Visual Analysis of the Use of Mixture Covariance Matrices in Face Recognition

Carlos E. Thomaz and Duncan F. Gillies

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

Abstract. The quadratic discriminant (QD) classifier has proved to be simple and effective in many pattern recognition problems. However, it requires the computation of the inverse of the sample group covariance matrix. In many biometric problems, such as face recognition, the number of training patterns is considerably smaller than the number of features, and therefore the covariance matrix is singular. Several studies have shown that the use of mixture covariance matrices defined as a combination between the sample group covariance matrices and, for instance, the pooled covariance matrix, not only overcomes the singularity and instability of the sample group covariance matrices but also improves the QD classifier performance. However, little attention has been paid to understanding what has happened with the final shape of these mixture covariance matrices. In this work, we visually analyse in the commonly used eigenfaces space the eigenvectors and eigenvalues of these covariance matrices, given by the three following approaches: maximum likelihood, maximum classification accuracy, and maximum entropy. Experiments using the two well-known ORL and FERET face databases have shown that the maximum entropy approach is the one that achieves a more intuitive visual performance and best classification accuracies, especially in face experiments where moderate changes in facial expressions, pose, and scale, occur.

1 Introduction

The quadratic discriminant (QD) classifier is one of the most popular parametric Bayesian classifiers. It requires the inverse of the sample group covariance matrices. Since in some applications, especially in face recognition, the number of training patterns per group is smaller than the number of features, the sample group covariance matrices become singular and the QD classifier cannot be used.

Several studies have shown [2, 3, 5, 6, 13] that by using a mixture covariance matrix defined as a combination between the ill-posed sample group covariance matrices and well-posed covariance matrices, such as the pooled covariance matrix, it is possi-

J. Kittler and M.S. Nixon (Eds.): AVBPA 2003, LNCS 2688, pp. 172-181, 2003.

ble to overcome the singularity and instability of the sample group covariance matrices and also improve the classification performance. However, given the high-dimensionality of these small sample size problems, little attention has been paid to understanding what has happened with the final shape of these mixture covariance matrices.

In this work, we visually analyse the eigenvectors and eigenvalues of the mixture covariance matrices in the well-known and commonly used eigenfaces space [9, 10]. These mixture covariance matrices are given by the combination of the sample group and pooled covariance information using three approaches: maximum likelihood, maximum classification accuracy, and maximum entropy. Experiments using the two well-known ORL and FERET face databases have shown that the maximum entropy approach is the one that not only preserves as much of the sample group covariance information as possible but also achieved the best classification accuracies.

2 The Quadratic Discriminant Classifier

The quadratic discriminant (QD) classifier is based on the *p*-multivariate Gaussian class-conditional probability densities.

Assuming the symmetrical or zero-one loss function, the Bayes QD rule [12] stipulates that an unknown pattern *x* should be assigned to the class *i* that minimises:

$$d_i(x) = \ln|S_i| + (x-\bar{x}_i)^T S_i^{-1}(x-\bar{x}_i) - 2\ln\pi_i\,, \tag{1}$$

where π_i is a prior probability associated with the *i*th class, $\bar{x}_i$ is the maximum likelihood estimate [14] of the corresponding true mean vector given by

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

and S_i is the maximum likelihood estimate of the respective true covariance matrix given by

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

where $x_{i,j}$ is the pattern j from class i and n_i is the number of training patterns from class i.

The use of the QD rule described in equation (1) is, however, especially problematic if S_i is a singular matrix, that is, if n_i is less than the dimension of the feature space p [8]. In fact, when the sample covariance matrix is singular the smallest ($p-n_i+1$) eigenvalues are estimated to be 0 and the corresponding eigenvectors are arbitrary, though constrained by orthogonality [5].

One straightforward method routinely applied to overcome the limited sample size problem on the QD classifier is to employ the Fisher's linear discriminant function (LD) classifier (e.g. [1]). The LD classifier is obtained by replacing the S_i in (1) with the pooled sample covariance matrix S_p defined as

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{N - g}, \tag{4}$$

where g is the number of classes and $N = n_1 + n_2 + \cdots + n_g$. However, S_p is theoretically a consistent estimator of the true covariance matrices Σ_i only when $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$.

3 Mixture Covariance Matrix

A less limited set of covariance estimators can be obtained by using a mixture covariance matrix defined as a linear or convex combination between the sample covariance matrix of each class S_i and the pooled covariance matrix S_p. It is given by

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

where the mixture parameter w_i takes on values $0 < w_i \leq 1$ and could be different for each class depending on the optimization technique used to solve the problem.

