Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

PERGAMON

Pattern Recognition 37 (2004) 1303–1306

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

Face recognition using partial least squares components

Jangsun Baek[a,*], Minsoo Kim[b]

[a] *Department of Statistics, Chonnam National University, Gwangju 500-757,South Korea*
[b] *AIPR Lab, CS Div., Korea Advanced Institute of Science and Technology, Daejon 305-701,South Korea*

Received 7 October 2003; accepted 28 October 2003

Abstract

The paper considers partial least squares (PLS) as a new dimension reduction technique for the feature vector to overcome the small sample size problem in face recognition. Principal component analysis (PCA), a conventional dimension reduction method, selects the components with maximum variability, irrespective of the class information. So PCA does not necessarily extract features that are important for the discrimination of classes. PLS, on the other hand, constructs the components so that the correlation between the class variable and themselves is maximized. Therefore PLS components are more predictive than PCA components in classification. The experimental results on Manchester and ORL databases show that PLS is to be preferred over PCA when classification is the goal and dimension reduction is needed.

Keywords: Partial least squares; Principal component analysis; Face recognition

1. Introduction

Many statistical pattern recognition methods often suffer from the small sample size problem [1]. This problem occurs when the sample size is small compared with the dimension of feature vector, which always appears in the face recognition applications. The statistical classification methods based on the population parameters of feature vector such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) may not perform well in face recognition because the number of training sample is not enough to correctly estimate the mean vector and the covariance matrix of high-dimensional feature vector.

Since the number of training sample is limited we usually try to reduce the dimension of feature vector. One of the most successful approaches to the reduction of the feature dimension in face recognition is based on principal component analysis (PCA) [2]. PCA is used to reduce the high-dimensional feature to only a few feature components which explain as much of the observed total feature variation as possible. This is achieved without regard to the variation of the observation's class or group. In contrast to PCA, partial least squares (PLS) chooses components so that the sample covariance between the group variable and a linear combination of the original feature vector is maximum. The PLS method is well suited for the prediction of regression models with many predictor variables [3] and extensively used in chemometrics. (*Journal of Chemometrics* has a lot of PLS applications, for example.) Recently it is also applied to the biometric data classification [4]. The statistical properties of PLS have been investigated by, for example, [3,5]. In Section 2, we describe PLS and compare it with PCA. We will illustrate a simulated example which shows why PLS components are much more reasonable than PCA components as new features of lower dimension. Real data experiments are carried out in Section 3 to show that the recognition error rates with PLS are lower than those with PCA.

2. Dimension reduction by PLS

The goal of dimension reduction methods is to reduce the high p-dimensional original face feature space to a lower

* Corresponding author. Tel.: +82-62-530-3446; fax: +82-62-530-3449.
E-mail addresses: jbaek@chonnam.ac.kr (J. Baek), mskim@ai.kaist.ac.kr (M. Kim).

0031-3203/$30.00 © 2003 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2003.10.014

K-dimensional component space ($K \ll p$). This is achieved by extracting K components in the feature space to optimize a defined objective criterion. We describe PLS in comparison with PCA, and show the superiority of the former over the latter by a simulated example. Let $\mathbf{X}$ be a standardized $n \times p$ matrix of n images and p original face features. That is, the features are standardized to have mean zero and standard deviation of one.

PCA constructs orthogonal linear combinations of the features to maximize the variance sequentially. The procedure is to find the weight vector $\mathbf{a}_k$ such that

$$\mathbf{a}_k = \arg\max_{\mathbf{a}'\mathbf{a}=1} \mathrm{Var}(\mathbf{Xa}) \quad \text{for} \quad k = 1, 2, \ldots, K$$

subject to the orthogonal constraint

$$\mathbf{a}_k'\mathbf{S}\mathbf{a}_j = 0 \quad \text{for all} \quad 1 \leqslant j < k,$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The solution of $\mathbf{a}_k$ turns out to be the eigenvector corresponding to the kth eigenvalue λ_k of $\mathbf{S}/(\mathrm{n}-1)$. The kth principal component is the linear combination of the features, $\mathbf{X}\mathbf{a}_k$.

Note that PCA may not yield components predictive of the class of image since it extracts components sequentially which maximize the total predictor (feature) variability, irrespective of how well the constructed components predict classes. For this reason, a different objective criterion for dimension reduction may be more appropriate for the class prediction.

