Jiří Grim

Multivariate statistical pattern recognition with nonreduced dimensionality

Kybernetika, Vol. 22 (1986), No. 2, 142--157

Persistent URL: http://dml.cz/dmlcz/125022

KYBERNETIKA — VOLUME 22 (1986), NUMBER 2

MULTIVARIATE STATISTICAL PATTERN RECOGNITION WITH NONREDUCED DIMENSIONALITY

JIŘÍ GRIM

A direct solution of multivariate classification problems is considered without preceding reduction of dimensionality. The unknown class-conditional distributions are approximated by finite mixtures of special type. For decision purposes the components of these mixtures can be reduced to functions defined on different subspaces. To optimize the choice of subspaces and of the related parameters maximum-likelihood principle is used. The corresponding m.-l. estimates are computed by the EM algorithm. In this way the feature selection problem can be solved independently for each component of the approximating mixtures without introducing any additional criteria.

1. INTRODUCTION

Considering the statistical approach to pattern recognition we assume that some objects described by real vectors

$$\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathscr{X} \subset \mathbb{R}_d$$

have to be classified with respect to a finite set of classes $\Omega = \{\omega_1, \omega_2, \ldots, \omega_J\}$. The objects are supposed to occur randomly according to some class-conditional probability distributions $P^*(\mathbf{x} \mid \omega)$ and the respective a priori probabilities $p^*(\omega)$, i.e. with the joint distribution

$$P^*(\mathbf{x}) = \sum_{\omega \in \Omega} P^*(\mathbf{x} \mid \omega)\, p^*(\omega)\,.$$

Given a vector $\mathbf{x} \in \mathscr{X}$ we can express the a posteriori probabilities of classes

$$p^*(\omega \mid \mathbf{x}) = \frac{P^*(\mathbf{x} \mid \omega)\, p^*(\omega)}{P^*(\mathbf{x})}, \quad \omega \in \Omega\,; \quad (P^*(\mathbf{x}) > 0) \tag{1.1}$$

A unique classification of the vector $\mathbf{x}$, if necessary, can be obtained e.g. by using the Bayes decision function

$$D\colon \mathscr{X} \to \Omega\,; \quad D(\mathbf{x}) = \omega'\colon P^*(\mathbf{x} \mid \omega')\, p^*(\omega') \geqq P^*(\mathbf{x} \mid \omega)\, p^*(\omega)\,; \quad \omega \in \Omega \tag{1.2}$$

which minimizes the probability of error. In this way the classification problem reduces to estimating the unknown component distributions $P^*(\boldsymbol{x} \mid \omega)\, p^*(\omega)$, $\omega \in \Omega$ from some available samples of independent observations;

$$(1.3) \qquad S_\omega = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_\omega}\} \subset \mathcal{X}\,; \quad \omega \in \Omega\,; \quad S_0 = \bigcup_{\omega \in \Omega} S_\omega\,; \quad N_0 = \sum_{\omega \in \Omega} N_\omega\,.$$

It is well known, however, that practical results of estimating multivariate distributions are mostly unsatisfactory. Loosely speaking, this is particularly due to the discrepancy between an increasing complexity of multivariate distributions and a relatively small size of the actually available (or manageable) data sets. Despite its obvious restrictive nature the standard way to avoid this difficulty is to reduce the dimensionality of the problem. For this reason, the statistical pattern recognition regularly involves different feature selection or feature extraction procedures as a first, more or less independent, step.

This paper describes a direct solution of multivariate classification problems without preceding reduction of dimensionality. The method is based on approximating unknown class-conditional distributions by finite mixtures of special type. The components of mixtures have the form of a product distribution common to all classes which is multiplied by a modifying parametric function defined on a subspace of $\mathcal{X}$. The subspace can be chosen independently for each component by means of a vector of binary parameters. As the common product distribution reduces both in the formula (1.1) for the a posteriori probabilities and in the Bayes decision rule (1.2), the class-conditional distributions may be actually replaced by mixtures of the modifying functions defined on different subspaces.

The term "approximating" is used throughout this paper to emphasize, that, unlike estimation problems, the form of the underlying probability distribution is not known. Let us recall that this is the case in almost all practical problems. In this respect the present paper is closely related to the previous work [5, 7] on developing flexible parametric models for approximation purposes.

To measure the quality of approximation, frequently [7, 9] the relative entropy is used

$$\mathcal{H}(P^*, P) = \mathsf{E}_{P^*}\left\{\log \frac{P^*(\boldsymbol{x})}{P(\boldsymbol{x})}\right\} \geqq 0$$

which is nonnegative and equals zero if and only if P equals P^* almost everywhere [8]. It follows that

$$Q(P) = \mathsf{E}_{P^*}\{\log P(\boldsymbol{x})\} \leqq \mathsf{E}_{P^*}\{\log P^*(\boldsymbol{x})\}$$

and therefore the left-hand part can be used as a criterion to be maximized by P. Obviously, if P^* is unknown, the log-likelihood function for P may be viewed as an estimate of $Q(P)$. For this reason, to optimize the parameters of approximating mixtures, maximum-likelihood criterion is used. The corresponding m.-l. estimates of parameters are computed by an iterative scheme called the EM algorithm [2].

In this way the selection of features is naturally included in estimating the conditional distributions without introducing any other independent criteria.

Finally, the application of the described method is illustrated by a numerical example. For this purpose an artificial decision problem was constructed with two classes of 16-dimensional binary vectors. The corresponding samples of the size $N = 6400$ were generated randomly according to mixtures of multivariate Bernoulli distributions. Also the parameters of mixtures were chosen randomly from some suitably specified intervals.

2. DECISION MODEL BASED ON MIXTURES OF SUBSPACE DEFINED COMPONENTS

In order to approximate unknown conditional probability distributions (density functions or discrete distributions) we use the following parametric model:

$$P(\boldsymbol{x} \mid \omega) = \sum_{m=1}^{M_\omega} w_m^\omega \, F_0(\boldsymbol{x} \mid \boldsymbol{b}_0) \, F(\boldsymbol{x} \mid \boldsymbol{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \boldsymbol{b}_0) \,, \quad \boldsymbol{x} \in \mathscr{X} \,. \tag{2.1}$$