Each mixture covariance matrix S_i^{mix} defined in equation (5) has the important property of admitting an inverse if the pooled estimate S_p does so [7]. 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. A number of optimisation techniques can be used to determine an appropriate blending of the S_i and S_p covariance matrices. In the next sub-sections, three possible approaches are briefly described.

3.1 Maximising the Likelihood

According to Hoffbeck and Landgrebe [6], the value of the mixture parameter w_i can be appropriately selected so that a best fit to the training patterns is achieved. Their technique is based on the leave-one-out-likelihood (LOOL) parameter estimation [8].

The strategy consists of evaluating several values of w_i over the optimisation grid $0 < w_i \leq 1$, and then choosing w_i that maximizes the average log likelihood of the corresponding p-multivariate normal density function, computed as follows:

$$LOOL_i(w_i) = \frac{1}{n_i}\sum_{v=1}^{n_i}\left[-\frac{1}{2}\ln\left|S_{i/v}^{mix}(w_i)\right| - \frac{1}{2}(x_{i,v} - \bar{x}_{i/v})^T\left(S_{i/v}^{mix}(w_i)\right)^{-1}(x_{i,v} - \bar{x}_{i/v})\right] \tag{6}$$

where the notation $/v$ represents the corresponding quantity with observation $x_{i,v}$ left out. Once the mixture parameter w_i is selected, the corresponding leave-one-out covariance estimate $S_i^{mix}(w_i)$ is calculated using all the n_i training observations and substituted for S_i into the QD rule defined in (1).

3.2 Maximising the Classification Accuracy

Another way of determining an appropriate value for the mixture parameter w_i described in equation (5) is based on the well-known Reguralized Discriminant Analysis classifier proposed by Friedman [5].

In this approach, all the mixture parameters w_i of each class are equal and selected to maximise the leave-one-out classification accuracy based on the QD rule defined in equation (1). The following classification rule is developed on the $N-1$ training observations exclusive of a particular observation $x_{i,v}$ and then used to classify $x_{i,v}$: Choose class k such that

$$d_k(x_{i,v}) = \min_{1 \le j \le g} d_j(x_{i,v}) \text{, with}$$
$$d_j(x_{i,v}) = \ln\left|S_{j/v}^{mix}(w)\right| + (x_{i,v} - \bar{x}_{j/v})^T \left(S_{j/v}^{mix}(w)\right)^{-1} (x_{i,v} - \bar{x}_{j/v}) - 2\ln\pi_j \,. \quad (7)$$

Each of the training observations is in turn held out and then classified in this manner. The resulting misclassification loss, i.e., the number of cases in which the observation left out is allocated to the wrong class, averaged over all the training observations is then used to choose the best mixture parameter w.

3.3 Maximising the Entropy

The covariance estimation problem can be viewed as a problem of estimating the parameters of Gaussian probability distributions under uncertainty. The maximum entropy criterion [4] maximises the uncertainty under incomplete information, and therefore may be a promising solution.

We have shown that, particularly in pre-processed or well-framed biometric recognition applications [3], in order to maximise the entropy given by the convex combination of S_i and S_p, we do not need to determine the best mixture parameter w_i but simply select the maximum variances of the corresponding matrices.

In the maximum entropy (ME) approach, the S_i^{mix} estimator is given by the following procedure:

1. Find the eigenvectors Φ_i^{me} of the covariance given by $S_i + S_p$.
2. Calculate the variance contribution of both S_i and S_p on the Φ_i^{me} basis, i.e.,

$$\begin{aligned} \Lambda_i^* &= diag[(\Phi_i^{me})^T S_i \Phi_i^{me}] = [\lambda_1^{i*}, \lambda_2^{i*}, ..., \lambda_p^{i*}] \\ \Lambda_p^* &= diag[(\Phi_i^{me})^T S_p \Phi_i^{me}] = [\lambda_1^{p*}, \lambda_2^{p*}, ..., \lambda_p^{p*}] \end{aligned} \quad (8)$$

3. Form a new variance matrix based on the largest values, that is,

$$\Lambda_i^{me} = diag[\max(\lambda_1^{i*}, \lambda_1^{p*}),..., \max(\lambda_p^{i*}, \lambda_p^{p*})] \tag{9}$$

4. Form the S_i^{mix} estimator

$$S_i^{mix} = \Phi_i^{me} \Lambda_i^{me} (\Phi_i^{me})^T . \tag{10}$$

It is important to note that the projection basis calculated on step 1 is not necessarily unique and was chosen only to simplify the procedure. In fact, it might be any orthonormal basis that diagonalises an unbiased mixture of S_i and S_p.

