K-MEDIANS CLUSTERING BASED l_1-PCA MODEL

Shu Yan LAM and Siu Kai CHOY

Department of Mathematics and Statistics, Hang Seng Management College

ABSTRACT

Principal Component Analysis (PCA) is one of the most widely used tools for the representation of high-dimensional data. Many different versions have been proposed to enhance the robustness of the model. Most of these ideas are not median based formulation, which is always a robust estimator in statistics. In this paper, we attempt to design a new median based PCA model based on *k*-medians clustering, for which each principal component is always the spatial median of the projected space. We prove that the proposed method converges. We also compare the proposed method with several state-of-the-art methods including l_1-PCA, RPCA and RPCA-OM. Experimental results show that the proposed *k*-medians clustering based PCA performs the best in many cases.

Index Terms— *k-medians*, Clustering, PCA, image reconstruction, dimensionality reduction

1. INTRODUCTION

Principal Component Analysis (PCA) [1-2] is one of the most prominent learning tools for linear data transformation techniques in image processing and machine learning. It has been widely used for the representation of high-dimensional data such as image data for appearance, shape, and visual tracking and is also popularly used as a preprocessing step to project high-dimensional data into a low-dimensional subspace. However, PCA also has limitations. Since large errors will dominate the mean square error (MSE), original PCA is prone to the presence of outliers that are significantly far away from the rest of the data points [3–5].

Many robust PCA methods have been proposed to handle the problem of outliers. One of the most extensive studies is the l_1 norm PCA (or l_1-PCA) model [6-10] as l_1 norm is a robust solver to the problem of outliers. Many different ways have been proposed to solve the l_1-PCA model such as linear programming [8]. However, most of these techniques do not preserve the rotation invariant property of the output projection matrix. Kwak reformulated the l_1-PCA model using the dual space technique, which can find a local optimal solution swiftly and well preserve the rotational invariant property of the output projection matrix [12]. Other than l_1-PCA model, Candes *et. al*. proposed the low-rank matrix factorization, namely robust PCA (RPCA) [13]. Its essential idea is to polish the input data matrix $\mathbf{X}$ in the sense of l_1 norm so that the rank of the matrix $\mathbf{X}$ reaches minimum. This method gives excellent results especially when the input data matrix is heavily corrupted by "salt and pepper" noise. More recently, Nie *et. al*. proposed a robust PCA based on an optimal mean formulation (RPCA-OM) [14]. The key contribution of this work is to find the correct mean of the input data matrix $\mathbf{X}$ by a nice optimization technique. This method outperforms many existing PCA methods.

Although most PCA methods perform excellently in many applications and are robust to outliers, they are not median based formulation. In this work, we attempt to design a new PCA model based on median modelling. Statistically, median plays an important role not only its robustness to outliers but also always serves as a good estimator [14]. The proposed median model is designed based on our mathematically proved result: the local optimal solutions of l_1-PCA model can be formulated as the local optimal solution of a two-group *k*-means clustering model [15]. Replacing the *k*-means clustering model by the *k*-medians clustering model would produce a median based PCA model. The *k*-medians clustering model finds the medians of cluster centers rather than means in *k*-means clustering model. We prove that the proposed method converges. We also compare the proposed method with several state-of-the-art methods including l_1-PCA, RPCA and RPCA-OM. Experimental results show that the proposed *k*-medians clustering based PCA performs among the best in two applications, namely image reconstruction and dimension reduction. The organization of this paper goes as follows. We will first review the l_1-PCA model and its equivalent *k*-means formulation. Then, the proposed *k*-medians clustering based PCA model will be introduced. Finally, the robustness of the proposed method is verified via experiments.

2. REVIEW OF l_1-PCA MODEL

In this section, we review the l_1-PCA model [12] and our proved result: l_1-PCA model can be reformulated as a two-group clustering model [16].

