

- from image sequences," in *Proc. IEEE Conf. Pattern Recognition, Image Processing*, 1981.
- [19] H. H. Nagel and B. Neumann, "On 3-D reconstruction from two perspective views," in *Proc. IJCAI*, 1981.
- [20] K. Prazdny, "Motion and structure from optical flow," in *Proc. 6th IJCAI*, 1979, pp. 702-704.
- [21] —, "Egomotion and relative depth map from optical flow," *Bio. Cybern.*, vol. 36, pp. 87-102, 1980.
- [22] —, "Determining the instantaneous direction of motion from optical flow generated by a curvilinearly moving observer," in *Proc. CGIP*, vol. 17, 1981, pp. 238-248.
- [23] —, "A simple method for recovering relative depth map in the case of a translating sensor," in *Proc. IJCAI*, 1981, pp. 698-699.
- [24] J. W. Roach and J. K. Aggarwal, "Determining the movement of objects from sequence of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 554-562, 1980.
- [25] R. Y. Tsai and T. S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces," in *Proc. PRIP*, 1982, pp. 112-118.
- [26] —, "Estimating three-dimensional motion parameters of a rigid planar patch," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1147-1152, Dec. 1981.
- [27] —, "Estimating three-dimensional motion parameters of a rigid planar patch," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 525-534, Aug. 1982.
- [28] —, "3-D time varying scene analysis," in *Proc. SPIE, Applications of Digital Image Processing IV*, vol. 359, San Diego, CA, Aug. 24-27, 1982.
- [29] S. Ullman, *The Interpretation of Visual Motion*. Cambridge, MA: MIT Press, 1979.
- [30] —, "The interpretation of structure from motion," in *Proc. Royal Soc. London*, vol. B, no. 203, pp. 405-426.

### A Posteriori Estimation of Correlated Jointly Gaussian Mean Vectors

MOSHÉ J. LASRY AND RICHARD M. STERN

**Abstract**—This paper describes the use of maximum *a posteriori* probability (MAP) techniques to estimate the mean values of features used in statistical pattern classification problems, when these mean feature values from the various decision classes are jointly Gaussian random vectors that are correlated across the decision classes. A set of mathematical formalisms is proposed and used to derive closed-form expressions for the estimates of the class-conditional mean vectors, and for the covariance matrix of the errors of these estimates. Finally, the performance of these algorithms is described for the simple case of a two-class one-feature pattern recognition problem, and compared to the performance of classical estimators that do not exploit the class-to-class correlations of the features' mean values.

**Index Terms**—Bayesian learning, dynamic speaker adaptation, maximum *a posteriori* probability, multivariate normal probability densities, speech recognition, statistical parameter estimation, statistical pattern recognition.

Manuscript received June 10, 1983; revised December 30, 1983. This work was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-81-K-1539.

The authors are with the Departments of Electrical Engineering and Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213.

## I. INTRODUCTION

Pattern classification systems frequently use parametric probability density functions (pdf's) to describe the objects to be classified. In many of these cases the exact statistical parameters specifying the data may vary for different data sets to be classified, and under these circumstances the classifier must estimate the statistical parameters of the probability densities as it performs the actual classifications.

Maximum *a posteriori* probability (MAP) estimation is a useful approach to the problem if the statistical fluctuations of the parameters to be estimated may be characterized as random variables with known *a priori* probability distributions. The MAP estimation procedure specifies in effect an optimal combination of this *a priori* knowledge of the parameters to be estimated with the *a posteriori* knowledge gained from observing labeled samples of the data to be classified.

In this paper we are concerned with classification problems in which the classifier uses Bayesian estimation methods to learn the mean vectors of normally distributed data to be classified, and in which the values of these mean vectors are correlated across the various decision classes. (We also assume that the probability distributions of the mean vectors across the various data sets to be classified are normally distributed.) In previous application of MAP estimation to learning the mean vectors of probability distributions from which samples were to be classified, the means for a given decision class had been assumed to be dependent only on observations from that decision class, and hence estimates of the mean vectors of the observations for that class would be updated only when a new sample from that particular class is presented to the classifier. In the present paper we propose an extension to the MAP procedure in which the mean vectors are assumed to be correlated across decision classes, and estimates of the means for *all* of the classes are updated after *any* new observation from *any* of the classes. This procedure, which we refer to as the extended MAP algorithm (EMAP), has been found useful in estimating statistical parameters for a feature-based speech recognition system [1], [5].

Our research on statistical estimation of correlated mean vectors was originally motivated by a desire to adapt our feature-based speech recognition system to the acoustical characteristics of individual speakers. In the simple application we considered, speaker adaptation consisted of adjusting the mean vectors of the various decision classes making use of the utterances input to the system by an individual speaker. We found empirically, for example, that if the system "knows" how a given speaker pronounces the letter M, it could infer some information about how the same speaker will pronounce the letter N, by making use of correlations of the mean values of the features for the letters M and N. Thus, the speed of adaptation of the classifier was increased by exploiting these correlations between the mean values of features describing the different classes of speech sounds to be recognized.

In the following sections we describe the mathematical formalisms used to derive a closed-form expression of the EMAP estimation algorithm in the case of multivariate Gaussian distributions. In Section II, we review the results obtained with classical application of the MAP estimation for updating the means of pdf's of normal distributions. In Section III, after defining a set of notations used to describe the possible interactions of the features within and across the various decision classes, we derive the EMAP estimation algorithm along with the estimated values of the class-conditional mean vectors and covariance matrices, and we describe how EMAP estimation has been successfully applied to perform speaker-adaptation on a feature-based speech recognition system. In Section IV, we examine the particular case of a two-class and one-feature problem to gain better insight about the performance and

the limitations of the EMAP estimation procedure. Finally, Section V summarizes the results obtained in this paper.

## II. REVIEW OF SELECTED CLASSICAL MAP RESULTS

The standard MAP estimation technique results from classical Bayesian hypothesis testing methods and has been extensively described in the pattern recognition literature [3]. We briefly summarize some relevant results for learning the means of normal densities.

For a given class  $\omega_k$  we assume a multivariate Gaussian conditional probability density function of feature values  $\mathbf{x}$

$$p(\mathbf{x}|\mu_k) \sim \mathcal{N}(\mu_k, \mathbf{S}_k) \quad (1)$$

where  $\mu_k$  is the mean vector to be estimated, and  $\mathbf{S}_k$  is the covariance matrix of the random vector  $\mathbf{x}$ .