Each component of this finite mixture consists of a common "background" distribution

$$F_0(\boldsymbol{x} \mid \boldsymbol{b}_0) = \prod_{i=1}^{d} f(x_i \mid b_{0i}) \,; \quad \boldsymbol{b}_0 = (b_{01}, b_{02}, \ldots, b_{0d}) \in \mathscr{B}^d \tag{2.2}$$

and a function

$$\begin{aligned} &F(\boldsymbol{x} \mid \boldsymbol{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \boldsymbol{b}_0) = \prod_{i=1}^{d} \left[\frac{f(x_i \mid b_{mi}^\omega)}{f(x_i \mid b_{0i})} \right]^{\varphi_{mi}^\omega} ; \quad \varphi_{mi}^\omega \in \{0, 1\} \,; \\ &\boldsymbol{b}_m^\omega = (b_{m1}^\omega, \ldots, b_{md}^\omega) \in \mathscr{B}^d \,; \quad \boldsymbol{\varphi}_m^\omega = (\varphi_{m1}^\omega, \ldots, \varphi_{md}^\omega) \in \{0, 1\}^d \end{aligned} \tag{2.3}$$

which is actually defined on a subspace

$$\mathscr{X}_m^\omega = \mathscr{X}_{i_1} \times \mathscr{X}_{i_2} \times \ldots \times \mathscr{X}_{i_l} \,; \quad \{i_1, i_2, \ldots, i_l\} \equiv \{1 \leqq i \leqq d : \varphi_{mi}^\omega = 1\} \,. \tag{2.4}$$

The univariate function f occurring in the formulas (2.2), (2.3) is assumed to be from a parametric family of probability distributions

$$\mathscr{F} = \{f(\xi \mid b) \,, \quad \xi \in \mathbb{R}; \, b \in \mathscr{B}\} \,; \quad (\mathscr{B} \subset \mathbb{R}_q) \tag{2.5}$$

with parameter b.

One can see that for any subspace $\mathscr{X}_m^\omega$ chosen by means of the binary parameters φ_{mi}^ω the components of the finite mixture (2.1) are valid probability distributions of product type:

$$F_0(\boldsymbol{x} \mid \boldsymbol{b}_0) \, F(\boldsymbol{x} \mid \boldsymbol{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \boldsymbol{b}_0) = \prod_{i=1}^{d} \left[f(x_i \mid b_{0i})^{1 - \varphi_{mi}^\omega} f(x_i \mid b_{mi}^\omega)^{\varphi_{mi}^\omega} \right] . \tag{2.6}$$

Particularly, setting some $\varphi_{mi}^\omega = 1$ we replace the function $f(x_i \mid b_{0i})$ in the product (2.6) by $f(x_i \mid b_{mi}^\omega)$ and introduce a new independent parameter b_{mi}^ω in the mixture (2.1). The actual number of the involved parameters can be therefore suitably specified

e.g. by the condition

$$\sum_{\omega\in\Omega}\sum_{i=1}^{d}\sum_{m=1}^{M_\omega}\varphi_{mi}^{\omega}=\varrho\ ;\quad (0\leqq\varrho\leqq d\sum_{\omega\in\Omega}M_\omega)\,.$$

It should be noted that the parametric model (2.1) generalizes the idea suggested in an earlier paper [6] where the background distribution F_0 was chosen as a constant or nearly constant function.

It is easy to see that in the formula (1.1) for a posteriori probabilities the background distribution F_0 may be reduced and the same holds for the inequality in the Bayes decision rule (1.2). Thus, for decision purposes, the different classes $\omega\in\Omega$ can be independently characterized on different subspaces of $\mathcal{X}$. Even more, the statistical properties of each class are expressed by a weighted sum of the component functions F, which may also be defined on different subspaces $\mathcal{X}_m^\omega$ (cf. (2.3), (2.4)). The parameters b_{0i} in (2.2) are necessary to define the functions (2.3) but, as it will be shown in Section 3, the background distribution F_0 need not be evaluated at all. This circumstance may become useful in case of an extremely high dimensionality.

The advantages of the parametric model (2.1) become more apparent when the description of objects is redundant but, as it is usually the case, the low informative variables are not identical in all classes. As an example let us consider the classification of hand-written numerals on a binary rastr. It is a difficult task to reduce the dimensionality of the corresponding binary space since the bits which are less relevant for one class of numerals may be highly informative for another one. Moreover, similar relations may arise between different variants of one and the same numeral. For the same reason the reduced description may cause a considerable increase of classification error. On the other side, using the parametric model (2.1), we are not confined to a single subset of variables. The subspaces $\mathcal{X}_m^\omega$ can be chosen independently to describe e.g. the typical representants of each class.

3. OPTIMIZATION OF PARAMETERS USING THE EM ALGORITHM

From the practical point of view otimization of the parametric model (2.1) is of fundamental meaning. In this section we describe a computationally efficient solution of this problem based on the EM algorithm.

As it appears the EM algorithm was developed in context of m.-l. estimating from incomplete data [2] and independently also as a method of identification of finite mixtures [4, 10]. In both fields the original justification of the algorithm was rather heuristic. The important proof of its convergence [4] has been presented first probably by Shlezinger [11] and later, apparently independently, by Baum et al. [1] and others (cf. Dempster et al. [2]). In the paper of Shlezinger the two basic steps of the EM algorithm are interpreted in context of pattern recognition as learning and self-learning respectively. A general and uniform formulation of the EM algorithm involving continuous case can be found in Dempster et al. [2]. This paper

clarifies also some aspects of convergence (see also Wu [12]) and shows many different application possibilities. In some cases [3] the EM algorithm can be used as a general optimization technique which is closely related to the gradient method. Applying the EM algorithm we follow the paper Grim [4] since in Dempster et al. [2] the corresponding modification is described rather schematically.

Let us suppose that for each class $\omega \in \Omega$ there is a set S_ω (cf. (1.3)) of independent observations which are identically distributed according to an unknown conditional distribution $P^*(\mathbf{x} \mid \omega)$. In order to optimize the approximating mixture (2.1) we maximize the corresponding global log-likelihood function. Using notation

$$P(\mathbf{x} \mid W_\omega, B_\omega, \Phi_\omega, \mathbf{b}_0) = F_0(\mathbf{x} \mid \mathbf{b}_0) \sum_{m=1}^{M_\omega} w_m^\omega F(\mathbf{x} \mid \mathbf{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \mathbf{b}_0) ;$$

$$W_\omega = (w_1^\omega, w_2^\omega, \ldots, w_{M_\omega}^\omega) ; \quad w_m^\omega \geqq 0 ; \quad \sum_{m=1}^{M_\omega} w_m^\omega = 1 ; \tag{3.1}$$

$$B_\omega = (\mathbf{b}_1^\omega, \mathbf{b}_2^\omega, \ldots, \mathbf{b}_{M_\omega}^\omega) ; \quad \Phi_\omega = (\boldsymbol{\varphi}_1^\omega, \boldsymbol{\varphi}_2^\omega, \ldots, \boldsymbol{\varphi}_{M_\omega}^\omega) ; \quad \omega \in \Omega$$

we can write

$$L_G = \frac{1}{N_0} \sum_{\omega \in \Omega} \sum_{\mathbf{x} \in S_\omega} \log \left[P(\mathbf{x} \mid W_\omega, B_\omega, \Phi_\omega, \mathbf{b}_0)\, p(\omega) \right] = \tag{3.2}$$

$$= \sum_{\omega \in \Omega} \frac{N_\omega}{N_0} \log p(\omega) + \sum_{\omega \in \Omega} \frac{N_\omega}{N_0} \frac{1}{N_\omega} \sum_{\mathbf{x} \in S_\omega} \log P(\mathbf{x} \mid W_\omega, B_\omega, \Phi_\omega, \mathbf{b}_0)$$

