KYBERNETIKA — VOLUME *18* (1982), NUMBER 3

ON NUMERICAL EVALUATION OF MAXIMUM-LIKELIHOOD ESTIMATES FOR FINITE MIXTURES OF DISTRIBUTIONS

JIŘÍ GRIM

The paper deals with estimation of finite distribution mixtures which are practically important in cluster analysis, pattern recognition and other fields. After a brief survey of existing methods attention is confined to maximum-likelihood estimates, especially to an iterative procedure frequently discussed in the recent literature. It is shown that this procedure in a general form converges monotonly to a possibly local maximum of likelihood function. Application of the general iterative procedure to a particular type of mixture is simplified and illustrated by several examples.

1. INTRODUCTION

Let us consider a parametric family

$$\mathscr{F} = \{f(\boldsymbol{x} \mid \boldsymbol{b}) : \boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}_d ; \quad \boldsymbol{b} \in \mathscr{B} \subset \mathbb{R}_q\} \tag{1.1}$$

of probability density functions $f(\boldsymbol{x} \mid \boldsymbol{b})$ defined on d-dimensional real vector space $\mathbb{R}_d$ and depending on a parameter $\boldsymbol{b}$ from a set $\mathscr{B} \subset \mathbb{R}_q$. We denote by $\mathscr{B}_M$ the M-fold cartesian product of $\mathscr{B}$ with itself and represent an element of $\mathscr{B}_M$ as column

$$\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M)^{\mathrm{T}} \in \mathscr{B}_M$$

Denoting further by

$$\mathscr{W}_M = \{\boldsymbol{W} = (w_1, \ldots, w_M)^{\mathrm{T}} \in \mathbb{R}_M : w_m \geqq 0; \sum_{m=1}^{M} w_m = 1\} \tag{1.2}$$

the set of all M-dimensional weight wectors $\boldsymbol{W}$, we define finite mixture as a probability density function of the form

$$f_M(\boldsymbol{x}) = f_M(\boldsymbol{x} \mid \boldsymbol{W}, \boldsymbol{B}) = \sum_{m=1}^{M} w_m f(\boldsymbol{x} \mid \boldsymbol{b}_m); \quad \boldsymbol{W} \in \mathscr{W}_M; \boldsymbol{B} \in \mathscr{B}_M . \tag{1.3}$$

(Analogously we obtain discrete finite mixture when the components $f(\boldsymbol{x} \mid \boldsymbol{b}_m)$ are discrete probability distributions). Note that the components $f(\boldsymbol{x} \mid \boldsymbol{b}_m)$ may be viewed

as conditional densities and their respective weights w_m as the corresponding a priori probabilities.

In order to identify an unknown mixture we have to estimate unknown parameters $\mathbf{W} \in \mathcal{W}_M$, $\mathbf{B} \in \mathcal{B}_M$ or in a more general case, also the number of components M. The available information, except eventual a priori knowledge, is usually represented by a sample of independent observations

$$\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \; ; \quad \mathbf{x}_n = (x_{n1}, \ldots, x_{nd})^T \in \mathbb{R}_d \tag{1.4}$$

which is supposed to be obtained by observing a d-dimensional random variable with an unknown density function $f^*(\mathbf{x})$.

Remark 1.1. In statistical considerations (parametric problem) the estimated function $f^*(\mathbf{x})$ is directly assumed to be of the form (1.3). In practical situations, however, this assumption may be only rarely justified. The unknown density is therefore rather approximated on a class of finite mixtures (approximation problem). This slight formulational difference may be meaningfull in some respects.

Estimation of finite mixtures is an old and difficult problem studied since 1894 (cf. [45]). There is a valuable survey paper by Isaenko and Urbakh [33] which includes nearly all important results and is frequently referred to in what follows. Also several of the references we cite contain extensive bibliographies (see e.g. [7], [17], [19], [30], [41], [43], [47], [61]).

Comparing theoretical and sample moments Pearson [45] first derived equations for the five unknown parameters in a mixture of two univariate normal densities. Method of moments was further developed and simplified (cf. [10], [12], [50], [51]), modified for other types of mixtures (cf. [6], [7], [52], [53]), combined with graphical techniques (cf. [5], [9], [55]), extended to mutidimensional case (cf. [13]) and compared with other methods (cf. [14], [19], [20], [57]).

More detailed discussion of these results may be found e.g. in [33]. Unfortunately, moment estimators are computationally complex especially in higher dimensions ($d > 1$) and for mixtures with more than two components. Also their sampling properties are not very good (cf. [14]).

Alternatively Doetsch [16] used Fourier transform to identify normal univariate components assuming exact knowledge of values of the decomposed mixture $f^*(\mathbf{x})$. This approach was further extended to other types of mixtures (cf. [41]) and generalized to multivariate case (cf. [54]). Instead of exactly known function $f^*(\mathbf{x})$ Stanat [54] used Fourier approximation to the sample distribution. However, it appears that statistical aspects of the problem are not sufficiently reflected by Doetsch's solution. The literature in monography [41] is almost nonoverlapping with that of [33].

Several experiments should also be mentioned in connection with application of minimum χ^2 (cf. [14], [19]) and Bayes estimators (cf. [35], [42]) to mixtures. Both methods appear to be intractable or computationally complex in higher di-

mensions ($d > 1$) except possibly in a discrete case. Similarly various graphical and semigraphical methods (cf. [5], [6]) are also difficult to extend to higher dimensions. We have not mentioned also some other special approaches (cf. [33]) but they all seem to be greatly inferior to maximum-likelihood method, which will be discussed in more details.

2. MAXIMUM LIKELIHOOD ESTIMATES FOR MIXTURES

In order to obtain maximum-likelihood estimates for a mixture (1.3) the corresponding likelihood function

$$(2.1) \qquad L(\mathbf{W}, \mathbf{B}) = L(\mathbf{W}, \mathbf{B} \mid \mathscr{S}) = \sum_{n=1}^{N} \ln f_M(\mathbf{x}_n) = \sum_{n=1}^{N} \ln \left[\sum_{m=1}^{M} w_m f(\mathbf{x}_n \mid \mathbf{b}_m) \right]$$

generated by a sample $\mathscr{S}$ is to be maximized with respect to parameters $\mathbf{W} \in \mathscr{W}_M$ and $\mathbf{B} \in \mathscr{B}_M$. Unfortunately likelihood equations obtained by setting derivatives of $L(\mathbf{W}, \mathbf{B})$ to zero seem to have no explicit solution in case of mixtures. Procedures for numerical evaluation of maximum-likelihood estimates were proposed for mixtures only recently probably under influence of modern computers. All these procedures are of iterative nature. Omitting standard approaches like steepest ascent and Newton-Raphson (cf. [15], [18]) we confine ourselves to an especially attractive iteration scheme repeatedly treated in recent literature (cf. [1], [4], [14], [17], [21], [26], [27], [28], [30] [33], [47], [56], [60], [61]). This scheme can be very simply programed and may be easily extended to higher dimensions and general mixtures of various types. To illustrate the main ideas of the method we write recurrent equations for parameters of a normal mixture:

$$(2.2) \qquad w_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n); \qquad p^{(t)}(m \mid \mathbf{x}_n) = \frac{w_m^{(t)} f(\mathbf{x}_n \mid \mathbf{c}_m^{(t)}, \mathbf{A}_m^{(t)})}{\sum_{j=1}^{M} w_j^{(t)} f(\mathbf{x}_n \mid \mathbf{c}_j^{(t)}, \mathbf{A}_j^{(t)})};$$

$$(2.3) \qquad \mathbf{c}_m^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n)} \sum_{n=1}^{N} \mathbf{x}_n \, p^{(t)}(m \mid \mathbf{x}_n);$$

$$(2.4) \qquad \mathbf{A}_m^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n)} \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{c}_m^{(t+1)})(\mathbf{x}_n - \mathbf{c}_m^{(t+1)})^{\mathrm{T}} \, p^{(t)}(m \mid \mathbf{x}_n);$$

$$m = 1, 2, \ldots, M; \quad t = 0, 1, \ldots$$

Here $w_m^{(t+1)}$, $\mathbf{c}_m^{(t+1)}$ $\mathbf{A}_m^{(t+1)}$ are weight, mean and covariance matrix of the m-th component respectively – after $(t + 1)$ iterations. Note that no stepsize is needed for computation and the constraints of the problem are automatically satisfied except for

possible singularity of matrices $\mathbf{A}_m^{(t+1)}$. Morover equations (2.3), (2.4) represent a natural "weighted" generalization of m.-l. estimates for single population.