In addition we assume the  $\mu_k$  is a random vector normally distributed around its mean value  $\mu_{0k}$ , with covariance matrix  $\mathbf{S}_{0k}$

$$p(\mu_k) \sim \mathcal{N}(\mu_{0k}, \mathbf{S}_{0k}).$$

We call  $\mathcal{X}_k$  the set of observations  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_k}\}$  obtained thus far from the Class  $\omega_k$ .

Under these assumptions, the MAP estimate of  $\mu_k$  is the value  $\hat{\mu}_k$  that maximizes the *a posteriori* pdf of  $\mu_k$  given the observations  $\mathbf{x}_1, \dots, \mathbf{x}_{n_k}$ :  $p(\mu_k|\mathbf{x}_1, \dots, \mathbf{x}_{n_k}, \omega_k)$ , and it is equal to [3]

$$\hat{\mu}_k = \mathbf{S}_{0k} \left( \mathbf{S}_{0k} + \frac{\mathbf{S}_k}{n} \right)^{-1} \mathbf{A}_k + \frac{\mathbf{S}_k}{n} \left( \mathbf{S}_{0k} + \frac{\mathbf{S}_k}{n} \right)^{-1} \mu_{0k}$$

where

$$\mathbf{A}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \mathbf{x}_j$$

is the sample mean of the observation vectors from Class  $\omega_k$ .

The MAP estimate  $\hat{\mu}_k$  is a linear combination of the *a priori* mean  $\mu_{0k}$  and the *a posteriori* sample mean  $\mathbf{A}_k$ . For  $n_k = 0$ ,  $\hat{\mu}_k$  is equal to  $\mu_{0k}$ . As  $n_k$  becomes large, the contribution of the *a priori* mean  $\mu_{0k}$  becomes small and  $\hat{\mu}_k$  asymptotes to the sample mean  $\mathbf{A}_k$ .

It can also be shown that given the observations  $\{\mathbf{x}_1, \dots, \mathbf{x}_{n_k}\}$ ,  $\mu_k$  is normally distributed around  $\hat{\mu}_k$  with a covariance matrix  $\mathbf{R}_k = \mathbf{S}_{0k} (\mathbf{S}_k + n_k \mathbf{S}_{0k})^{-1} \mathbf{S}_k$ , so that

$$p(\mu_k|\mathcal{X}_k) \sim \mathcal{N}(\hat{\mu}_k, \mathbf{R}_k). \quad (2)$$

The matrix  $\mathbf{R}_k$  is equal to the expected value of  $(\mu_k - \hat{\mu}_k)(\mu_k - \hat{\mu}_k)^t$  and can be viewed as the correlation matrix of the errors [6]. Thus, the diagonal terms of  $\mathbf{R}_k$  represent the mean-square error of the feature mean values for Class  $\omega_k$ .

Given (1) and (2), it can be proved [3] that

$$p(\mathbf{x}|\mathcal{X}_k) \sim \mathcal{N}(\hat{\mu}_k \mathbf{S}_k + \mathbf{R}_k) \quad (3)$$

which is the updated version of the class-conditional probability density function.<sup>1</sup>

We note that the results derived in this section are based on the assumption that the class-conditional covariance matrices are independent of the class-conditional mean vectors. An optimal procedure has been derived for the cases where this assumption is not valid and is reported in [4]. In this procedure, the covariance matrices as well as the means of normal densities are estimated on the basis of observed labeled samples, making use of the properties of Wishart distributions [2].

<sup>1</sup>This result can be intuitively understood if we view an observation vector  $\mathbf{x}$  as the sum of two random vectors, the mean vector  $\mu_k$  distributed around its estimate  $\hat{\mu}_k$  with a covariance matrix  $\mathbf{R}_k$  and an independent random vector  $\mathbf{y}$  with  $p(\mathbf{y} \sim \mathcal{N}(0, \mathbf{S}_k))$ . We can then use the fact that the sum of two independent normally distributed random vectors is a random vector whose mean is the sum of the means and whose covariance matrix is the sum of the covariance matrices.

$$\begin{aligned} \mu_1 &= \begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix} \\ \mu_2 &= \begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix} \end{aligned} \quad \mu = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix}$$

Fig. 1. Composite structure of the generalized mean vector in the case of a two-class problem with two features.  $\mu_1$  is the mean vector of the features for Class 1,  $\mu_2$  is the mean vector of the features for Class 2, and  $\mu$  is the generalized mean vector.

## III. THE EXTENDED MAP ESTIMATION PROCEDURE

In the EMAP estimation procedure, the statistical values of the features for all the decision classes are updated simultaneously rather than on a class-by-class basis, making use of the correlations of feature mean values across the different classes. We begin this section by defining a set of notational conventions in order to capture and represent the dependencies between statistical feature values and decision classes. Next, we obtain closed form expressions for the extended MAP estimate of the mean vectors of the features for all decision classes, and the mean-square error of these estimates. Finally, we briefly describe how this method has been applied to the problem of speaker-adaptation in a speech recognition system for classification of isolated English letters.

### A. Notational Conventions

We consider a pattern recognition problem with  $C$  decision classes and  $D$  features. We assume that a set  $\mathcal{X}_k$  of  $n_k$  observations  $\mathcal{X}_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \dots, \mathbf{x}_{k,n_k}\}$  belonging to Class  $\omega_k$  has been input to the classifier, where  $\mathbf{x}_{k,j}$  is the  $j$ th observation obtained from Class  $\omega_k$ . Finally, we assume that for a given class  $\omega_k$ , the random vector,  $\mathbf{x}_{k,j}$  composed of  $D$  feature values is normally distributed around a mean vector  $\mu_k$  with a covariance matrix  $\mathbf{S}_k$

$$p(\mathbf{x}_{k,j}|\mu_k) \sim \mathcal{N}(\mu_k, \mathbf{S}_k).$$

With these notational conventions, the MAP estimates of the mean vectors are obtained by maximizing the overall probability

$$p(\mu_1, \mu_2, \dots, \mu_C | \mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_C).$$

