Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Pattern Recognition Letters 24 (2003) 2159–2165

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Using mixture covariance matrices to improve face and facial expression recognitions

Carlos E. Thomaz [a,*], Duncan F. Gillies [a], Raul Q. Feitosa [b]

[a] *Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2BZ, UK*
[b] *Department of Electrical Engineering, Catholic University of Rio de Janeiro, r. Marques de Sao Vicente 225, Rio de Janeiro 22453-900, Brazil*

Abstract

In several pattern recognition problems, particularly in image recognition ones, there are often a large number of features available, but the number of training examples for each pattern is significantly less than the dimension of the feature space. This statement implies that the sample group covariance matrices often used in the Gaussian maximum probability classifier are singular. A common solution to this problem is to assume that all groups have equal covariance matrices and to use as their estimates the pooled covariance matrix calculated from the whole training set. This paper uses an alternative estimate for the sample group covariance matrices, here called the mixture covariance, given by an appropriate linear combination of the sample group and pooled covariance matrices. Experiments were carried out to evaluate the performance of this method in two biometric classification applications: face and facial expression. The average recognition rates obtained by using the mixture covariance matrices were higher than the usual estimates.

Keywords: Face recognition; Facial expression recognition; Gaussian maximum probability classifier; Mixture covariance matrix

1. Introduction

A critical issue for the Gaussian maximum probability classifier is the inverse of the sample group covariance matrices. Since in practice these matrices are not known, estimates must be computed based on the observations (pattern examples) available in a training set. In some applications, however, there are often a large number of features available, but the number of training examples for each group is limited and significantly less than the dimension of the feature space. This implies that the sample group covariance matrices will be singular.

This problem, which is called a "small sample size problem" (Fukunaga, 1990), is quite common in pattern recognition, particularly in image recognition where the number of features is very large. One way to overcome this problem is to assume that all groups have equal covariance matrices and to use as their estimates the weighting average of each sample group covariance matrix,

[*] Corresponding author.
E-mail addresses: cet@doc.ic.ac.uk (C.E. Thomaz), dfg@doc.ic.ac.uk (D.F. Gillies), raul@ele.puc-rio.br (R.Q. Feitosa).

0167-8655/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0167-8655(03)00085-0

given by the pooled covariance matrix calculated from the whole training set.

The aim of this work is to investigate another estimate for the sample group covariance matrices, here called mixture covariance matrices, given by a linear combination of the sample group covariance matrix and the pooled covariance matrix. The mixture covariance matrices are based on the Hoffbeck and Landgrebe (1996) approach and have the property of having the same rank as the pooled estimate, while allowing a different estimate for each group. Thus, the mixture estimate may result in higher accuracy.

In order to evaluate this approach, two biometric applications were considered: face recognition and facial expression recognition. The evaluation used different image databases for each application. A probabilistic model was used to combine the well-known dimensionality reduction technique called principal component analysis (PCA) and the Gaussian maximum probability classifier, and in this way we could investigate the performance of the mixture covariance matrices on the recognition tasks referred to above. Experiments carried out show that the mixture covariance estimates attained the best performance in both applications.

2. Dimensionality reduction

In biometric applications, the number of training samples is limited and usually significantly less than the number of pixels of each image. Thus the high-dimensional space is very sparsely represented; making the parameter estimation quite difficult—a problem that is called the curse of dimensionality (e.g., Bishop, 1997).

One of the most successful approaches to the problem of creating a low dimensional image representation is based on PCA. PCA generates a set of orthonormal basis vectors, known as principal components, which minimizes the mean square reconstruction error and describes major variations in the whole training set considered. Instead of analysing the maximum probability classifier directly on the face or facial expression images, PCA is applied first to provide dimensionality reduction. Many researchers have confirmed that the PCA representation has good generalization ability especially when the distributions of each class are separated by the mean difference (Kirby and Sirovich, 1990; Turk and Pentland, 1991; Zhao et al., 1998; Liu and Wechsler, 2000). However, even after reduction the feature space is still often of higher dimension than the number of training samples.

