Small Sample Problem in Bayes Plug-in Classifier for Image Recognition

Carlos E. Thomaz 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

Raul Q. Feitosa
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

The Bayes plug-in classifier has been successfully applied to discriminate high dimensional data. This classifier is based on similarity measures that involve the inverse of the sample group covariance matrices. These matrices, however, are singular in "small sample size" problems. Therefore, other methods of covariance estimation have been proposed where the sample group covariance estimate is replaced by covariance matrices of various forms. In this paper, a new covariance estimator is proposed and compared with two covariance estimators known as RDA and LOOC. The new estimator does not require an optimisation procedure, but an eigenvector-eigenvalue ordering process to select information from the projected sample group covariance matrices whenever possible and the pooled covariance otherwise. The effectiveness of the method is shown by experimental results carried out on face and facial expression recognition, using different databases for each application.

1. Introduction

Image pattern recognition problems, especially face and facial expression ones, are examples of "small sample size" problems. In such applications there are a large number of features available but the number of training samples for each pattern is considerably less than the dimension of the feature space.

The Bayes plug-in classifier has been successfully applied to discriminate high dimensional data [2,7,10,11]. This classifier is based on similarity measures that involve the inverse of the true covariance matrix of each class. 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 estimating true covariance matrices is the maximum likelihood estimator defined by the corresponding sample group covariance matrices. However, in "small sample size" applications the sample group covariance matrices are singular.

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. The decision concerning whether to choose the sample group covariance matrices or the pooled covariance matrix represents a limited set of estimates for the true covariance matrices [3]. Therefore, other approaches have been applied not only to overcome the small size effect but also to provide higher classification accuracy.

In this paper, a new covariance estimator is proposed and compared with two covariance estimators known as RDA [3] and LOOC [7]. Experiments were carried out to evaluate these approaches on face and facial expression recognition, using different databases for each application. The effectiveness of the new covariance method is shown by the results.

2. The Bayes Plug-in Classifier

The Bayes plug-in classifier, also called the Gaussian maximum likelihood classifier, is based on the *p*-multivariate normal or Gaussian class-conditional probability densities.

Assuming that all of the *g* groups or classes have the same prior probabilities, the optimal Bayes classification rule may be specified as: Assign pattern *x* to class *i* if

$$d_i(x) = \min_{1 \le j \le g} [\ln\left|\Sigma_j\right| + (x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j)] = \min_{1 \le j \le g} d_j(x) \qquad \textbf{(1)}$$

where μ_j and Σ_j are the true class *j* population mean vector and covariance matrix. The Bayes classification described in (1) is also known as the quadratic discriminant rule (QD).

In practice, however, the true values of the mean and covariance matrix are seldom known and must be replaced by their respective estimates calculated from the training samples available. The mean is estimated by the usual sample mean $\bar{x}_i$ which is the maximum likelihood estimator of μ_i. The covariance matrix is commonly estimated by the sample group covariance matrix S_i which is the unbiased maximum likelihood estimator of Σ_i.

From replacing the true values of the mean and covariance matrix in (1) by their respective estimates ("plug-in"), the QD rule can be rewritten as: Assign pattern *x* to class *i* that *minimizes* the generalized distance between *x* and $\bar{x}_i$

$$d_i(x) = \ln\left|S_i\right| + (x - \bar{x}_i)^T S_i^{-1}(x - \bar{x}_i) \, . \qquad (2)$$

2.1. "Small Sample Size" Problems

It is well known that the misclassification rate defined in (2) approaches the optimal rate obtained by equation (1) only when the sample sizes in the training set approach infinity [1].

In fact, the performance of (2) can be seriously degraded in small samples due to the instability of the sample estimators [9]. For *p*-dimensional patterns the use of S_i is especially problematic if less than *p* + 1 training observations from each class are available, that is, the sample group covariance matrix is singular if the number of observations of each group is less than the dimension of the feature space.