Kwak reformulated the l_1-PCA model using the dual space technique [12] and implicitly implies that the projection vectors are the mean of the signed data. Given the data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n]$ and the centroid $\boldsymbol{\mu}$, the l_1-PCA model minimizes the following objective function

$$\max_{\|\mathbf{u_j}\|_2=1} \sum_{i=1}^{n} |(\mathbf{x_i}-\boldsymbol{\mu})^T\mathbf{u_j}| \quad \text{for } j = 1,2 \dots D. \tag{1}$$

$\mathbf{u_j}$ is the jth principal component vector. Its dual space formulation is

$$\max_{\|\mathbf{u_j}\|_2=1} \sum_{i=1}^{n} |(\mathbf{x_i}-\boldsymbol{\mu})^T\mathbf{u_j}| = \max_{\|\mathbf{u_j}\|_2=1} \max_{s_i\in\{-1,1\}} \sum_{i=1}^{n} s_i(\mathbf{x_i}-\boldsymbol{\mu})^T\mathbf{u_j}. \tag{2}$$

The local optimal solution of this model is then given by

$$\mathbf{u_j} = \frac{\sum_{i=1}^{n} s_i(\mathbf{x_i}-\boldsymbol{\mu})}{\|\sum_{i=1}^{n} s_i(\mathbf{x_i}-\boldsymbol{\mu})\|} = \frac{\frac{1}{n}\sum_{i=1}^{n} s_i(\mathbf{x_i}-\boldsymbol{\mu})}{\left\|\frac{1}{n}\sum_{i=1}^{n} s_i(\mathbf{x_i}-\boldsymbol{\mu})\right\|}, \tag{3}$$

$$\text{and } s_i = \text{sign}\left((\mathbf{x_i}-\boldsymbol{\mu})^T\mathbf{u_j}\right) \quad \text{for } i = 1,2,\dots n. \tag{4}$$

The numerator of the right hand side of Equation (4) is the mean of the signed data $[\mathbf{x_1}-\boldsymbol{\mu}, \mathbf{x_2}-\boldsymbol{\mu}, \dots \mathbf{x_n}-\boldsymbol{\mu}]$.

The mean of the signed data implies that the l_1-PCA model is equivalent to a two-group clustering model [16]. That is, the l_1 −PCA model can be formulated as:

$$\min_{\mathbf{c_k}} \sum_{k=1}^{2}\sum_{i=1}^{2n} I(\mathbf{y_i}\in C_k)\|\mathbf{y_i}-\mathbf{c_k}\|^2, \tag{5}$$

where $\mathbf{c_1} = -\mathbf{c_2}$, $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \dots \mathbf{y_{2n}}] = [\mathbf{x_1}-\boldsymbol{\mu}, \dots \mathbf{x_n}-\boldsymbol{\mu}, -(\mathbf{x_1}-\boldsymbol{\mu}), \dots -(\mathbf{x_n}-\boldsymbol{\mu})]$, $I(\mathbf{y_i}\in C_k)$ is the indicator function and $C_k = \{\mathbf{y_i}: \|\mathbf{y_i}-\mathbf{c_k}\| \le \|\mathbf{y_i}-\mathbf{c_j}\| \text{ for } j = 1,2\}$. $I(\mathbf{y_i}\in C_k) = 1$ if $\mathbf{y_i}\in C_k$ and $I(\mathbf{y_i}\in C_k) = 0$. The principal component is then obtained by normalizing $\mathbf{c}$, $\mathbf{u_j} = \frac{\mathbf{c_1}}{\|\mathbf{c_1}\|}$.

3. K-MEDIANS BASED PCA

In this section, we use *k-medians* clustering model to develop a new PCA model. *k-medians* clustering model has been developed as a substitute of *k*-means clustering model to find cluster centers [17-20]. It has been well-tested that *k*-medians clustering model is much more powerful than *k*-means clustering model especially when outliers are present.