Usually the probabilities $p(\omega)$ may be estimated by the respective relative frequencies. However, sometimes the a priori probabilities are not related with the respective samplesizes N_ω. For this reason we confine ourselves to the second part of the function (3.2) and replace the relative frequences N_ω/N_0 by input parameters $p(\omega)$. Using symbols $\mathbf{W} = \{W_\omega, \omega \in \Omega\}$, $\mathbf{B} = \{B_\omega, \omega \in \Omega\}$, $\boldsymbol{\Phi} = \{\Phi_\omega, \omega \in \Omega\}$ we denote

$$L(\mathbf{W}, \mathbf{B}, \boldsymbol{\Phi}, \mathbf{b}_0) = \sum_{\omega \in \Omega} \frac{p(\omega)}{N_\omega} \sum_{\mathbf{x} \in S_\omega} \log P(\mathbf{x} \mid W_\omega, B_\omega, \Phi_\omega, \mathbf{b}_0) = \tag{3.3}$$

$$= \sum_{\omega \in \Omega} \frac{p(\omega)}{N_\omega} \sum_{\mathbf{x} \in S_\omega} \log \left[\sum_{m=1}^{M_\omega} w_m^\omega F_0(\mathbf{x} \mid \mathbf{b}_0) F(\mathbf{x} \mid \mathbf{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \mathbf{b}_0) \right]$$

In the case of the log-likelihood function (3.3) the two fundamental steps of the EM algorithm may be specified as follows [2, 4]:

Expectation step: Given the parameters $\mathbf{W}$, $\mathbf{B}$, $\boldsymbol{\Phi}$, $\mathbf{b}_0$ compute the a posteriori probabilities

$$p(m \mid \mathbf{x}, \omega) = \frac{w_m^\omega F(\mathbf{x} \mid \mathbf{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \mathbf{b}_0)}{\sum_{j=1}^{M_\omega} w_j^\omega F(\mathbf{x} \mid \mathbf{b}_j^\omega, \boldsymbol{\varphi}_j^\omega, \mathbf{b}_0)} ; \quad \begin{matrix} m = 1, 2, \ldots, M_\omega ; \\ \mathbf{x} \in S_\omega ; \quad \omega \in \Omega ; \end{matrix} \tag{3.4}$$

to determine the conditional expectation

$$\mathscr{L}(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{b}_0) = \sum_{\omega \in \Omega} \frac{p(\omega)}{N_\omega} \sum_{\boldsymbol{x} \in S_\omega} \{ \sum_{m=1}^{M_\omega} p(m \mid \boldsymbol{x}, \omega) \log [w_m^\omega F_0(\boldsymbol{x} \mid \boldsymbol{b}_0) \cdot \cdot F(\boldsymbol{x} \mid \boldsymbol{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \boldsymbol{b}_0)] \} \tag{3.5}$$

Maximization step: Under fixed weights (3.4) compute the new values of $\boldsymbol{W}$, $\boldsymbol{B}$, $\boldsymbol{\Phi}$ and $\boldsymbol{b}_0$ by maximizing the function $\mathscr{L}$:

$$('\boldsymbol{W}, '\boldsymbol{B}, '\boldsymbol{\Phi}, '\boldsymbol{b}_0) = \arg\max_{\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{b}_0} \{\mathscr{L}(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{b}_0)\} \tag{3.6}$$

As it follows from the relation (3.6), the EM algorithm transforms the original problem to a repeated maximization of the function (3.5) which may be viewed as a weighted version of L. Obviously, the application of the EM algorithm is efficient only if we derive a simple explicit solution of the relation (3.6). For this purpose we use first the substitution (2.6):

$$\mathscr{L}(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{b}_0) = \sum_{\omega \in \Omega} p(\omega) \sum_{m=1}^{M_\omega} \left[\frac{1}{N_\omega} \sum_{\boldsymbol{x} \in S_\omega} p(m \mid \boldsymbol{x}, \omega) \right] \log w_m^\omega +$$

$$+ \sum_{i=1}^{d} \sum_{\omega \in \Omega} \frac{p(\omega)}{N_\omega} \sum_{m=1}^{M_\omega} \varphi_{mi}^\omega \sum_{\boldsymbol{x} \in S_\omega} p(m \mid \boldsymbol{x}, \omega) \log f(x_i \mid b_{mi}^\omega) +$$

$$+ \sum_{i=1}^{d} \sum_{\omega \in \Omega} \frac{p(\omega)}{N_\omega} \sum_{m=1}^{M_\omega} (1 - \varphi_{mi}^\omega) \sum_{\boldsymbol{x} \in S_\omega} p(m \mid \boldsymbol{x}, \omega) \log f(x_i \mid b_{0i})$$

Further, denoting

$$'w_m^\omega = \frac{1}{N_\omega} \sum_{\boldsymbol{x} \in S_\omega} p(m \mid \boldsymbol{x}, \omega) ; \quad v(\boldsymbol{x} \mid m, \omega) = \frac{p(m \mid \boldsymbol{x}, \omega)}{\sum_{\boldsymbol{y} \in S_\omega} p(m \mid \boldsymbol{y}, \omega)} ; \tag{3.7}$$

$$m = 1, 2, \ldots, M_\omega ; \quad \omega \in \Omega ,$$

we can write

$$\mathscr{L}(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{b}_0) = \sum_{\omega \in \Omega} p(\omega) \sum_{m=1}^{M_\omega} {}'w_m^\omega \log w_m^\omega + \tag{3.8}$$

$$+ \sum_{i=1}^{d} \sum_{\omega \in \Omega} p(\omega) \sum_{m=1}^{M_\omega} \varphi_{mi}^\omega \, {}'w_m^\omega \sum_{\boldsymbol{x} \in S_\omega} v(\boldsymbol{x} \mid m, \omega) \log f(x_i \mid b_{mi}^\omega) +$$

$$+ \sum_{i=1}^{d} \sum_{\omega \in \Omega} p(\omega) \sum_{m=1}^{M_\omega} (1 - \varphi_{mi}^\omega) \, {}'w_m^\omega \sum_{\boldsymbol{x} \in S_\omega} v(\boldsymbol{x} \mid m, \omega) \log f(x_i \mid b_{0i})$$