It appears that Hasselblad [26] first recognized computational advantages of the above procedure, though originally in a form restricted to grouped data. By Hosmer [30]: "Iterative m.-l. estimates were proposed by Hasselblad and subsequently have been looked at by Day, Hosmer and Wolfe". Similarly Cohen [11] recalls that "... the more general case $M \geqq 3$ dealt with by Hasselblad seems to have received little if any previous attention".

When omitting the iteration index (t) equations (2.2)–(2.4) may be easily obtained by algebraically rearranging the corresponding likelihood equations. Using this heuristic idea Hasselblad [26] derived first an iteration scheme for univariate normal mixture of M components and later ([27], [28]) a general "successive substitutions" procedure for mixtures from exponential family. In the same way Behboodian [4] obtained Eq. (2.2)–(2.4) for univariate – and Day [14] for multivariate normal mixtures. Day considered only two components with common covariance matrix and pointed out the existence of singular solutions with general normal mixtures. By Wolfe [61] Eq. (2.2)–(2.4) are optimal "... in the limiting case of very widely separated components ..." and may be viewed as a special case of the method of scoring [36]. However there is no exact evidence of convergence properties in papers [26], [28], [4], [61].

In a recent paper Peters and Walker [47] studied Eq. (2.2)–(2.4) as a special case of a more general "deflected gradient type" iterative procedure converging locally in certain sense. Like Hasselblad [28] they observed in experiments that the convergence is monotone, i.e. that the likelihood function is actually increased at each iteration of Eq. (2.2)–(2.4), but they were unable to prove it.

Finally an outstanding paper of Shlezinger [56] should be mentioned in which a general form of the above iterative procedure is suggested applicable for any type of multivariate finite mixture. Also the monotone convergence of this procedure to some possibly local maximum is proved in full generality. Shlezinger's paper clarifies the underlying principle of the procedure which may be independently modified for a more general purpose (cf. [22]). The main results of the paper [56] are presented in Sections 3 and 4 though in a different way. In Section 5 an implicit relation occurring in Shlezinger's original procedure is generally solved for a class of mixtures. Examples of use of the general procedure for special types of mixtures are presented in Section 6.

3. GENERAL ITERATIVE PROCEDURE

To characterize the principle of the procedure we first express likelihood function (2.1) in a special form (cf. [56]). Defining the value of indetifinite expression $0 \ln 0$ by

$$0 \ln 0 = \lim_{\zeta \to 0_+} \zeta \ln \zeta = 0 \tag{3.1}$$

and denoting

$$\boldsymbol{P} = [p(m \mid \boldsymbol{x}_n)]_{m=1}^{M}{}_{n=1}^{N}\,; \quad p(m \mid \boldsymbol{x}_n) = \frac{w_m f(\boldsymbol{x}_n \mid \boldsymbol{b}_m)}{f_M(\boldsymbol{x}_n)}\,. \tag{3.2}$$

we can write

$$L(\boldsymbol{W}, \boldsymbol{B}) = \sum_{m=1}^{M} [\sum_{n=1}^{N} p(m \mid \boldsymbol{x}_n)] \ln w_m + \sum_{m=1}^{M} [\sum_{n=1}^{N} p(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}_m)] - \tag{3.3}$$

$$- \sum_{n=1}^{N} \sum_{m=1}^{M} p(m \mid \boldsymbol{x}_n) \ln p(m \mid \boldsymbol{x}_n)$$

whenever it holds

$$f_M(\boldsymbol{x}_n) > 0\,; \quad n = 1, 2, \ldots, N\,. \tag{3.4}$$

In the course of iterative process the $M \times N$ stochastic matrix $\boldsymbol{P}$ may be viewed as an internal dependent parameter. At each iteration expression (3.3) is first maximized under fixed values of $p(m \mid \boldsymbol{x}_n)$ with regard to the parameters $\boldsymbol{W} \in \mathscr{W}_M$ and $\boldsymbol{B} \in \mathscr{B}_M$ and next the matrix $\boldsymbol{P}$ is recomputed for new values of $\boldsymbol{W}$ and $\boldsymbol{B}$. In the following we describe the general iterative procedure more precisely:

Step 0. Choose initial values $\boldsymbol{W}^{(0)} \in \mathscr{W}_M$, $\boldsymbol{B}^{(0)} \in \mathscr{B}_M$ such that the inequalities

$$f_M^{(0)}(\boldsymbol{x}_n) = \sum_{m=1}^{M} w_m^{(0)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(0)}) > 0\,; \quad n = 1, \ldots, N \tag{3.5}$$

are satisfied and compute the matrix of parameters

$$\boldsymbol{P}^{(0)} = [p^{(0)}(m \mid \boldsymbol{x}_n)]_{m=1}^{M}{}_{n=1}^{N}\,; \quad p^{(0)}(m \mid \boldsymbol{x}_n) = \frac{w_m^{(0)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(0)})}{f_M^{(0)}(\boldsymbol{x}_n)} \tag{3.6}$$

Step 1. For a given matrix $\boldsymbol{P}^{(t)} = [p^{(t)}(m \mid \boldsymbol{x}_n)]_{m=1}^{M}{}_{n=1}^{N}$, $(t = 0, 1, 2, \ldots)$ compute new parameters $\boldsymbol{W}^{(t+1)} \in \mathscr{W}_M$, $\boldsymbol{B}^{(t+1)} \in \mathscr{B}_M$ by formulas

$$w_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n)\,; \tag{3.7}$$

$$\boldsymbol{b}_m^{(t+1)} = \arg\max_{\boldsymbol{b} \in \mathscr{B}} \{\sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b})\}\,; \quad m = 1, 2, \ldots, M \tag{3.8}$$

Step 2. Using parameters $\boldsymbol{W}^{(t+1)}$, $\boldsymbol{B}^{(t+1)}$ compute the corresponding matrix $\boldsymbol{P}^{(t+1)}$

$$\boldsymbol{P}^{(t+1)} = [p^{(t+1)}(m \mid \boldsymbol{x}_n)]_{m=1}^{M}{}_{n=1}^{N}\,; \quad p^{(t+1)}(m \mid \boldsymbol{x}_n) = \frac{w_m^{(t+1)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t+1)})}{f_M^{(t+1)}(\boldsymbol{x}_n)} \tag{3.9}$$

and continue by Step 1.

Remark 3.1. Using notation (3.8) we assume that the parameter $\boldsymbol{b}_m^{(t+1)} \in \mathscr{B}$ corresponds to a finite maximum of the parenthesized expression, i.e.

$$\infty > \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t+1)}) \geqq \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b})\,; \quad \boldsymbol{b} \in \mathscr{B} \tag{3.10}$$

Further we assume that $\boldsymbol{b}_m^{(t+1)} \in \mathscr{B}$ is uniquely chosen when the maximum is not unique.