One method routinely applied to overcome the "small sample size" problem and consequently deal with the singularity and instability of the S_i is to employ the so-called linear discriminant rule (LD) which is obtained by replacing the S_i in (2) with the pooled sample covariance matrix

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{N - g}, \qquad (3)$$

where n_i is the number of training observations from class *i*, *g* is the number of groups or classes and $N = n_1 + n_2 + \cdots + n_g$.

Since more observations are taken to calculate the pooled covariance matrix, S_p will potentially have a higher rank than S_i (and be eventually full rank). Theoretically, however, S_p is a consistent estimator of the true covariance matrices Σ_i only when Σ_1=Σ_2=…=Σ_g.

3. Covariance Estimators

The sample group covariance matrices S_i (or QD classifier) and the pooled covariance matrix S_p (or LD classifier) represent a limited set of estimates for the true covariance matrices Σ_i, particularly in small sample size problems. Therefore, other estimators have been applied not only to overcome these problems but also to provide higher classification accuracy. In the next sub-sections, the Friedman's RDA [3] and the Hoffbeck's LOOC [7] methods are briefly described. A new covariance estimator is proposed in section 4.

3.1. Friedman's RDA Method

The Friedman's RDA method is basically a two-dimensional optimisation method that shrinks both the S_i towards S_p and also the eigenvalues of the S_i towards equality by blending the first shrinkage with multiples of the identity matrix.

In this context, the sample covariance matrices S_i of the discriminant rule defined in (2) are replaced by the following covariance estimator

$$\begin{aligned} S_i^{rda}(\lambda,\gamma) &= (1-\gamma)S_i^{rda}(\lambda) + \gamma\left(\frac{tr(S_i^{rda}(\lambda))}{p}\right)I, \\ S_i^{rda}(\lambda) &= \frac{(1-\lambda)(n_i - 1)S_i + \lambda(N-g)S_p}{(1-\lambda)n_i + \lambda N} \end{aligned} \quad , \qquad (4)$$

where the notation "tr" denotes the trace of a matrix. The RDA mixing parameters λ and γ are restricted to the range 0 to 1 (optimisation grid) and are selected to maximise the leave-one-out classification accuracy based on the rule defined in (2).

Although the RDA method is theoretically a well-established approach, it has practical drawbacks. Despite the substantial amount of computation saved by taking advantage of matrix updating formulas [3], RDA is a computationally intensive method. For each point on the two-dimensional optimisation grid, RDA requires the evaluation of the proposed estimates of every class. In situations where a large number of *g* groups is considered, the RDA seems to be unfeasible. In addition, as RDA maximises the classification accuracy calculating all covariance estimates simultaneously, it is restricted to using the same value of the mixing parameters for all the classes. These same values may not be optimal for all classes.

3.2. Hoffbeck's LOOC Method

In practical situations, it seems appropriate to allow covariance matrices to be estimated by distinct mixing parameters. Hoffbeck [7] has proposed a leave-one-out covariance estimator (LOOC) that depends only on covariance estimates of single classes.

In LOOC each covariance estimate is optimised independently and a separate mixing parameter is computed for each class based on the corresponding likelihood information. The idea is to examine pair-wise mixtures of the sample group covariance estimates S_i

and the unweighted common covariance estimate S, defined as

$$S = \frac{1}{g}\sum_{i=1}^{g} S_i \,, \tag{5}$$

together with their diagonal forms. The LOOC estimator has the following form:

$$S_i^{looc}(\alpha_i) = \begin{cases} (1-\alpha_i)\mathrm{diag}(S_i) + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\mathrm{diag}(S) & 2 < \alpha_i \le 3 \end{cases} \tag{6}$$

The mixing or shrinkage parameter α_i determines which covariance estimate or mixture of covariance estimates is selected. The strategy consists of evaluating several values of α_i over the optimisation grid $0 \le \alpha_i \le 3$, and then choosing α_i that maximizes the average log likelihood of each corresponding *p*-variate normal density function [7].

The computation of the LOOC estimate requires only one density function be evaluated for each point on the α_i one-dimensional optimisation grid, but also involves calculating the inverse and determinant of the (*p* by *p*) matrix $S_i^{looc}(\alpha_i)$ for each training observation belonging to the *i*th class. Analogously, Hoffbeck has reduced the LOOC required computation by considering valid approximations of the covariance estimates and using rank-one updating formulas [7]. Therefore, the final form of LOOC requires less computation than RDA estimator.