The main idea of the maximum entropy approach is to expand in a straight-forward way the S_i smaller and consequently less reliable eigenvalues while trying to keep most of its larger eigenvalues unchanged.

4 Experiments

In order to investigate and visualise the different approaches of blending the sample group and pooled covariance matrices, two experiments using the two well-known ORL (http://www.uk.research.att.com/facedatabase.html) and FERET Face Databases [11] were performed.

The experiments were carried out as follows. First the face images from the original vector space are projected to a lower dimensional space (face subspace) using Principal Component Analysis (PCA) [9, 10] and then classified using the pooled covariance matrix and the three mixture covariance approaches described in the previous sections. Each experiment was repeated 25 times using several eigenfaces. Distinct training and test sets were randomly drawn, and the mean and standard deviation of the recognition rate, as well as the mean of the likelihood and classification accuracy mixture parameters, were calculated. Then, based on the best classification accuracy of the several PCA features used, the number of eigenfaces to visualise and calculate the covariance eigenvectors and eigenvalues on the face subspace was determined. The best classification results were obtained by using respectively 40 and 50 eigenfaces (which we call most effective eigenfaces) for the ORL and FERET databases.

The ORL face experiments were computed using for each individual 5 images to train and 5 images to test. In the FERET experiments, sets containing 4 "frontal b series" images for each of 200 total subjects were considered. Each image set is composed of a regular facial expression (referred as "ba" images in the FERET database), an alternative expression ("bj" images), and two symmetric images ("be" and "bf" images) taken with 15 degrees pose angle effects. The FERET training and test sets were composed of 3 and 1 images respectively. Since in all applications the same number of training examples per subject was considered, the prior probabilities were assumed equal for all classes and recognition tasks. For implementation convenience all ORL and FERET images were first resized to 64x64 and 96x64 pixels. The mixture parameter range was taken to be $[0.1, 0.2, ..., 1.0]$ for both w_i likelihood and w classification accuracy optimisations.

5 Results

Figures 1 and 2 present the visual analysis of two examples of the ORL and FERET covariance blending experiments using the respective most effective eigenfaces. These examples were chosen based on the closeness of the likelihood and classification mixture parameters to their respective mean values. The test average recognition rates (with standard deviations) over the different PCA dimensions for the ORL and FERET databases are also shown in Tables 1 and 2, respectively.

The visual results of Figures 1a-1b, and 2a-2b, can be described as follows. The first image row corresponds to the training images of a specific subject, and the second and third following rows correspond to the eigenvectors (in descending ordering of eigenvalues, from left to right) of the respective sample group and pooled covariance matrices transformed back to the image space by using the corresponding most effective eigenfaces. Accordingly, the fourth, fifth and sixth image rows correspond to the eigenvectors of the maximum likelihood (ML), maximum classification accuracy (MC), and maximum entropy (ME) mixture covariance matrices. The numbers below each image row describe the magnitude of the eigenvalue of each covariance eigenvector with its corresponding percentage of total variance shown in parentheses.

Since only 5 images of each individual were used to form the ORL training set, the results of Figures 1a and 1b relative to the sample group covariance estimates were limited, in terms of total variation within the subject's images, to the first 4 eigenvectors. The remaining eigenvectors (only 4 shown) are arbitrary, constrained to the orthogonality assumption on the face space, and should be replaced or modified using the pooled information. As can be seen on Figures 1a and 1b, the mixture covariance matrices that preserve as much of the sample group covariance information as possible were the ones blended using the maximum entropy approach. It is important to note that although the percentage of total variation of each eigenvalue was different due to the use of the pooled information, the first eigenvectors and eigenvalues of the maximum entropy covariance matrices are quite similar to the respective sample group covariance ones. In terms of how accurate the mixture covariance results were to the choice of the training and test sets (shown in Table 1), it is fair to say that the performance of the maximum entropy approach was better than the pooled estimate and slightly better than the other two mixture covariance estimates.

Figures 2a and 2b show the results of the FERET experiments. Analogously to the ORL experiments, the sample group covariance information became limited to the first 2 eigenvectors. The remaining eigenvectors (only 3 shown) represent no subject variation at all and are arbitrary, constrained to the orthogonality assumption on the face space. As can be observed, the visual results of the mixture covariance estimates seem to be more related to the pooled information than the sample group one. Again, the maximum entropy approach was the one that preserved as much of the sample group covariance information on the covariance matrices blending as possible. However, in this application where the face images are well-framed thus favouring the pooled covariance matrix, there is no significant visual or classification performance improvement (shown in Table 2) in using mixture covariance matrices.