Remark 3.2. Observe that if $f_M^{(t)}(\boldsymbol{x}_n) > 0$ then $p^{(t)}(m_0 \mid \boldsymbol{x}_n) > 0$ for some m_0, $(1 \leqq m_0 \leqq M)$. Further, considering Eq. (3.7), (3.10) we can write

$$p^{(t)}(m_0 \mid \boldsymbol{x}_n) > 0 \Rightarrow w_{m_0}^{(t+1)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_{m_0}^{(t+1)}) > 0 \Rightarrow \tag{3.11}$$
$$\Rightarrow f_M^{(t+1)}(\boldsymbol{x}_n) > 0 \ (\Rightarrow p^{(t+1)}(m_0 \mid \boldsymbol{x}_n) > 0)$$

and condition (3.5) is automatically satisfied in further iterations.

Note that the M-tuple of equations (3.8) represent a special case of the following more general equation.

$$\boldsymbol{B}^{(t+1)} = \arg\max_{\boldsymbol{B} \in \mathscr{B}_M} \{ \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}_m) \} \tag{3.12}$$

characterized by the inequality

$$\infty > \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t+1)}) \geqq \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}_m) ; \tag{3.13}$$
$$\boldsymbol{B} \in \mathscr{B}_M$$

Obviously, when the parameters $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M$ are independent, then (3.12) reduces to (3.8).

Note further, that expression maximized in (3.8) or (3.12) may be multiplied by any positive constant. Thus for $w_m^{(t)} > 0$ Eq. (3.8) is equivalent to

$$\boldsymbol{b}_m^{(t+1)} = \arg\max_{\boldsymbol{b} \in \mathscr{B}} \left\{ \sum_{n=1}^{N} \frac{f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t)})}{f_M^{(t)}(\boldsymbol{x}_n)} \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) \right\} \tag{3.14}$$

which may be more practical since it is formally applicable even if $w_m^{(t)} = 0$. By appropriate norming both expressions in (3.8) and (3.14) may be transformed to the form

$$\mathscr{L}(\boldsymbol{b}) = \mathscr{L}(\boldsymbol{b} \mid \boldsymbol{V}, \mathscr{S}) = \sum_{n=1}^{N} v_n \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) ; \quad \boldsymbol{b} \in \mathscr{B} ; \quad \boldsymbol{V} \in \mathscr{W}_N \tag{3.15}$$

which will be called weighted likelihood function for obvious analogy.

To simplify notation we introduce mappings

$$\text{a)} \ \Phi_0 : \mathscr{W}_M \times \mathscr{B}_M \to \mathscr{P}_{MN} ; \quad \boldsymbol{P}^{(t)} = \Phi_0(\boldsymbol{W}^{(t)}, \boldsymbol{B}^{(t)}) \tag{3.16}$$
$$\text{b)} \ \Phi_1 : \mathscr{P}_{MN} \to \mathscr{W}_M ; \quad \boldsymbol{W}^{(t+1)} = \Phi_1(\boldsymbol{P}^{(t)})$$
$$\text{c)} \ \Phi_2 : \mathscr{P}_{MN} \to \mathscr{B}_M ; \quad \boldsymbol{B}^{(t+1)} = \Phi_2(\boldsymbol{P}^{(t)})$$

defined by Eq. (3.9), (3.7) and (3.8) respectively, whereby $\mathscr{P}_{MN} \subset \mathbb{R}_{MN}$ denotes the set of all $M \times N$ stochastic matrices. By iterating first and second steps of the above procedure we obtain sequences

$$\dot{\boldsymbol{W}} = \{\boldsymbol{W}^{(t)}\}_{t=0}^{\infty} ; \quad \dot{\boldsymbol{B}} = \{\boldsymbol{B}^{(t)}\}_{t=0}^{\infty} ; \quad \dot{\boldsymbol{P}} = \{\boldsymbol{P}^{(t)}\}_{t=0}^{\infty}$$

interrelated through equations (3.16). It is easy to see that if the sequences $\dot{\boldsymbol{W}}$, $\dot{\boldsymbol{B}}$, $\dot{\boldsymbol{P}}$ converge, i.e.

$$a)\ \lim_{t\to\infty} \boldsymbol{W}^{(t)} = \boldsymbol{W}^* \in \mathscr{W}_M\ ;\quad b)\ \lim_{t\to\infty} \boldsymbol{B}^{(t)} = \boldsymbol{B}^* \in \mathscr{B}_M\ ;\quad c)\ \lim_{t\to\infty} \boldsymbol{P}^{(t)} = \boldsymbol{P}^* \in \mathscr{P}_{MN}$$

and the mapping Φ_2 is continuous at the point $\boldsymbol{P}^*$ then parameters $\boldsymbol{W}^*$, $\boldsymbol{B}^*$ represent fixed point of the procedure, i.e. they satisfy equations

$$(3.17)\qquad a)\ \boldsymbol{W}^* = \Phi_1(\boldsymbol{P}^*)\ ;\quad b)\ \boldsymbol{B}^* = \Phi_2(\boldsymbol{P}^*)\ ;\quad c)\ \boldsymbol{P}^* = \Phi_0(\boldsymbol{W}^*, \boldsymbol{B}^*)$$

We conclude this Section by proving that fixed points defined by Eq. (3.17) satisfy corresponding likelihood equations provided that desirable derivatives exist. In certain sense this is a counterpart of the well known fact, that iterative equations analogous to (3.16) may be obtained just by rewritting the likelihood equations. Without any loss of generality only nonzero weights $w_m^* > 0$ are considered in the theorem since otherwise related parameters may be ignored.

Theorem 3.1. Let $[\boldsymbol{W}^*, \boldsymbol{B}^*] \in \mathscr{W}_M \times \mathscr{B}_M$ be a fixed point of the iterative procedure with all the weights w_m^*, $(m = 1, 2, \ldots, M)$ positive. Then the parameters $\boldsymbol{W}^*$, $\boldsymbol{B}^*$ satisfy necessary conditions for a maximum of likelihood function $L(\boldsymbol{W}, \boldsymbol{B})$ provided that desirable derivatives exist.

Proof. Using a Lagrangian multiplier λ we form the function

$$(3.18)\qquad L'(\boldsymbol{W}, \boldsymbol{B}) = \sum_{n=1}^{N} \ln\Big[\sum_{m=1}^{M} w_m f(\boldsymbol{x}_n \mid \boldsymbol{b}_m)\Big] + \lambda\Big(1 - \sum_{m=1}^{M} w_m\Big)$$

Necessary conditions for a maximum of $L'(\boldsymbol{W}, \boldsymbol{B})$ are expressed by equations

$$(3.19)\qquad \frac{\partial L'}{\partial w_m} = \sum_{n=1}^{N} \frac{f(\boldsymbol{x}_n \mid \boldsymbol{b}_m)}{\sum_{j=1}^{M} w_j f(\boldsymbol{x}_n \mid \boldsymbol{b}_j)} - \lambda = 0\ ,\quad m = 1, \ldots, M\ ;$$

$$(3.20)\qquad \operatorname{grad}_{\boldsymbol{b}_m} L' = \sum_{n=1}^{N} \frac{w_m \operatorname{grad}_{\boldsymbol{b}_m} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m)}{\sum_{j=1}^{M} w_j f(\boldsymbol{x}_n \mid \boldsymbol{b}_j)} = 0\ ,\quad m = 1, \ldots, M\ ;$$

Note first that multiplying equations (3.19) by corresponding weights w_m and summing across m we obtain $\lambda = N$. To prove the Theorem we use equations for fixed point (3.17) in the original notation (3.7)–(3.9). Substituting from (3.9) c) we can write (3.17) a) in the form

$$(3.21)\qquad w_m^* \left(\sum_{n=1}^{N} \frac{f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^*)}{\sum_{j=1}^{M} w_j^* f(\boldsymbol{x}_n \mid \boldsymbol{b}_j^*)} - N \right) = 0\ ;\quad m = 1, \ldots, M\ ;$$

It follows that $\boldsymbol{W}^*$ and $\boldsymbol{B}^*$ satisfy Eq. (3.19) since w_m^* is positive. Further it may be

seen that equations (3.17) b) imply necessary conditions

$$\text{grad}_{\boldsymbol{b}} \{ \sum_{n=1}^{N} p^*(m \mid \boldsymbol{x}_n) \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) \}|_{\boldsymbol{b}=\boldsymbol{b}_m^*} = 0 ; \quad m = 1, \ldots, M \tag{3.22}$$

which may be rewritten in the form

$$\sum_{n=1}^{N} \frac{w_m^* \, \text{grad}_{\boldsymbol{b}} f(\boldsymbol{x}_n \mid \boldsymbol{b})|_{\boldsymbol{b}=\boldsymbol{b}_m^*}}{\sum_{j=1}^{M} w_j^* f(\boldsymbol{x}_n \mid \boldsymbol{b}_j^*)} = 0 ; \quad m = 1, \ldots, M . \tag{3.23}$$

The proof is complete since by (3.23) the parameters $\boldsymbol{W}^*$, $\boldsymbol{B}^*$ satisfy also Eq. (3.20). □

4. CONVERGENCE PROPERTIES

In this section we use repeatedly a discrete version of an inequality known in information theory (cf. [39]). For the sake of completeness we bring the proof here since it is brief.