The originality of *k-medians* clustering model is to replace the squared l_2 norm in the *k*-means clustering model by l_2 norm. This leads to the *k-medians* clustering model

$$\min_{\mathbf{c_k}} \sum_{k=1}^{2}\sum_{i=1}^{2n} I(\mathbf{y_i}\in C_k)\|\mathbf{y_i}-\mathbf{c_k}\|. \tag{6}$$

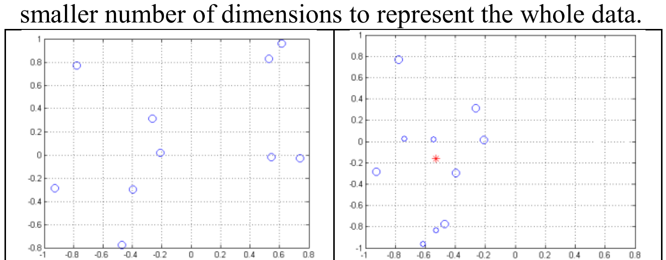
This proposed model is to find the spatial median $\mathbf{c}$ of the signed data $g_i(\mathbf{x_i}-\boldsymbol{\mu})$ where $\mathbf{c} = \mathbf{c_1} = -\mathbf{c_2}$ and $g_i = 1 \text{ or } -1$. This is justified as below.

$$\sum_{k=1}^{2}\sum_{i=1}^{2n} I(\mathbf{y_i}\in C_k)\|\mathbf{y_i}-\mathbf{c_k}\|$$

$$= \sum_{i=1}^{2n} I(\mathbf{y_i}\in C_1)\|\mathbf{y_i}-\mathbf{c}\| + \sum_{i=1}^{2n} I(\mathbf{y_i}\in C_2)\|\mathbf{y_i}+\mathbf{c}\|$$

$$= \sum_{i=1}^{2n}\|s_i\mathbf{y_i}-\mathbf{c}\| = 2\sum_{i=1}^{n}\|g_i(\mathbf{x_i}-\boldsymbol{\mu})-\mathbf{c}\|.$$

where $s_i = 1 \text{ or } -1$. The second equality holds because $\mathbf{y_i} = \mathbf{y}_{n+i}$. The third equality holds because $g_i = 1 \text{ or } -1$ and $\mathbf{Y}$ is the signed data $\mathbf{X}$. In robust statistics, the above model is to find the spatial median of $g_i(\mathbf{x_i}-\boldsymbol{\mu})$ [5]. It has been proved that this spatial median model is robust to the presence of outliers and it has 50% breakdown point property [21,22]. That is, the cluster centers would not be dragged far away from the majority even if half of the data are corrupted. This spatial median model has an implicit meaning that $\mathbf{c}$ is the "closest" coordinate vector to all data points $(\mathbf{x_i}-\boldsymbol{\mu})$ flipping to same half-plane. This is further illustrated via Figure 1. 10 two-dimensional random points with centroid 0 are generated. They are shown in Figure 1(a) The signed data points together with the spatial median (marked as red *) are shown in Figure 1(b). The sign function g_i is to flip all data points to the same half-plane. The spatial median is the "closest" coordinate vector to all these signed data points in the same half-plane. Projecting the data onto the direction $\mathbf{c}$ means projecting all data points onto the line, which is always "closest" to all data points. This property allows the projected data "looks like" the original data. In image reconstruction and dimensionality reduction applications, this feature allows the use of a smaller number of dimensions to represent the whole data.



(a) Original Data | (b) Signed data with spatial median

Figure 1. Illustration of the proposed idea.

To regularize the origin of the l_2 norm so that it is differentiable, we modify the proposed model as

$$\min_{\mathbf{c_k}} \sum_{k=1}^{2}\sum_{i=1}^{2n} I(\mathbf{y_i}\in C_k)\|\mathbf{y_i}-\mathbf{c_k}\|_\epsilon, \tag{7}$$

where $\|\mathbf{x}\|_\epsilon = \sqrt{\epsilon + \||\mathbf{x}|\|^2}$. ϵ is served as a regularization parameter smoothing the l_2 norm function. In all our experiments, $\epsilon = 10^{-4}$. Alterative updating scheme is used to find the local optimal solution for the cluster centers $\mathbf{c_k}$. By taking the first derivative of this equation, the cluster center must satisfy the following equation

$$\sum_{i=1}^{2n} I(\mathbf{y_i}\in C_k)\frac{\mathbf{c_k}-\mathbf{y_i}}{\|\mathbf{c_k}-\mathbf{y_i}\|_\epsilon} = 0 \quad \text{for } k = 1,2. \tag{8}$$