3. Maximum probability classifier

The basic problem in the decision–theoretic methods for pattern recognition consists of finding a set of g discriminant functions $d_1(x), d_2(x), \dots, d_g(x)$, where g is the number of groups or classes, with the decision rule such that if the p-dimensional pattern vector x belongs to the class i $(1 \leqslant i \leqslant g)$, then $d_i(x) \geqslant d_j(x)$, for all $i \neq j$ and $1 \leqslant j \leqslant g$.

The Bayes classifier designed to maximize the total probability of correct classification, where equal prior probabilities for all groups are assumed, corresponds to a set of discriminant functions equal to the corresponding probability density functions, that is, $d_i(x) = f_i(x)$ for all classes (Johnson and Wichern, 1998). The most common probability density function applied to pattern recognition systems is based on the Gaussian multivariate distribution

$$d_i(x) = f_i(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i)\right], \tag{1}$$

where μ_i and Σ_i are the class i population mean vector and covariance matrix respectively. The notation "$|\cdot|$" denotes the determinant of a matrix.

In practice, however, the true values of the mean and the covariance matrix are seldom known and must be estimated from training samples. The mean is estimated by the usual sample mean

$$\mu_i \equiv \bar{x}_i = \frac{1}{k_i}\sum_{j=1}^{k_i} x_{i,j}, \tag{2}$$

where $x_{i,j}$ is observation j from class i, and k_i is the number of training observations from class i. The covariance matrix is commonly estimated by the sample group covariance matrix defined as

$$\Sigma_i \equiv S_i = \frac{1}{(k_i - 1)} \sum_{j=1}^{k_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^{\mathrm{T}}. \tag{3}$$

If we replace the true values of the mean and the covariance matrix in Eq. (1) by their respective estimates, the Bayes decision rule achieves optimal classification accuracy only when the number of training samples increases toward infinity (e.g., Hoffbeck and Landgrebe, 1996). In fact for p-dimensional patterns the sample covariance matrix is singular if less than $p+1$ independent training examples from each class i are available, that is, the sample covariance matrix cannot be calculated if k_i is less than the dimension of the feature space.

One method routinely applied to solve this problem is to assume that all classes have equal covariance matrices, and to use as their estimates the pooled covariance matrix. This covariance matrix is a weighting average of each sample group covariance matrix and, assuming that all classes have the same number of training observations, is given by

$$S_{\text{pooled}} = \frac{1}{g} \sum_{i=1}^{g} S_i. \tag{4}$$

Since more observations are taken to calculate the pooled covariance matrix S_{pooled}, this one will potentially have a higher rank than S_i and will be eventually full rank. Although the pooled estimate does provide a solution for the algebraic problem arising from the insufficient number of training observations in each group, S_{pooled} is theoretically a consistent estimator of the true covariance matrices only when $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$.

4. Mixture covariance matrix

The choice between the sample group covariance matrix and the pooled covariance one represents a limited set of estimates for the true covariance matrix. A more flexible set can be obtained using the mixture covariance matrix.

4.1. Definition

The mixture covariance matrix is a linear combination between the pooled covariance matrix S_{pooled} and the sample covariance matrix of one class S_i. It is given by

$$S_i^{\text{mix}}(w_i) = w_i S_{\text{pooled}} + (1 - w_i) S_i, \tag{5}$$

where the mixture parameter w_i takes on values $0 < w_i \leqslant 1$ and is different for each class. This parameter controls the degree of shrinkage of the sample group covariance matrix estimates toward the pooled one.

Fig. 1 illustrates the geometric idea of the mixture covariance matrix on a two-dimensional feature space containing three hypothetical classes. The constant probability densities contours of S_i and S_{pooled} are represented by the dashed and dotted gray ellipses respectively. The mixture covariance estimates assume that the ellipses corresponding to the true covariance matrices are placed somewhere in between S_i and S_{pooled} contours, as shown by the solid black ellipses.