Lemma 4.1. Any two discrete probability distributions $\boldsymbol{a} = (a_1, \ldots, a_M)^T \in \mathscr{W}_M$, $\boldsymbol{b} = (b_1, \ldots, b_M)^T \in \mathscr{W}_M$ satisfy the inequality

$$\sum_{m=1}^{M} a_m \ln \frac{a_m}{b_m} \geqq 0 \tag{4.1}$$

whereby the equal sign in (4.1) holds if and only if $\boldsymbol{a} = \boldsymbol{b}$.

Proof. Using convention (3.1) we may assume $a_m > 0$ for all $m = 1, 2, \ldots, M$, without any loss of generality. If we denote

$$\alpha_m = \frac{b_m}{a_m} ; \quad m = 1, 2, \ldots, M \tag{4.2}$$

then by Jensen's inequality, we obtain relation

$$\sum_{m=1}^{M} a_m \ln \alpha_m \leqq \ln \left[\sum_{m=1}^{M} a_m \alpha_m \right] = 0 \tag{4.3}$$

which may be rewritten in the form (4.1). Since logarithm is a strictly concave function, the left-hand side of (4.3) equals zero only if $\alpha_1 = \alpha_2 = \ldots = \alpha_M$, i.e. only if $\boldsymbol{a} = \boldsymbol{b}$. The proof is complete since the left-hand side of (4.1) is zero if $\boldsymbol{a} = \boldsymbol{b}$. □

The most interesting property of the iterative procedure is its monotone convergence more precisely expressed in the following theorem.

Theorem 4.1. Let $\dot{\boldsymbol{W}}$, $\dot{\boldsymbol{B}}$, $\dot{\boldsymbol{P}}$ be sequences produced by the iterative procedure (cf. (3.16)). Then the sequence

$$\dot{L} = \{L^{(t)}\}_{t=0}^{\infty} ; \quad L^{(t)} = L(\boldsymbol{W}^{(t)}, \boldsymbol{B}^{(t)}) \tag{4.4}$$

is nondecreasing, i.e.

$$L^{(t+1)} - L^{(t)} \geqq 0; \quad t = 0, 1, \ldots \tag{4.5}$$

whereby the left-hand side of the inequality (4.5) is strictly positive if $\boldsymbol{W}^{(t+1)} \neq \boldsymbol{W}^{(t)}$ or $\boldsymbol{P}^{(t)} \neq \boldsymbol{P}^{(t+1)}$. Further if $L^{(t+1)} = L^{(t)}$ then $\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)}$ and $\boldsymbol{P}^{(t+1)} = \boldsymbol{P}^{(t)}$.

Proof. We first express the difference $L^{(t+1)} - L^{(t)}$. Employing (3.3) we get

$$L^{(t+1)} - L^{(t)} = \sum_{n=1}^{N} f_M^{(t+1)}(\boldsymbol{x}_n) - \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln w_m^{(t)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t)}) + \tag{4.6}$$

$$+ \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln p^{(t)}(m \mid \boldsymbol{x}_n).$$

By simultaneous adding and subtracting the expression

$$\sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln w_m^{(t+1)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t+1)}) \tag{4.7}$$

we may rewrite equation (4.6) in the form

$$L^{(t+1)} - L^{(t)} = \sum_{n=1}^{N} [\sum_{m=1}^{M} p^{(t)}(m \mid \boldsymbol{x}_n)] \ln f_M^{(t+1)}(\boldsymbol{x}_n) + \tag{4.8}$$

$$+ \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln \frac{w_m^{(t+1)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t+1)})}{w_m^{(t)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t)})} + \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln \frac{p^{(t)}(m \mid \boldsymbol{x}_n)}{w_m^{(t+1)} f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t+1)})}.$$

Finally, using formulas (3.7) and (3.9), we obtain

$$L^{(t+1)} - L^{(t)} = \sum_{m=1}^{M} \sum_{n=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_n) \ln \frac{f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t+1)})}{f(\boldsymbol{x}_n \mid \boldsymbol{b}_m^{(t)})} + \tag{4.9}$$

$$+ N \sum_{m=1}^{M} w_m^{(t+1)} \ln \frac{w_m^{(t+1)}}{w_m^{(t)}} + \sum_{n=1}^{N} \sum_{m=1}^{M} p^{(t)}(m \mid \boldsymbol{x}_n) \ln \frac{p^{(t)}(m \mid \boldsymbol{x}_n)}{p^{(t+1)}(m \mid \boldsymbol{x}_n)}.$$

Note that the value of (4.7) is finite by relations (3.10) and (3.11). In Eq. (4.9) the first term on the right is nonnegative by the inequality (3.10) and hence we can write

$$L^{(t+1)} - L^{(t)} \geqq N \sum_{m=1}^{M} w_m^{(t+1)} \ln \frac{w_m^{(t+1)}}{w_m^{(t)}} + \tag{4.10}$$

$$+ \sum_{n=1}^{N} \sum_{m=1}^{M} p^{(t)}(m \mid \boldsymbol{x}_n) \ln \frac{p^{(t)}(m \mid \boldsymbol{x}_n)}{p^{(t+1)}(m \mid \boldsymbol{x}_n)}.$$

The proof is complete since the assertion of Theorem 4.1 is now an immediate consequence of Lemma 4.1 and the above inequality (4.10). □

Remark 4.1. It should be noted that Theorem 4.1 remains valid even if the inequality (3.10) holds only for $\boldsymbol{b} = \boldsymbol{b}_m^{(t)}$, i.e. if the parameter $\boldsymbol{b}_m^{(t+1)}$ in Eq. (3.8) does not maximize but only increase the parenthesized expression. In this case however Theorem 3.1 need not hold.

As it follows from Eq. (4.9) the total increment of the maximized likelihood function consists of three nonnegative parts. The first one given by the first sum in (4.9) is achieved in the Step 1 (Eq. (3.8)) while computing the new parameters $\boldsymbol{b}_m^{(t+1)}$. The second and third part represent a measure of dissimilarity between $\boldsymbol{W}^{(t+1)}$ and $\boldsymbol{W}^{(t)}$ and between $\boldsymbol{P}^{(t+1)}$ and $\boldsymbol{P}^{(t)}$ respectively, which is known as relative entropy or minimum discrimination information (cf. [39]).

It is easy to see that if $L(\boldsymbol{W}, \boldsymbol{B})$ is a bounded function on $\mathscr{W}_M \times \mathscr{B}_M$ then the sequence $\dot{L}$ converges since it is nondecreasing by Theorem 4.1. Unfortunately likelihood function is not bounded in the important case of normal mixture since covariance matrices may become singular. However, the sequence $\dot{L}$ converges mostly also in case of normal mixtures whereby singular solutions cause difficulties only when initial estimates are poor and/or sample size is small (cf. [18], [30]).