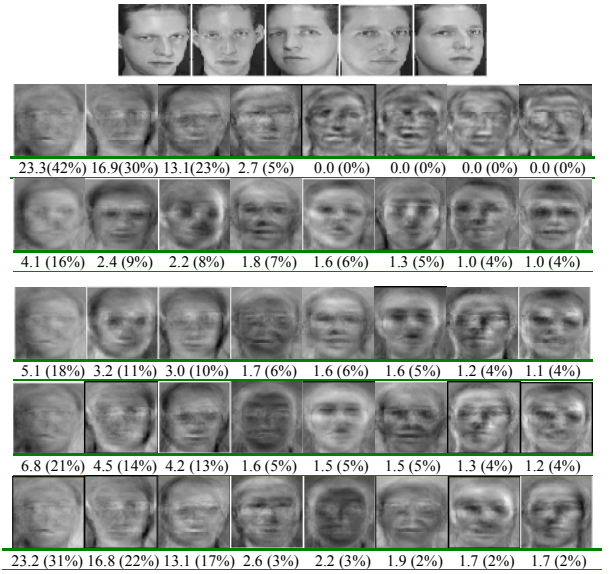

Fig. 1a. Visual analysis of a ORL subject (top-down): image examples of a subject, image eigenvectors with eigenvalues of the corresponding sample group covariance matrix; pooled covariance matrix; ML covariance (w=0.9 and w_{avg}=0.92); MC covariance (w=0.8 and w_{avg}=0.61), and ME covariance

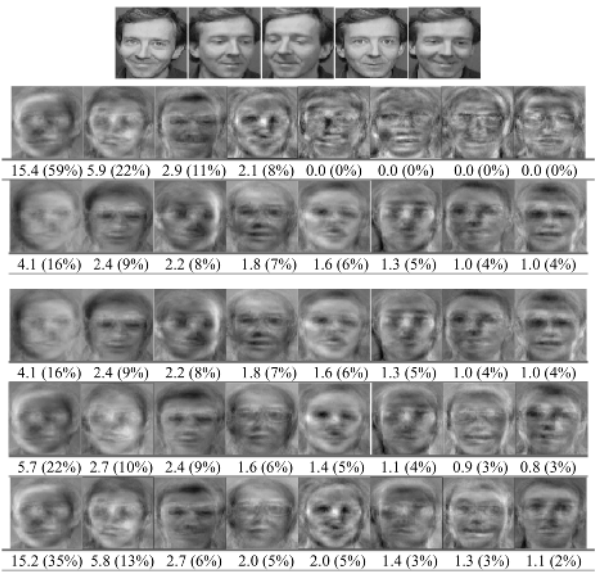

Fig. 1b. Analogous to Fig.1a but with ML covariance parameter w=1.0 (w_{avg}=0.83)

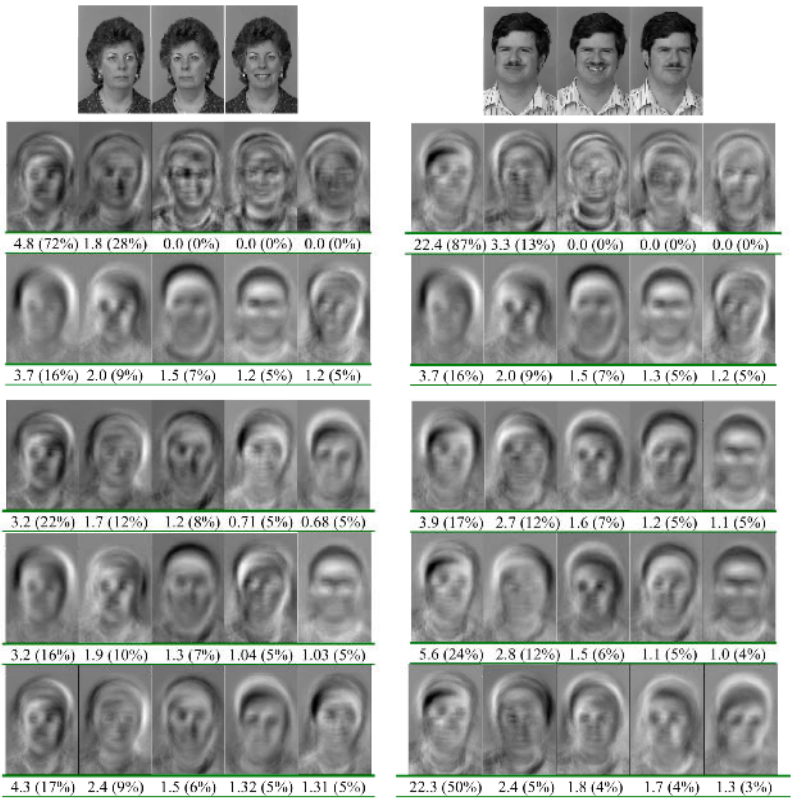

Fig. 2a. Visual analysis of a FERET subject (top-down): image examples of a subject, image eigenvectors with eigenvalues of the corresponding sample group covariance matrix; pooled covariance matrix; ML covariance (w=0.5 and w_{avg}=0.62); MC covariance (w=0.8 and w_{avg}=0.85), and ME covariance