When the mean vectors of the different classes are independent random vectors, the above expression is equal to

$$p(\mu_1, \mu_2, \dots, \mu_C | \mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_C) = \prod_{k=1}^C p(\mu_k | \mathcal{X}_k)$$

and the estimates of the mean vectors are obtained as described in the previous section. In the following we assume that the mean vector  $\mu_k$  of Class  $\omega_k$  is a Gaussian random vector with a known probability density function  $\mathcal{N}(\mu_{0k}, \mathbf{S}_{0k})$ , and that all the  $\mu_k$ 's are jointly Gaussian random vectors with cross-covariance matrix equal to  $\mathbf{B}_{k,r} = E(\mu_k - \mu_{0k})(\mu_r - \mu_{0r})^t$ . These assumptions can be expressed in a simple way by defining a set of *generalized notations*, as follows.<sup>2</sup>

We define  $\mu = (\mu_1, \mu_2, \dots, \mu_C)^t$  as the generalized mean vector (Fig. 1). Given that the  $\mu_k$ 's are jointly Gaussian random vectors of size  $D$ , we can view  $\mu$  as a random vector of size  $CD$ , normally distributed around its mean value  $\mu_0 = (\mu_{01}, \mu_{02}, \dots, \mu_{0C})^t$ , with a  $CD \times CD$  covariance matrix  $\mathbf{S}_0$ , which we refer to as the *meancrosscovariance* matrix.

$\mathbf{S}_0$ , which is equal to the expected value of  $(\mu - \mu_0)(\mu - \mu_0)^t$ , can be intuitively viewed as a composite of  $C^2$  blocks of size  $D \times D$ . The  $(k, r)$ th block of  $\mathbf{S}_0$ , equal to the expected value

<sup>2</sup>The term *generalized* is used to refer to mathematical objects such as vectors or matrices that contain information about all the decision classes.

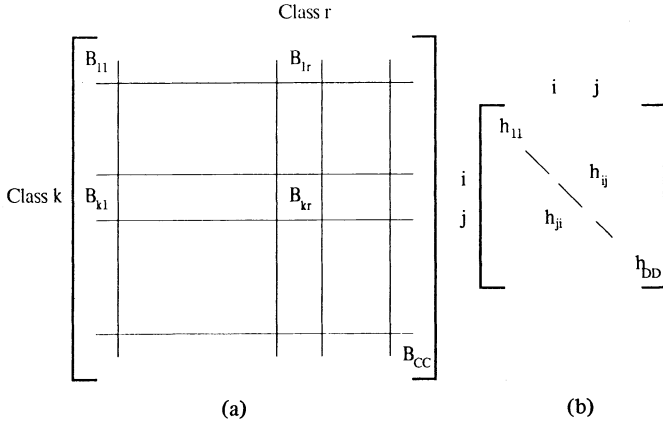


Fig. 2. Composite structure of the meancrosscovariance matrix. (a) The  $CD \times CD$  meancrosscovariance matrix is composed of  $C^2$  blocks of size  $D \times D$ . Block  $B_{k,r}$  is a  $D \times D$  matrix which represents the covariance between classes  $k$  and  $r$ . (b) the element  $h_{i,j}$  in  $B_{k,r}$  is the covariance between the mean of feature  $i$  in Class  $k$  and the mean of feature  $j$  in Class  $r$ .

of  $(\mu_k - \mu_{0k})(\mu_r - \mu_{0r})^t$  is the cross-covariance matrix of  $\mu_k$  and  $\mu_r$  and has been previously defined as  $B_{k,r}$ . As shown in Fig. 2, the  $(d_1, d_2)$ th element of the  $(k, r)$ th block represents the "influence" of the mean value of Feature  $d_1$  for Class  $k$  on the mean value of Feature  $d_2$  for Class  $r$ .

In the EMAP estimation procedure, we wish to evaluate and maximize

$$\begin{aligned} p(\mu_1, \mu_2, \dots, \mu_C | \mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_C) \\ = p(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_C | \mu_1, \mu_2, \dots, \mu_C) \\ \cdot \frac{p(\mu_1, \mu_2, \dots, \mu_C)}{p(\mathcal{X}_1, \dots, \mathcal{X}_C)} \end{aligned}$$

Using the notational conventions defined above, we can write

$$\begin{aligned} p(\mu_1, \mu_2, \dots, \mu_C) = p(\mu) = (2\pi)^{-CD/2} \det(\mathcal{S}_0)^{-1/2} \\ \cdot \exp\left(-\frac{1}{2}(\mu - \mu_0)^t \mathcal{S}_0^{-1} (\mu - \mu_0)\right). \end{aligned}$$

Since we assume the observations to be statistically independent, and since the pdf of the observations from a given class is independent of the means of other classes, we obtain

$$p(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_C | \mu_1, \mu_2, \dots, \mu_C) = \prod_{k=1}^C p(\mathcal{X}_k | \mu_k)$$

and

$$\begin{aligned} p(\mathcal{X}_k | \mu_k) = \prod_{j=1}^{n_k} (2\pi)^{-D/2} \det(\mathcal{S}_k)^{-1/2} \\ \exp\left(-\frac{1}{2}(\mathbf{x}_{k,j} - \mu_k)^t \mathcal{S}_k^{-1} (\mathbf{x}_{k,j} - \mu_k)\right). \end{aligned}$$

Therefore, the joint *a posteriori* pdf of the mean vectors can be expressed as

$$\begin{aligned} p(\mu | \mathcal{X}_1, \dots, \mathcal{X}_C) = \alpha \exp\left(-\frac{1}{2}(\mu - \mu_0)^t \mathcal{S}_0^{-1} (\mu - \mu_0)\right) \\ \prod_{k=1}^C \prod_{j=1}^{n_k} \exp\left(-\frac{1}{2}(\mathbf{x}_{k,j} - \mu_k)^t \mathcal{S}_k^{-1} (\mathbf{x}_{k,j} - \mu_k)\right) \end{aligned}$$

where  $\alpha$  is a constant that depends on the determinants of the various matrices and on the observations but is independent of the mean vectors, and is the previously defined generalized mean vector.