It can be seen [4] that for any fixed binary parameters φ_{mi}^ω (and under fixed weights $v(\boldsymbol{x} \mid m, \omega)$) the function (3.8) is maximized by $\boldsymbol{W} = '\boldsymbol{W}$ (cf. (3.7)), $\boldsymbol{B} = '\boldsymbol{B}$ and $\boldsymbol{b}_0 = {}'\boldsymbol{b}_0$ where

$$'b_{mi}^\omega = \arg\max_{b \in \mathscr{B}} \{ \sum_{\boldsymbol{x} \in S_\omega} v(\boldsymbol{x} \mid m, \omega) \log f(x_i \mid b) \} \tag{3.9}$$

$$m = 1, 2, \ldots, M_\omega ; \quad i = 1, 2, \ldots, d ; \quad \omega \in \Omega$$

$$(3.10)\quad {}'b_{0i} = \arg\max_{b\in\mathcal{B}} \left\{ \frac{1}{\varkappa_{0i}} \sum_{\omega\in\Omega} p(\omega) \sum_{m=1}^{M_\omega} (1 - \varphi_{mi}^{\omega})\, {}'w_m^{\omega} \sum_{\boldsymbol{x}\in S_\omega} v(\boldsymbol{x} \mid m, \omega) \log f(x_i \mid b) \right\}$$

and $\varkappa_{0i}$ are the corresponding norming coefficients:

$$\varkappa_{0i} = \sum_{\omega\in\Omega} p(\omega) \sum_{m=1}^{M_\omega} (1 - \varphi_{mi}^{\omega})\, {}'w_m^{\omega}\,; \quad (\varkappa_{0i} > 0)\,; \quad i = 1, 2, \ldots, d\,.$$

For $\varkappa_{0i} = 0$ the parameters ${}'b_{0i}$ can be chosen arbitrarily since it is not included in (3.8). Consequently, the following inequality is satisfied

$$(3.11)\qquad \mathscr{L}(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{b}_0) \leqq \mathscr{L}({}'\boldsymbol{W}, {}'\boldsymbol{B}, \boldsymbol{\Phi}, {}'\boldsymbol{b}_0)\,.$$

Now, making substitutions $\boldsymbol{W} = {}'\boldsymbol{W}$, $\boldsymbol{B} = {}'\boldsymbol{B}$, $\boldsymbol{b}_0 = {}'\boldsymbol{b}_0$ in the formula (3.8) and introducing the quantities

$$ {}'q_{mi}^{\omega} = p(\omega)\, {}'w_m^{\omega} \sum_{\boldsymbol{x}\in S_\omega} v(\boldsymbol{x} \mid m, \omega) \log \frac{f(x_i \mid {}'b_{mi}^{\omega})}{f(x_i \mid {}'b_{0i})}\,;$$

$$m = 1, 2, \ldots, M_\omega\,; \quad i = 1, 2, \ldots, d\,; \quad \omega \in \Omega$$

we obtain

$$(3.12)\quad \mathscr{L}({}'\boldsymbol{W}, {}'\boldsymbol{B}, \boldsymbol{\Phi}, {}'\boldsymbol{b}_0) = \sum_{\omega\in\Omega} p(\omega) \sum_{m=1}^{M_\omega} {}'w_m^{\omega} \log {}'w_m^{\omega} + \sum_{i=1}^{d} \sum_{\omega\in\Omega} p(\omega) \sum_{m=1}^{M_\omega} \varphi_{mi}^{\omega}\, {}'q_{mi}^{\omega} +$$

$$+ \sum_{i=1}^{d} \sum_{\omega\in\Omega} p(\omega) \sum_{m=1}^{M_\omega} {}'w_m^{\omega} \sum_{\boldsymbol{x}\in S_\omega} v(\boldsymbol{x} \mid m, \omega) \log f(x_i \mid {}'b_{0i})\,.$$

If we order the terms ${}'q_{mi}^{\omega}$ in a descending way

$$\{{}'q_{m_k i_k}^{\theta_k}\}_{k=1}^{K}\,; \quad {}'q_{m_k i_k}^{\theta_k} \geqq {}'q_{m_{k+1} i_{k+1}}^{\theta_{k+1}}\,; \quad K = d \sum_{\omega\in\Omega} M_\omega\,;$$

and if we set

$${}'\varphi_{m_k i_k}^{\theta_k} = \begin{cases} 1, & k = 1, 2, \ldots, r\,; \qquad \theta_k \in \Omega\,; \\ 0, & k = r+1, \ldots, K\,; \quad 1 \leqq i_k \leqq d\,; \quad 1 \leqq m_k \leqq M_{\theta_k}\,; \end{cases}$$

then the parameters ${}'\varphi_{mi}^{\omega}$ satisfy the inequality

$$(3.13)\qquad \mathscr{L}({}'\boldsymbol{W}, {}'\boldsymbol{B}, \boldsymbol{\Phi}, {}'\boldsymbol{b}_0) \leqq \mathscr{L}({}'\boldsymbol{W}, {}'\boldsymbol{B}, {}'\boldsymbol{\Phi}, {}'\boldsymbol{b}_0)\,.$$

The relations (3.11) and (3.13) already imply that the maximized function (3.3) is nondecreasing at each iteration of the relations (3.4)–(3.6), (cf. the generalized EM algorithm [2]). As it will be shown in Section 4, in some important cases the parameters ${}'b_{0i}$ can be expressed as linear combinations of ${}'b_{mi}^{\omega}$:

$$(3.14)\qquad {}'b_{0i} = \frac{1}{\varkappa_{0i}} \sum_{\omega\in\Omega} p(\omega) \sum_{m=1}^{M_\omega} (1 - \varphi_{mi}^{\omega})\, {}'w_m^{\omega}\, {}'b_{mi}^{\omega}\,; \quad i = 1, \ldots, d\,.$$

In view of the simplicity of this formula it could be feasible to compute the parameters ${}'\boldsymbol{\Phi}$, ${}'\boldsymbol{b}_0$ which locally maximize the expression (3.12) under fixed parameters ${}'\boldsymbol{W}$, ${}'\boldsymbol{B}$ (cf. Sec. 4).

Remark 3.1. It can be seen that, without formal difficulties, the univariate function f in the relations (3.9), (3.10) may be chosen from different families for each index i.

In this way e.g. discrete and continuous variables may occur simultaneously in the vector $\boldsymbol{x}$.

Let us note finally that some problems arising in estimation theory are less important in the context of approximating. Thus, the existence of local maxima of the function L merely implies different approximation possibilities of different quality. Similarly the choice of the number of components in (2.1) influences only the quality of approximation. The frequently discussed slow convergence of the EM algorithm in the final stages of computation is also of little importance since the corresponding changes of the criterion are usually negligible.

4. APPLICATION TO PARTICULAR TYPES OF MIXTURES

Using the results of Section 3 we summarize first the EM algorithm in more detail. With regard to the particular choice of mixtures in this section we use the formula (3.14) instead of the general relation (3.10).

