Information Processing Letters 115 (2015) 812–816

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Combined supervised information with PCA via discriminative component selection

Sheng Huang [a,b,*], Dan Yang [b], Ge Yongxin [b], Xiaohong Zhang [a,b]

[a] Ministry of Education, Key Laboratory of Dependable Service Computing in Cyber Physical Society, Chonqing, 400044, China
[b] School of Software Engineering at Chongqing University, Chonqing, 400044, China

ARTICLE INFO

Article history:
Received 29 August 2013
Received in revised form 13 June 2015
Accepted 13 June 2015
Available online 18 June 2015
Communicated by S.M. Yiu

Keywords:
Algorithms
Design of algorithms
Principal Component Analysis (PCA)
Face recognition
Linear Discriminant Analysis (LDA)
Multivariate statistic analysis

ABSTRACT

Principal Component Analysis (PCA) is a classical multivariate statistical algorithm for data analysis. Its goal is to extract principal features or properties from data, and to represent them as a set of new orthogonal variables called principal components. Although PCA has obtained extensive successes across almost all the scientific disciplines, it is clear that PCA cannot incorporate the supervised information such as class labels. In order to overcome this limitation, we present a novel methodology to combine the supervised information with PCA by discriminatively selecting the components. Our method use the fisher criterion to evaluate the discriminative abilities of bases of original PCA and find the first n best ones to yield the new PCA projections. Clearly, the proposed method is general to all PCA family algorithms and even can be applied to other unsupervised multivariate statistical algorithms. Furthermore, another desirable advantage of our method is that it doesn't break the original structure of the PCA components and thereby keeps their visual interpretability. As two examples, we apply our method to incorporate the supervise information with PCA and Robust Sparse PCA (RSPCA) to improve their discriminative abilities. Experimental results on two popular databases demonstrate the effectiveness of our method.

1. Introduction

Principal Component Analysis (PCA) [1,2] is probably the most popular multivariate statistical technique for data processing and dimensionality reduction, and has wide range of applications almost throughout all the scientific disciplines. Essentially, PCA aims at learning a subspace spanned by a set of mutual orthogonal bases called Principal Components (PCs) along which data variance can be maximally preserved. In such PCA subspace, the structure of the input data can be effectively captured.

Although PCA has obtained extensive successes [1–7], PCA has an obvious drawback that it cannot utilize class labels to improve its discriminative ability and further benefit the solutions to the supervised issues. Currently, there are two common ways to address this problem. The first one is to perform another supervised projection, such as Linear Discriminant Analysis (LDA) [8] and Locality Preserving Projections (LPP) [9], in PCA space. The second way is to put the class label to the end of sample vector as the additional dimension [5]. Although many studies have proved that the previous approaches can significantly improve the discriminative ability of PCA, these approaches clearly break the original structures of the components. However, preserving the original structure of components is very important, since the components are visually interpretable and represent some physic meanings in some

* Corresponding author at: School of Software Engineering at Chongqing University, Chonqing, 400044, China.
E-mail address: huangsheng@cqu.edu.cn (S. Huang).

http://dx.doi.org/10.1016/j.ipl.2015.06.010
0020-0190/

specific problems [4]. The reason why the PCs are interpretable is mainly due to the fact that each component is generally a linear combination of input observations (variables) [4,6]. For example, in biology, each involved variables may be corresponding to a specific gene. Thus, a PC may indicate a gene sequence in this case. Moreover, due to the improvements of the L1-norm based mathematical works [10], improving the interpretability of PCA component also becomes a recent hot topic in machine learning field [4,6,7].

In this letter, we intend to incorporate the supervised information with PCA by selecting the discriminative PCs based on the class labels. The main advantage of this method is that it doesn't suffer the problem as the previous approaches do. Thus, the physic meaning of the component can be kept. In almost all PCA applications, PCs corresponding to the larger eigenvalues are selected for constructing the final PCA projection. This is because the larger eigenvalue indicates more information preserved along the directions of relevant components. However, for a supervised problem, such as face recognition and gene expression classification, discriminative information preservation is more meaningful than whole information preservation. Therefore, we can improve PCA via re-ranking PCs according to their discriminant abilities. Furthermore, the amount of PCs is adequate to provide such potential, since it is equal to the rank of covariance matrix of data. Motivated by the successes of LDA and fisher score based feature selection [8,11–13], the well known fisher criterion [8] is adopted to evaluate the discriminative ability of each PC. After evaluation, each PC will achieve a confidence called *fisher score* and this confidence indicates its discriminant ability. Due to the mutually orthogonality of PCs, PCs can be considered as independent with each other. Thus, we can directly sort the PCs based on the fisher scores and select the d most discriminative PCs to yield the new PCA projections. We apply our method to PCA and a very recent PCA algorithm named Robust Sparse PCA (RSPCA) [4] to evaluate the effectiveness of our method. The experimental results from two popular databases demonstrate that we present a remarkable improvement to the PCA algorithms.