The above expression may be rewritten as follows:

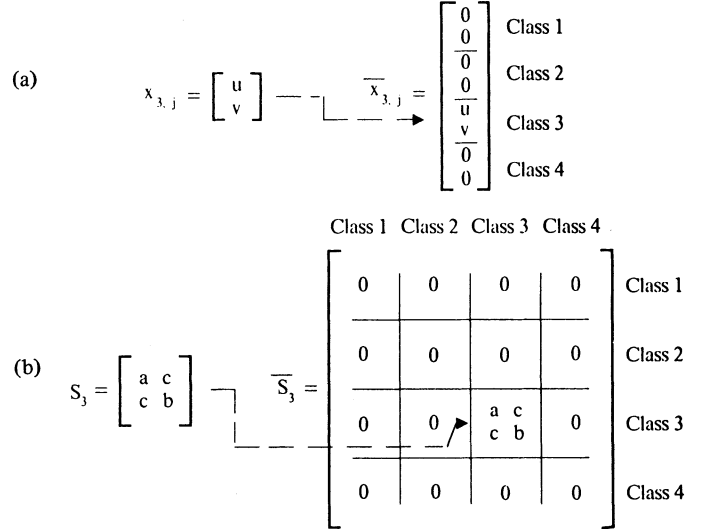


Fig. 3. The  $CD$ -dimensional extension of vectors and matrices. (a)  $\bar{\mathbf{x}}_{3j}$  is the  $CD$ -dimensional extension of the  $j$ th observation vector from Class 3, which is referred to as  $\mathbf{x}_{3j}$  in the text. (b)  $\bar{\mathcal{S}}_3$  is the  $CD \times CD$  dimensional extension of  $\mathcal{S}_3$ , the covariance matrix of the data from Class 3.

$$\begin{aligned} p(\mu | \mathcal{X}_1, \dots, \mathcal{X}_C) = \alpha \exp\left(-\frac{1}{2}(\mu - \mu_0)^t \mathcal{S}_0^{-1} (\mu - \mu_0)\right) \\ - \frac{1}{2} \sum_{k=1}^C \sum_{j=1}^{n_k} (\mathbf{x}_{k,j} - \mu_k)^t \mathcal{S}_k^{-1} (\mathbf{x}_{k,j} - \mu_k). \end{aligned} \quad (4)$$

The vectors and matrices inside the exponential are of different size.  $\mathcal{S}_0^{-1}$  is a  $CD \times CD$  matrix while each of the matrices of type  $\mathcal{S}_k^{-1}$  is only of size  $D \times D$ . Similarly,  $\mu$  and  $\mu_0$  are of length  $CD$  while the  $\mathbf{x}_{k,j}$ 's and the  $\mu_k$ 's are of length  $D$ . In order to facilitate subsequent computations, we avoid this nonhomogeneity of dimensionalities by padding the  $\mathbf{x}_k$ 's,  $\mu_k$ 's, and  $\mathcal{S}_k$ 's with appropriate vectors and matrices containing zeros. Specifically, we define

- $\bar{\mathbf{x}}_{k,j}$  as the  $CD$ -dimensional extension of  $\mathbf{x}_{k,j}$
- $\bar{\mu}_k$  as the  $CD$ -dimensional extension of  $\mu_k$
- $\bar{\mathcal{S}}_k$  as the  $CD \times CD$  extension of  $\mathcal{S}_k$
- $\bar{\mathcal{Q}}_k$  as the  $CD \times CD$  extension of  $\mathcal{Q}_k$ , where  $\mathcal{Q}_k = \mathcal{S}_k^{-1}$  is defined as the inverse covariance matrix of Class  $\omega_k$ .

If  $(\bar{\mathbf{x}}_{k,j})_i$  is the  $i$ th component of  $\bar{\mathbf{x}}_{k,j}$ , we define

$$(\bar{\mathbf{x}}_{k,j})_i = \begin{cases} (\mathbf{x}_{k,j})_{i-(k-1)D} & \text{if } (k-1)D < i \leq kD \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, if  $(\bar{\mathcal{S}}_k)_{ij}$  is the  $(ij)$ th element of  $\bar{\mathcal{S}}_k$  we define

$$(\bar{\mathcal{S}}_k)_{ij} = \begin{cases} (\mathcal{S}_k)_{i-(k-1)D, j-(k-1)D} & \text{if } (k-1)D < ij \leq kD \\ 0 & \text{otherwise.} \end{cases}$$

These notational conventions are illustrated in Fig. 3. The extended vector  $\bar{\mu}_k$  is obtained from  $\mu_k$  in the same fashion as  $\bar{\mathbf{x}}_{k,j}$  is obtained from  $\mathbf{x}_{k,j}$  and  $\bar{\mathcal{Q}}_k$  is obtained from  $\mathcal{Q}_k$  in the same fashion as  $\bar{\mathcal{S}}_k$  is obtained from  $\mathcal{S}_k$ . It is important to note that  $\bar{\mathcal{S}}_k$  is not invertible since it is block-diagonal and some of its diagonal blocks are equal to the null matrix. Therefore  $\bar{\mathcal{Q}}_k$  does not refer to the inverse of  $\bar{\mathcal{S}}_k$  but rather to the  $CD$ -dimensional extension of  $\mathcal{Q}_k = \mathcal{S}_k^{-1}$ .

In addition, we define  $\bar{\mathcal{S}} = \sum_{k=1}^C \bar{\mathcal{S}}_k$ , and  $\bar{\mathcal{N}}$  as a  $CD \times CD$  diagonal matrix where the diagonal terms in the  $k$ th  $D \times D$  diagonal block are equal to  $n_k$  (the number of observations from

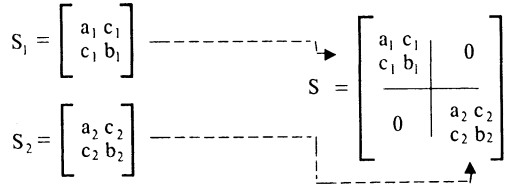


Fig. 4. Composite structure of the generalized covariance matrix in the case of a two-class problem with two features.  $S_1$  and  $S_2$  are the covariance matrices of the data from Classes  $\omega_1$  and  $\omega_2$ , and  $S$  is the generalized covariance matrix of the data.

Class  $\omega_k$ ). We call  $S$  the generalized covariance matrix (Fig. 4), and  $N$  the generalized number of observations matrix.