Step 1: Given the parameters $\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{b}_0$ compute the weights:

$$p(m \mid \boldsymbol{x}, \omega) = \frac{w_m^\omega F(\boldsymbol{x} \mid \boldsymbol{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \boldsymbol{b}_0)}{\sum_{j=1}^{M_\omega} w_j^\omega F(\boldsymbol{x} \mid \boldsymbol{b}_j^\omega, \boldsymbol{\varphi}_j^\omega, \boldsymbol{b}_0)}; \tag{4.1}$$

$$v(\boldsymbol{x} \mid m, \omega) = \frac{p(m \mid \boldsymbol{x}, \omega)}{\sum_{\boldsymbol{y} \in S_\omega} p(m \mid \boldsymbol{y}, \omega)}; \quad \begin{array}{l} m = 1, 2, \ldots, M_\omega; \\ \boldsymbol{x} \in S_\omega; \quad \omega \in \Omega. \end{array} \tag{4.2}$$

Step 2: Under fixed weights (4.1), (4.2) compute new values of $\boldsymbol{W}$ and $\boldsymbol{B}$ by the formulas

$$'w_m^\omega = \frac{1}{N_\omega} \sum_{\boldsymbol{x} \in S_\omega} p(m \mid \boldsymbol{x}, \omega); \quad \begin{array}{l} m = 1, 2, \ldots, M_\omega; \\ i = 1, 2, \ldots, d; \quad \omega \in \Omega \end{array} \tag{4.3}$$

$$'b_{mi}^\omega = \arg\max_{b \in \mathcal{B}} \{ \sum_{\boldsymbol{x} \in S_\omega} v'(\boldsymbol{x} \mid m, \omega) \log f(x_i \mid b) \} \tag{4.4}$$

Step 3: Given the parameters $'\boldsymbol{W}, '\boldsymbol{B}$ and $\boldsymbol{\Phi}$ compute a new value of $\boldsymbol{b}_0$ by eqs.

$$\begin{aligned} '\boldsymbol{b}_{0i} &= \frac{1}{\varkappa_{0i}} \sum_{\omega \in \Omega} p(\omega) \sum_{m=1}^{M_\omega} (1 - \varphi_{mi}^\omega) \, '\boldsymbol{w}_m^\omega \, '\boldsymbol{b}_{mi}^\omega; \quad (\varkappa_{0i} > 0) \\ \varkappa_{0i} &= \sum_{\omega \in \Omega} p(\omega) \sum_{m=1}^{M_\omega} (1 - \varphi_{mi}^\omega) \, '\boldsymbol{w}_m^\omega; \quad i = 1, 2, \ldots, d. \end{aligned} \tag{4.5}$$

For $\varkappa_{0i} = 0$ set $'b_{0i} = b_{0i}$.

Step 4: Using the parameters $'\boldsymbol{W}, '\boldsymbol{B}, '\boldsymbol{b}_0$ and the weights (4.2) compute the quantities

$$'q_{mi}^\omega = p(\omega) \, 'w_m^\omega \sum_{\boldsymbol{x} \in S_\omega} v'(\boldsymbol{x} \mid m, \omega) \log \frac{f(x_i \mid '\boldsymbol{b}_{mi}^\omega)}{f(x_i \mid '\boldsymbol{b}_{0i})}; \quad \begin{array}{l} m = 1, 2, \ldots, M_\omega; \\ i = 1, 2, \ldots, d; \quad \omega \in \Omega, \end{array} \tag{4.6}$$

find a monotone order

$$(4.7) \qquad {}'q^{\theta_1}_{m_1 i_1} \geqq {}'q^{\theta_2}_{m_2 i_2} \geqq \ldots \geqq {}'q^{\theta_r}_{m_r i_r} \geqq \ldots ; \quad \begin{matrix} 1 \leqq m_k \leqq M_{\theta_k} ; \\ 1 \leqq i_k \leqq d ; \quad \theta_k \in \Omega \end{matrix}$$

and define

$$(4.8) \qquad {}'\varphi^{\theta_k}_{m_k i_k} = \begin{cases} 1; & k = 1, 2, \ldots, r ; \\ 0; & k = r+1, \ldots, K ; \end{cases} \quad (K = d \sum_{\omega \in \Omega} M_\omega)$$

If ${}'\boldsymbol{\Phi} \neq \boldsymbol{\Phi}$ continue by the Step 3 using the new parameter ${}'\boldsymbol{\Phi}$. Otherwise continue by the Step 1 with $\boldsymbol{W} = {}'\boldsymbol{W}$, $\boldsymbol{B} = {}'\boldsymbol{B}$, $\boldsymbol{\Phi} = {}'\boldsymbol{\Phi}$ and $\boldsymbol{b}_0 = {}'\boldsymbol{b}_0$.

Remark 4.1. It may be useful, e.g. for the sake of easy implementation, to apply the algorithm above under a given fixed background distribution. In this case it suffices to replace eq. (4.5) by ${}'\boldsymbol{b}_0 = \boldsymbol{b}_0$.

Remark 4.2. The relations (4.6)–(4.8) could be modified to select optimal features in usual sense. Instead of the quantities ${}'q^\omega_{mi}$ we would order the partial sums

$${}'Q_i = \sum_{\omega \in \Omega} \sum_{m=1}^{M_\omega} {}'q^\omega_{mi} ; \quad i = 1, 2, \ldots, d$$

to define the optimal subspace at each iteration.

To apply the described algorithm to a particular type of mixture we have to specify only the relation (4.4) and the formula (4.6). Let us consider first the case of normal mixture.

A. Normal mixture

Let $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathscr{X} \equiv \mathbb{R}_d$ be real vectors and $\mathscr{F} = \{f(\xi \mid c, a), \ \xi \in \mathbb{R}; \ c, a \in \mathbb{R}\}$ be the class of univariate normal densities

$$(4.9) \qquad f(\xi \mid c, a) = \frac{1}{\sqrt{(2\pi a^2)}} \exp\left\{ -\frac{1}{2}\left(\frac{\xi - c}{a}\right)^2 \right\} ; \quad \xi \in \mathbb{R}$$

with a pair of parameters $c, a \in \mathbb{R}$ standing for $b \in \mathscr{B}$ (cf. (2.5)). By the formulas (2.2), (2.3) we have

$$F_0(\boldsymbol{x} \mid \boldsymbol{c}_0, \boldsymbol{a}_0) = \prod_{i=1}^{d} \left[\frac{1}{\sqrt{(2\pi a_{0i}^2)}} \exp\left\{ -\frac{1}{2}\left(\frac{x_i - c_{0i}}{a_{0i}}\right)^2 \right\} \right] ; \quad \boldsymbol{x}, \boldsymbol{c}_0, \boldsymbol{a}_0 \in \mathbb{R}_d$$

$$F(\boldsymbol{x} \mid \boldsymbol{c}^\omega_m, \boldsymbol{a}^\omega_m, \boldsymbol{\varphi}^\omega_m, \boldsymbol{c}_0, \boldsymbol{a}_0) = \prod_{i=1}^{d} \left[\left| \frac{a_{0i}}{a^\omega_{mi}} \right| \exp\left\{ -\left(\frac{1}{2} \frac{x_i - c^\omega_{mi}}{a^\omega_{mi}} \right)^2 + \frac{1}{2}\left(\frac{x_i - c_{0i}}{a_{0i}}\right)^2 \right\} \right]^{\varphi^\omega_{mi}}$$

$$c^\omega_{mi}, a^\omega_{mi} \in \mathbb{R} ; \quad m = 1, 2, \ldots, M_\omega ; \quad \omega \in \Omega .$$

The components of the approximating mixture are therefore normal densities with diagonal covariance matrices.