This leads to the following update equation

$$\mathbf{c}_k^t = \frac{\sum_{i=1}^{2n} I(\mathbf{y_i}\in C_k^{t-1})\frac{\mathbf{y_i}}{\|\mathbf{c}_k^{t-1}-\mathbf{y_i}\|_\epsilon}}{\sum_{i=1}^{2n} I(\mathbf{y_i}\in C_k^{t-1})\frac{1}{\|\mathbf{c}_k^{t-1}-\mathbf{y_i}\|_\epsilon}} \tag{9}$$

for k = 1,2.

The principal component can then be obtained by the normalization $\mathbf{u} = \frac{\mathbf{c}_1}{\|\mathbf{c}_1\|}$. Later, we will show that $\mathbf{c}_1 = -\mathbf{c}_2$ upon convergence of the proposed algorithm.

Input: Centralized Data matrix $\mathbf{X_c} \in \mathrm{R}^{d\times n}$, a small positive value τ and an initial guess $\mathbf{C} = [\mathbf{c}_1^T \; -\mathbf{c}_1^T]^T$.
Output: An unit vector $\mathbf{u} \in \mathrm{R}^{d\times 1}$
1. Construct a new data matrix $\mathbf{Y} = [\mathbf{x}_1^c, \mathbf{x}_2^c, \ldots \mathbf{x}_n^c, -\mathbf{x}_1^c, -\mathbf{x}_2^c, \ldots -\mathbf{x}_n^c]$.
2. Compute the two indicator functions $\mathrm{I}(\mathbf{y}_i \in \mathrm{C}_k)$
3. Update the cluster centers according to Equation (9) until convergence
4. If the l_2 norm difference of the cluster centers between two consecutive iterations is smaller than τ, output the result as $\mathbf{u} = \mathbf{c}_1/\|\mathbf{c}_1\|$. Otherwise, go to Step 2.

Table 1. Algorithm for producing a principal component.

Table 1 shows the algorithm of the proposed method, which first principal component of the data. The initial guess $\mathbf{c}_1$ is set as $\lambda\mathbf{u}$, where $\mathbf{u}$ is the singular vector corresponding to largest singular value λ of the centralized data. For the p-th principal component ($p \geq 2$), the data $\mathbf{X}_c(p) = [\mathbf{x}_1^c(p), \mathbf{x}_2^c(p), \ldots, \mathbf{x}_n^c(p)]$ is projected to a subspace by the following equation

$$\mathbf{x}_i^c(p) = (\mathbf{I} - \mathbf{u}(p-1)\mathbf{u}(p-1)^T)\mathbf{x}_i^c(p-1), \quad (10)$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{u}(p-1)$ is the $(p-1)$-th principal component and $\mathbf{x}_i^c(0) = \mathbf{x}_i^c$. Then, the algorithm is applied to the data set $\mathbf{X}_c(p)$. The output $\mathbf{u}(p)$ of the algorithm is the p-th principal component and it is orthonormal to other principal components.

The following lemma show that the equality $\mathbf{c}_1 = -\mathbf{c}_2$ must hold upon convergence of the algorithm.

