Separate-Group Covariance Estimation With Insufficient Data for Object Recognition

*Carlos Eduardo Thomaz[1], Raul Queiroz Feitosa[2], Álvaro Veiga[3]

[1,2,3]Catholic University of Rio de Janeiro
Department of Electrical Engineering
r. Marquês de São Vicente 225,22453-900, Rio de Janeiro, Brazil

[2]Department of Computer Engineering
University of Rio de Janeiro
r. São Francisco Xavier, 524, 20559-900, Rio de Janeiro, Brazil

{cethomaz,raul,alvf}@ele.puc-rio.br

Abstract. Many similarity measures used for classification involve the inverse of the group covariance matrices. However, the number of observations available in the training set for each group is, in many cases, significantly inferior to the dimension of the feature space, what implies that the sample covariance matrix is singular. A common solution to this problem is to assume the same covariance matrix for all groups using the pooled covariance matrix computed from the whole training set. This paper investigates an alternative estimate for the group covariance matrices, called Mixed Covariance, given by a linear combination of the sample group and pooled covariance matrices. This estimate has the same rank of the pooled covariance matrix without assuming equal covariance for all groups. Experiments were carried out to evaluate the performance associated with the proposed estimate in two automatic recognition applications: face and facial expression. The average recognition rates obtained by using the mixed covariance were higher than the usual sample group and pooled covariance estimates.

1. Introduction

Many similarity measures used for classification involve the inverse of the group covariance matrices. Since in practical cases these matrices are not known, estimates must be computed based on the patterns available in a training set. The usual choice for the estimate of the covariance matrices is the sample group covariance. However, the number of training examples for each group is, in many cases, significantly less than the dimension of the feature space. This implies that the sample covariance matrix will be singular.

A common solution to this problem is to assume that all populations have the same covariance matrix and to use the pooled covariance estimate computed from the whole training set. The resulting matrix will have the same rank as the data matrix of the training set.

This paper investigates a new estimate for the group covariance matrices, called mixed covariance, given by a linear combination of the sample group and the pooled covariance estimates. It has the property of having the same rank as the pooled estimate, while allowing a different estimate for each group, what may imply in a better modeling of the population involved in the problem.

In order to evaluate the proposed approach two pattern recognition applications were considered: automatic face recognition and automatic facial expression recognition. The evaluation used different image database for each application and two dimensionality reduction techniques were applied: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The mixed covariance estimate attained the best performance in the experiments for both applications.

2. Maximum Probability Classifiers

The basic problem in the decision-theoretic methods for pattern recognition consists of finding a set of g discriminant functions $d_1(\mathbf{x})$, $d_2(\mathbf{x})$, ..., $d_g(\mathbf{x})$, (g is the number of groups) with the property that if the n-dimensional pattern vector $\mathbf{x}$ belongs to the group π_i ($1 \le i \le g$) then $d_i(\mathbf{x}) \ge d_j(\mathbf{x})$, for all $i \ne j$, $1 \le i \le g$.

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 respective probability density functions, that is, $d_i(\mathbf{x}) = f_i(\mathbf{x})$ for all classes.

Many proposed pattern recognition systems assume that the population of all groups can be properly modeled by a multivariate normal distribution [13]. Its density function can be expressed by:

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

for $1 \le i \le g$, where μ_i and Σ_i are respectively the mean and the covariance of group π_i. Since those values are seldom available, estimates must be provided. This works

* C.E. Thomaz (cet@doc.ic.ac.uk) is currently starting a Ph.D. in Computer Science (Visual Information Processing) at Imperial College, London.

focus on the usual sample estimate for the mean and on three estimates for the covariance, as described below.

2.1. Sample Group Covariance Matrix

The most commonly used estimate for Σ_i is the Sample Group Covariance matrix defined by:

$$\mathbf{S}_i = \frac{1}{k_i - 1} \sum_i (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i), \text{ for } 1 \le i \le g \quad (2)$$

where $\mathbf{x}_{ij}$ are the training examples of group π_i, k_i is the number of them, and $\bar{\mathbf{x}}_i$ is the corresponding sample mean. By using this estimate, equation (1) takes the form:

$$d_{sample\,i}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{S}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)\right], \quad (3)$$

The matrix $\mathbf{S}_i$ will be singular if k_i is less than the dimension of the feature space.