It should be pointed out that the convergence of $\dot{L}$ does not generally imply that of the sequences $\dot{\boldsymbol{W}}$, $\dot{\boldsymbol{B}}$ and $\dot{\boldsymbol{P}}$. Without making additional assumptions we can prove the following assertion.

Theorem 4.2. If the sequence $\dot{L}$ has a finite limit, i.e.

$$\lim_{t\to\infty} L^{(t)} = L^* ; \quad L^* < \infty \tag{4.11}$$

then sequences $\dot{\boldsymbol{W}}$, $\dot{\boldsymbol{P}}$ satisfy the following necessary conditions of convergence

$$\lim_{t\to\infty} \left\| \boldsymbol{W}^{(t+1)} - \boldsymbol{W}^{(t)} \right\| = 0 ; \quad \lim_{t\to\infty} \left\| \boldsymbol{P}^{(t+1)} - \boldsymbol{P}^{(t)} \right\| = 0 \tag{4.12}$$

where $\|\cdot\|$ denotes usual Euclidean norm.

Proof. To prove the Theorem we make use of the inequality

$$\sum_{m=1}^{M} a_m \ln \frac{a_m}{b_m} \geqq \tfrac{1}{4} \Big[\sum_{m=1}^{M} |a_m - b_m| \Big]^2 \geqq \tfrac{1}{4} \|\boldsymbol{a} - \boldsymbol{b}\|^2 ; \quad \boldsymbol{a}, \boldsymbol{b} \in \mathscr{W}_M \tag{4.13}$$

derived by Kullback [38]. Applying (4.13) to (4.10) we obtain

$$L^{(t+1)} - L^{(t)} \geqq \tfrac{1}{4} \left\| \boldsymbol{W}^{(t+1)} - \boldsymbol{W}^{(t)} \right\|^2 + \tfrac{1}{4} \left\| \boldsymbol{P}^{(t+1)} - \boldsymbol{P}^{(t)} \right\|^2 . \tag{4.14}$$

The proof is complete since the left-hand side of (4.14) tends to zero by Eq. (4.11). □

5. EXPLICIT SOLUTION OF THE *STEP* 1

Recall that the iterative procedure includes implicit relation (3.8), (resp. (3.12)) which is to be solved for all components at each iteration. Obviously an efficient use of the procedure is possible only if there is a simple solution of this relation. It is easily verified that in case of normal mixture Eq. (3.8) is equivalent to (2.3) and (2.4). Considering these equations we see that the weighted likelihood function in (3.8) is maximized by correspondingly weighted m.-l. estimates for a single normal population.

This analogy suggests possibility of a more general solution first noticed by Hasselblad [28]. In the following Theorem the generalized solution is obtained for m.-l. estimates which are additive with regard to the sample in certain sense.

Theorem 5.1. Assume that for any sample $\mathscr{S}$ the likelihood function

$$L(\boldsymbol{b} \mid \mathscr{S}) = \sum_{n=1}^{N} \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) ; \quad \boldsymbol{b} \in \mathscr{B} \subset \mathbb{R}_q \tag{5.1}$$

is maximized by $\hat{\boldsymbol{b}}(\mathscr{S}) \in \mathscr{B}$,

$$\hat{\boldsymbol{b}}(\mathscr{S}) = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\beta}(\boldsymbol{x}_n) ; \quad \boldsymbol{\beta}(\boldsymbol{x}_n) = (\beta_1(\boldsymbol{x}_n), \ldots, \beta_q(\boldsymbol{x}_n))^{\mathrm{T}} \in \mathbb{R}_q \tag{5.2}$$

where $\boldsymbol{\beta}(\cdot)$ is a vector of real functions defined on $\mathbb{R}_d$, i.e. we can write

$$L(\hat{\boldsymbol{b}}(\mathscr{S}) \mid \mathscr{S}) \geqq L(\boldsymbol{b} \mid \mathscr{S}), \quad \text{for all} \quad \boldsymbol{b} \in \mathscr{B} \tag{5.3}$$

Then, given a vector $\boldsymbol{V}^* = (v_1^*, \ldots, v_N^*) \in \mathscr{W}_N$ we may form weighted likelihood function (cf. (3.15)).

$$\mathscr{L}(\boldsymbol{b} \mid \mathscr{S}, \boldsymbol{V}^*) = \sum_{n=1}^{N} v_n^* \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) ; \quad \boldsymbol{b} \in \mathscr{B} \tag{5.4}$$

which is maximized by $\boldsymbol{b}^*(\mathscr{S}) \in \mathscr{B}$,

$$\boldsymbol{b}^*(\mathscr{S}) = \sum_{n=1}^{N} v_n^* \boldsymbol{\beta}(\boldsymbol{x}_n) \tag{5.5}$$

provided that the functions $f(\boldsymbol{x}_n \mid \boldsymbol{b})$; $n = 1, \ldots, N$ are continuous with respect to $\boldsymbol{b}$ at $\boldsymbol{b} = \boldsymbol{b}^*(\mathscr{S})$.

Proof. For any given weight vector $\boldsymbol{V}^* \in \mathscr{W}_N$ and an integer $k > N$ we can choose integers $k_1, k_2, \ldots, k_N$ with the property

$$\sum_{n=1}^{N} k_n = k ; \quad k_n \geqq 0 ; \quad \left| \frac{k_n}{k} - v_n^* \right| \leqq \frac{1}{k} ; \quad n = 1, 2, \ldots, N \tag{5.6}$$

Using the above integers we define a weight vector

$$\boldsymbol{V}^{(k)} = (v_1^{(k)}, \ldots, v_N^{(k)})^{\mathrm{T}} \in \mathscr{W}_N \quad v_n^{(k)} = \frac{k_n}{k} \tag{5.7}$$

and applying vectors from $\mathscr{S}$ repeatedly we construct an artificial sample

$$\mathscr{S}^{(k)} = \{\boldsymbol{y}_1(1), \ldots, \boldsymbol{y}_1(k_1), \ldots, \boldsymbol{y}_N(1), \ldots, \boldsymbol{y}_N(k_N)\} ; \tag{5.8}$$

$$\boldsymbol{y}_n(l) = \boldsymbol{x}_n \in \mathscr{S} ; \quad l = 1, \ldots, k_n ; \quad n = 1, 2, \ldots, N$$

Now, by assumption of the Theorem, the likelihood function

$$L(\boldsymbol{b} \mid \mathscr{S}^{(k)}) = \sum_{n=1}^{N} k_n \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) = k \sum_{n=1}^{N} v_n^{(k)} \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) \tag{5.9}$$

is maximized by $\hat{\boldsymbol{b}}(\mathscr{S}^{(k)}) \in \mathscr{B}$,

$$\hat{\boldsymbol{b}}(\mathscr{S}^{(k)}) = \frac{1}{k} \sum_{n=1}^{N} k_n \boldsymbol{\beta}(\boldsymbol{x}_n) = \sum_{n=1}^{N} v_n^{(k)} \boldsymbol{\beta}(\boldsymbol{x}_n) . \tag{5.10}$$

Thus, for any $k > N$ it follows

$$\sum_{n=1}^{N} v_n^{(k)} \ln f(\boldsymbol{x}_n \mid \hat{\boldsymbol{b}}(\mathscr{S}^{(k)})) \geqq \sum_{n=1}^{N} v_n^{(k)} \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) \quad \text{for all} \quad \boldsymbol{b} \in \mathscr{B} \tag{5.11}$$

It can be seen that

$$\lim_{k \to \infty} \boldsymbol{V}^{(k)} = \boldsymbol{V}^* \in \mathscr{W}_N ; \quad \lim_{k \to \infty} \hat{\boldsymbol{b}}(\mathscr{S}^{(k)}) = \boldsymbol{b}^*(\mathscr{S}) . \tag{5.12}$$

Letting $k \to \infty$ in (5.11) and considering the assumed continuity of $f(\boldsymbol{x}_n \mid \boldsymbol{b})$ at $\boldsymbol{b}^*(\mathscr{S})$ we obtain the inequality

$$\sum_{n=1}^{N} v_n^* \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}^*(\mathscr{S})) \geqq \sum_{n=1}^{N} v_n^* \ln f(\boldsymbol{x}_n \mid \boldsymbol{b}) ; \quad \boldsymbol{b} \in \mathscr{B} \tag{5.13}$$

which completes the proof. □

Remark 5.1. Note that m.-l. estimate in (5.2) may be supposed in a more general form

$$\hat{\boldsymbol{b}}(\mathscr{S}) = \psi \left(\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\beta}(\boldsymbol{x}_n) \right) \tag{5.14}$$

where $\psi(\cdot)$ is a continuous function.

