Using Mixture Covariance Matrices to Improve Face and Facial Expression Recognitions

Carlos E. Thomaz[1], Duncan F. Gillies[1] and Raul Q. Feitosa[2]

[1] Imperial College of Science Technology and Medicine, Department of Computing,
180 Queen's Gate, London SW7 2BZ, United Kingdom
{cet, dfg}@doc.ic.ac.uk
[2] Catholic University of Rio de Janeiro, Department of Electrical Engineering,
r. Marques de Sao Vicente 225, Rio de Janeiro 22453-900, Brazil
raul@ele.puc-rio.br

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 samples 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 associated with this estimate in two recognition applications: face and facial expression. The average recognition rates obtained by using the mixture covariance matrices were higher than the usual estimates.

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 (patterns) available in a training set. In some applications, however, there are often a large number of features available, but the number of training samples 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" [5], 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, given by the pooled covariance matrix calculated from the whole training set.

This paper uses another estimate for the sample group covariance matrices [4], 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 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 pattern recognition applications were considered: face recognition and facial expression recognition. The evaluation used different image databases for each application and a probabilistic model combines the well-known dimensionality reduction technique called Principal Component Analysis (PCA) and the Gaussian maximum probability classifier to investigate the mixture covariance matrices on the referred recognition tasks. Experiments carried out show that the mixture covariance estimates attained the best performance in both applications considered.

2 Dimensionality Reduction

One of the most successful approaches to the problem of creating a low dimensional image representation is based on Principal Component Analysis (PCA). PCA generates a set of orthonomal basis vectors, known as principal components, that minimizes the mean square reconstruction error and describe major variations in the whole training set considered.

The reasoning behind applying PCA first for dimensionality reduction instead of analyzing the maximum probability classifier directly on the face or facial expression images is based on the original high-dimensional space. As the number of training samples is limited and significantly less than the number of pixels of each image, the high-dimensional space is indeed sparsely represented, making the parameter estimation quite complicated – this behaviour is called the curse of dimensionality [8]. Furthermore, 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 [1,6,7,9].

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**), ..., 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 $(1 \le i \le g)$, then $d_i(\mathbf{x}) \ge d_j(\mathbf{x})$, for all $i \ne j$ and $1 \le j \le \mathrm{g}$.

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, $d_{\mathrm{i}}(\mathbf{x})=f_{\mathrm{i}}(\mathbf{x})$ for all classes [8]. The most common probability density function applied to pattern recognition systems is based on the Gaussian multivariate distribution

$$d_i(x) = f_i(x \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\boldsymbol{\pi})^{p/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(x-\boldsymbol{\mu}_i) \right], \qquad \textbf{(1)}$$

where μ_{i} and Σ_{i} are the class i population mean vector and covariance matrix. Usually 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

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

From replacing the true values of the mean and the covariance matrix in (1) by their respective estimates, the Bayes decision rule achieves optimal classification accuracy only when the number of training samples increases toward infinity [4]. In fact for p-dimensional patterns the sample covariance matrix is singular if less than $p + 1$ training samples from each class i are available, that is, the sample covariance matrix can not 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

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

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 samples in each group, assuming equal covariance for all groups may bring about distortions in the modeling of the classification problem and consequently lower accuracy.

4 Mixture Covariance Matrix

The choice between the sample group covariance matrix and the pooled covariance one represents a restrictive set of estimates for the true covariance matrix. A less limited 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 each class S_i. It is given by

$$Smix_i(w_i) = w_i S_{pooled} + (1 - w_i) S_i \ . \tag{5}$$

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

Each $Smix_i$ matrix has the important property of admitting an inverse if the pooled estimate S_{pooled} does so [2]. This implies that if the pooled estimate is non-singular and the mixture parameter takes on values $w_i > 0$, then the $Smix_i$ will be non-singular.

Then the remaining question is: what is the value of the w_i that gives a relevant linear mixture between the pooled and sample covariance 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 [4], 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.

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

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

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

The notation $\backslash r$ indicates the corresponding quantity is calculated with the r^{th} observation from class i removed. Following the same idea, the sample covariance matrix and the pooled covariance matrix of class i without sample r are

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

$$S_{pooled_{i \backslash r}} = \frac{1}{g} \left[\sum_{j=1}^{g} S_j - S_i + S_{i \backslash r} \right]. \qquad \textbf{(8)}$$