4. A New Covariance Estimator

Friedman's RDA and Hoffbeck's LOOC approaches described in the previous section, and several other similar methods [5,6,11] not described in this report, optimised linear combinations of the sample group covariance matrices and, for instance, the pooled covariance matrix. Not only does this overcome the "small sample size" problem but it also achieves better classification accuracy than LD and standard QD classifiers.

In situations, however, where S_i are singular, such approaches may lead to inconvenient biasing mixtures. This statement, which is better explained in the following sub-section, forms the basis of the new covariance estimator idea, called Covariance Projection Ordering method.

4.1. Covariance Projection Ordering Method

The Covariance Projection Ordering estimator (COPO) examines the combination of the sample group covariance matrices and the pooled covariance matrix in the QD classifiers using their spectral decomposition representations. This new estimator has the property of having the same rank as the pooled estimate, while allowing a different estimate for each group.

First, in order to understand the aforementioned inconvenient biasing mixtures, let a matrix S_i^{mix} be given by the following linear combination:

$$S_i^{mix} = aS_i + bS_p \,, \tag{7}$$

where the mixing parameters a and b are positive constants, and the pooled covariance matrix S_p is a non-singular matrix. The S_i^{mix} eigenvectors and eigenvalues are given by the matrices Φ_i^{mix} and Λ_i^{mix}, respectively. From the covariance spectral decomposition formula [4], it is possible to write

$$(\Phi_i^{mix})^T S_i^{mix} \Phi_i^{mix} = \Lambda_i^{mix} = \begin{bmatrix} \lambda_1^{mix} & & & 0 \\ & \lambda_2^{mix} & & \\ & & \ddots & \\ 0 & & & \lambda_p^{mix} \end{bmatrix} = diag[\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_p^{mix}] \tag{8}$$

where $\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_p^{mix}$ are the S_i^{mix} eigenvalues and *p* is the dimension of the measurement space considered. Using the information provided by equation (7), equation (8) can be rewritten as:

$$\begin{aligned} (\Phi_i^{mix})^T S_i^{mix} \Phi_i^{mix} &= diag[\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_p^{mix}] \\ &= (\Phi_i^{mix})^T \left[aS_i + bS_p \right] \Phi_i^{mix} \\ &= a(\Phi_i^{mix})^T S_i \Phi_i^{mix} + b(\Phi_i^{mix})^T S_p \Phi_i^{mix} \\ &= a\Lambda^{i*} + b\Lambda^{p*} \\ &= diag[a\lambda_1^{i*} + b\lambda_1^{p*}, a\lambda_2^{i*} + b\lambda_2^{p*}, ..., a\lambda_p^{i*} + b\lambda_p^{p*}] \end{aligned} \tag{9}$$

where $\lambda_1^{i*}, \lambda_2^{i*}, ..., \lambda_p^{i*}$ and $\lambda_1^{p*}, \lambda_2^{p*}, ..., \lambda_p^{p*}$ are the corresponding spread values of sample group covariance and pooled covariance matrices spanned by the S_i^{mix} eigenvectors matrix Φ_i^{mix}. Then, the discriminant score of the QD rule in spectral decomposition form becomes

$$d_i(x) = \sum_{k=1}^{p} \ln\left(a\lambda_k^{i*} + b\lambda_k^{p*}\right) + \sum_{k=1}^{p} \frac{[(\phi_{ik}^{mix})^T (x - \bar{x}_i)]^2}{a\lambda_k^{i*} + b\lambda_k^{p*}} \,, \tag{10}$$

where ϕ_{ik}^{mix} is the corresponding *k*-th eigenvector of the matrix S_i^{mix}.

