In addition, some training images and test images for each person have white or black dots added as noise. We randomly added the noise to 30% of the pictures. The positions of outliers are randomly distributed. The location of noise is random, and the ratio of the noise pixels to the number of image pixels is set from 0.05 to 0.35. Each image is in grayscale and normalized to \( 32 \times 32 \) [6]. Fig. 2 shows some samples with or without occlusion of one person from the CMU-PIE face database.

The AR facial database [37] contains 3120 color images corresponding to 126 people’s faces. The images were divided into two sessions. Each session consists of 13 images, which is seven frontal face images and six images with occlusion, such as scarf or glasses. Each session presents the variation of expression, illumination condition, and occlusion. In our experiment, we select randomly 13 images of everyone for a total of 1605 images as the training set and the remaining
Fig. 3. Images of one person under two different sessions from the AR database.

for testing. We randomly select the training and test images of each person to add occlusion as that in the CMU-PIE database. We manually cropped the face data portion and resized it to 50 × 40 [12]. We display some samples in Fig. 3.

The Extended Yale B face database [38] includes 38 individuals with 64 frontal-face images of each individual; a person’s 64 images were taken from five different angles and divided into five subsets; we choose 2414 pictures from it. In the experiment, 32 images of each individual are used for the training data, and the remaining images are used for testing. We select randomly training and test images of each person to add occlusion as that in the CMU-PIE database. Each image is in grayscale and normalized to 32 × 32 [32].

A. Recognition Rate Comparison for PCA/2DPCA

Figs. 4(a), 5(a), and 6(a) show the recognition performance with different feature numbers for the C-PCA and C-AnglePCA compared with other 1-D methods. Similarly, the recognition performance with different feature numbers for the GC-2DPCA compared with other 2-D methods is presented in Figs. 4(b), 5(b), and 6(b). Table I lists the average recognition rate (and the corresponding standard deviation) on three databases for the 2-D methods. To validate whether our method is better than the others with respect to the mean value in Table I, we check how significant these results are by employing the t-test in Table II. For the sake of simplicity, we only perform t-tests on the recognition rate of both the F-2DPCA and GC-2DPCA. We set the default significance level as $\alpha = 0.05$, the elements of Table II represent the $h$(sig), where $h$ is a Boolean-type variable, indicating whether to accept the null hypothesis, and sig is the probability that the null hypothesis can be accepted. When sig is a small probability, the null hypothesis is questioned. The t-tests’ experiments are performed by using the ttest2 function in MATLAB. In the experiment, the $p$ in GC-2DPCA is the optimal value in each dataset in the range of $[0.5, 1, 1.5]$.

Figs. 4(a), 5(a), and 6(a) show that our method is easily extended to the existing PCA models. For example, the result of the C-AnglePCA is superior to the original AnglePCA. From the results of Figs. 4–6, we can find that our methods are generally excellent to the existing state-of-the-art algorithms. Even if we add some noise to the three public datasets, our methods maintain better properties of the traditional PCA, and keep the “close” data point close after dimensionality reduction.

Table I shows that our approach achieves the best recognition accuracy compared to the others. The main reason is that the GC-2DPCA employs the $l_{2,1}$-norm as the distance metric in the object function of the 2DPCA, and it solves the optimal solution using a nongreedy algorithm which has a closed-form solution in each iteration and local convergence. Our approaches also well characterize the geometric structure of the data by considering the similarity of the data. Overall, our experiments show that nongreedy strategies are indeed better than greedy strategies.

The null hypothesis is that the average recognition accuracies of the GC-2DPCA and F-2DPCA are from the same distribution at a 0.05 significance level. From Table II, $h = 1$, thereby indicating that the null hypothesis is rejected at the 0.05 significance level. In addition, the possibility of the original hypothesis being established (sig = 0.0005) is very small on CMU PIE, and the other two datasets are the same. It is shown that the result of the algorithm GC-2DPCA can be statistically determined to be larger than the result of the F-2DPCA (i.e., the comparison of the mean values of the two groups of data is meaningful in Table I). Combined with Table I, our method can be proved to have better performance than other benchmarks.