6. APPLICATION TO SPECIAL TYPES OF MIXTURES

Applying the iterative procedure in a particular case we need to specify essentially only the general implicit relation (3.8) or eventually (3.12). In what follows we employ Theorem 5.1 or solve the implicit relation directly. The examples presented illustrate, that virtually any type of mixture may be identified.

a) Normal mixture

Weighted likelihood function corresponding to normal components

$$f(\boldsymbol{x} \mid \boldsymbol{c}_m, \boldsymbol{A}_m) = \frac{1}{\sqrt{((2\pi)^d \det \boldsymbol{A}_m)}} \exp \{ -\tfrac{1}{2} (\boldsymbol{x}_n - \boldsymbol{c}_m)^{\mathrm{T}} \boldsymbol{A}_m^{-1} (\boldsymbol{x}_n - \boldsymbol{c}_m) \} ; \tag{6.1}$$

$$(m = 1, \ldots, M)$$

is given (cf. (3.15), (3.8)) by

$$\mathscr{L}(\boldsymbol{c}_m, \boldsymbol{A}_m) = \sum_{n=1}^{N} v_n \ln \left[\frac{1}{\sqrt{((2\pi)^d \det \boldsymbol{A}_m)}} \exp \{ -\tfrac{1}{2} (\boldsymbol{x}_n - \boldsymbol{c}_m)^{\mathrm{T}} \boldsymbol{A}_m^{-1} (\boldsymbol{x}_n - \boldsymbol{c}_m) \} \right]; \tag{6.2}$$

$$v_n = p^{(t)}(m \mid \boldsymbol{x}_n) / \sum_{i=1}^{N} p^{(t)}(m \mid \boldsymbol{x}_i) .$$

If $\mathbf{A}_m$ is a positive definite matrix then, by Theorem 5.1, expression (6.2) as a function of $\mathbf{c}_m$ is maximized by

$$\mathbf{c}_m^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n)} \sum_{n=1}^{N} \mathbf{x}_n \, p^{(t)}(m \mid \mathbf{x}_n) . \tag{6.3}$$

Similarly for a fixed vector $\mathbf{c}_m \in \mathbb{R}_d$ the function $\mathscr{L}(\mathbf{c}_m, \mathbf{A}_m)$ is maximized by

$$\mathbf{A}_m^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n)} \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{c}_m)(\mathbf{x}_n - \mathbf{c}_m)^{\mathrm{T}} \, p^{(t)}(m \mid \mathbf{x}_n) . \tag{6.4}$$

Combining (6.3) and (6.4) we obtain the well known equations (2.3), (2.4).

Note that proving formula (6.4) we should consider possible occurrence of singular matrices in Eq. (5.2), (5.5) and (5.10). The proof of Theorem 5.1 may be easily modified if $\mathbf{A}_m^{(t+1)}$ obtained in (6.4) is a positive definite matrix. On the other side a singular matrix $\mathbf{A}_m^{(t+1)}$ corresponds to a singular point of likelihood function. In this case the computation must be stopped and repeated with other initial values. In approximation problems (cf. Remark 1.1) the singular component may be omitted.

In case of equal covariance matrices $\mathbf{A}_1 = \mathbf{A}_2 = \ldots = \mathbf{A}_M = \mathbf{A}$ we have to use Eq. (3.12) instead of (3.8). Consequently we obtain weighted likelihood function

$$\mathscr{L}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M, \mathbf{A}) = \tag{6.5}$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{p^{(t)}(m \mid \mathbf{x}_n)}{N} \ln \left[\frac{1}{\sqrt{((2\pi)^d \det \mathbf{A})}} \exp \left\{ -\tfrac{1}{2} (\mathbf{x}_n - \mathbf{c}_m)^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x}_n - \mathbf{c}_n) \right\} \right].$$

Again from Theorem 5.1 it follows that (6.5) as a function of $\mathbf{A}$ is maximized by

$$\mathbf{A}^{(t+1)} = \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{p^{(t)}(m \mid \mathbf{x}_n)}{N} (\mathbf{x}_n - \mathbf{c}_m)(\mathbf{x}_n - \mathbf{c}_m)^{\mathrm{T}} = \tag{6.6}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} - \sum_{m=1}^{M} w_m^{(t+1)} c_m c_m^{\mathrm{T}}$$

where the optimal vectors $\mathbf{c}_m$ may be substituted from (6.3).

b) Laplace mixture

Assuming components in form of Laplace densities

$$f(\mathbf{x} \mid \mathbf{a}_m, \mathbf{c}_m) = \prod_{\alpha=1}^{d} \left[\frac{1}{2a_{m\alpha}} \exp \left\{ - \frac{|x_\alpha - c_{m\alpha}|}{a_{m\alpha}} \right\} \right]; \quad \mathbf{x} \in \mathbb{R}_d; \; a_{m\alpha} > 0; \; \mathbf{a}_m, \mathbf{c}_m \in \mathbb{R}_d \tag{6.7}$$

we obtain weighted likelihood function

$$\mathscr{L}(\mathbf{c}_m, \mathbf{a}_m) = \sum_{n=1}^{N} v_n \ln \left[\prod_{\alpha=1}^{d} \left(\frac{1}{2a_{m\alpha}} \exp \left\{ \frac{-|x_{n\alpha} - c_{m\alpha}|}{a_{m\alpha}} \right\} \right) \right] = \tag{6.8}$$

$$= \sum_{\alpha=1} \left\{ - \ln 2a_{m\alpha} - \sum_{n=1}^{N} v_n \frac{|x_{n\alpha} - c_{m\alpha}|}{a_{m\alpha}} \right\}; \quad v_n = \frac{p^{(t)}(m \mid \mathbf{x}_n)}{\sum_{i=1}^{N} p^{(t)}(m \mid \mathbf{x}_i)}$$

which is maximized by

$$(6.9) \quad a_{m\alpha}^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n)} \sum_{n=1}^{N} |x_{n\alpha} - c_{m\alpha}| \, p^{(t)}(m \mid \mathbf{x}_n); \quad \alpha = 1, 2, \ldots, d;$$

for any fixed vector $\mathbf{c}_m \in \mathcal{R}_d$. Further if

$$(6.10) \qquad x_{i_1\alpha} \leqq x_{i_2\alpha} \leqq \ldots \leqq x_{i_N\alpha}$$

is the order of numbers $x_{1\alpha}, x_{2\alpha}, \ldots, x_{N\alpha}$ then the optimal vector $\mathbf{c}_m^{(t+1)}$ may be defined by inequalities

$$(6.11) \quad c_{m\alpha}^{(t+1)} = x_{i_k\alpha}; \quad k : \sum_{j=1}^{k-1} p^{(t)}(n \mid \mathbf{x}_{i_j}) < \tfrac{1}{2} \sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n) \leqq \sum_{j=1}^{k} p^{(t)}(m \mid \mathbf{x}_{i_j}).$$

c) Uniform mixture

We define multivariate uniform probability density by Eq.