2.2. Pooled Covariance Matrix

One way to get around this problem is to assume that all groups have equal covariance matrices, and to use as its estimate the weighted average of each sample group covariance matrix, given by

$$\mathbf{S}_{pooled} = \frac{(k_1 - 1)\mathbf{S}_1 + (k_2 - 1)\mathbf{S}_2 + \cdots + (k_g - 1)\mathbf{S}_g}{k_1 + k_2 + \cdots + k_g - g}. \quad (4)$$

By introducing the equal covariance assumption in equation (1), and after some simplifications [13], the following set of discriminant functions can be derived:

$$d_{pooled\,i}(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}_{pooled}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i), \text{ for } 1 \le i \le g \quad (5)$$

Since more observations are taken to calculate $\mathbf{S}_{pooled}$, it will pottentially have a higher rank than $\mathbf{S}_i$ and will be eventually full rank. Although the pooled estimate does provide a solution for the algebric problem arising from the insufficient number of training patterns in each group, assuming equal covariance for all groups may bring about distortions in the modeling of the recognition problem.

2.3. Mixed Sample Covariance Matrix

The Mixed Covariance Matrix is a tradeoff between $\mathbf{S}_{pooled}$ and $\mathbf{S}_i$. It is given by

$$\mathbf{S}_{mix\,i} = a\mathbf{S}_{pooled} + (1 - a)\mathbf{S}_i, \text{ where } 0 < a < 1. \quad (6)$$

Figure 1 gives the geometric interpretation of the proposed Mixed Sample Covariance Matrix. The ellipsoids correspond to the contour of the constant density for three groups. The dashed gray lines represent the different sample group covariance estimates, while the pooled estimate is represented in dotted gray lines. The proposed mixed sample estimates assume that the ellipsoid corresponding to the true covariance is placed somewhere between both ellipsoids, as shown by the solid black lines.

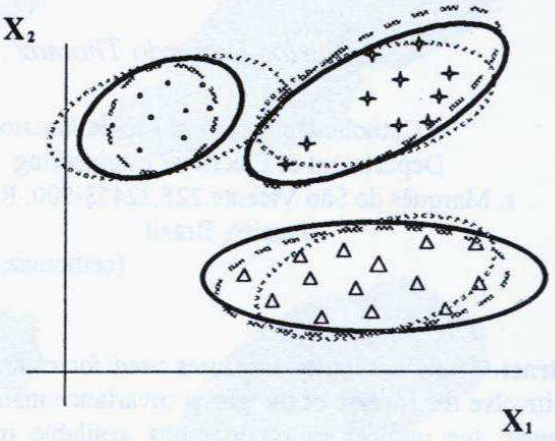

Figure 1: Geometric interpretation of $\mathbf{S}_i$, $\mathbf{S}_{pooled}$ and $\mathbf{S}_{mix}$.

Each $\mathbf{S}_{mix\,i}$ matrix has the important property of admitting an inverse if $\mathbf{S}_{pooled}$ does. Let the dimension of the feature space n be such that $\mathbf{S}_{pooled}$ is full rank (invertible) and $\mathbf{S}_i$ is not full rank (non invertible). Thus $\mathbf{S}_{pooled}$ and $\mathbf{S}_i$ are respectively positive definite and positive semi-definite matrices. Since $0<a<1$, a and $(1-a)$ are both positive numbers. Therefore, $a\,\mathbf{S}_{pooled}$ and $(1-a)\mathbf{S}_i$ are still positive definite and positive semi-definite matrices. For a matrix **A** positive definite and a matrix **B** positive semi-definite the next inequality is valid [5]

$$\det(\mathbf{A} + \mathbf{B}) \ge \det(\mathbf{A}) \quad (7)$$

Hence,

$$\det(\mathbf{S}_{mix\,i}) = \det[a\mathbf{S}_{pooled} + (1 - a)\mathbf{S}_i] \ge a\det(\mathbf{S}_{pooled}) > 0 \quad (8)$$

Since $a > 0$, this implies that $\mathbf{S}_{mix}$ will be non singular. By using $\mathbf{S}_{mix}$ in the place of the group covariance matrix, equation (1) takes the form:

$$d_{mix\,i}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{S}_{mix\,i}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}_{mix\,i}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)\right]. \quad (9)$$

3. Projection spaces in recognition

One of the most successful approaches to the problem of creating a low dimensional image representation is based on Principal Component Analysis (PCA). It was firstly proposed by Sirovich and Kirby [7] for representing face images. After the classical work of Turk and Pentland [8], many extensions were proposed [1,2,10,11,12,15]. Swets and Weng called the selected eigenfaces the Most Expressive Features (MEF) [3], since they give the minimum mean square reconstruction error and describe the major variations in the set of samples.

While the goal of PCA is to minimize the reconstruction error, the Linear Discriminant Analysis (LDA) provides a procedure to determine a set of axes whose projections of different groups have the maximum separa-

tion. It can be shown [3,13] that the discriminants axes are given by the eigenvectors of $\mathbf{C} = \mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ is the sample within groups covariance matrix, and $\mathbf{B}$ is the sample between groups covariance matrix.