Fig. 2b. Analogous to Fig. 2a but with the ML covariance parameter w=0.9 (w_{avg}=0.95)

Table 1. ORL classification results

Eigenfaces	Pooled	Smix - ML	Smix - MC	Smix - ME
10	88.4% (1.4%)	91.9% (1.6%)	93.8% (1.7%)	93.5% (1.5%)
20	91.8% (1.8%)	94.4% (1.7%)	94.7% (1.4%)	95.2% (1.8%)
40	95.4% (1.5%)	96.2% (1.5%)	96.5% (1.6%)	96.7% (1.5%)
60	95.0% (1.6%)	95.7% (1.5%)	95.4% (1.6%)	95.9% (1.6%)
80	94.6% (1.9%)	94.9% (1.7%)	94.7% (1.9%)	94.8% (1.7%)

Table 2. FERET classification results.

Eigenfaces	Pooled	Smix - ML	Smix - MC	Smix - ME
10	94.9% (1.1%)	94.7% (1.4%)	95.3% (1.1%)	95.3% (1.2%)
30	96.8% (0.8%)	96.6% (1.1%)	97.0% (0.9%)	97.2% (1.0%)
50	96.9% (0.8%)	96.7% (1.1%)	97.3% (1.0%)	97.8% (0.9%)
70	96.7% (0.9%)	96.5% (0.9%)	96.9% (0.9%)	97.3% (0.9%)

6 Conclusion

In this paper, a visual study of three mixture covariance matrix approaches for the quadratic discriminant (QD) classifier has been undertaken in the context of face recognition. This analysis allows a better understanding of the importance and applicability of blending the sample group covariance matrices and the pooled covariance one in small sample size, high-dimensional problems. The maximum entropy approach that preserves as much of the sample group covariance information as possible achieved a more intuitive visual performance and the best classification accuracies, especially in face experiments where moderate changes in facial expressions, pose, and scale, occurred.

The QD classifier is a simple and powerful classifier in eigenface recognition. The experiments performed show clearly that the use of mixture covariance matrices is worth consideration for improving recognition, especially when implemented by using the straight-forward maximum entropy approach.

Acknowledgment

The first author was partially supported by the Brazilian Government Agency CAPES under grant No. 1168/99-1. Also, portions of the research in this paper use the FERET database of facial images collected under the FERET program.

References

[1] C. Liu and H. Wechsler, "Probabilistic Reasoning Models for Face Recognition". In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'98, Santa Barbara, California, USA, June 1998.

[2] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa, "Using Mixture Covariance Matrices to Improve Face and Facial Expression Recognitions", in Proc. of *3rd International Conference of Audio- and Video-Based Biometric Person Authentication AVBPA'01*, Springer-Verlag LNCS 2091, pp. 71-77, Halmstad, Sweden, June 2001.

[3] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa. A New Quadratic Classifier applied to Biometric Recognition. In proceedings of the Post-ECCV Workshop on Biometric Authentication, Springer-Verlag LNCS 2359, pp. 186-196, Copenhagen, Denmark, June 2002.

[4] E. T. Jaynes, "On the rationale of maximum-entropy methods", *Proceedings of the IEEE*, vol. 70, pp. 939-952, 1982.

[5] J.H. Friedman, "Reguralized Discriminant Analysis", *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-175, March 1989.

[6] 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.

[7] J.R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. Chichester: John Wiley & Sons Ltd., 1999.
[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second edition. Boston: Academic Press, 1990.
[9] 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.
[10] M. Turk and A. Pentland, "Eigenfaces for Recognition, "Journal of Cognitive Neuroscience, Vol. 3, pp. 72-85, 1991.
[11] P. J. Phillips, H. Wechsler, J. Huang and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms", *Image and Vision Computing Journal*, vol. 16, no. 5, pp. 295-306, 1998.
[12] R.A. Johnson R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*, by Prentice-Hall, Inc., 3d. edition, 1992.
[13] T. Greene and W.S. Rayens, "Covariance pooling and stabilization for classification", *Computational Statistics & Data Analysis*, vol. 11, pp. 17-42, 1991.
[14] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, second edition. New York: John Wiley & Sons, 1984.