Finally, we define  $A_k = (1/n_k) \sum_{j=1}^{n_k} \mathbf{x}_{k,j}$  as the sample mean (average) or the observation from Class  $\omega_k$  and  $\bar{A}_k$  as its  $CD$ -dimensional extension.

### B. Derivation of the Extended MAP Estimate

Using the above notational conventions, we can now write

$$(\mathbf{x}_{k,j} - \boldsymbol{\mu}_k)^t \mathbf{S}_k^{-1} (\mathbf{x}_{k,j} - \boldsymbol{\mu}_k) = (\bar{\mathbf{x}}_{k,j} - \boldsymbol{\mu}_k)^t \bar{\mathbf{Q}}_k (\bar{\mathbf{x}}_{k,j} - \bar{\boldsymbol{\mu}}_k).$$

Summing over  $j$ , and collecting like terms, we obtain

$$\begin{aligned} & \sum_{j=1}^{n_k} (\mathbf{x}_{k,j} - \boldsymbol{\mu}_k)^t \mathbf{S}_k^{-1} (\mathbf{x}_{k,j} - \boldsymbol{\mu}_k) \\ &= \left( \sum_{j=1}^{n_k} \bar{\mathbf{x}}_{k,j} \bar{\mathbf{Q}}_k \bar{\mathbf{x}}_{k,j} \right) - 2n_k \bar{\boldsymbol{\mu}}_k^t \bar{\mathbf{Q}}_k \bar{A}_k + n_k \bar{\boldsymbol{\mu}}_k^t \bar{\mathbf{Q}}_k \bar{\boldsymbol{\mu}}_k. \end{aligned} \quad (5)$$

For a given  $k$ ,  $\bar{\mathbf{Q}}_k$  is a  $CD \times CD$  matrix that contains only zeros except in the  $k$ th diagonal block, in which it is equal to the inverse covariance matrix of Class  $\omega_k$ . Therefore,

$$\sum_{k=1}^C n_k \bar{\boldsymbol{\mu}}_k^t \bar{\mathbf{Q}}_k \bar{A}_k = \boldsymbol{\mu}^t \mathbf{N} \mathbf{S}^{-1} \mathbf{A}$$

and

$$\sum_{k=1}^C n_k \bar{\boldsymbol{\mu}}_k^t \bar{\mathbf{Q}}_k \bar{A}_k = \boldsymbol{\mu}^t \mathbf{N} \mathbf{S}^{-1}$$

where  $\mathbf{A} = \sum_{k=1}^C \bar{A}_k$ . ( $\mathbf{A}$  is the generalized average observation vector.) Equation (5) can now be written as

$$\begin{aligned} & \sum_{k=1}^C \sum_{j=1}^{n_k} (\mathbf{x}_{k,j} - \boldsymbol{\mu}_k)^t \mathbf{S}_k^{-1} (\mathbf{x}_{k,j} - \boldsymbol{\mu}_k) \\ &= \sum_{k=1}^C \sum_{j=1}^{n_k} \bar{\mathbf{x}}_{k,j} \bar{\mathbf{Q}}_k \bar{\mathbf{x}}_{k,j} - 2\boldsymbol{\mu}^t \mathbf{N} \mathbf{S}^{-1} \mathbf{A} + \boldsymbol{\mu}^t \mathbf{N} \mathbf{S}^{-1} \boldsymbol{\mu} \end{aligned}$$

and (4) becomes

$$\begin{aligned} p(\boldsymbol{\mu} | \mathcal{X}_1, \dots, \mathcal{X}_C) &= \beta \exp \left( -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^t \mathbf{S}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right. \\ & \quad \left. + \boldsymbol{\mu}^t (\mathbf{N} \mathbf{S}^{-1}) \mathbf{A} - \frac{1}{2} \boldsymbol{\mu}^t (\mathbf{N} \mathbf{S}^{-1}) \boldsymbol{\mu} \right). \end{aligned} \quad (6)$$

The parameter  $\beta$  depends on the  $\alpha$  from (4) and on the observations, but it is independent of  $\boldsymbol{\mu}$ . If we now keep only the terms that depend on  $\boldsymbol{\mu}$  under the exponential and incorporate the other terms in a new parameter  $\gamma$ , (6) can be written as

$$\begin{aligned} p(\boldsymbol{\mu} | \mathcal{X}_1, \dots, \mathcal{X}_C) &= \gamma \exp \left( -\frac{1}{2} (\boldsymbol{\mu}^t (\mathbf{S}_0^{-1} + \mathbf{N} \mathbf{S}^{-1}) \boldsymbol{\mu} \right. \\ & \quad \left. - 2\boldsymbol{\mu}^t (\mathbf{S}_0^{-1} \boldsymbol{\mu}_0 + \mathbf{N} \mathbf{S}^{-1} \mathbf{A})) \right). \end{aligned} \quad (7)$$

This pdf can be rewritten as

$$p(\boldsymbol{\mu} | \mathcal{X}_1, \dots, \mathcal{X}_C) = \delta \exp \left( -\frac{1}{2} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^t \mathbf{H} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \right)$$

where the values of  $\mathbf{H}$  and  $\hat{\boldsymbol{\mu}}$  are obtained by equating the corresponding coefficients:

$$\mathbf{H} = \mathbf{S}_0^{-1} + \mathbf{N} \mathbf{S}^{-1} \quad (8)$$

$$\mathbf{H} \hat{\boldsymbol{\mu}} = \mathbf{S}_0^{-1} \boldsymbol{\mu}_0 + \mathbf{N} \mathbf{S}^{-1} \mathbf{A}. \quad (9)$$

Combining (8) and (9), and using basic properties of matrix inversion and multiplication, we finally obtain

$$\hat{\boldsymbol{\mu}} = \mathbf{S} (\mathbf{S} + \mathbf{S}_0 \mathbf{N})^{-1} \boldsymbol{\mu}_0 + \mathbf{S}_0 (\mathbf{S} + \mathbf{N} \mathbf{S}_0)^{-1} \mathbf{N} \mathbf{A}. \quad (10)$$

Furthermore, the *a posteriori* probability density function of the generalized mean vector is equal to

$$p(\boldsymbol{\mu} | \mathcal{X}_1, \dots, \mathcal{X}_C) = \delta \exp \left( -\frac{1}{2} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^t \mathbf{K}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \right) \quad (11)$$

where

$$\mathbf{K} = \mathbf{H}^{-1} = (\mathbf{S}_0^{-1} + \mathbf{N} \mathbf{S}^{-1})^{-1} = \mathbf{S}_0 (\mathbf{S} + \mathbf{N} \mathbf{S}_0)^{-1} \mathbf{S} \quad (12)$$

and  $\hat{\boldsymbol{\mu}}$  is given by (10).