Lemma: Consider the following update equation

$$\mathbf{c}_k^t = \frac{1}{\sum_{i=1}^n \mathrm{I}(\mathbf{y}_{i,k} \in \mathrm{C}_k^{t-1})} \sum_{i=1}^n \mathrm{I}(\mathbf{y}_{i,k} \in \mathrm{C}_k^{t-1}) \frac{\mathbf{y}_{i,k}}{\|\mathbf{y}_{i,k} - \mathbf{c}_k^{t-1}\|}, \quad \text{for } k = 1,2 \quad (11)$$

where $\mathbf{y}_{i,k} \in \mathrm{C}_k^{t-1}$, $\mathbf{c}_k^t$ is the cluster center $\mathbf{c}_k$ and C_k^{t-1} is the set C_k at $t-$th iteration. If $\mathbf{c}_1^{t-1} = -\mathbf{c}_2^{t-1}$, $\mathbf{c}_1^t = -\mathbf{c}_2^t$.

Proof: $\mathbf{Y}$ is the union of $\mathbf{X}$ and $-\mathbf{X}$. As $\mathbf{c}_1^{t-1} = -\mathbf{c}_2^{t-1}$, $\mathrm{C}_1^{t-1} = -\mathrm{C}_2^{t-1}$. This is justified as below. $\mathbf{y}_{i,1}$ is closer to $\mathbf{c}_1^{t-1}$ than $\mathbf{c}_2^{t-1} = -\mathbf{c}_1^{t-1}$. Then, $-\mathbf{y}_{i,1} = \mathbf{y}_{i,2}$ must be closer to $\mathbf{c}_2^{t-1} = -\mathbf{c}_1^{t-1}$ than $\mathbf{c}_1^{t-1}$. This implies that C_2 is the negative of C_1.

$$\begin{aligned}
\mathbf{c}_1(t) &= \frac{1}{\sum_{i=1}^n \mathrm{I}(\mathbf{y}_{i,1} \in \mathrm{C}_1^{t-1})} \sum_{i=1}^n \mathrm{I}(\mathbf{y}_{i,1} \in \mathrm{C}_1^{t-1}) \frac{\mathbf{y}_{i,1}}{\|\mathbf{y}_{i,1} - \mathbf{c}_1^{t-1}\|} \\
&= \frac{-1}{\sum_{i=1}^n \mathrm{I}(-\mathbf{y}_{i,1} \in -\mathrm{C}_1^{t-1})} \sum_{i=1}^n \mathrm{I}(-\mathbf{y}_{i,1} \in -\mathrm{C}_1^{t-1}) \frac{-\mathbf{y}_{i,1}}{\|-\mathbf{y}_{i,1} + \mathbf{c}_1^{t-1}\|} \\
&= \frac{-1}{\sum_{i=1}^n \mathrm{I}(\mathbf{y}_{i,2} \in \mathrm{C}_2^{t-1})} \sum_{i=1}^n \mathrm{I}(\mathbf{y}_{i,2} \in \mathrm{C}_2^{t-1}) \frac{\mathbf{y}_{i,2}}{\|\mathbf{y}_{i,2} - \mathbf{c}_2^{t-1}\|} \\
&= -\mathbf{c}_2(t).
\end{aligned} \quad (12)$$

□

As this lemma is true for all $t \geq 1$, the relationship $\mathbf{c}_1 = -\mathbf{c}_2$ will stay true for any initial guess satisfying this condition until convergence. The following theorem proves that the proposed clustering algorithm converges.

Theorem: The proposed clustering algorithm converges.

Proof: Define

$$\mathrm{J}(\mathrm{C}_k, \mathrm{c}_k) = \sum_{k=1}^{2} \sum_{i=1}^{2n} \mathrm{I}(\mathbf{y}_i \in \mathrm{C}_k) \|\mathbf{y}_i - \mathbf{c}_k\|_\epsilon.$$