2. Methodology

We begin by introducing some notions. The $d \times n$ matrix $W = [w_1, \cdots, w_n]$ presents the whole PCA projection where the d-dimensional column vector w_i denotes the ith basis of PCA (Principal Component). Matrix $X = [x_1, \cdots, x_n] \subset \mathcal{R}^m$ is the sample matrix and the class labels are denoted as a vector $C = [1, 2, \cdots, p]$. Matrix $X_c, c \in C$ denotes the subset whose samples are belonging to class c. Matrix $Y = [y_1, \cdots, y_i, \cdots, y_p] \subset \mathcal{R}^m, i \in C$ denotes the mean space of samples where y_i is the mean of samples belonging to the class i.

Motivated by the successes of LDA and the fisher score based feature selections [8,11,12], we evaluate the discriminant ability of each PC using fisher criterion. The idea of Fisher criterion is derived from LDA. It measures the discriminative ability of each PC by computing the ratio of the trace of its between-class scatter matrix to the trace of its within-class scatter matrix. And this ratio is the so-called fisher score. Since the projected samples on each PC are all scalars, the between-class scatter matrix actually is the variance of the means of different classes, and the within-class scatter matrix is actually the sum of the variances of the homogenous samples. Therefore, the basis evaluation function is finally formulated as follows

$$\begin{aligned}\mathcal{F}(w_i) &= \frac{\sigma({w_i}^T Y)}{\sum_{c \in C} n_c \cdot \sigma({w_i}^T X_c) + \epsilon} \\ &= \frac{{w_i}^T (Y - \bar{Y})(Y - \bar{Y})^T w_i}{\sum_{c \in C} n_c {w_i}^T (X_c - \bar{X}_c)(X_c - \bar{X}_c)^T w_i + \epsilon}\end{aligned} \tag{1}$$

where $\sigma(t)$ is the variance of t and n_c indicates the sample number of class c. Matrix $\bar{Y}$ has the same size as matrix Y and each column is the column mean of the matrix Y. Similarly, matrix $\bar{X}_c$ is a same size matrix whose column is the column mean of the matrix X_c. ϵ is a very small positive constant for avoiding diving by zero.

The greater value of the numerator of Equation (1) indicates the larger distance between each two classes. The smaller value of the denominator of Equation (1) indicates the smaller distance between each two homogenous samples. Thus, it can be easily deduced that the larger fisher score means the stronger discriminating power of the relevant component. Moreover, according to Equation (1), clearly, Fisher criterion and LDA share the same objective function. The only difference between them is that the projection matrix W in LDA is treated as a variable while the projection matrix W in Fisher criterion is considered as an input. Therefore, LDA learns its own projections while Fisher criterion is deemed as a kind of evaluation metric which is employed to evaluate the discriminating power for a given projection (base).

The principal components can be considered as mutually independent under PCA framework, since they are orthogonal with each other. Thus, the most d-dimensional discriminative PCA projection is constructed by the components corresponding to the first d largest fisher scores. The detail of how to select discriminative components is described in Algorithm 1.

Algorithm 1 Selecting Discriminative Components.

Require:
The training data X;
The sample class labels L;
The original $d \times n$ PCA projections $W = [w_1, \cdots, w_n]$; The amount of selected discriminative components m where $m \leq n$;
Ensure:
The output $d \times m$ discriminative PCA projections D;
1: Define a temporary array F to store the fisher scores.
2: **for** each $i \in [1, n]$ **do**
3: Calculate the fisher score f of the ith component by Equation (1) with parameters of w_i, X and L;
4: Put the fisher score f into the ith entry of array F;
5: **end for**
6: Descendingly sort the fisher score array F, $[F, index] = SORT(F)$ where *index* indicates the new index of array after the sorting;
7: Re-rank the PCA projections W based on *index*, $W = W(index)$;
8: Put the first m components of re-ranked PCA projections into discriminative PCA projections D, $D = W(:, 1:m)$;
9: **return** D;