As can be observed, the discriminant score described in equation (10) considers the dispersions of sample group covariance matrices spanned by all the S_i^{mix} eigenvectors. Therefore, in problems where the group sample sizes n_i are small compared with the dimension of the feature space *p*, the corresponding ($p - n_i + 1$) lower dispersion values are often estimated to be 0 or approximately 0, indicating that these values are not reliable. In this way, a linear combination as defined in equation (7) of the sample group covariance matrix and the pooled covariance in a subspace where the former is poorly represented seems to be not convenient. Other covariance estimators have used the same parameters *a* and *b* defined in equation (7) for the whole feature space and consequently have not addressed this problem.

The COPO estimator is a simple approach to overcome this problem. Basically, the idea is to use all the sample group covariance information available whenever possible and the pooled covariance information otherwise. Regarding equations (7) and (9), this idea can be derived as follows:

$$\begin{aligned} S_i^{copo} &= \sum_{k=1}^{p} \lambda_k^{copo} \phi_{ik}^{copo} (\phi_{ik}^{copo})^T \text{ , where} \\ \lambda_k^{copo} &= \begin{cases} \lambda_k^{i*} & \text{if } 1 \le k \le rank(S_i), \\ \lambda_k^{p*} & \text{otherwise,} \end{cases} \end{aligned} \qquad \textbf{(11)}$$

and ϕ_{ik}^{copo} is the corresponding *k*-th eigenvector of the matrix given by $S_i + S_p$ ordered in λ_k^{i*} decreasing values. Then the discriminant scored described in equation (10) becomes:

$$d_i(x) = \sum_{k=1}^{r} \ln \lambda_k^{i*} + \sum_{k=r+1}^{p} \ln \lambda_k^{p*} + \sum_{k=1}^{r} \frac{[(\phi_{ik}^{copo})^T (x - \bar{x}_i)]^2}{\lambda_k^{i*}} + \sum_{k=r+1}^{p} \frac{[(\phi_{ik}^{copo})^T (x - \bar{x}_i)]^2}{\lambda_k^{p*}} \qquad \textbf{(12)}$$

where $r = rank(S_i)$.

The COPO estimator provides a new combination of the sample group covariance matrices and the pooled covariance matrix in such a way that this combination is strongly related to the rank of S_i or, equivalently, to the number of training samples n_i . It can be viewed as a *p*-dimensional non-singular approximation of an *r*-dimensional singular matrix.

The COPO method does not require an optimisation procedure, but an eigenvector-eigenvalue ordering process to select information from the projected sample group covariance matrices whenever possible and the pooled covariance otherwise. Therefore, the computational issues regarding the Covariance Projection Ordering approach is less severe than the Friedman's RDA and Hoffbeck's LOOC approaches. In addition, the COPO method is not restricted to use the same covariance combination for all classes, allowing covariance matrices to be distinctly estimated.

5. Experiments and Results

In order to evaluate the Covariance Projection Ordering (COPO) approach, two image recognition applications were considered: face recognition and facial expression recognition. In the face recognition experiments the ORL Face Database was used, which contains ten images for each of 40 individuals, a total of 400 images. The Tohoku University has provided the database for the facial expression experiment. This database [8] 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.

5.1. Experiments

The experiments were carried out as follows. First the well-known dimensionality reduction technique called Principal Component Analysis (PCA) [12] reduced the dimensionality of the original images and secondly the Bayes plug-in classifier using one of the five covariance estimators was applied: 1) Sample group covariance (Sgroup); 2) Pooled covariance (Spooled); 3) Covariance projection ordering (Scopo); 4) Friedman's covariance (Srda); 5) Hoffbeck's covariance (Slooc).

Each experiment was repeated 25 times using several PCA dimensions. Distinct training and test sets were randomly drawn, and the mean and the standard deviation of the recognition rate were calculated. The face recognition classification was computed using for each individual in the ORL database 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 RDA optimisation grid was taken to be the outer product of $\lambda = [0, 0.125, 0.354, 0.650, 1.0]$ and $\gamma = [0, 0.25, 0.5, 0.75, 1.0]$, identically to the Friedman's work [3]. Analogously, the size of the LOOC mixture parameter [7] was $\alpha_i = [0, 0.25, 0.5, ..., 2.75, 3.0]$.