$$(6.12) \quad f(\mathbf{x} \mid \mathbf{a}_m, \mathbf{b}_m) = \prod_{\alpha=1}^{d} \left[\frac{\varphi(x_\alpha \mid a_{m\alpha}, b_{m\alpha})}{(b_{m\alpha} - a_{m\alpha})} \right]; \quad \begin{matrix} \mathbf{x} \in \mathcal{R}_d \\ b_{m\alpha} > a_{m\alpha}; \quad \mathbf{a}_m, \mathbf{b}_m \in \mathcal{R}_d \end{matrix}$$

where $\varphi(\xi \mid a_{m\alpha}, b_{m\alpha})$ is characteristic function of the interval $\langle a_{m\alpha}, b_{m\alpha} \rangle$

$$(6.13) \qquad \varphi(\xi \mid a_{m\alpha}, b_{m\alpha}) = \begin{cases} 1; & \xi \in \langle a_{m\alpha}, b_{m\alpha} \rangle \\ 0; & \xi \notin \langle a_{m\alpha}, b_{m\alpha} \rangle \end{cases}$$

The corresponding weighted likelihood function

$$\mathscr{L}(\mathbf{a}_m, \mathbf{b}_n) = \sum_{n=1}^{N} v_n \ln \left[\prod_{\alpha=1}^{d} \frac{\varphi(x_{n\alpha} \mid a_{m\alpha}, b_{m\alpha})}{(b_{m\alpha} - a_{m\alpha})} \right] =$$

$$= \sum_{\alpha=1}^{d} \{ - \ln(b_{m\alpha} - a_{m\alpha}) + \sum_{n=1}^{N} v_n \ln \varphi(x_{n\alpha} \mid a_{m\alpha}, b_{m\alpha}) \}; \quad v_n = \frac{p^{(t)}(m \mid \mathbf{x}_n)}{\sum_{i=1}^{N} p^{(t)}(m \mid \mathbf{x}_i)}$$

is maximized by

$$(6.15) \qquad a_{m\alpha}^{(t+1)} = \min_{n : p^{(t)}(m|\mathbf{x}_n) > 0} \{x_{n\alpha}\}; \quad b_{m\alpha}^{(t+1)} = \max_{n : p^{(t)}(m|\mathbf{x}_n) > 0} \{x_{n\alpha}\}.$$

Note that in case of uniform mixture the property (3.11) undesirably increases the meaning of initial parameters. Roughly speaking the solution obtained will not differ very much from the initial parameters.

d) Bernoulli mixture

Consider a discrete mixture the components of which are multivariate Bernoulli distributions

$$(6.16) \qquad f(\mathbf{x} \mid \boldsymbol{\theta}_m) = \prod_{\alpha=1}^{d} \theta_{m\alpha}^{x_\alpha}(1 - \theta_{m\alpha})^{1-x_\alpha} ; \quad x_\alpha \in \{0, 1\} ; \quad 0 \leqq \theta_{m\alpha} \leqq 1$$

defined on the set of d-dimensional binary vectors. The corresponding weighted likelihood function is given by

$$(6.17) \qquad \mathscr{L}(\boldsymbol{\theta}_m) = \sum_{n=1}^{N} v_n \ln \Big[\prod_{\alpha=1}^{d} \theta_{m\alpha}^{x_{n\alpha}}(1 - \theta_{m\alpha})^{1-x_{n\alpha}} \Big] =$$

$$= \sum_{\alpha=1}^{d} \Big\{ \Big[\sum_{n=1}^{N} v_n x_{n\alpha} \Big] \ln \theta_{m\alpha} + \Big[1 - \sum_{n=1}^{N} v_n x_{n\alpha} \Big] \ln (1 - \theta_{m\alpha}) \Big\} ; \quad v_n = \frac{p^{(t)}(m \mid \mathbf{x}_n)}{\sum_{i=1}^{N} p^{(t)}(m \mid \mathbf{x}_i)} .$$

Using Theorem 5.1 or directly Lemma 4.1 we obtain the following explicit form of the relation (3.8)

$$(6.18) \qquad \theta_{m\alpha}^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}_n)} \sum_{n=1}^{N} x_{n\alpha} \cdot p^{(t)}(m \mid \mathbf{x}_n) ; \; \alpha = 1, 2, \ldots, d .$$

7. CONCLUSION

General iterative procedure considered in the present paper proved to be a feasible method for obtaining m.-l. estimates for finite mixtures of various types. Computational properties of this procedure have been tested practically by several authors with satisfactory results (cf. [28], [61], [47], [21]). In this connection there is a frequently discussed problem of multiple solutions. Analyzing a particular set of data one can find all the local maxima of likelihood function by repeating the computation from enough different starting points. Usually the over-all maximum is used to determine the best solution ([4], [61]) but other rules may be preferable when the sample size is small (cf. [3]). Note that related problem of the proper choice of initial values need not be as crucial as sometimes supposed (cf. [14], [61]) since the iterative process converges to a local maximum nearly always.

Recall that the number of components M is not estimated by the iterative procedure. In the practically important approximation problems (cf. Remark 1.1) there is a possibility to delete "superfluous" components in the course of computation. A question arises if appropriate tests could be developed for this purpose possibly in a way suggested by Theorem 5.1.

ACKNOWLEDGEMENT

Special thanks are due to my scientific supervisor doc. Dr. Ing. Jiří Beneš, DrSc. for his helpful commets and kind support during preparation of this paper.

(Received October 30, 1981.)

REFERENCES

[1] С. А. Айвазян, З. И. Бежаева, О. В. Староверов: Классификация многомерных наблюдений (Classification of Multivariate Observations). Статистика, Москва 1974.

[2] Н. Н. Апраушева: Алгоритм расщепления смеси нормальных классов (Algorithm for resolution of a mixture of normal classes). Сб. Программы и алгоритмы (1976), 68.

[3] Н. Н. Апраушева: Определение числа классов в задачах классификации I (Determination of the number of classes in classification problems I). Известия АН СССР - Тех. кибернетика (1981), 3, 71—77.

[4] J. Behboodian: On a mixture of normal distributions. Biometrika *57* (1970), 1, 215—217.

[5] C. G. Bhattacharya: A simple method of resolution of a distribution into Gaussian components. Biometrics *23* (1967), 115—137.

[6] W. R. Blischke: Moment estimators for the parameters of a mixture of two binomial distributions. Ann. Math. Statist. *33* (1962), 2, 444—454.

[7] W. R. Blischke: Mixtures of distributions. In: Classical and Contagious Discrete Distributions (G. P. Patil, ed.), Pergamon Press, New York 1963.

[8] W. R. Blischke: Estimating the parameters of mixtures of binomial distributions. J. Amer. Statist. Assoc. *59* (1964), 306, 510—528.

[9] C. Bürrau: The half-invariants of the sum of two typical laws of errors with an application to the problem of dissecting a frequency curve into components. Scand. Actuar. J. *17* (1934), 1, 1—5.

[10] C. V. L. Charlier: Researches into the theory of probability. Meddelanden fran Lunds Astron. Observ. (1906) Sec. 2, Bd. 1.

[11] A. C. Cohen: Discussion of "Estimation of parameters for a mixture of normal distributions" by Victor Hasselblad. Technometrics *8* (1966), 3, 445—446.

[12] A. C. Cohen: Estimation in mixtures of two normal distributions. Technometrics *9* (1967), 15—28.

[13] P. W. Cooper: Some topics on nonsupervised adaptive detection for multivariate normal distributions. In: Computer and Information Sciences — II. (J. T. Tou, ed.), Academic Press, New York 1967.

[14] N. E. Day: Estimating the components of a mixture of normal distributions. Biometrika *56* (1969), 463—474.

[15] N. P. Dick, D. C. Bowden: Maximum likelihood estimation for mixtures of two normal distributions. Biometrics *29* (1973), 4, 781—790.