Then the average log likelihood of the excluded samples can be written as follows:

$$\bar{L}_i(w_i) = \frac{1}{k_i} \left[\sum_{r=1}^{k_i} \ln\left[f\left(x_{i,r} \mid \bar{x}_{i \backslash r}, Smix_{i \backslash r}(w_i) \right) \right] \right], \qquad \textbf{(9)}$$

where $f\left(x_{i,r} \mid \bar{x}_{i \backslash r}, Smix_{i \backslash r}(w_i) \right)$ is the Gaussian probability function defined in (1) with $\bar{x}_{i \backslash r}$ mean vector and $Smix_{i \backslash r}(w_i)$ covariance matrix defined as

$$Smix_{i \backslash r}(w_i) = w_i S_{pooled_{i \backslash r}} + (1 - w_i) S_{i \backslash r} \,. \qquad \textbf{(10)}$$

This approach, if implemented in a straightforward way, would require computing the inverse and determinant of the $Smix_{i\backslash r}(w_i)$ for each training sample. As the $Smix_{i\backslash r}(w_i)$ is a *p* by *p* matrix and *p* is typically a large number, this computation would be quite expensive. Hoffbeck and Landgrebe [4], using the Sherman-Morrison-Woodbury formula [3], have showed that it is possible to significantly reduce the required computation by writing the mixture covariance matrix in a form as follows:

$$\ln\left[f\left(x_{i,r} \mid \bar{x}_{i\backslash r}, Smix_{i\backslash r}(w_i)\right)\right] = -\frac{1}{2}\ln\left[\left|Q\right|(1-vd)\right] - \frac{1}{2}\left(\frac{k_i}{k_i-1}\right)^2\left[\frac{d}{1-vd}\right], \qquad \textbf{(11)}$$

where

$$Q = \left[(1-w_i)\frac{(k_i-1)}{(k_i-2)} + w_i\frac{1}{g(k_i-2)}\right]S_i + w_i S_{pooled}\,, \qquad \textbf{(12)}$$

$$v = \frac{k_i}{(k_i-1)(k_i-2)}\left[1 - w_i\frac{(g-1)}{g}\right], \qquad \textbf{(13)}$$

$$d = (x_{i,r} - \bar{x}_i)Q^{-1}(x_{i,r} - \bar{x}_i)^T\,. \qquad \textbf{(14)}$$

Once the mixture parameter w_i that maximizes the average log likelihood is selected, the proposed covariance matrix estimate is calculated using all the training samples and replaced into the maximum probability classifier defined in (1).

5 Experiments

Two experiments with two different databases were performed.

In the face recognition experiment the ORL Face Database 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. 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 64x64 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}, S_{pooled} and $Smix_{\mathrm{i}}$) was applied. Each experiment was repeated 25 times using several PCA dimensions. Distinct training and testing 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 5 images to train and 5 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 ($0 < w_i \leq 1$) optimization range was taken to be 20, that is w_i = [0.05, 0.10, 0.15, …, 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.

Since only 5 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 4 PCA components. Table 1 shows that in all but one experiment the *Smix* estimate led to higher accuracy than did both the pooled covariance and 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 *Smix* standard deviations were similar to the other two covariance estimates.

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 *Smix* estimate showed to be as sensitive to the choice of the training and test sets as the other two estimates.

PCA	**Sgroup**		**Spooled**		**Smix**	
Components	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 1. Face Recognition Results

PCA	**Sgroup**		**Spooled**		**Smix**	
Components	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 2. Facial Expression Recognition Results

7 Conclusion

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 property of having 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 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 accuracy. Regarding the sensitiveness to the choice of the training and test sets, the mixture covariance matrices presented similar performance to the other two usual estimates.

References

1. C. Liu and H. Wechsler, “Learning the Face Space – Representation and Recognition”. In Proc. 15th Int’l Conference on Pattern Recognition, ICPR’2000, Barcelona, Spain, September 2000.
2. C.E. Thomaz, R.Q. Feitosa, A. Veiga, “Separate-Group Covariance Estimation with Insufficient Data for Object Recognition”. In Proc. Fifth All-Ukrainian International Conference, pp. 21-24, Ukraine, 1999.
3. G.H. Golub and C.F. Van Loan, *Matrix Computations*, second edition. Baltimore: Johns Hopkins Univ. Press, 1989.
4. J.P. Hoffbeck and D.A. Landgrebe, "Covariance Matrix Estimation and Classification with Limited Training Data", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 7, July 1996.
5. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second edition. Boston: Academic Press, 1990.
6. M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 12, No. 1, Jan. 1990.
7. M. Turk and A. Pentland, “Eigenfaces for Recognition, “Journal of Cognitive Neuroscience, Vol. 3, pp. 72-85, 1991.
8. R.A. Johnson R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*, by Prentice-Hall, Inc., 3d. edition, 1992.
9. W. Zhao, R. Chellappa and A. Krishnaswamy, “Discriminant Analysis of Principal Components for Face Recognition,” Proc. 2nd International Conference on Automatic Face and Gesture Recognition, pp. 336-341, Japan, April, 1998.