The function $f(\mathbf{c}_k) = \|\mathbf{y}_i - \mathbf{c}_k\|_\epsilon$ is convex. So, $\mathbf{c}_k$ which satisfies Equation (8) is the global minimizer of $\mathrm{J}(\mathrm{C}_k, \mathrm{c}_k)$. Thus, $\mathrm{J}(\mathrm{C}_k^{t-1}, \mathrm{c}_k^{t-1}) \geq \mathrm{J}(\mathrm{C}_k^{t-1}, \mathrm{c}_k^t)$. Obviously, $\mathrm{J}(\mathrm{C}_k^{t-1}, \mathrm{c}_k^t) \geq \mathrm{J}(\mathrm{C}_k^t, \mathrm{c}_k^t)$. Thus, $\mathrm{J}(\mathrm{C}_k^{t-1}, \mathrm{c}_k^{t-1}) \geq \mathrm{J}(\mathrm{C}_k^t, \mathrm{c}_k^t)$. The objective function is bounded below. The method converges. □

6. EXPERIMENTAL RESULTS

In this section, we verify the robustness of the proposed method and compare its performance with the original PCA (Orig-PCA) and three state-of-the-art PCAs: l_1-PCA, RPCA [13] and RPCA − OM [14] in two different applications: image reconstruction and dimension reduction. We implemented the l_1-PCA method while the programming codes of the other two methods are downloaded from the websites [23,24]. We used the default settings of the downloaded codes. As RPCA-OM is a relatively time consuming method, we only consider the projection onto 14 different dimensions for experimental evaluations, which Nie *et. al.* used a similar strategy [14]. These dimensions are 5, 10, …65, 70. However, even under this setting, RPCA-OM can take over an hour to obtain the 14 projection matrices. The stopping criterion of the l_1-PCA method and the proposed method was set if the difference between norms of projection matrix $\mathbf{U}$ in consecutive iterations was less than 10^{-4}. The data mean μ of the l_1-PCA and proposed method was set as the spatial median of the input data $\mathbf{X}$. The l_1-PCA used the singular vector corresponding to largest singular value of the data as the initial guess. All the methods were tested under the Matlab R2010b. Intel® Core ™ i5-3450 CPU @3.10GHz 3.10GHz 8.00 GB 64 bit Windows 8.1.

Four publicly available datasets were selected for performance evaluation. The basic information of these four datasets are summarized in the following table.

Database Name	Dimensions	Number of samples
JAFFE	91x76	213
Yale Face	64×64	165
Umist	79x65	575
TDT2 Document Database	500	2000

Table 2. Basic information of the selected databases.

A. Face Reconstruction

The average reconstruction error is defined by the averaged distance between an original unoccluded image and the reconstructed image as below:

$$\mathrm{err}(\mathrm{m}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left\|\left(\mathbf{x}_{\mathrm{i}}^{\mathrm{orig}} - \boldsymbol{\mu}\right) - \sum_{\mathrm{j}=1}^{\mathrm{m}}\mathbf{u}_{\mathrm{j}}\mathbf{u}_{\mathrm{j}}^{\mathrm{T}}(\mathbf{x}_{\mathrm{i}} - \boldsymbol{\mu})\right\|, \qquad (13)$$

where $\boldsymbol{\mu}$ is the centroid, $\mathbf{x}_{\mathrm{i}}^{\mathrm{orig}}$ and $\mathbf{x}_i$ are the original and the corresponding occluded image respectively and m is the number of principal components. The range of m is from 1 to 70. We simulate the outliers as below:

(1) Corrupting Image as outliers: we randomly corrupted images by randomly blocking part of the images by 30×30 blocks, which are black and white dots.

(2) Adding dummy images as outliers: We added dummy images to the image databases. The dummy images are images from other databases or images with purely black and white dots.

Figure 2 shows the reconstruction errors of different PCA methods. When the number of principal components is greater than 30, the difference among different methods becomes significant and the proposed method and RPCA perform better than other methods and achieve the smallest average reconstruction errors. In the Yale face database, the proposed method apparently performs better than the second best method - RPCA.

B. Dimension Reduction

The dimensions of all samples are reduced using PCA methods and classifications were performed in the reduced subspace [25]. Prototype $\mathbf{P}$ for each class and the probe $\mathbf{T}$ are projected onto $\mathbf{U}$. The class is found to minimize the distance $\epsilon = ||\mathbf{U}(\mathbf{T} - \mathbf{P})||$.