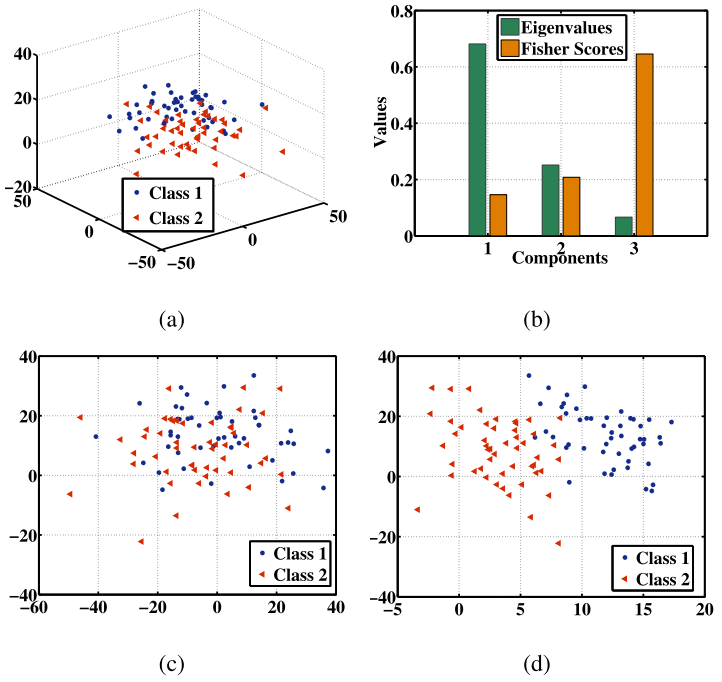

Fig. 1. The toy example for illustrating our works, (a) the samples of the two classes in sample space, (b) the fisher scores and eigenvalues corresponding to the different PCA components, (c) the samples in the original PCA space and (d) the samples in the new PCA space which is optimized by our method.

3. Toy examples

We conduct a toy experiment in a small toy dataset to illustrate our proposed method. This dataset is constructed by us and it contains two classes with fifty samples for each. Fig. 1(c) presents the distribution of the samples in the original PCA space while Fig. 1(d) presents the distribution of the samples in the optimized PCA space whose bases are discriminatively selected by our method. It is clear from these two figures that the optimized PCA space can better separate the samples. Furthermore, we also present the eigenvalue and fisher score of the components in Fig. 1(b). The previous observations all demonstrate that the discriminative ability of component is independent to the eigenvalue and a supervised component selection can improve the discriminative ability of PCA algorithms.

4. Experiments

Two popular face databases, ORL [14] and PIE [15], are used to evaluate the effectiveness of our method. The ORL database contains 400 images from 40 subjects [14]. Each subject has ten images acquired at different time. In this database, the subjects have varying facial expressions and facial details. We resize the face images to size 32×32 pixels. The PIE face database [15] contains 68 individual with 41368 face images as a whole. In this paper, we selected a subset (C27), containing 3060 images of 68 individuals (each individual has 45 images). The C27 subset involves variations in illumination,facial expression and the size of each image is 64×64 pixels.

In these experiments, the Nearest Neighbor (NN) Classifier is used as the default classifier and all the recognition rate means the top recognition accuracy. We use cross validation to evaluate our method on both PIE and ORL databases according to their sample number of each subject. In this letter, we take PCA and robust sparse PCA (RSPCA) as two cases and apply our method to improve their discriminative abilities. For convenience to talk, we name the new PCA with our improvement *Discriminative PCA* (DPCA) and name robust sparse PCA with our improvement *Discriminative Robust Sparse PCA* (DRSPCA).

Table 1 and Table 2 respectively present the recognition results of PCA and RSPCA before and after using our method on ORL and PIE face databases. From the observations, it is clear that our method can effectively improve the discriminative abilities of both these two PCA algorithms. With our improvement, PCA improves at least 1.75%, 2.42% accuracies and RSPCA improves at least 1.25%, 1.67% accuracies on ORL and PIE database respectively under all three different cross validation schemes. From the results, another interesting observed phenomenon is that the performance of PCA is much better than the one of RSPCA. We mainly attribute this to the different intentions of PCA of RSPCA. PCA is developed for pruning redundancy of information while RSPCA is developed for improving the robustness of PCA. In other words, all these algorithms have not considered the discriminating powers of projections. Therefore, although RSPCA is the enhanced version of PCA, RSPCA is still not better at face recognition in which the discriminating powers of projections play an important role. Moreover, RSPCA is a robust version of PCA. It may excessively emphasize the sparsity of the base which can lead to more losses of discriminative information.