The implicit relation (4.4) is transformed by the substitution (4.9) to the form

$$(4.10)\quad ('c_{mi}^{\omega}, 'a_{mi}^{\omega}) = \arg\max_{c,a\in R}\left\{\sum_{\mathbf{x}\in S_\omega} v(\mathbf{x}\mid m,\omega)\log\left[\frac{1}{\sqrt{(2\pi a^2)}}\exp\left\{-\frac{1}{2}\left(\frac{x_i-c}{a}\right)^2\right\}\right]\right\}$$

where the parenthesized sum represents a weighted version of the usual log-likelihood function for normal density. It can be shown [4] that this expression is maximized by an analogously weighted version of the related m.-l. estimate:

$$'c_{mi}^{\omega} = \sum_{\mathbf{x}\in S_\omega} x_i\, v(\mathbf{x}\mid m,\omega)\,;\quad i = 1, 2, \ldots, d\,;$$

$$('a_{mi}^{\omega})^2 = \sum_{\mathbf{x}\in S_\omega}(x_i - 'c_{mi}^{\omega})^2\, v(\mathbf{x}\mid m,\omega)\,;\quad m = 1, 2, \ldots, M_\omega\,;\quad \omega\in\Omega\,.$$

Similarly we would obtain an explicit solution of the relation (3.10) which can be rewritten in the form

$$'c_{0i} = \frac{1}{\varkappa_{0i}}\sum_{\omega\in\Omega} p'(\omega)\sum_{m=1}^{M_\omega}(1-\varphi_{mi}^{\omega})\,'w_m^{\omega}\,'c_{mi}^{\omega}\,;\quad i = 1, 2, \ldots, d\,;$$

$$('a_{0i})^2 = \frac{1}{\varkappa_{0i}}\sum_{\omega\in\Omega} p'(\omega)\sum_{m=1}^{M_\omega}(1-\varphi_{mi}^{\omega})\,'w_m^{\omega}\left[('a_{mi}^{\omega})^2 + ('c_{mi}^{\omega})^2 - ('c_{0i})^2\right].$$

Finally, after substitution (4.9), the formula (4.6) can be simplified as follows

$$'q_{mi}^{\omega} = 'w_m^{\omega}\, p(\omega)\log\frac{(a_{0i})^2}{(a_{mi}^{\omega})^2}\,.$$

B. Bernoulli mixture

Considering binary vectors $\mathbf{x} = (x_1, \ldots, x_d)\in\mathcal{X}\equiv\{0,1\}^d$; $x_i\in\{0,1\}$ we shall assume the univariate probability distributions $f\in\mathcal{F}$ in the Bernoulli form

$$(4.11)\qquad f(\xi\mid b) = b^{\xi}(1-b)^{1-\xi},\quad \xi\in\{0,1\}\,;\quad 0\leqq b\leqq 1\,.$$

According to the eqs. (2.2), (2.3) we obtain

$$(4.12)\qquad F_0(\mathbf{x}\mid \mathbf{b}_0) = \prod_{i=1}^{d} b_{0i}^{x_i}(1-b_{0i})^{1-x_i},\quad \mathbf{x}\in\mathcal{X}\,;$$

$$(4.13)\qquad F(\mathbf{x}\mid \mathbf{b}_m^{\omega}, \boldsymbol{\varphi}_m^{\omega}, \mathbf{b}_0) = \prod_{i=1}^{d}\left[\left(\frac{b_{mi}^{\omega}}{b_{0i}}\right)^{x_i}\left(\frac{1-b_{mi}^{\omega}}{1-b_{0i}}\right)^{1-x_i}\right]^{\varphi_{mi}^{\omega}}.$$

Thus the components of the approximating mixture (2.1) are multivariate Bernoulli distributions (cf. (2.6)):

$$F_0(\mathbf{x}\mid \mathbf{b}_0)\,F(\mathbf{x}\mid \mathbf{b}_m^{\omega}, \boldsymbol{\varphi}_m^{\omega}, \mathbf{b}_0) =$$

$$= \prod_{i=1}^{d}\left[(b_{mi}^{\omega})^{x_i}(1-b_{mi}^{\omega})^{1-x_i}\right]^{\varphi_{mi}^{\omega}}\left[b_{0i}^{x_i}(1-b_{0i})^{1-x_i}\right]^{1-\varphi_{mi}^{\omega}}.$$

Substituting (4.11) in the formula (4.4) we can write

$$'b_{mi}^{\omega} = \arg\max_{0 \leqq b \leqq 1} \{[\sum_{x \in S_\omega} x_i v('x \mid m, \omega)] \log b + [1 - \sum_{x \in S_\omega} x_i v(x \mid m, \omega)] \log (1 - b)\} .$$

Again, it can be shown that the explicit solution of this relation is given by the weighted sum [4]

$$'b_{mi}^{\omega} = \sum_{x \in S_\omega} x_i v(x \mid m, \omega) . \tag{4.14}$$

After substitution (4.11) in (3.10) we can derive an analogous explicit solution for b_{0i} which can be rewritten in the form (4.5). The substitutions (4.11) and (4.14) in the formula (4.6) yield:

$$'q_{mi}^{\omega} = p(\omega)'w_m^{\omega}\left['b_{mi}^{\omega} \log \frac{'b_{mi}^{\omega}}{'b_{0i}} + (1 - 'b_{mi}^{\omega}) \log \frac{(1 - 'b_{mi}^{\omega})}{(1 - 'b_{0i})}\right]. \tag{4.15}$$

The last parenthesized expression is known as the relative entropy which may be viewed as a measure of dissimilarity between the two distributions $('b_{mi}^{\omega}; 1 - 'b_{mi}^{\omega})$ and $('b_{0i}; 1 - 'b_{0i})$. The subspaces chosen for different components by (4.7) and (4.8) are therefore the most informative in this sense.