B. Reconstruction Error Comparison for 2DPCA

Fig. 7 presents the convergence curve of the GC-2DPCA on three databases. It can be seen that our method can monotonically decrease the value of the objective function (12) in each iteration. To be fair, we set the number of projection matrices 25 and set the number of iterations as 50.
Table III lists the average time consumption of the 2-D methods on the three databases. From it, we can see that the running time of the GC-2DPCA is similar to those of the 2DPCA-L1-nongreedy. Although our method is slower than the classic 2DPCA, it is better than other methods with respect to the run consumption time. Table III and Fig. 7 illustrate that our proposed method is fast and robust.

In Table III, we find the following results.
1) The operation of the GC-2DPCA is more time consuming than the traditional 2DPCA. The 2DPCA only needs one-step decomposition, but the GC-2DPCA has a closed-form solution and local convergence using a nongreedy algorithm.

2) In all nongreedy strategies, the 2DPCA-L1-nongreedy is better than the GC-2DPCA. That is because both algorithms make full use of the row information of the image matrix. Meanwhile, the GC-2DPCA considers more optimization strategies than the 2DPCA-L1-nongreedy.

3) In all nongreedy strategies, the GC-2DPCA is better than the F-2DPCA. The better performance is attributed to the faster calculation of the vector form compared to the matrix form.

Figs. 8–10 present the reconstruction error of the seven methods in the ten experiments for the three datasets. Table IV lists the average reconstruction error and the corresponding standard deviation of the 2-D methods on three databases. The three datasets have random noise added 10 times each, and each algorithm is run on these datasets. From these results, it can be seen that our method is generally superior to the other methods.

From Table IV, we can observe the following.

1) The 2DPCA is inferior overall to the other six methods. The main reason is that the 2DPCA employs the squared $F$-norm as a distance metric. The robustness of the 2DPCA is affected. The GC-2DPCA algorithm we proposed performs better than the other methods among the datasets.

2) Our method is overall superior to the F-2DPCA. It is likely due to the image’s covariance matrix, while F-2DPCA only directly uses the image’s covariance matrix.

3) Combined with Table I, when we set $p = 1$, the GC-2DPCA is overall superior to the F-2DPCA for both recognition and reconstruction tasks. It is because it can further obtain robust and efficient solutions according to the characteristics of the data. We also discussed earlier that most existing $l_{2,1}$-norm-based 2DPCA methods are special cases of our GC-2DPCA.
Table V shows the significant of the results in Table IV by employing t-tests. The null hypothesis is that the average reconstruction errors of the GC-2DPCA and F-2DPCA are from the same distribution at a 0.05 significance level. From Table V, we can observe that: $t = 1$, sig = $8.4e-12$, thereby indicating the null hypothesis is rejected at the 0.05 significance level. It is shown that the results of the algorithm GC-2DPCA can be statistically determined to be larger than the results of the F-2DPCA (i.e., the comparison of the means of the data of the two groups is meaningful in Table IV). The results of the algorithm GC-2DPCA and the algorithm 2DPCA-L1 have the same situation on CMU PIE and AR but have not Extended Yale B, where the results of the GC-2DPCA and 2DPCA-L1 are very similar. We believe this is because the L1-based method has some robustness advantages, especially with the Extended Yale B dataset. The t-test further proves that our method is superior to the other benchmarks in Table IV.

In this article, we first obtain the reconstruction weight and then obtain the recognition weight by the transformation formula in Algorithm 1, namely, first order. If we change the order of the solution, namely, second order, then the recognition performance remains almost the same as that in Table VI, but the reconstruction performance is degraded. The reason is that the latter depends on (2), and the transformation formula will cause (2) and (8) to be inconsistent, which will inevitably affect performance. Algorithm 2 has the same situation. Therefore, we adopt the first order.