5.2. Results

Tables 1 and 2 present the training and test average recognition rates (with standard deviations) of the ORL and Tohoku face and facial expression databases, respectively, over the different PCA dimensions. Also presented are the mean of the optimised RDA and LOOC parameters. For the ORL face database, only 6 LOOC parameters corresponding to the subjects 1, 5, 10, 20, 30 and 40 are shown. The notation "-" in the Sgroup rows indicate that the sample group covariance were singular and could not classify the samples.

Table 1 shows that on the training set and for less than 20 PCA components the Scopo estimator led to higher face recognition classification accuracy than the linear covariance estimator (Spooled) and both optimised quadratic discriminant estimators (Srda and Slooc). For the test samples, the Srda and Slooc estimators often outperformed the Scopo in lower dimensional space, but these performances deteriorated when the dimensionality increased, particularly the Slooc ones. It seems that in higher dimensional space, when the Sgroup estimate became extremely poorly represented, the RDA and LOOC parameters did not counteract the Sgroup mixing singularity effect. The Scopo estimator achieved the best recognition rate – 96.4% – for all PCA components considered. In terms of how sensitive the covariance results were to the choice of training and test sets, the covariance estimators similarly had the same performances, particularly in high dimensional space.

	PCAs				
	4	10	20	40	60
Training					
Sgroup	99.5(0.4)	-	-	-	-
Spooled	73.3(3.1)	96.6(1.2)	99.2(0.6)	100.0(0.0)	100.0(0.0)
Scopo	97.0(1.1)	99.9(0.2)	100.0(0.0)	100.0(0.0)	100.0(0.0)
Srda	81.2(2.8)	99.5(0.7)	99.9(0.2)	100.0(0.0)	100.0(0.0)
Slooc	89.4(1.9)	98.9(0.7)	99.6(0.4)	99.8(0.3)	99.9(0.2)
Test					
Sgroup	51.6(4.4)	-	-	-	-
Spooled	59.5(3.0)	88.4(1.4)	91.8(1.8)	95.4(1.5)	95.0(1.6)
Scopo	69.8(3.4)	90.2(2.5)	94.0(1.9)	96.4(1.6)	95.9(1.5)
Srda	64.7(3.9)	92.4(1.9)	94.0(1.4)	96.0(1.7)	95.6(1.6)
Slooc	70.1(3.1)	90.8(2.2)	93.5(2.2)	93.0(1.8)	92.0(1.8)
RDA					
λ	0.1	0.1	0.2	0.3	0.3
γ	0.0	0.0	0.1	0.2	0.3
LOOC					
α1	1.6	2.0	2.6	2.9	3.0
α5	1.3	1.6	1.6	1.7	1.8
α10	2.3	2.4	2.2	2.8	2.9
α20	1.6	1.6	1.9	2.3	2.7
α30	1.5	1.5	1.6	1.6	1.8
α40	1.4	1.6	1.8	2.1	2.5

Table 1. ORL face database results.

	PCAs				
	10	30	50	70	100
Training					
Sgroup	76.3(3.6)	-	-	-	-
Spooled	49.6(3.9)	91.9(1.7)	97.6(1.0)	99.6(0.5)	100.0(0.0)
Scopo	66.6(3.2)	95.8(1.6)	99.2(0.8)	100.0(0.0)	100.0(0.0)
Srda	75.0(6.7)	96.7(2.9)	98.5(1.0)	99.2(1.0)	99.9(0.2)
Slooc	51.4(4.9)	91.0(4.1)	95.8(2.0)	98.8(1.3)	99.9(0.3)
Test					
Sgroup	38.8(5.6)	-	-	-	-
Spooled	26.5(6.8)	70.1(7.8)	79.4(5.8)	83.9(7.0)	84.4(6.5)
Scopo	31.5(5.8)	68.3(5.5)	79.5(5.8)	85.0(7.0)	84.1(6.0)
Srda	37.8(5.9)	73.0(7.4)	80.1(6.2)	79.9(8.7)	81.3(6.7)
Slooc	26.3(5.3)	65.2(5.6)	71.2(8.2)	79.9(8.7)	87.2(5.8)
RDA					
λ	0.0	0.4	0.8	0.7	0.7
γ	0.0	0.0	0.0	0.1	0.3
LOOC					
α1	2.3	0.6	0.9	2.9	3.0
α2	2.4	1.4	2.3	2.9	2.9
α3	2.8	1.0	1.7	2.8	3.0
α4	2.8	2.3	2.1	3.0	3.0
α5	2.8	0.6	0.9	2.3	3.0
α6	2.6	0.5	1.1	2.8	3.0

Table 2. Tohoku facial expression database results.