5. NUMERICAL EXAMPLE

An artificial decision problem was constructed with two equiprobable classes $\Omega = \{\omega_1, \omega_2\}$; $p(\omega_1) = p(\omega_2) = \frac{1}{2}$ and the conditional distributions defined as multivariate Bernoulli mixtures:

$$P(x \mid \omega) = \sum_{m=1}^{M_\omega} w_m^{\omega} \prod_{i=1}^{d} (b_{mi}^{\omega})^{x_i} (1 - b_{mi}^{\omega})^{1 - x_i} ; \quad x \in \{0, 1\}^d ; \quad \omega \in \Omega . \tag{5.1}$$

Table 1. Original parameters of mixtures $P(x \mid \omega_1)$, $P(x \mid \omega_2)$

	w_{1-3}	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
		b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}
Class 1	·3210	·790	·773	·062	·378	·241	·159	·808	·535
		·842	·444	·236	·346	·324	·524	·476	·447
	·3850	·277	·317	·764	·159	·848	·298	·394	·451
		·568	·672	·596	·797	·748	·474	·109	·085
	·2940	·488	·896	·701	·302	·906	·601	·430	·850
		·922	·598	·118	·648	·584	·598	·323	·836
Class 2	·4050	·734	·503	·592	·619	·536	·279	·287·	·898
		·612	·164	·266	·419	·294	·926	·943	·078
	·2890	·086	·292	·939	·189	·775	·715	·488	·359
		·257	·783	·845	·158	·061	·773	·095	·727
	·3060	·118	·126	·884	·923	·634	·885	·800	·833
		·089	·308	·898	·599	·061	·676	·158	·489

The number of components in mixtures (5.1) was taken to be $M_{\omega_1} = M_{\omega_2} = 3$. The corresponding parameters $\boldsymbol{W}, \boldsymbol{B}$ (see Table 1) were chosen randomly from the following intervals

$$w_1^\omega, w_2^\omega \in (0{\cdot}267, 0{\cdot}467) ; \quad w_3^\omega = 1 - w_1^\omega - w_2^\omega ; \quad b_{mi}^\omega \in (0{\cdot}05, 0{\cdot}95) ;$$

$$\omega \in \Omega ; \quad m = 1, 2, 3 ; \quad i = 1, 2, \ldots, 16 .$$

The short interval for weights was specified to obtain approximately equally significant components.

Next, two samples S_{ω_1}, S_{ω_2} of size $N_{\omega_1} = N_{\omega_2} = 6400$ were generated randomly according to the respective mixtures defined by the Table 1. To verify the actual statistical properties of the generated data we estimated the original parameters from the samples S_{ω_1}, S_{ω_2}. Using the EM algorithm we obtained m.-l. estimates (cf. Table 2) which show some small deviations from the Table 1. As the primary

Table 2. Estimated parameters of mixtures $P(\boldsymbol{x} \mid \omega_1)$, $P(\boldsymbol{x} \mid \omega_2)$.

	w_{1-3}	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
		b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}
Class 1	·3234	·795	·750	·066	·391	·261	·160	·809	·508
		*·846	·462	·240	·334	*·332	·528	*·466	·452
	·3741	·259	·302	·763	·165	·848	·285	·381	·426
		*·551	·678	·596	·807	*·744	·474	*·107	·090
	·3025	·481	·896	·714	·283	·905	·566	·434	·845
		*·923	·616	·134	·636	*·587	·600	*·325	·838
Class 2	·3994	·734	·503	·592	·619	·536	·279	·287	·898
		*·601	·162	·272	·419	*·300	·921	*·946	·088
	·2911	·092	·281	·948	·169	·765	·703	·485	·354
		*·251	·800	·850	·157	*·066	·778	*·086	·737
	·3095	·109	·120	·895	·935	·650	·871	·803	·833
		*·088	·330	·890	·583	*·060	·671	*·155	·492

objective of the present paper is to compare the results of different approaches, we used the entries of the Table 2 as true parameters in further computations to reduce random influences.

The classification problem obtained by this "plug-in" method was characterized by an error matrix based on the Bayes decision function (1.4):

$$E(\omega' \mid \omega) = \sum_{\boldsymbol{x} \in \mathcal{X}} \delta(D(\boldsymbol{x}), \omega')\, P^*(\boldsymbol{x} \mid \omega)\, p^*(\omega) ; \quad \omega, \omega' \in \Omega .$$

Here $E(\omega' \mid \omega)$ is the probability that a vector $\boldsymbol{x} \in \mathcal{X}$ from a class ω will be classified into ω' (see Table 3). Obviously, the sum of nondiagonal elements is the resulting probability of error $PE = E(\omega_1 \mid \omega_2) + E(\omega_2 \mid \omega_1) = 0{\cdot}072$.

For the sake of comparison, first an optimal 3-dimensional subspace of the original

sample space was determined. By evaluating all possibilities we found that the projections of the given conditional distributions into the subspace $\mathscr{X}_9 \times \mathscr{X}_{13} \times \mathscr{X}_{15}$ (defined by the respective entries of Table 2) yield the minimal probability of error $PE = 0{\cdot}221$ (cf. Table 4).

Table 3. Error matrix for Table 2, $PE = 0{\cdot}072$.

Class	1	2
1	0·463	0·037
2	0·035	0·465

Table 4. Error matrix for the 3-dimensional marginals, $PE = 0{\cdot}221$.

Class	1	2
1	0·380	0·120
2	0·101	0·399

Further, the parametric model (2.1) of comparable complexity was optimized in two modifications, under fixed and optimized background distribution respectively (cf. Remark 4.1). Let us recall that for decision purposes we need only the weight vectors W_ω and the functions $F(\boldsymbol{x} \mid \boldsymbol{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \boldsymbol{b}_0)$; $m = 1, \ldots, M_\omega$; $\omega \in \Omega$. If we denote

$$\gamma_m^\omega = w_m^\omega \prod_{i=1}^{d} \left(\frac{1 - b_{mi}^\omega}{1 - b_{0i}}\right)^{\varphi_{mi}{}^\omega}; \quad \alpha_{mi}^\omega = \frac{b_{mi}^\omega(1 - b_{0i})}{b_{0i}(1 - b_{mi}^\omega)};$$

$$i = 1, 2, \ldots, d; \quad m = 1, 2, \ldots, M_\omega; \quad \omega \in \Omega$$

then we can write (cf. (4.13)):

$$w_m^\omega F(\boldsymbol{x} \mid \boldsymbol{b}_m^\omega, \boldsymbol{\varphi}_m^\omega, \boldsymbol{b}_0) = \gamma_m^\omega \prod_{i=1}^{d} (\alpha_{mi}^\omega)^{\varphi_{mi}{}^\omega x_i}$$

The total number of independent quantitative parameters included in the final decision model (without considering the binary matrix $\boldsymbol{\Phi}$) is then given by the formula

$$r = \sum_{\omega \in \Omega} (M_\omega + \sum_{m=1}^{M_\omega} \sum_{i=1}^{d} \varphi_{mi}^\omega). \tag{5.2}$$

Thus, in both modifications we used the value $r = 22$ which corresponds to the previous 3-dimensional projection.