In many applications, like these considered in this work, the number of patterns available in the sample (K) is less than the number of features and $\mathbf{W}$ is singular. As Swets and Weng [3] observed, this problem can be circumvent by first applying the PCA to the entire sample and then choosing p, the number of Most Expressive Features, such that $p+g \leq K$. The LDA procedure is then applied on the sample projected on the MEF subspace, where the matrix $\mathbf{W}$ will be full rank. The resulting g-1 axes define the basis of a new subspace, called the Most Discriminant Features space (MDF) [3].

4. Experiments

With the evaluation purpose six distinct recognition systems were built. Each system is characterized by one out of two basis (MEF and MDF) and one out of three estimates for the covariance matrices ($\mathbf{S}_i$, $\mathbf{S}_{pooled}$,$\mathbf{S}_{mix\ i}$). In all systems the maximum probability classifier was applied, each time using one of the three covariance estimates. The linear combination factor a corresponding to the mixed sample covariance matrix proposed assumed the following values: 0.1, 0.3, 0.5, 0.7 and 0.9.

4.1. Database

The experiments to evaluate the classification schemes for the face recognition problem make use of the ORL Face Database [4,14], containing ten images for each of 40 individuals, a total of 400 images.

The facial expression database was provided by the Tohoku University [6,9]. It is composed of 193 images of expressions posed by nine Japanese females. Each person posed three or four examples of each six fundamental facial expressions: angry, disgusting, fear, happy, sad and surprised, as define in [9]. The database has at least 29 images for each fundamental facial expression. For implementation convenience all images were first resized to 64x64 pixels.

4.2. Training and Testing Sets

The face recognition classifiers were implemented using for each individual 5 images to train and 5 images to test. For the MEF's computation, the PCA eigenvectors corresponding to the top 70 eigenvalues were kept. The Experiments have shown that the use of more than 70 components brings no performance improvement.

Since the number of MDF's is limited by the number of groups and there are 40 people to recognize, the MDF features took values in the range [1, 39]. The MDF's were computed based on the first 50 MEF's. No improvement was observed in the experiments by using more than 50 MEF's in the computation of the MDF's.

For the facial expression recognition problem, 29 images of each fundamental facial expression were used. The training image set included a total of 120 images, consisting of 20 images of each facial expression, and the testing set contains the remaining 54 expression images. Our experiments on the MEF subspace used till 65 principal components and the computation of the MDF's was based on the first 55 MEF's. The number of MDF's was limited by 5, since there are 6 fundamental facial expression to recognize.

5. Results

The main results of the experiments are summarized in figures 2, 3 and 4. Figure 2 shows the average recognition rate for each covariance matrix estimate as a function of the number of MEF components. The six curves represent the performance of the $\mathbf{S}_{pooled}$ estimate against the five $\mathbf{S}_{mix}$ estimates corresponding to a linear combination factor a equals to 0.1, 0.3, 0.5, 0.7 and 0.9. Since only 5 images of each individual were used to form the training set, the results relative to the sample group covariance estimate were limited to 4 MEF components and, therefore, does not appear on the graph of figure 2. It shows that the $\mathbf{S}_{mix}$ produced a better performance than the $\mathbf{S}_{pooled}$ estimate, for all number of MEF components and for all values of the linear combination factor a considered in these experiments. The best recognition rate for the $\mathbf{S}_{mix}$ estimate was 96.88% with 40 MEF components and a equals to 0.7. The performance is similar to the best results reported in previous works [4,14] which used the same database.

Figure 3 shows the average recognition rate for each covariance estimate as a function of the number of MDF components. Again, because of the training size of each group, the sample group covariance results are not shown. No evident superiority of any estimate is shown in figure 3. Depending on the value of the linear combination factor a and on the number of components, $\mathbf{S}_{mix}$ may have a better or worse performance than $\mathbf{S}_{pooled}$. The best recognition rate – 96.52% – for all the MDF components considered is reached by the $\mathbf{S}_{mix}$ estimate with the linear factor a equals to 0.7 for 39 MDF components.

One of the results of the facial expression recognition is summarized in figure 4. In this case, seven curves representing the performance of the $\mathbf{S}_{group}$, $\mathbf{S}_{pooled}$, and the five $\mathbf{S}_{mix}$ estimates (for a = 0.1;0.3;0.5;0.7;0.9) are presented. From figure 4 it can be observed that for more than 20 MEF components, in which case the $\mathbf{S}_{group}$ becomes singular, the $\mathbf{S}_{mix}$ estimate reaches better recognition rates than the $\mathbf{S}_{pooled}$ estimate for all values of a considered in the experiment. The best recognition rate of these experiments – 85.19% – was obtained by the $\mathbf{S}_{mix}$ estimate with a linear combination factor a equals to 0.3 and for 65 MEF components. This performance is similar to the best results reported so far [6,9] for the same database.