The objective criterion for constructing components in PLS is to sequentially maximize the covariance between the class variable and a linear combination of the features. Suppose there are $G+1$ classes (persons: $0, 1, \ldots, G$) to be recognized. We define G-dimensional random vector $\mathbf{y} = (y_1, y_2, \ldots, y_G)'$, where $y_i = 1$ and $y_j = 0$ for all $j \neq i$ when the face image belongs to the class $i-1$, $i = 1, 2, \ldots, G$, and $y_i = 1$ for all $i = 1, 2, \ldots, G$, when the image belongs to the class G. Then we can get the class vector observations $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ from the training sample of images to construct the $n \times G$ class matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]'$. PLS is to find the weight vector $\mathbf{b}_k$ such that

$$\mathbf{b}_k = \arg\max_{\mathbf{b}'\mathbf{b}=1,\ \mathbf{c}'\mathbf{c}=1} \mathrm{Cov}^2(\mathbf{Xb}, \mathbf{Yc}) \quad \text{for} \quad k = 1, 2, \ldots, K$$

subject to the orthogonality constraint

$$\mathbf{b}_k'\mathbf{S}\mathbf{b}_j = 0 \quad \text{for all} \quad 1 \leqslant j < k,$$

where $\mathbf{b}$, $\mathbf{c}$ are unit vectors, and $\mathbf{S} = \mathbf{X}'\mathbf{X}$, The procedure is called the multivariate PLS. The kth PLS component is the linear combination of the original features, $\mathbf{X}\mathbf{b}_k$. Since PLS extracts the components to maximize the correlation between the component and the class variable, PLS is expected to be more predictive than PCA. PCA can attain similar performance only when the selected principal components are fortunately in the direction which is predictive of the class. PLS component scores can be calculated by standard statistical packages, for example, SAS. We give an example which shows why PLS selects better components of low dimension for classification than PCA does as follows.

Suppose we have two-dimensional feature vector $X = (X_1, X_2)'$ and there are two classes to be classified. We assume also that the feature vector for each class follows two-dimensional multivariate normal distribution with different mean vector and the same covariance matrix. More specifically, the first and second class feature distribution is $MVN_2(\mu_1, \Sigma), MVN_2(\mu_2, \Sigma)$, with

$$\mu_1 = (1, 2)', \quad \mu_2 = (2, 1)', \quad \Sigma = \begin{pmatrix} 1/4 & 1/8 \\ 1/8 & 1/4 \end{pmatrix},$$

respectively. That is, the correlation between X_1 and X_2 is 0.5. We generated 50 feature vectors randomly from each distribution. Therefore we have in total 100 observations in the training sample. Fig. 1 is the scatter plot of the randomly generated observations with its class identification. Suppose we want to reduce the dimension of the original two-dimensional feature vector X to construct an univariate new feature. Then PCA is the selection of a new coordinate system obtained by rotating the original system with X_1 and X_2 as the coordinate axes. The first new axis represents the direction with maximum variability, which is PCA_1 of Fig. 1, and the resulting first principal component scores are evaluated by projecting the original data on PCA_1.

In PLS, on the other hand, we should find a new axis on which the correlation between the class variable and the linear combination of the original features is maximized. In order to maximize the correlation, the new axis must be set on the direction which is able to predict the class of the observations well. Therefore we get PLS_1 in Fig. 1 as the axis for the first PLS component. Fig. 2(a) contains the box plots of the first PCA component scores for each class. It is shown that PCA_1 can hardly classify the class. Fig. 2(b)

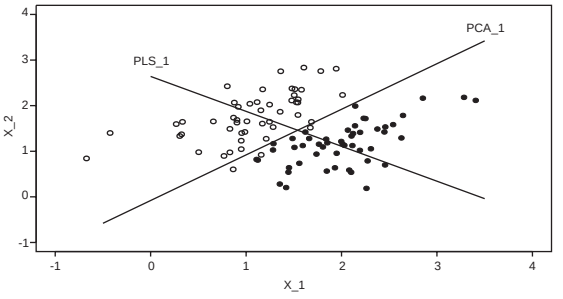

Fig. 1. The axes of PCA_1 and PLS_1 of the randomly generated two-dimensional observations with its class identification. Empty circles are from $MVN_2(\mu_1, \Sigma)$, and filled ones are from $MVN_2(\mu_2, \Sigma)$.

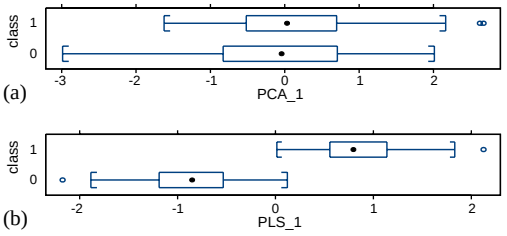

Fig. 2. The box plots of PCA_1 and PLS_1 component scores for two classes (class 0 is $MVN_2(\mu_1, \Sigma)$ and class 1 is $MVN_2(\mu_2, \Sigma)$): (a) PCA_1 and (b) PLS_1.

contains the box plots of the first PLS component scores for each class. We can see that there is a clear enough distinction between the distributions of PLS component for two classes to discriminate. The LDA with PCA and PLS scores gives us 0.48 and 0.03 as the leave-one-out cross-validation misclassification rates, respectively. The higher error rate of PCA scores in this example is inevitable because the first PCA component is selected to explain the total feature variability as maximum as possible, irrespective of how well the constructed component predicts the class.