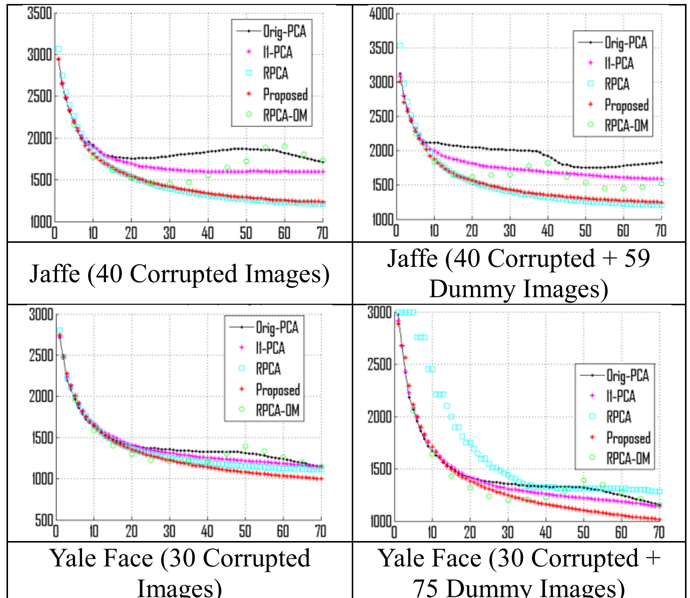
(1) Face Databases (JAFFE, Yale Face and Umist): We randomly selected two images from each class for testing and the rest were used for training. This is repeated five times. 60 dummy images were added to all databases as outliers. The dummy images are images from other databases or images with purely black and white dots.

(2) TDT2 Database: We selected the top 10 categories for our experimental evaluation. Each document is represented as a normalized term-frequency vector, with top 200 words selected according to mutual information. We randomly selected 500 documents for training and the rest were used for testing. This is repeated 20 times.

For the classification rates of each PCA methods, we use the area under curves (AUC) and the maximum classification rates as experimental evaluations. They are shown in Table. AUC refers to the overall classification rate of the PCA method. A large AUC implies a better performance. The best performance is in bold face type. Apparently, the proposed method performed the best in terms of AUC in all face databases (JAFFE, Yale Face and Umist). This implies that the projection obtained by the proposed PCA often gives better classification results. For the TDT2 database, the maximum classification rate of the proposed method is the best, which is almost 1% better than the second best method $\mathrm{RPCA-OM}$.

7. CONCLUSIONS

In this work, we proposed a new PCA method based on k-medians clustering model. Its essential idea is to project the input data matrix in the sense of spatial median. We have proved that the proposed two group k-medians clustering model converges. We compared the performance of the proposed method with several state-of-the-arts methods in two different applications, namely, image reconstruction and dimension reduction. Experimental results show that the proposed method performed among the best in many cases.



Jaffe (40 Corrupted Images) | Jaffe (40 Corrupted + 59 Dummy Images)

Yale Face (30 Corrupted Images) | Yale Face (30 Corrupted + 75 Dummy Images)

Figure 2. Average reconstruction errors of different PCA methods.

		Orig-PCA	l_1-PCA	RPCA	RPCA − OM	L2-norm
Yale Face	AUC	60.3067	60.4633	60.3067	60.4167	**60.8500**
	Max	0.9267	**0.9333**	0.9267	**0.9333**	0.9267
Jaffe	AUC	66.5000	66.8200	66.5000	66.3150	**67.0400**
	Max	0.9800	0.9900	0.9800	0.9900	0.9900
Umist	AUC	67.6700	67.6300	67.6700	66.9550	**67.8350**
	Max	1	1	1	1	1
TDT	AUC	63.9468	**64.0828**	40.1933	63.7253	63.8168
	Max	0.9539	0.9607	0.6371	0.9567	**0.9641**

Table 3. Area under the curves (AUCs) and maximum recognition rates of different PCA methods.