In the first modification the fixed parameters b_{0i} were set equal to the relative marginal frequences of the whole data set $S_0 = S_{\omega_1} \cup S_{\omega_2}$ (cf. Table 5). The optimal parameters obtained under these conditions are displayed in the Table 5. The corresponding error matrix (cf. Table 6) implies the probability of error $PE = 0{\cdot}139$. Having included the background distribution into optimization we obtained the parameters displayed in the Table 7. The probability of error was $PE = 0{\cdot}111$ (cf. Table 8). Thus, introducing the two modifications of the parametric model (2.1) we were able to reduce the probability of error $PE = 0{\cdot}221$ arising in the optimal 3-dimensional subspace to the values $PE = 0{\cdot}139$ and $PE = 0{\cdot}111$ respectively.

All the results introduced in this section (Tables 2–8) were obtained by the EM algorithm (4.1)–(4.8) and partly verified by repeated computations with different

Table 5. Optimal subset of parameters (fixed background, $r = 22$).

	w_{1-3}	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
		b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}
Class 1	·3366	— ·999	·909 —	— ·090	— —	— —	— —	— —	— —
	·1546	·913 —	— —	·001 —	— —	— —	— —	— —	— —
	·5088	— —	— —	— —	— —	— ·736	— —	— —	— —
Class 2	·3418	— —	— ·128	— —	— —	— —	— —	— ·999	— ·061
	·5738	·104 ·152	·187 —	·929 ·876	— —	— ·052	·796 —	— —	— —
	·0844	— —	— —	— —	— —	— —	— —	— —	— —
Background distribution		·427 ·550	·473 ·495	·656 ·486	·430 ·498	·657 ·360	·458 ·666	·520 ·370	·649 ·418

Table 6. Error matrix for Table 5, $PE = 0{\cdot}139$.

Class	1	2
1	0·441	0·059
2	0·080	0·420

Table 7. Optimal subset of parameters (optimized background, $r = 20$).

	w_{1-3}	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
		b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}
Class 1	·6426	— —	— —	— —	·158 ·738	— ·709	— —	— —	— —
	·3574	— —	— —	·242 —	— —	— —	— —	— —	— —
	·0000	— —	— —	— —	— —	— —	— —	— —	— —
Class 2	·3917	— —	— ·155	— —	— —	— —	— ·927	— ·947	·902 ·088
	·6083	·109 ·167	·198 —	·921 ·865	— —	— ·064	·784 —	— —	— —
	·0000	— —	— —	— —	— —	— —	— —	— —	— —
Background distribution		·566 ·717	·593 ·577	·643 ·321	·558 ·384	·657 ·302	·315 ·602	·520 ·229	·588 ·498

randomly chosen starting points. One iteration of the algorithm required about 2 minutes CPU time on the IBM 370/135 and 30 iterations were needed to obtain the parameters for one mixture in the Table 2. However, after 10–15 iterations the accuracy of estimates is already relatively high (see also Grim [5]). In view of

Table 8. Error matrix for Table 7, $PE = 0{\cdot}111$.

Class	1	2
1	0·431	0·069
2	0·042	0·458

experiments performed earlier with similar data [5] it appears that the results in Table 2 correspond to one unique maximum of the log-likelihood function. The choice of the starting point is more crucial if one optimizes an incomplete parameter set since the initial set of variables is very stable and usually does not change very much. For this reason, to obtain the parameters displayed in the Tables 5 and 7, we started the computations by the complete Table 2 and reduced the number of parameters gradually.

6. CONCLUDING REMARKS

An interesting feature of the parametric model (2.1) is its simple applicability to incomplete data vectors. As the components of mixtures are of product form we can obtain any marginal distribution by omitting appropriate terms. Thus, in case of incomplete observations, we can evaluate the corresponding marginals of conditional distributions and the a posteriori probabilities without replacing the missing values by some estimates. In this way we can adapt also the algorithm in Section 4 to enable the estimation of the mixture (2.1) from incomplete data. Denoting $S\omega_i$, $(S\omega i \subset S\omega)$ the subset of vectors for which the ith coordinate is specified we introduce differentiated weights (cf. (4.2))

$$v_i(\mathbf{x} \mid m, \omega) = \frac{p(m \mid \mathbf{x}, \omega)}{\sum\limits_{\mathbf{y} \in S_{\omega i}} p(m \mid \mathbf{y}, \omega)}; \quad \mathbf{x} \in S_{\omega i}; \quad i = 1, 2, \ldots, d;$$

and modify the related equations (4.4), (4.6):

$$'b_{mi}^{\omega} = \arg\max_{b \in \mathcal{B}} \Big\{ \sum_{\mathbf{x} \in S_{\omega i}} v_i(\mathbf{x} \mid m, \omega) \log f(x_i \mid b) \Big\}$$

$$'q_{mi}^{\omega} = p(\omega)\, 'w_m^{\omega} \sum_{\mathbf{x} \in S_{\omega i}} v_i(\mathbf{x} \mid m, \omega) \log \frac{f(x_i \mid 'b_{mi}^{\omega})}{f(x_i \mid 'b_{0i})}.$$

One can verify that the arguments of Section 3 apply to this version of the algorithm without essential changes.

The idea of a removable background distribution could be applied also to more sophisticated models based on dependence structures [7]. However, the resulting procedure would be probably less efficient because of the arising algorithmical complexities.

(Received April 25, 1985.)

REFERENCES

[1] L. E. Baum, T. Petrie, G. Soules and N. Weiss: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. *41* (1970), 164—171.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B *39* (1977), 1—38.

[3] J. Grim: An algorithm for maximizing a finite sum of positive functions and its application to cluster analysis. Problems Control Inform. Theory *10* (1981), 427—437.

[4] J. Grim: On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions. Kybernetika *18* (1982), 173—190.

[5] J. Grim: Application of finite mixtures to multivariate statistical pattern recognition. In: Proceedings of DIANA Conf. held in Liblice near Prague, September 27—October 1, 1982.

[6] J. Grim: Design and optimization of multilevel homogeneous structures for multivariate pattern recognition. In: Proc. Fourth Formator Symposium (J. Beneš, L. Bakule, eds.), Academia, Prague 1983, pp. 223—240.

[7] J. Grim: On structural approximating multivariate discrete probability distributions. Kybernetika *20* (1984), 1—17.

[8] S. Kullback: Information Theory and Statistics. Dover, New York 1968.

[9] M. A. G. Mattoso Maia and M. C. Fairhurst: On the use of I-divergence for generating distribution approximations. IEEE Trans. Pattern Anal. and Mach. Intel. *PAMI-5* (1983), 661—664.

[10] R. A. Redner and H. F. Walker: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review *26* (1984), 195—239.

[11] M. I. Shlezinger: Relation between learning and self-learning in pattern recognition (in Russian). Kibernetika (Kiev) (1968), 2, 81—88.

[12] C. F. J. Wu: On the convergence properties of the EM algorithm. Ann. Statist. *11* (1983), 95—103.

Ing. Jiří Grim, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.