3. Experimental results

In this section we will present the results of a face recognition experiment by PCA and PLS dimension reduction, using two small face databases, the Manchester database with 30 individuals and 10 images per person, and the Olivetti Research Laboratory (ORL) database with 10 different images for 40 individuals. Therefore the sample sizes for the Manchester and ORL data are 300 and 400, respectively. In both databases, the individuals show different facial expressions on all of their images (happy, sad, surprised, etc.). All images are grayscale. The resolution of the Manchester data in this experiment is 16×16, and that of the ORL data is 28×23, which were averaged on each 32×32 pixels of the original 512×512 pixels, and on each 4×4 pixels of the original 112×92 pixels, respectively. The features are grayscale intensities, and the dimension of the feature vector is 256 (16×16) and 644 (28×23) for the Manchester and ORL databases, respectively.

In this experiment, we first reduce the dimension by PCA and PLS so that the high dimension of p features is reduced to a lower dimension of K components. Here we consider K from 20 to 100 by 10 for the Manchester data, and from 20 to 50 by 5 for the ORL data. Since the reduced dimension is now low ($K < n$), we apply conventional classifiers such as LDA and QDA to the K components data. Then the misclassification rate is calculated by the leave-one-out cross-validation. That is, one of the n images is left out for the test and a classification function is obtained based on the remaining $n - 1$ images. The classification function is then used to predict the class of the left out image. This procedure is repeated for each of the n images in the training data set to get the misclassification rate. The misclassification rates of classifier LDA and QDA with different reduced K-dimensional PCA and PLS components are shown in Tables 1 and 2

Table 1
The misclassification rates for Manchester data (16×16 300 images)

		Number of components (K)								
Classifier	Method	20	30	40	50	60	70	80	90	100
LDA	PCA	0.217	0.176	0.173	0.157	0.153	0.147	0.160	0.143	0.143
	PLS	0.153	0.103	0.077	0.057	0.037	0.040	0.037	0.043	0.017
QDA	PCA	0.223	0.233	0.243	0.230	0.243	0.287	0.277	0.287	0.310
	PLS	0.203	0.180	0.197	0.187	0.183	0.217	0.233	0.263	0.293

Table 2
The misclassification rates for ORL data (28×23 400 images)

		Number of components (K)						
Classifier	Method	20	25	30	35	40	45	50
LDA	PCA	0.035	0.023	0.018	0.015	0.010	0.013	0.015
	PLS	0.025	0.020	0.008	0.005	0.008	0.005	0.005
QDA	PCA	0.050	0.048	0.045	0.043	0.043	0.045	0.048
	PLS	0.038	0.030	0.025	0.030	0.030	0.030	0.030

for the Manchester and ORL databases, respectively. The LDA performance is better than that of QDA in both databases since large number of parameter estimates of covariance matrices in QDA have poor precision compared with LDA in small sample. In Table 1, the lowest LDA error rate for PCA is 0.143 with $K = 90$ components, but it can be reached with less than $K = 30$ components for PLS. Table 2 of the ORL data also shows the lower error rates of PLS. Tables 1 and 2 indicate that PLS outperforms PCA for both classifiers since the misclassification rate of PLS is always lower than that of PCA for each reduced dimension K.

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Boston, 1990.
[2] M. Turk, A. Pentland, Eigen faces for recognition, J. Cognitive Neurosci. 3 (1991) 71–86.
[3] P.M. Garthwaite, An interpretation of partial least squares, J. Am. Stat. Assoc. 89 (1994) 122–127.
[4] D.V. Nguyen, D.M. Rocke, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics 18 (2002) 39–50.
[5] S. Helland, T. Almøy, Comparison of prediction methods when only a few components are relevant, J. Am. Stat. Assoc. 89 (1994) 583–591.

About the Author—JANGSUN BAEK received the B.S. and M.S. degrees in Applied Statistics at Yonsei University, Seoul, South Korea, in 1981 and 1984, respectively, and the Ph.D. degree in Statistics at Texas A& M University, College Station, Texas, in 1991. From 1991 to 1993 he was a postdoctoral fellow at the Department of Statistical Science, Southern Methodist University. In 1993, he joined the faculty of the Department of Statistics, Chonnam National University, Gwangju, South Korea. His research interests include pattern recognition, multivariate statistics, nonparametric function estimation.

About the Author—MINSOO KIM received his B.S., M.S. and Ph.D. degrees in Computer Science and Statistics at Chonnam National University, Gwangju, South Korea, in 1994, 1996 and 2000, respectively. From 2000 to 2002, he was a postdoctoral fellow at Chonnam National University. He has been working as a postdoctoral fellow at Korea Advanced Institute of Science and Technology (KAIST) since 2003. His research interests include pattern recognition, wavelet, multivariate statistics.