Since  $\hat{\boldsymbol{\mu}}$  is the value of  $\boldsymbol{\mu}$  that maximizes the *a posteriori* density  $p(\boldsymbol{\mu} | \mathcal{X}_1, \dots, \mathcal{X}_C)$ , it is the MAP estimate of the generalized mean vector  $\boldsymbol{\mu}$ .

The expressions obtained for  $\hat{\boldsymbol{\mu}}$  and  $\mathbf{K}$  make intuitive sense:  $\hat{\boldsymbol{\mu}}$  is a combination of the *a priori* generalized mean vector and the average observation vector. When the number of observations becomes large, the norm of the matrix  $\mathbf{N}$  becomes large too, and the contribution of the observations predominates the contribution of the *a priori* mean. When there are no observations, the matrix  $\mathbf{N}$  contains only zeros and (10) yields  $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_0$ . Similarly, the covariance matrix  $\mathbf{K}$ , given by (12), is equal to the meancrosscovariance matrix  $\mathbf{S}_0$  when no observations have been obtained (i.e.,  $\mathbf{N} = 0$ ) and its norm decreases when the number of observations increases, so that the distribution of  $\boldsymbol{\mu}$  becomes "tighter" around its estimate  $\hat{\boldsymbol{\mu}}$  as more observations are obtained.

### C. Performance of the Extended MAP Estimate

From (11), we learn that the *a posteriori* pdf of the random vector  $\boldsymbol{\mu}$  given the sets of observations  $(\mathcal{X}_1, \dots, \mathcal{X}_C)$  is Gaussian with mean  $\hat{\boldsymbol{\mu}}$  and covariance matrix  $\mathbf{K} = \mathbf{S}_0 (\mathbf{S} + \mathbf{N} \mathbf{S}_0)^{-1} \mathbf{S}$ . Because the mode of a Gaussian pdf is also its mean,  $\hat{\boldsymbol{\mu}}$  can also be interpreted as the minimum mean-square estimate (MMSE) of  $\boldsymbol{\mu}$  and the correlation matrix of the errors for that estimate is equal to

$$E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^t = \mathbf{R}.$$

If we consider the vector  $\hat{\boldsymbol{\mu}}$  as a composite of  $C$  vectors  $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_C$  of size  $D$ , and the matrix  $\mathbf{K}$  as a composite of  $C^2$  blocks of size  $D \times D$ , then if  $\mathbf{R}_k$  is the  $(k, k)$ th block of  $\mathbf{K}$

$$\mathbf{R}_k = E(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^t$$

and

$$p(\boldsymbol{\mu}_k | \mathcal{X}_1, \dots, \mathcal{X}_C) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_k, \mathbf{R}_k).$$

$\mathbf{R}_k$  can be viewed as the correlation matrix of the errors of the random vector  $\boldsymbol{\mu}_k$ , or as the conditional variance of  $\boldsymbol{\mu}_k$  given the sets of observations  $\mathcal{X}_1, \dots, \mathcal{X}_C$ . The  $j$ th diagonal term of  $\mathbf{R}_k$  is the mean-square error of the class-conditional mean of Feature  $j$  in Class  $\omega_k$ . Hence the performance of the extended MAP estimate of the mean vectors, including the advantage obtained by considering the samples from other correlated decision classes can easily be obtained by examining the diagonal elements of the matrix  $\mathbf{K}$ .

#### D. Application to Pattern Classification Problems

The MAP estimation procedure can be successfully applied to a number of pattern classification problems for which the mean vectors characterizing the various decision classes are random and jointly Gaussian. In feature-based speech recognition systems, for example, the mean values of the features used for classification of unknown sounds may vary greatly from one speaker to another [5]. If the pdf's of the class-conditional mean vectors over the entire population of speakers are known in such cases, we can use the EMAP estimation procedure to adjust them to the characteristics of individual speakers. After the recognition system classifies each incoming utterance it receives feedback from the user indicating which class has been input. The system uses this labeled observation as well as the previous ones to update the mean values of all features for all decision classes. The modified values of the pdf's of all decision classes are then used in performing subsequent classifications.

Specifically we assume that  $p(\mathbf{x}_k | \boldsymbol{\mu}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{S}_k)$ , and that  $p(\boldsymbol{\mu}_k | \mathcal{X}_1, \dots, \mathcal{X}_C) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_k, \mathbf{R}_k)$  as discussed in Section II.

We then obtain, as in (3),

$$p(\mathbf{x}_k | \mathcal{X}_1, \dots, \mathcal{X}_C) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_k, \mathbf{S}_k + \mathbf{R}_k)$$

where  $\mathbf{x}_k$  is a random observation from Class  $\omega_k$ . The matrix  $\mathbf{K}$  previously defined and its diagonal blocks, the  $\mathbf{R}_k$ 's, have to be recomputed after each new observation since the obtention of more accurate estimates of the class-conditional mean vectors reduces the contribution of imperfect knowledge of those mean vectors to the covariance of the observations used in performing the classifications. In particular, when the number of observations is very large, the matrices  $\mathbf{R}_k$  approach the null matrix so that  $\mathbf{S}_k + \mathbf{R}_k \approx \mathbf{S}_k$ .

Speech recognition is a good application for EMAP estimation since there exist high correlations between the pronunciation of different sounds by a given speaker. A preliminary application of the EMAP estimation procedure to the isolated-letter recognition system FEATURE [5] produced a decrease in the classification error rate of the English alphabet from 12.5 percent to approximately 6 percent after four presentations of each letter from a new speaker, and decreases in error rates from 14 percent to less than 5 percent for the confusable sets of letters {M, N} and {B, D, P, T}.