Each S_i^{mix} matrix has the important property of admitting an inverse if the pooled estimate S_{pooled} does so (Magnus and Neudecker, 1999, pp. 21–22). This implies that if the pooled estimate is non-singular and the mixture parameter takes on values $w_i > 0$, then the S_i^{mix} will be non-singular.

Therefore the remaining question is: what is the value of the w_i that gives a relevant linear mixture between the pooled and sample covariance

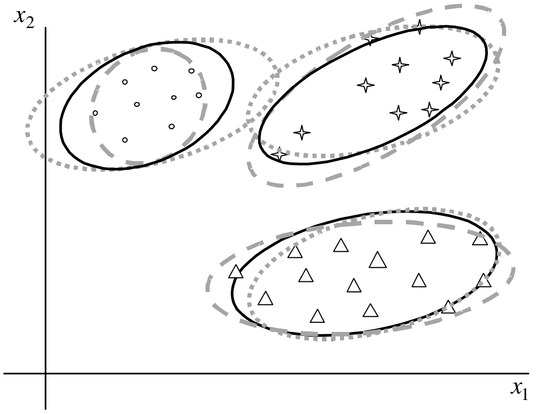

Fig. 1. Geometric idea of the mixture covariance matrix.

estimates? A method that determines an appropriate value of the mixture parameter is described in the next section.

4.2. The mixture parameter

According to Hoffbeck and Landgrebe (1996), the value of the mixture parameter w_i can be appropriately selected so that a best fit to the training samples is achieved. Their technique is based on the leave-one-out-likelihood (L) parameter estimation (Fukunaga, 1990).

In the L method, one observation of the class i training set is removed and the mean and covariance matrix from the remaining $k_i - 1$ examples is estimated. After that the likelihood of the excluded sample is calculated given the previous mean and covariance matrix estimates. This operation is repeated a further $k_i - 1$ times and the average log likelihood is computed over all the k_i observations. The strategy is to evaluate several different values of w_i in the range $0 < w_i \leqslant 1$, and then choose w_i that maximizes the average log likelihood.

The mean of class i without observation r may be computed as

$$\bar{x}_{i\backslash r} = \frac{1}{(k_i - 1)}\left[\left(\sum_{j=1}^{k_i} x_{i,j}\right) - x_{i,r}\right]. \tag{6}$$

The notation $i \setminus r$ conforms to the Hoffbeck and Landgrebe (1996) work. It indicates that the corresponding quantity is calculated with the rth observation from class i removed. Following the same idea, the sample covariance matrix and the pooled covariance matrix of class i without observation r are

$$S_{i\backslash r} = \frac{1}{(k_i - 2)}\left[\left(\sum_{j=1}^{k_i} (x_{i,j} - \bar{x}_{i\backslash r})(x_{i,j} - \bar{x}_{i\backslash r})^{\mathrm{T}}\right) - (x_{i,r} - \bar{x}_{i\backslash r})(x_{i,r} - \bar{x}_{i\backslash r})^{\mathrm{T}}\right], \tag{7}$$

$$S_{\mathrm{pooled}_{i\backslash r}} = \frac{1}{g}\left[\left(\sum_{j=1}^{g} S_j\right) - S_i + S_{i\backslash r}\right]. \tag{8}$$

Thus the average log likelihood of the excluded observations can be calculated as follows:

$$\overline{L}_i(w_i) = \frac{1}{k_i}\left[\sum_{r=1}^{k_i} \ln\left[f\left(x_{i,r}|\bar{x}_{i\backslash r}, S_{i\backslash r}^{\mathrm{mix}}(w_i)\right)\right]\right], \tag{9}$$

where $f(x_{i,r}|\bar{x}_{i\backslash r}, S_{i\backslash r}^{\mathrm{mix}}(w_i))$ is the Gaussian probability function defined in Eq. (1) with $\bar{x}_{i\backslash r}$ mean vector and $S_{i\backslash r}^{\mathrm{mix}}(w_i)$ covariance matrix defined as

$$S_{i\backslash r}^{\mathrm{mix}}(w_i) = w_i S_{\mathrm{pooled}_{i\backslash r}} + (1 - w_i) S_{i\backslash r}. \tag{10}$$