The results of the Tohoku facial expression recognition are presented in table 2. For more than 50 PCA components on the training set, the Scopo estimator performed as well or better than all the covariance estimators considered. Regarding the test samples, however, there is no overall dominance of any covariance estimator. In lower dimension space, Srda led to higher classification accuracies, followed by Scopo, Spooled and Slooc. On the other hand, when the dimensionality increased and the true covariance matrices became apparently equal and highly ellipsoidal, Srda performed poorly while Scopo, Spooled and Slooc improved. In the highest dimensional space the LOOC optimisation, which considers the diagonal form of the pooled estimate, took advantage of the equal-ellipsoidal behaviour (for more than 70 PCAs all α_i parameters are bigger than the value 2) achieving the best recognition rate – 87.2% – for all PCA components calculated. In this recognition application, all the computed covariance estimators were quite sensitive to the choice of the training and test sets.

6. Conclusion

In this paper, alternative optimised Bayes plug-in covariance estimators available in statistical pattern recognition were described, with regard to the difficulties caused by small sample sizes, and a new covariance estimator was proposed. Experiments were carried out to evaluate these approaches on two real data recognition tasks: face and facial expression recognition. These experiments confirmed that choosing an intermediate estimator between the linear and quadratic classifiers improve the classification accuracy in settings for which samples sizes are small and number of parameters or features is large.

The new covariance estimator, called Covariance Projection Ordering method (COPO), has proved to be a powerful technique in small sample size image recognition problems, especially when concerns about computational costs exist. The new estimator does not require an optimisation procedure, but an eigenvector-eigenvalue ordering processing to select information from the projected sample group covariance matrices whenever possible and the pooled covariance otherwise. This estimator can be viewed as a non-singular approximation of a singular covariance matrix.

The above results are encouraging and comparisons between estimators like RDA and LOOC have to be analysed utilising other pattern recognition problems.

Acknowledgment

The first author was partially supported by the Brazilian Government Agency CAPES under grant No. 1168/99-1.

References

[1] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, second edition. New York: John Wiley & Sons, 1984.

[2] C.Liu and H. Wechsler, “Probabilistic reasoning models for face recognition”, in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, California, USA, June 23-25, 1998.

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

[4] K. Fukunaga, Introduction to Statistical Pattern Recognition, second edition. Boston: Academic Press, 1990.

[5] T. Greene and W.S. Rayens, “Partially pooled covariance matrix estimation in discriminant analysis”, Communications in Statistics-Theory and Methods, vol. 18, no. 10, pp. 3679-3702, 1989.

[6] T. Greene and W.S. Rayens, “Covariance pooling and stabilization for classification”, Computational Statistics & Data Analysis, vol. 11, pp. 17-42, 1991.

[7] J.P. Hoffbeck and D.A. Landgrebe, “Covariance Matrix Estimation and Classification With Limited Training Data”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 7, pp. 763-767, July 1996.

[8] M. J. Lyons, J. Budynek, and S. Akamatsu, “Automatic Classification of Single Facial Images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1357-1362, December 1999.

[9] F. O’Sullivan, “A Statistical Perspective on Ill-Posed Inverse Problems”, Statistical Science, vol. 1, pp. 502-527, 1986.

[10] 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.

[11] S. Tadjudin and D.A. Landgrebe, “Covariance Estimation With Limited Training Samples”, IEEE Transactions on Geoscience and Remote Sensing, vol. 37, no. 4, July 1999.

[12] M. Turk and A. Pentland, “Eigenfaces for Recognition”, Journal of Cognitive Neuroscience, vol. 3, pp. 72-85, 1991.