SIAM J. APPL. MATH.
Vol. 35, No. 2, September 1978

0036-1399/78/3502-0017 $01.00/0

AN ITERATIVE PROCEDURE FOR OBTAINING MAXIMUM-LIKELIHOOD ESTIMATES OF THE PARAMETERS FOR A MIXTURE OF NORMAL DISTRIBUTIONS*

B. CHARLES PETERS, JR.† AND HOMER F. WALKER‡

Abstract. This paper addresses the problem of obtaining numerically maximum-likelihood estimates of the parameters for a mixture of normal distributions. In recent literature, a certain successive-approximations procedure, based on the likelihood equations, was shown empirically to be effective in numerically approximating such maximum-likelihood estimates; however, the reliability of this procedure was not established theoretically. Here, we introduce a general iterative procedure, of the generalized steepest-ascent (deflected-gradient) type, which is just the procedure known in the literature when the step-size is taken to be 1. We show that, with probability 1 as the sample size grows large, this procedure converges locally to the strongly consistent maximum-likelihood estimate whenever the step-size lies between 0 and 2. We also show that the step-size which yields optimal local convergence rates for large samples is determined in a sense by the "separation" of the component normal densities and is bounded below by a number between 1 and 2.

1. Introduction. Let x be an n-dimensional random variable whose density function p is a convex combination of normal densities, i.e.,

$$p(x)=\sum_{i=1}^{m}\alpha_i^0 p_i(x) \quad \text{for } x\in\mathbb{R}^n,$$

where

$$\alpha_i^0>0, \qquad \sum_{i=1}^{m}\alpha_i^0=1,$$

and

$$p_i(x)=\frac{1}{(2\pi)^{n/2}|\Sigma_i^0|^{1/2}}\exp\left[-\tfrac{1}{2}(x-\mu_i^0)^T\,\Sigma_i^{0^{-1}}(x-\mu_i^0)\right].$$

If $\{x_k\}$, $k=1,\cdots,N$, $\subseteq\mathbb{R}^n$ is an independent sample of observations on x, then a *maximum-likelihood estimate* of the parameters $\{\alpha_i^0,\mu_i^0,\Sigma_i^0\}_{i=1,\cdots,m}$ is a choice of parameters $\{\alpha_i,\mu_i,\Sigma_i\}_{i=1,\cdots,m}$ which locally maximizes the *log-likelihood function*

$$L=\sum_{k=1}^{N}\log p(x_k),$$

in which p is evaluated with the true parameters $\{\alpha_i^0,\mu_i^0,\Sigma_i^0\}_{i,\cdots,m}$ replaced by the estimate $\{\alpha_i,\mu_i,\Sigma_i\}_{i=1,\cdots,m}$. (In the following, it is usually clear from the context which parameters are used in evaluating the density functions p_i and p. Therefore, these parameters are explicitly pointed out only when some ambiguity exists.) We admit local maxima of L as maximum-likelihood estimates in order to avoid difficulties presented by the fact that L has no global maximum. It is observed below that this creates no problems when one is concerned with consistent maximum-likelihood estimates.

* Received by the editors June 26, 1976, and in final revised form October 28, 1977.

† Department of Mathematics, University of Houston, Houston Texas 77004. This author was a National Research Council Research Associate at Lyndon B. Johnson Space Center during the preparation of this work.

‡ Department of Mathematics, University of Houston, Houston, Texas 77004. This work was supported in part by The National Aeronautics and Space Administration under Contract JSC-NAS-9-12777.

Clearly, L is a differentiable function of the parameters to be estimated. Equating to zero the partial derivatives of L with respect to these parameters, one obtains, after a straightforward calculation, the following necessary condition for a maximum-likelihood estimate:

$$\alpha_i = \frac{\alpha_i}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}, \tag{1a}$$

$$\mu_i = \left\{\frac{1}{N} \sum_{k=1}^{N} x_k \frac{p_i(x_k)}{p(x_k)}\right\} \Big/ \left\{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right\}, \qquad i = 1, \cdots, m, \tag{1b}$$

$$\Sigma_i = \left\{\frac{1}{N} \sum_{k=1}^{N} (x_k - \mu_i)(x_k - \mu_i)^T \frac{p_i(x_k)}{p(x_k)}\right\} \Big/ \left\{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right\}. \tag{1c}$$

These are known as the *likelihood equations.* It follows from (1a) that the denominators in (1b) and (1c) are equal to 1 at a maximum-likelihood estimate and, hence, their presence appears somewhat superfluous. However, these denominators play a crucial role in establishing the convergence of the iterative procedure described below.