As Hoffbeck and Landgrebe (1996) pointed out, this approach, if implemented in a straightforward way, would require computing the inverse and determinant of the $S_{i\backslash r}^{\mathrm{mix}}(w_i)$ for each training sample. Since the $S_{i\backslash r}^{\mathrm{mix}}(w_i)$ is a p by p matrix and p is typically a large number, this computation would be quite expensive. However, they showed that it is possible to significantly reduce the required computation by using the Sherman–Morrison–Woodbury formula (Golub and Van Loan, 1989, p. 51) given by

$$(A + uu^{\mathrm{T}})^{-1} = A^{-1} - \frac{A^{-1}uu^{\mathrm{T}}A^{-1}}{1 + u^{\mathrm{T}}A^{-1}u}. \tag{11}$$

where A is a n by n matrix and u is a n by 1 vector. This allowed them to write the log likelihood of the excluded samples in an analogous form as follows:

$$\ln\left[f\left(x_{i,r}|\bar{x}_{i\backslash r}, S_{i\backslash r}^{\mathrm{mix}}(w_i)\right)\right] = -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\left[|Q|(1 - vd)\right] - \frac{1}{2}\left(\frac{k_i}{k_i - 1}\right)^2\left[\frac{d}{1 - vd}\right], \tag{12}$$

where

$$Q = \left[(1 - w_i)\frac{(k_i - 1)}{(k_i - 2)} + w_i\frac{1}{g(k_i - 2)}\right]S_i + w_i S_{\mathrm{pooled}}, \tag{13}$$

$$v = \frac{k_i}{(k_i - 1)(k_i - 2)}\left[1 - w_i\frac{(g - 1)}{g}\right], \tag{14}$$

$$d = (x_{i,r} - \bar{x}_i)^{\mathrm{T}} Q^{-1} (x_{i,r} - \bar{x}_i). \tag{15}$$

Finally, when the parameter w_i is selected, the mixture covariance matrix estimate defined in Eq. (5) is calculated using all the training examples and replaced into the maximum probability classifier.

5. Experiments

Two biometric experiments with two different databases were performed.

In the face recognition experiment the olivetti face database (ORL) containing ten images for each of 40 individuals, a total of 400 images, were used. The Tohoku University has provided the database for the facial expression experiment. This database is composed of 193 images of expressions posed by nine Japanese females (Lyons et al., 1999). Each person posed three or four examples of each six fundamental facial expression: anger, disgust, fear, happiness, sadness and surprise. The database has at least 29 images for each fundamental facial expression. For implementation convenience all images were first resized to 64×64 pixels.

The experiments were carried out as follows. First PCA reduces the dimensionality of the original images and secondly the Gaussian maximum probability classifier using one out of the three covariance estimates S_i (or Sgroup), S_{pooled} (or Spooled) and S_i^{mix} (or Smix) was applied. Each experiment was repeated 25 times using several PCA dimensions. Distinct training and test sets were randomly drawn, and the mean and standard deviation of the recognition rate were calculated.

The face recognition classification was computed using for each individual five images to train and five images to test. In the facial expression recognition, the training and test sets were respectively composed of 20 and 9 images. The size of the mixture parameter ($0 < w_i \leqslant 1$) optimisation range was taken to be 20, that is $w_i = [0.05, 0.10, 0.15, \ldots, 1]$.

6. Results

Tables 1 and 2 present the training and test average recognition rates (with standard deviations) of the face and facial expression databases, respectively, over the different PCA dimensions. Also the optimised mixture parameters w_i over the common PCA components of both applications are shown in Table 3.