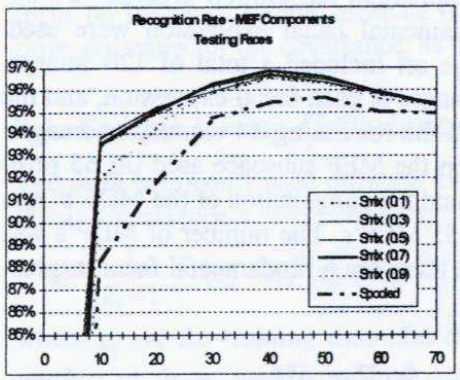

Figure 2: Face Recognition

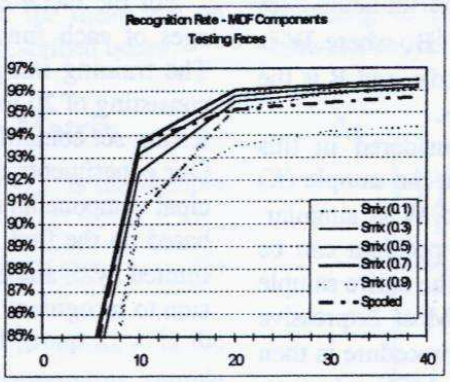

Figure 3: Face Recognition

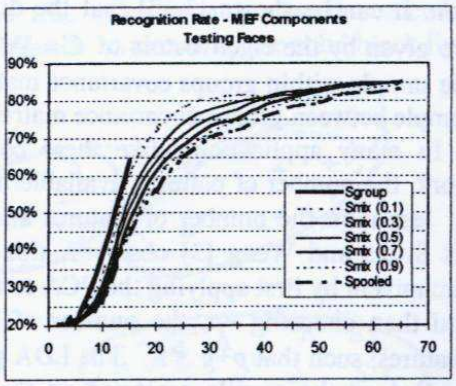

Figure 4: Facial Expression Recognition

There is no clear superiority of any covariance estimator for the facial expression recognition as a function of the number of MDF components. Therefore, these results are not shown.

6. Conclusion

This work proposed a new estimate for the covariance matrix for object recognition applications, called mixed covariance matrix. The new estimate has the same rank of the data matrix and is, therefore, invertible even in the cases where the usual sample group estimate does not admit an inverse due to an insufficient number of patterns for each group in the training set. Behind this advantage, the proposed estimate does not assume equal covariance matrices for all groups. This allows a better representation for the population of each group.

Extensive experiments were carried out to evaluate this approach on two recognition tasks: face recognition and facial expression recognition. A maximum probability classifier was built using the proposed estimate, the usual sample group and pooled estimates. In both tasks the best recognition performance was reached by the mixed group covariance estimate, especially when the MEF projection was used.

Acknowledgments

The first author was partially supported by the Brazilian Government Agencies CNPq and CAPES.

References

1. A. Pentland, B. Moghaddam, and T. Strainer, "View-based and modular eigenspaces for face recognition", CVPR, June, 1994.
2. A. Pentland, et al., "Experiments with Eigenfaces", International Joint Conference on Artificial Intelligence, Chamberry, France, August ,1993.
3. D. L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieva," IEEE Ttrans. on PAMI, Vol. 18, pp. 831-836, Aug. 1996.
4. F. Samaria and A. Harter, "Parametrisation of a stochastic model for human face identification", Proc. 2nd IEEE workshop on Applications of Computer Vision, 1994.
5. J.R. Magnus and H. Neudecker, "Matrix Differencial Calculus with Applications in Statistics and Econometrics", Wiley and Sons Limited, pg. 21-21, 1995.
6. M.J. Lyons, J.Budynek, and S.Akamatsu, "Automatic Classification of Single Facial Images", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1357-1362, December 1999.
7. 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.
8. M. Turk and A. Pentland, "Eigenfaces for Recognition, "Journal of Cognitive Neuroscience, Vol. 3, pp. 72-85, 1991.
9. P. Ekman and W.V.Friesen, "Pictures of Facial Affect", Human Interaction Laboratory, Univ. of California Medical Center, San Francisco, 1976.
10. P. J. B. Hancock, A. M. Burton and V. Bruce, "Face processing: human perception and principal components analysis", Memory and Cognition, 1996.
11. P. J. B. Hancock, V. Bruce and A. M. Burton, "Testing Principal Component Representations for Faces", Proc. of 4th Neural Computation and Psychology Workshop, 1997.
12. P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs FisherFace: Recognition Using Class Specific Linear Projection", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, July 1997.
13. R.A. Johnson R.A. and D.W. Wichern, "Applied Multivariate Statistical Analysis", by Prentice-Hall, Inc., 3d. edition, 1992.
14. S.Lawrence, C. L. Giles, A.C. Tsoi and A. D. Back, "Face Recognition: A Convolutional Neural-Network Approach", IEEE Trans. Neural Networks, vol. 8, no. 1, pp. 98-113, January 1997.
15. 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.