#### IV. PERFORMANCE OF THE EXTENDED MAP ESTIMATE FOR A SIMPLE TWO-CLASS ONE-FEATURE PROBLEM

In this section, we consider a simple problem with two decision classes  $\omega_1$  and  $\omega_2$  and only one feature, and we derive the closed form expression of the pdf's of the means of that feature for  $\omega_1$  and  $\omega_2$ ,  $\mu_1$  and  $\mu_2$ , respectively. The optimal estimate for this simple case can be described analytically in an intuitively pleasing fashion, and it provides additional insight into the conditions for which the additional information provided by observations from other decision classes is most useful.

We assume that  $\mu_1$  and  $\mu_2$  are correlated jointly Gaussian random variables with means  $\mu_{01}$  and  $\mu_{02}$ , variances  $\sigma_{01}^2$  and  $\sigma_{02}^2$  and cross-covariance  $\rho\sigma_{01}\sigma_{02}$ , where  $\rho$  is the correlation coefficient. We also assume that the conditional pdf's of the observations from Class  $\omega_1$  and Class  $\omega_2$  given  $\mu_1$  and  $\mu_2$  are Gaussian with variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively. Thus, if  $x$  and  $y$  are random observations from  $\omega_1$  and  $\omega_2$ , respectively, we have

$$p(x|\mu_1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad p(y|\mu_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$p(\mu_1) \sim \mathcal{N}(\mu_{01}, \sigma_{01}^2) \quad p(\mu_2) \sim \mathcal{N}(\mu_{02}, \sigma_{02}^2)$$

$$E(\mu_1 - \mu_{01})(\mu_2 - \mu_{02}) = \rho\sigma_{01}\sigma_{02}.$$

Finally, we assume that  $n_1$  observations  $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$  have been obtained from Class  $\omega_1$  and that  $n_2$  observations  $\mathbf{y}_1, \dots, \mathbf{y}_{n_2}$

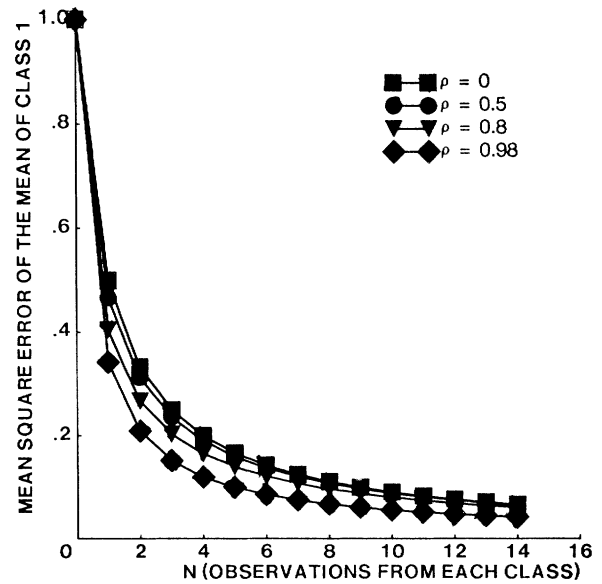


Fig. 5. Evolution of the mean-squared error of an estimated mean in the case of a two-class problem with one feature, as a function of the number of observations  $n$  and the *a priori* class-to-class correlation of the feature means,  $\rho$ . In this example it is assumed that  $n_1 = n_2 = n$  and that  $\sigma_1 = \sigma_2 = \sigma_{01} = \sigma_{02} = 1$ .

have been obtained from Class  $\omega_2$ . If we replace  $\mathbf{S}, \mathbf{S}_0, \mathbf{N}$ , and  $\mathbf{A}$  in (10) by their actual values, we obtain the estimates  $\hat{\mu}_1$  and  $\hat{\mu}_2$  of  $\mu_1$  and  $\mu_2$ :

$$\hat{\mu}_1 = \frac{n_1\sigma_{01}^2(\sigma_2^2 + n_2\sigma_{02}^2(1 - \rho^2))A_1}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2} + \frac{\sigma_1^2(\sigma_2^2 + n_2\sigma_{02}^2)\mu_{01}}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2} + \frac{n_2\rho\sigma_1^2\sigma_{01}\sigma_{02}(A_2 - \mu_{02})}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2}. \quad (13)$$

$$\hat{\mu}_2 = \frac{n_2\sigma_{02}^2(\sigma_1^2 + n_1\sigma_{01}^2(1 - \rho^2))A_2}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2} + \frac{\sigma_2^2(\sigma_1^2 + n_1\sigma_{01}^2)\mu_{02}}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2} + \frac{n_1\rho\sigma_2^2\sigma_{01}\sigma_{02}(A_1 - \mu_{01})}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2}. \quad (14)$$

The three terms of the above equations for  $\hat{\mu}_k$  ( $k = 1, 2$ ) represent the contributions of the observations from Class  $\omega_k$ , the *a priori* mean of Class  $\omega_k$ , and the observations from the other class, respectively. From the covariance matrix  $\mathbf{K}$  of the generalized mean vector, we can obtain the mean-square errors  $r_1^2$  and  $r_2^2$  of  $\hat{\mu}_1$  and  $\hat{\mu}_2$ , where

$$r_1^2 = E(\hat{\mu}_1 - \mu_1)^2 = \frac{\sigma_1^2\sigma_{01}^2(\sigma_2^2 + n_2\sigma_{02}^2(1 - \rho^2))}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2} \quad (15)$$

$$r_2^2 = E(\hat{\mu}_2 - \mu_2)^2 = \frac{\sigma_2^2\sigma_{02}^2(\sigma_1^2 + n_1\sigma_{01}^2(1 - \rho^2))}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1n_2\rho^2\sigma_{01}^2\sigma_{02}^2}. \quad (16)$$

We restrict our attention to the parameters of Class  $\omega_1$ , i.e.,  $\hat{\mu}_1$  and  $r_1^2$ , but the same type of analysis may be performed for the parameters of Class  $\omega_2$ .