Since only five images of each individual were used to form the face recognition training set, the results relative to the sample group covariance estimate were limited to four PCA components. Table 1 shows that in all but one experiment the S_i^{mix} (or Smix) estimate led to higher accuracy than either the pooled or the sample group covariance matrices. In terms of how sensitive the mixture covariance results were to the choice of the training and test sets, it is fair to say that the S_i^{mix} standard deviations were similar to the pooled estimate.

Table 2 shows the results of the facial expression recognition. For more than 20 components when the sample group covariance estimate became singular, the mixture covariance estimate reached higher recognition rates than the pooled covariance estimate. Again, regarding the computed standard deviations, the S_i^{mix} estimate showed to be as sensitive to the choice of the training and test sets as the other two estimates.

Table 1
Face recognition results

PCA components	Sgroup		Spooled		Smix	
	Training	Test	Training	Test	Training	Test
4	99.5 (0.4)	51.6 (4.4)	73.3 (3.1)	59.5 (3.0)	90.1 (2.1)	70.8 (3.2)
10			96.6 (1.2)	88.4 (1.4)	99.4 (0.5)	92.0 (1.5)
20			99.2 (0.6)	91.8 (1.8)	100.0 (0.1)	94.5 (1.7)
30			99.9 (0.2)	94.7 (1.7)	100.0 (0.0)	95.9 (1.5)
40			100.0 (0.0)	95.4 (1.5)	100.0 (0.0)	96.2 (1.6)
50			100.0 (0.0)	95.7 (1.2)	100.0 (0.0)	96.4 (1.5)
60			100.0 (0.0)	95.0 (1.6)	100.0 (0.0)	95.8 (1.6)
70			100.0 (0.0)	94.9 (1.6)	100.0 (0.0)	95.4 (1.6)

Table 2
Facial expression recognition results

PCA components	Sgroup		Spooled		Smix	
	Training	Test	Training	Test	Training	Test
5	41.5 (4.2)	20.6 (3.9)	32.3 (3.0)	21.6 (3.8)	34.9 (3.3)	21.3 (4.1)
10	76.3 (3.6)	38.8 (5.6)	49.6 (3.9)	26.5 (6.8)	58.5 (3.7)	27.9 (5.6)
15	99.7 (0.5)	64.3 (6.4)	69.1 (3.6)	44.4 (5.3)	82.9 (2.9)	49.7 (7.7)
20			81.2 (2.6)	55.9 (7.7)	91.4 (2.8)	61.3 (7.1)
25			86.9 (2.8)	64.9 (6.9)	94.8 (2.2)	68.3 (5.1)
30			91.9 (1.7)	70.1 (7.8)	96.8 (1.3)	72.3 (6.2)
35			94.3 (1.7)	72.0 (7.4)	97.7 (1.1)	75.6 (5.5)
40			95.9 (1.4)	75.6 (7.1)	98.3 (1.1)	77.2 (5.7)
45			96.7 (1.3)	78.4 (6.5)	98.6 (0.8)	79.1 (5.4)
50			97.6 (1.0)	79.4 (5.8)	99.2 (0.7)	81.0 (6.6)
55			98.5 (0.9)	81.6 (6.6)	99.5 (0.6)	82.8 (6.3)
60			99.1 (0.8)	82.1 (5.9)	99.6 (0.6)	83.6 (7.2)
65			99.5 (0.6)	83.3 (5.5)	99.8 (0.4)	84.5 (6.2)

Table 3
The average (with standard deviations) of the optimum mixture parameters

PCA components	Linear mixture parameter	
	Face	Facial expression
10	0.58 (0.25)	0.76 (0.19)
20	0.65 (0.21)	0.49 (0.15)
30	0.71 (0.18)	0.56 (0.15)
40	0.77 (0.16)	0.67 (0.15)
50	0.82 (0.13)	0.77 (0.11)
60	0.85 (0.11)	0.85 (0.09)