Table 1
Recognition performance comparison (in percents) using ORL database.

Cross-validations	Mean recognition rate ± standard deviation		
	Two-fold	Three-fold	Five-fold
PCA	85.25% ± 0.35%	89.72% ± 1.27%	91.25% ± 3.19%
DPCA	87.75% ± 3.89%	92.23% ± 2.10%	93.00% ± 2.44%
Improvement	2.50%	2.51%	1.75%
RSPCA	82.00% ± 1.41%	89.72% ± 2.10%	92.25% ± 2.71%
DRSPCA	85.25% ± 2.47%	91.67% ± 3.63%	93.50% ± 2.85%
Improvement	3.25%	1.91%	1.25%

Table 2
Recognition performance comparison (in percents) using PIE database.

Cross-validations	Mean recognition rate ± standard deviation		
	Three-fold	Five-fold	Nine-fold
PCA	93.17% ± 3.46%	95.62% ± 5.52%	95.49% ± 8.16%
DPCA	97.42% ± 1.68%	98.04% ± 2.93%	97.91% ± 4.52%
Improvement	4.25%	2.42%	2.42%
RSPCA	90.16% ± 3.16%	94.41% ± 6.78%	94.77% ± 9.12%
DRSPCA	94.44% ± 5.45%	96.08% ± 4.99%	96.73% ± 6.15%
Improvement	4.28%	1.67%	1.96%

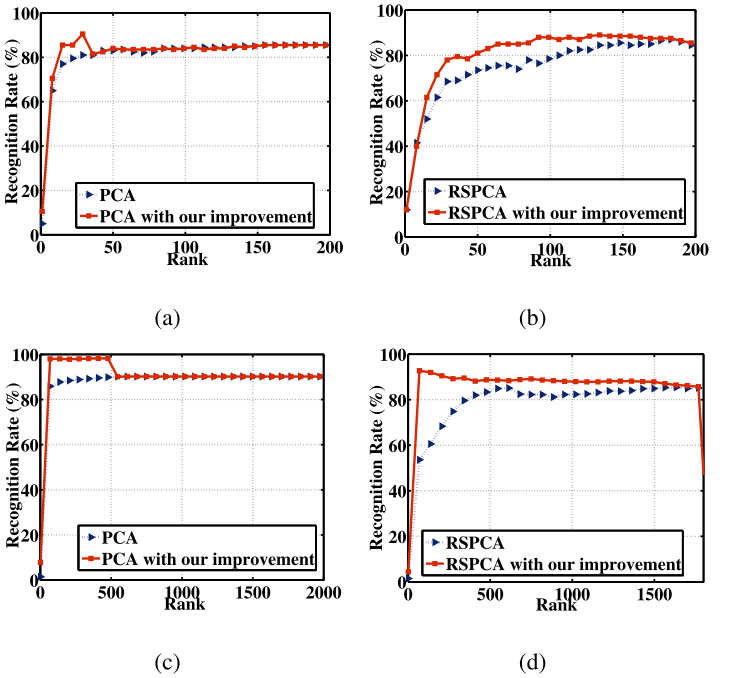

Fig. 2. The recognition rate versus the retained dimensions, (a) the results of PCA and DPCA on ORL database (5 trains), (b) the results of RSPCA and DRSPCA on ORL database (5 trains), (c) the results of PCA and DPCA on PIE database (30 trains) and (d) the results of RSPCA and DRSPCA on PIE database (30 trains).

We also conduct some experiments to plot the relationships between recognition rate and retained dimension of the PCA algorithms. As Fig. 2 shows, the PCA algorithms with our improvement consistently keep on top in all cases in comparison with their original version. Furthermore, with our improvement, PCA algorithms can more easily obtain the best recognition rate with retaining less dimension. This phenomenon verifies that the discriminative PCA projections can be constructed by a few of PCs and most of PCs are actually the redundancies for discriminating. Sometime, these redundancies may even corrupt the discriminating power of PCA. And we think this may be the main reason why the original PCA algorithms do not have the peaks of discriminating performance as same as their improved versions.