### VI. Conclusion

In this article, we proposed two novel PCA and 2DPCA dimensionality reduction models, namely, the C-PCA and GC-2DPCA, for data classification and representation. They use the centered weights to measure the similarity of the data points so that the structural information of the data can be preserved. In addition, the GC-2DPCA model retains the rotational invariance and obtains generalized $\ell_2,1$-norm performance. The C-PCA and C-AnglePCA show that our method is easily extended to corresponding upgraded versions of the existing methods. To handle the GC-2DPCA model, we proposed an efficient iteration algorithm, which has a closed-form solution at each iteration. The experimental results on the CMU-PIE, Extended Yale B, and AR databases have illustrated the robustness and effectiveness of our proposed methods.

### ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the Associate Editor for their constructive comments and suggestions. They would also like to thank Q. Wang for providing the code of Angle-PCA and F-2DPCA.
REFERENCES


Gongzhu Zhou (S’18) was born in Chuzhou, China, in 1995. He received the bachelor’s degree in computational mathematics from the Tianjin University of Science and Technology, Tianjin, China, in June 2017. He is currently pursuing the master’s degree in computer technology with the College of Intelligence and Computing, Tianjin University, Tianjin, China.

His current research interests include dimensionality reduction, pattern recognition, and machine learning.

Guangquan Xu (M’18) received the Ph.D. degree in computer application technology from Tianjin University, Tianjin, China, in 2008. He is a Full Professor with the Tianjin Key Laboratory of Advanced Networking, College of Intelligence and Computing, Tianjin University. He is also a Distinguished Professor with the International Business School (Big Data School), Qingdao Huaqian University, Qingdao, China. His current research interests include cyber security and trust management.

Prof. Xu is a member of CCF and ACM.
Jianye Hao received the B.E. degree in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2008, and the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Associate Professor with the School of Software, Tianjin University, Tianjin, China. He was a Postdoctoral Research Fellow with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, and Pillar of Information System Technology and Design (ISTD), Singapore University of Technology and Design (SUTD), Singapore, from 2013 to 2015. He was also a Lecturer with ISTD, SUTD, from 2014 to 2015. He was a Visiting Research Fellow with the National University of Singapore, Singapore, and the University of Wollongong, Wollongong, NSW, Australia, in 2011 and 2014, respectively. His current research interests include artificial intelligence, multiagent systems, machine learning, and game theory.

Shizhan Chen (M’10) received the B.S., M.S., and Ph.D. degrees in computer science and technology from Tianjin University, Tianjin, China, in 1998, 2004, and 2010, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, Tianjin University. He is also leading a research project supported by the National Natural Science Foundation of China. His current research interests include service computing and mobile Internet. Dr. Chen is a reviewer of some international conferences and journals.

Junfeng Xu received the Ph.D. degree in computer science and technology from the Dalian University of Technology, Dalian, China, under the supervision of Prof. K. Li. He is an Associate Research Fellow of cyber security and cyber threat detection with the China Information Technology Security Evaluation Center, Beijing, China. His current research interests include mobile computing security and Internet of Things (IoT) security, study of reverse engineering and analysis of mobile apps, IoT firmware and IoT apps, and online network monitoring.

Xi Zheng (M’12) received the bachelor’s degree in computer information system from Fudan University, Shanghai, China, and the master’s degree in computer and information science from the University of New South Wales, Sydney, NSW, Australia. He is a Software Engineer with the University of Texas at Austin, Austin, TX, USA, and a Chief Solution Architect with Menulog, Sydney, NSW, Australia, from 2005 to 2012 (the company sold for $8.55 billion). He is currently an Assistant Professor/Lecturer of software engineering with Macquarie University, Macquarie Park, NSW, Australia. He has published over 40 high-quality publications in top journals and conferences, such as PerCOM, ICSE, WWW Journal, the IEEE INTERNET OF THINGS JOURNAL, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE SYSTEMS JOURNAL, and the ACM Transactions on Embedded Computing Systems. His current research interests include service computing, Internet of Things security, and reliability analysis.

Mr. Zheng was a recipient of the Best Paper in Australian Distributed Computing and Doctoral Conference in 2017, and the Deakin Research Outstanding Award in 2016. He is a reviewer for top journals and conferences, such as IEEE Systems Journal, ACM Transactions on Design Automation of Electronic Systems, Pervasive and Mobile Computing, IEEE TRANSACTIONS ON CLOUD COMPUTING, and PerCOM.