Another result revealed by these experiments is related to the optimum mixture parameters w_i. Table 3 shows the average (with standard deviations) of the selected mixture parameter w_i over the common face and facial expression PCA components. It can be seen that as the dimension of the feature space increases, the average and standard deviation of the mixture parameter w_i in all but one experiment increases and decreases respectively, making the mixture covariance of each class (S_i^{mix}) more similar to the pooled covariance (S_{pooled}) than the sample group one (S_i). Although this behaviour depends on the applications considered, it suggests that in both pre-processed image classification tasks the sparseness of the sample group covariance matrix could influence its linear combination to the pooled covariance matrix. In other words, it seems that when the group sample sizes are small compared with the dimension of the feature space, the pooled information is more reliable than that provided sparsely by each group. Research is currently being done in order to understand and prove this behaviour under certain constraints (Thomaz et al., 2001, 2002).

7. Conclusions

This paper used an estimate for the sample group covariance matrices, here called mixture covariance matrices, given by an appropriate linear combination of the sample group covariance matrix and the pooled covariance one. The mixture covariance matrices have the same rank as the pooled estimate, while allowing a different estimate for each group.

Extensive experiments were carried out to evaluate this approach on two biometric recognition tasks: face recognition and facial expression recognition. A Gaussian maximum probability classifier was built using the mixture estimate and the typical sample group and pooled estimates. In both tasks the mixture covariance estimate achieved the highest classification performance. Regarding the sensitivity to the choice of the training and test sets, the mixture covariance matrices gave a similar performance to the other two usual estimates.

The experiments were carried out using well-framed images without normalisation, in order to compare the performance of the different covariance estimators. We would expect the similar comparative results, perhaps with an overall improvement in performance, would be obtained by using techniques such as normalisation or by incorporating other image features.

The results presented in this work suggested that in both pre-processed image classification tasks the sparseness of the sample group covariance matrix could influence its linear combination to the pooled covariance matrix. Further work is being undertaken to study this relationship.

Acknowledgements

The authors would like to thank the referees for their constructive comments which helped to improve this paper. The first author was partially supported by the Brazilian Government Agency CAPES under the grant no. 1168/99-1.

References

Bishop, C.M., 1997. Neural Networks for Pattern Recognition. Oxford University Press, New York.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, Boston.

Golub, G.H., Van Loan, C.F., 1989. Matrix Computations. Johns Hopkins University Press, Baltimore.

Hoffbeck, J.P., Landgrebe, D.A., 1996. Covariance matrix estimation and classification with limited training data. IEEE Trans. Pattern Anal. Machine Intell. 18 (7), 763–767.

Johnson, R.A., Wichern, D.W., 1998. Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey.

Kirby, M., Sirovich, L., 1990. Application of the Karhunen–Loeve procedure for the characterization of human faces. IEEE Trans. Pattern Anal. Machine Intell. 12 (1), 103–108.

Liu, C., Wechsler, H., 2000. Learning the face space—representation and recognition. In: Proc. 15th Internat. Conf. on Pattern Recognition, ICPR'2000.

Lyons, M.J., Budynek, J., Akamatsu, S., 1999. Automatic classification of single facial images. IEEE Trans. Pattern Anal. Machine Intell. 21 (12), 1357–1362.

Magnus, J.R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics. John Wiley & Sons Ltd., Chichester.

Thomaz, C.E., Gillies, D.F., Feitosa, R.Q., 2001. Small sample problem in bayes plug-in classifier for image recognition. In: Proc. Internat. Conf. on Image and Vision Computing New Zealand. pp. 295–300.

Thomaz, C.E., Gillies, D.F., Feitosa, R.Q., 2002. A new quadratic classifier applied to biometric recognition. In: Proc. Post-ECCV Workshop on Biometric Authentication, Springer-Verlag LNCS 2359, Copenhagen, Denmark, pp. 186–196.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cognitive Neurosci. 3, 71–86.

Zhao, W., Chellappa, R., Krishnaswamy, A., 1998. Discriminant analysis of principal components for face recognition. In: Proc. 2nd Internat. Conf. on Automatic Face and Gesture Recognition. pp. 336–341.