We draw the top six bases of PCA algorithms before and after component selection in Fig. 3. This experiment is conducted on ORL database. 200 samples with five sam-

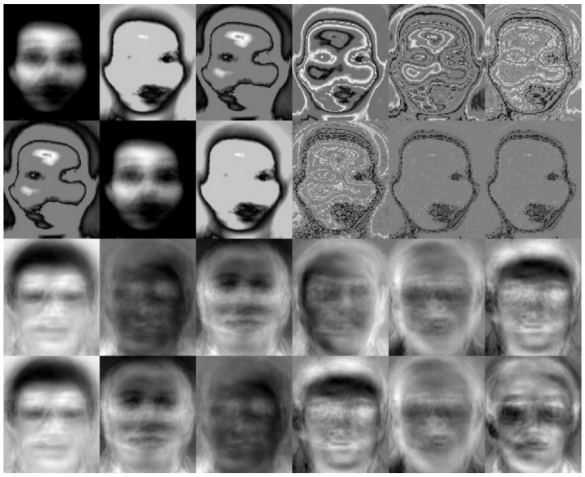

Fig. 3. The first and second rows are the top six bases of RSPCA and DRSPCA respectively, the third and forth rows are the top six bases of PCA and DPCA.

ples for each subject are used to training the PCA algorithms. Clearly, the top six bases corresponding to the top six largest eigenvalues are different to the top six most discriminative bases. It verifies the assumption we raised previously that the largest eigenvalue doesn't always mean the best discriminating power.

5. Conclusion

In this letter, we present a novel methodology to incorporate the supervised information with PCA algorithms. Different to the traditional ways, our method utilize the class labels to select discriminative components from whole PCs and yield them as a new PCA projections. Motivated by the successes of LDA and fisher score based feature selection, we use the fisher criterion to evaluate the discriminative ability of each component. After evaluation, each component will obtain a confidence named fisher score which indicates the discriminating power of the component. Therefore, we can re-rank these components according to these fisher scores and select the most discriminative components. The main advantage of our method is that it doesn't break the original structures of components. For this reason, the semantics of the components can be kept. Apparently, our method is not only general to PCA algorithms but also general to the other unsupervised multivariate statistical algorithms.

Acknowledgements

This work has been supported by the Fundamental Research Funds for the Central Universities (No. CDJXS11181162). The authors would thank Dr. Menglin Jiang and Dr. JingJing Liu for their useful suggestions.

References

[1] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2005.
[2] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.
[3] A. Eftekhari, M. Forouzanfar, H. Abrishami Moghaddam, J. Alirezaie, Block-wise 2D kernel PCA/LDA for face recognition, Inf. Process. Lett. 110 (17) (2010) 761–766.
[4] D. Meng, Q. Zhao, Z. Xu, Improve robustness of sparse PCA by L_1-norm maximization, Pattern Recognit. 45 (1) (2012) 487–497.
[5] S. Chen, T. Sun, Class-information-incorporated principal component analysis, Neurocomputing 69 (1) (2005) 216–223.
[6] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.
[7] R. Zass, A. Shashua, Nonnegative sparse PCA, in: Advances in Neural Information Processing Systems, NIPS, 2006, pp. 1561–1568.
[8] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[9] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[10] E.J. Candes, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies?, IEEE Trans. Inf. Theory 52 (12) (2006) 5406–5425.
[11] K. Tsuda, M. Kawanabe, K.-R. Müller, Clustering with the fisher score, in: Advances in Neural Information Processing Systems, NIPS, 2002, pp. 729–736.
[12] Q. Gu, Z. Li, J. Han, Generalized fisher score for feature selection, in: Proceedings of the Conference Annual Conference on Uncertainty in Artificial Intelligence, UAI, Corvallis, Oregon, 2011, pp. 266–273.
[13] S. Huang, M. Elhoseiny, A. Elgammal, D. Yang, Improving nonnegative matrix factorization via ranking its bases, in: IEEE International Conference on Image Processing, ICIP, IEEE, 2014, pp. 5951–5955.
[14] F.S. Samaria, A.C. Harter, Parameterisation of a stochastic model for human face identification, in: IEEE Workshop on Applications of Computer Vision, WACV, IEEE, 1994, pp. 138–142.
[15] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression (PIE) database, in: IEEE International Conference on Automatic Face and Gesture Recognition, FG, IEEE, 2002, pp. 46–51.