[16] G. Doetsch: Zerlegung einer Funktion in Gaussche Fehlerkurven und zeitliche Zurückverfolgung eines Temperaturzustandes. Math. Z. *41* (1936), 283—318.

[17] R. O. Duda, P. E. Hart: Pattern Classification and Scene Analysis. John Wiley, New York—London 1973.

[18] E. B. Fowlkes: Some methods for studying the mixture of two normal (lognormal) distributions. J. Amer. Statist. Assoc. *74* (1979), 367, 561—575.

[19] J. G. Fryer, C. A. Roberston: A comparison of some methods of estimating mixed normal distributions. Biometrika *59* (1972), 639—648.

[20] N. T. Gridgeman: A comparison of two methods of analysis of normal distributions. Technometrics *12* (1970), 4, 832—833.

[21] J. Grim: Metody shlukové analýzy a jejich využití při zpětnovazebním řízení velkých systémů (Methods of cluster analysis and their application for feedback control of large systems). Dissertation, Institute of Information Theory and Automation, Prague 1979.

[22] J. Grim: An algorithm for maximizing a finite sum of positive functions and its application to cluster analysis. Problems of Control and Information Theory *10* (1981), 6, 427—437.

[23] E. J. Gumbel: La dissection d'une repartition. Annales de l'Université de Lyon *3* (1939), 39—51.

[24] A. K. Gupta, T. Miyawaki: On uniform mixture model. Biometrical J. *20* (1978), 631—637.

[25] L. F. Guseman, J. R. Walton: Methods for estimating proportions of convex combinations of normals using linear feature selection. Comm. Statist. A — Theory Methods *A7* (1978), 1439—1450.

[26] V. Hasselblad: Estimation of parameters for a mixture of normal distributions. Technometrics *8* (1966), 431—444.

[27] V. Hasselblad: Finite mixtures of distributions from the exponential family. Ph. D. Dissertation University of California, Los Angeles 1967.

[28] V. Hasselblad: Estimation of finite mixtures of distributions from the exponential family. J. Amer. Statist. Assoc. *64* (1969), 328, 1459—1471.

[29] B. M. Hill: Information for estimating the proportions in mixtures of exponential and normal distributions. J. Amer. Statist. Assoc. *58* (1963), 918—932.

[30] D. W. Hosmer: A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of samples. Biometrics *29* (1973), 761—770.

[31] D. W. Hosmer: A use of mixtures of two normal distributions in a classification problem. J. Statist. Comput. Simulation *6* (1978), 384, 281—294.

[32] D. W. Hosmer, N. P. Dick: Information and mixtures of two normal distributions. J. Statist. Comput. Simulation *6* (1977), 137—148.

[33] О. К. Исаенко, К.Ю. Урбах: Разделение смесей распределений вероятностей на их составляющие (Decomposition of mixtures of probability distributions into their components). Теория вероятностей и математическая статистика, теор. кибернетика, том 13, 37—58, ВИНИТИ, Москва 1976.

[34] I. R. James: Estimation of the mixing proportion in a mixture of two normal distributions from simple rapid measurements. Biometrics *34* (1978), 2, 265—275.

[35] E. John: Bayesian estimation of mixture distributions. Ann. Math. Statist. *39* (1968), 4, 1289—1302.

[36] B. K. Kale: On the solution of likelihood equations by iteration processes: The multiparametric case. Biometrika *49* (1962), 479—486.

[37] R. Kanno: Estimation of parameters for a mixture of two normal distributions. Rep. Statist. Appl. Res. JUSE *22* (1975), 4, 1—15.

[38] S. Kullback: An information-theoretic derivation of certain limit relations for a stationary Markov Chain. SIAM J. Control *4* (1966), 3, 454—459.

[39] S. Kullback: Information Theory and Statistics. Wiley, New York—Dover 1968.

[40] P. D. M. Macdonald: Estimation of finite distribution mixtures. In: Applied Statistics (R. P. Gupta, ed.), North-Holland 1975.

[41] P. Medgyessy: Decomposition of Superpositions of Density Functions and Discrete Distributions. Akadémiai Kiadó, Budapest 1977.

[42] G. Meeden: Bayes estimation of the mixing distributions, the discrete case. Ann. Math. Statist. *43* (1972), 6, 1993—1999.

[43] W. Molenaar: Survey of estimation methods for a mixture of two normal distributions. Statist. Neerlandica *19* (1965), 4, 249—265.

[44] G. D. Murray, D. M. Titterington: Estimation problems with data from a mixture. Appl. Statist. *27* (1978), 3, 325—334.
[45] K. Pearson: Contributions to the mathematical theory of evolution 1: Dissection of frequency curves. Philos. Trans. Roy. Soc. London Ser. *A 185* (1894), 71—110.
[46] B. C. Peters, W. A. Coberly: The numerical evaluation of the maximum-likelihood estimate of mixture proportions. Comm. Statist. A — Theory Methods *A5* (1976), 12, 1127—1135.
[47] B. C. Peters, H. F. Walker: An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. SIAM J. Appl. Math. *35* (1978), 2, 362—378.
[48] B. C. Peters, H. F. Walker: The numerical evaluation of the maximum-likelihood estimate of a subset of mixture proportions. SIAM J. Appl. Math. *35* (1978), 3, 447—452.
[49] J. G. Postaire, C. P. A. Vasseur: An approximate solution to normal mixture identification with application to unsupervised pattern classification. IEEE Trans. on Pattern Analysis & Machine Intelligence *PAMI-3* (1981), 2, 163—179.
[50] R. E. Quandt, J. B. Ramsey: Estimating mixtures of normal distributions and switching regressions. J. Amer. Statist. Assoc. *73* (1978), 364, 730—752.
[51] C. R. Rao: Advanced Statistical Methods in Biometric Research, John Wiley and Sons, New York 1952.
[52] P. R. Rider: The method of moments applied to a mixture of two exponential distributions. Ann. Math. Statist. *32* (1961), 1, 143—147.
[53] W. Schilling: A frequency distribution represented as the sum of two Poisson distributions. J. Amer. Statist. Assoc. *42* (1947), 407—424.
[54] D. F. Stanat: Unsupervised learning of mixtures of probability functions. In: Pattern Recognition (L. Kanal, ed.), Thompson Book Co., Washington D. C. 1968, 357—389.
[55] B. Strömgren: Tables and diagrams for dissecting a frequency curve into components by the half-invariant method. Scand. Actuar. J. *17* (1934), 1, 7—54.
[56] М. И. Шлезингер: Взаимосвязь обучения и самообучения в распознавании образов (Relation between learning and self-learning in pattern recognition). Кибернетика (Киев) (1968), 2, 81—88.
[57] W. Y. Tan, W. C. Chang: Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. J. Amer. Statist. Assoc. *67* (1972), 339, 702—708.
[58] H. F. Walker: Estimating the proportions of two populations in a mixture using linear maps. Comm. Statist. A — Theory Methods *A9* (1980), 8, 837—849.
[59] J. H. Wolfe: A computer program for the maximum likelihood analysis of types. (Technical Bulletin 65—15), U.S. Naval Personnel Research Activity, San Diego 1965.
[60] J. H. Wolfe: NORMIX: computational methods for estimating the parameters of multivariate normal mixtures of distributions. (Research Memorandum SRM 68—2), U.S. Naval Personnel Research Activity, San Diego 1967.
[61] J. H. Wolfe: Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research *5* (1970), July, 329—350.
[62] S. J. Yakowitz: Unsupervised learning and the identification of finite mixtures. IEEE Trans. Inform. Theory *IT-16* (1970), 5, 330—338.
[63] T. Y. Young, G. Coraluppi: Stochastic estimation of a mixture of normal density functions using an information criterion. IEEE Trans. Inform. Theory *IT-16* (1970), 258—263.

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.