8. ACKNOWLEDGEMENTS

This work was partially supported by the RGC funding UGC/FDS14/E02/14.

9. REFERENCES

[1] K. Pearson, “On Lines and Planes of Closest Fit to Systems of Points in Space,” *London Edinburgh and Dublin Philosophical Magazine and Journal of Science*, Vol. 2, pp. 559–572, 1901.

[2] H. Hotelling, “Analysis of a Complex of Statistical Variables into Principal Components,” *Journal of Educational Psychology*, Vol. 24, pp. 417–441, 1933.

[3] N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, and K. Cohen, “Robust Principal Component Analysis for Functional Data,” *Test*, Vol. 8, No. 1, pp. 1–73, 1999.

[4] R. A. Maronna, “Principal Components and Orthogonal Regression Based on Robust Scales,” *Technometrics*, Vol. 47, pp. 264–273, 2005.

[5] A. Baccini, P. Besse, and A. Falguerolles, “A L_1-norm PCA and a Heuristic Approach,” *Ordinal Symbolic Data Analysis*, pp. 359–368, 1996.

[6] Q. Ke and T. Kanade, “Robust L_1-norm Factorization in the Presence of Outliers and Missing Data by Alternative Convex Programming,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 739–746, 2005.

[7] R. Dahyot, P. Charbonnier, and F. Heitz, “Robust Visual Recognition of Colour Images,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1685–1690, 2000.

[8] J. P. Brooksa, J.H. Duláb and E.L. Boonea, “A Pure L1-norm Principal Component Analysis,” *Computational Statistics & Data Analysis*, Vol. 61, pp. 83–98, 2010.

[9] F. Nie, H. Huang, C. Ding, D. Luo and H. Wang, "Robust Principal Component Analysis with Non-greedy l_1 -norm Maximization," *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, pp. 1433-1438, 2011.

[10] A. Ng, “Feature Selection, L_1 Versus L_2 Regularization and Rotational Invariance,” *in Proc. International Conference on Machine Learning*, pp. 78–86, 2004.

[11] C. Ding, D. Zhou, X. He, and H. Zha, “R1-PCA: Rotational Invariant L_1 -norm Principal Component Analysis for Robust Subspace Factorization,” *International Conference on Machine Learning*, pp. 281–288, 2006.

[12] N. Kwak, "Principal Component Analysis Based on l_1-norm Maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, pp. 1672 – 1680, 2008.

[13] E. Candes, X. Li, Y. Ma and John Wright, "Robust Principal Component Analysis?" *Journal of the ACM*, Vol. 58, Issues 3, Article No. 11, 2011.

[14] F. Nie, J. Yuan and H. Huang, "Optimal Mean Robust Principal Component Analysis ," *International Conference on Machine Learning*, pp. 1062-1070, 2014.

[15] Peter J. Huber, *Robust Statistics*, Wiley; 2 edition, 2009.

[16] B. S. Y. Lam and S. K. Choy, “Reformulating the l_1-PCA Model As a Two Group Clustering Problem, *Electronic Letters*, under review.

[17] G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, 2007.

[18] B. Mirkin, *Clustering: A Data Recovery Approach*, Chapman & Hall/CRC Computer Science & Data Analysis, 2012.

[19] C. Ding and X. He, "K-means Clustering via Principal Component Analysis," *International Conference on Machine Learning*, pp. 225-232, 2004.

[20] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, 1st edition, 2005.

[21] C. G. Small, "A Survey of Multidimensional Medians," *International Statistical Review / Revue Internationale de Statistique*, Vol. 58, pp. 263-277, 1990.

[22] J. B. S. Haldane, "Note on the Median of a Multivariate Distribution," *Biometrika* Vol. 35, pp. 414-417, 1948.

[23]http://perception.csl.illinois.edu/matrix-rank/sample_code.html

[24] http://www.escience.cn/people/fpnie/papers.html

[25] X. Wang, X. Tang, "A unified framework for subspace face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 26, No. 9, pp. 1222-1228, 2004.