### A. Analysis of the Expression of $\hat{\mu}_1$

As shown above,  $\hat{\mu}_1$  is the sum of three terms and is of the form

$$\hat{\mu}_1 = \alpha A_1 + (1 - \alpha)\mu_{01} + \beta(A_2 - \mu_{02})$$

where

$$\alpha = \frac{n_1 \sigma_{01}^2 (\sigma_2^2 + n_2 \sigma_{02}^2 (1 - \rho^2))}{(\sigma_1^2 + n_1 \sigma_{01}^2)(\sigma_2^2 + n_2 \sigma_{02}^2) - n_1 n_2 \rho^2 \sigma_{01}^2 \sigma_{02}^2}$$

$$\beta = \frac{n_2 \rho \sigma_1^2 \sigma_{01} \sigma_{02}}{(\sigma_1^2 + n_1 \sigma_{01}^2)(\sigma_2^2 + n_2 \sigma_{02}^2) - n_1 n_2 \rho^2 \sigma_{01}^2 \sigma_{02}^2}$$

For  $n_1 = 0$ ,  $\alpha = 0$  and the first term of  $\hat{\mu}_1$  (contribution of the observations from  $\omega_1$ ) is equal to 0 as could be expected. When  $n_1$  becomes very large,  $\alpha$  tends to 1 and  $\beta$  tends to 0, so that the contribution of the second and third terms of  $\hat{\mu}_1$  becomes negligible. Finally, if  $\rho = 0$ ,  $\beta$  is equal to 0 and there is no contribution of the observations from Class  $\omega_2$ ; this could also be predicted since in that case the random variables  $\mu_1$  and  $\mu_2$  are independent.  $\beta$  increases with  $|\rho|$ , so that the contribution of the observations from Class  $\omega_2$  increases with the amount of correlation of  $\mu_1$  and  $\mu_2$ .

### B. Analysis of the Mean-Squared Error

A simple way to evaluate the performance of the EMAP estimation procedure is by analyzing the expression of the mean-square error between the random variable  $\mu_1$  and its estimate  $\hat{\mu}_1$ .

If no observations have been obtained from Class  $\omega_1$ , (i.e.,  $n_1 = 0$ ), we note from (15) that the mean-squared error of  $\hat{\mu}_1$  is equal to

$$r_1^2 = E(\hat{\mu}_1 - \mu_1)^2 = \frac{\sigma_{01}^2 (\sigma_2^2 + n_2 \sigma_{02}^2 (1 - \rho^2))}{\sigma_2^2 + n_2 \sigma_{02}^2}. \quad (17)$$

From (17), we see that  $r_1^2$  is always smaller than  $\sigma_{01}^2$ , the *a priori* variance of  $\omega_1$  and is a decreasing function  $n_2$ . As  $n_2$  becomes large,  $r_1^2$  approaches its lower bound of  $\sigma_{01}^2 (1 - \rho^2)$ . As could be expected, this lower bound decreases when the cross-covariance between  $\mu_1$  and  $\mu_2$  increases.

Finally, let us consider the case where  $n_1 = n_2 = n$ ,  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , and  $\sigma_{01}^2 = \sigma_{02}^2 = \sigma_0^2$ . In that case

$$r_1^2 = \sigma_0^2 \frac{\sigma^2 (\sigma^2 + n \sigma_0^2 (1 - \rho^2))}{(\sigma^2 + n \sigma_0^2)^2 - n^2 \rho^2 \sigma_0^4}. \quad (18)$$

Fig. 5 shows the evolution of  $r_1^2$  as a function of  $n$  for different values of  $\rho$  when  $\sigma^2 = \sigma_0^2 = 1$ .

We can see that the expected mean-square error asymptotes to 0 in all curves as  $n$  becomes large. We can also see that the most significant improvement of the mean square error for small  $n$  is obtained in the case of highly correlated mean values. We note from (16) that for a given  $n$ ,  $r_1^2$  is an increasing function of  $\sigma^2/\sigma_0^2$ . Therefore, the most advantageous conditions for the use of the extended MAP estimate occur when 1) the correlation of the mean values across classes is high, 2) the ratio of the variance of the data to the variance of the mean,  $\sigma/\sigma_0$  is low, and 3) the number of observations is small.

## V. SUMMARY AND CONCLUSIONS

In this paper we reviewed the classical MAP estimation procedure for updating the probability density functions of Gaussian random mean vectors from a set of labeled observations. We extended the procedure so that it could take into consideration not only the feature-to-feature correlations within a decision class but also the correlations of the features' means from

one class to another. We formally evaluated this procedure for a simple two-class and one-feature case. We showed that the mean-squared error of estimates of the mean vectors is always smaller when the class-to-class correlations are taken into account, and that the greatest improvement afforded by the exploitation of class-to-class correlations is obtained when the number of observed samples is small, the class-to-class correlations of the means are high, and the ratio of the variance of the data to the variance of the mean is large.

The formulation of the estimation procedure was simplified by the use of a set of notational conventions that capture the covariances of the feature mean values within a given class as well as the cross-covariances between the mean vectors of different classes.

### ACKNOWLEDGMENT

The authors wish to thank Dr. R. A. Cole for the initial motivation for this research, as well as general help and advice on feature-based speech recognition, Dr. B. V. K. V. Kumar for useful discussions, and Dr. D. R. Reddy for providing a stimulating working environment and computing resources.

### REFERENCES

- [1] R. A. Cole, R. M. Stern, M. S. Philips, S. M. Brill, A. P. Pilant, and P. Specker, "Feature-based speaker-independent recognition of isolated English letters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, 1983, pp. 731-733.
- [2] H. Cramer, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1946.
- [3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
- [4] D. G. Keehn, "A note on learning for Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 126-132, 1965.
- [5] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for isolated letter recognition using MAP estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, 1983, pp. 734-737.
- [6] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1973, pp. 60-85.

## Estimation of Error Rates in Classification of Distorted Imagery

M. J. LAHART

**Abstract**—This correspondence considers the problem of matching image data to a large library of objects when the image is distorted. Two types of distortions are considered: blur-type, in which a transfer function is applied to Fourier components of the image, and scale-type, in which each Fourier component is mapped into another. The objects of the library are assumed to be normally distributed in an appropriate feature space. Approximate expressions are developed for classification error rates as a function of noise. The error rates they predict are compared with those from classification of artificial data, generated by a Gaussian random number generator, and with error rates from classification of actual data. It is demonstrated that, for classification purposes, distortions can be characterized by a small number of parameters.

**Index Terms**—Image classification, image matching, feature extraction, pattern classification, pattern recognition.

Manuscript received May 26, 1983; revised August 22, 1983.

The author was with the U.S. Naval Research Laboratory, Washington, DC 20375. He is now with the National Institutes of Health, Bethesda, MD 20209.