A number of authors have investigated solutions of the likelihood equations and the consistency of maximum-likelihood estimates in general. (See, for example, Cramér [3], Huzurbazar [8], Wald [13], Chanda [2], Aitchison and Silvey [1], and the discussion in Zacks [15].) For completeness we have included in Appendix A a brief proof of a multidimensional analogue of Cramér's result to the effect that, loosely speaking, there is a unique solution of the likelihood equations which tends with probability 1 to the true parameters as the sample size N approaches infinity. Furthermore, this solution is a maximum-likelihood estimate, indeed the unique strongly consistent maximum-likelihood estimate. More precisely, if certain regularity conditions on the derivatives of the density function with respect of the parameters are satisfied and the information matrix is positive-definite, then with probability 1, for any sufficiently small neighborhood of the true parameters, there is for sufficiently large N a unique solution of the likelihood equations in that neighborhood and this solution is a maximum-likelihood estimate. This note is addressed to the problem of determining this strongly consistent maximum-likelihood estimate by successive approximations.

The likelihood equations, as written, suggest the following iterative procedure for obtaining a solution: Beginning with some set of starting values, obtain successive approximations to a solution by inserting the preceding approximations in the expressions on the right-hand sides of (1a), (1b), and (1c). This scheme is attractive for its relative ease of implementation, and we discuss below the findings of several authors concerning its use in obtaining maximum-likelihood estimates. For a discussion of other methods of determining maximum-likelihood estimates, see Kale [9] and Wolfe [14] as well as the authors given below.

Empirical studies of Day [4], Duda and Hart [5], and Hasselblad [6] suggest that this scheme is convergent and that convergence is particularly fast when the component normal densities in p are "widely separated" in a certain sense. Unfortunately, the likelihood equations have many solutions in general, and the iterates may converge to solutions, including "singular solutions" (see [5]), which are not the strongly consistent maximum-likelihood estimate if care is not taken in the choice of starting values. No theoretical evidence of convergence is given in [4], [5], or [6]. Peters and Coberly [12] have proved that, if all of the parameters μ_i and Σ_i are held fixed, then the iterative procedure suggested by the equation (1a) alone converges locally to a maximum-likelihood estimate of the parameters $\alpha_i, i = 1, \cdots, m$. (An iterative procedure is said to *converge locally* to a limit if the iterates converge to that limit whenever the starting values are sufficiently near that limit.) They also report on numerical studies in which the computational feasibility of this procedure is demonstrated.

In the following, we present a general iterative procedure for determining the strongly consistent maximum-likelihood estimate, of which the above procedure is a special case. Indeed, our procedure is a generalized steepest-ascent (deflected-gradient) method, and the above procedure is obtained when the step-size is taken to be 1. We show that, with probability 1 as the sample size grows large, this procedure converges locally to the strongly consistent maximum-likelihood estimate whenever the step-size is between 0 and 2. Furthermore, the value of the step-size which yields optimal local convergence rates is bounded from below by a number which always lies between 1 and 2. In fact, this optimal step-size lies near 1 if the component populations are "widely separated" in a certain sense and cannot be much smaller than 2 if two or more of the component populations have nearly identical means and covariance matrices. We also prove that, if the covariance matrices Σ_i are held fixed, then the restricted iterative procedure for the parameters α_i and μ_i has these local convergence properties with probability 1 whenever the sample size is at least $m(n+1)$. We conclude by comparing this procedure to other numerical methods for determining maximum-likelihood estimates.

2. The general iterative procedure. In order to minimize notational difficulties, we introduce several vector spaces and give useful representations of their elements. For each i, $1 \leqq i \leqq m$, α_i, μ_i, and Σ_i are elements of the vector spaces $\mathbb{R}^1$, $\mathbb{R}^n$, and the set of all real, symmetric $n \times n$ matrices, respectively. We denote by $\mathscr{A}$, $\mathscr{M}$, and $\mathscr{S}$ the respective m-fold direct sums of these spaces with themselves, and we represent elements of $\mathscr{A}$, $\mathscr{M}$, and $\mathscr{S}$ as columns

$$\bar{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} \in \mathscr{A}, \quad \bar{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} \in \mathscr{M}, \quad \bar{\Sigma} = \begin{pmatrix} \Sigma_1 \\ \vdots \\ \Sigma_m \end{pmatrix} \in \mathscr{S}.$$

It will be convenient to adopt the following notational equivalence for elements of the direct sum $\mathscr{A} \oplus \mathscr{M} \oplus \mathscr{S}$.

$$\Theta = \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \\ \mu_1 \\ \vdots \\ \mu_m \\ \Sigma_1 \\ \vdots \\ \Sigma_m \end{pmatrix}$$

If, for $i=1,\cdots,m$ and $\Theta\in\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$, we denote

$$A_i(\Theta)=\frac{\alpha_i}{N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)},$$

$$M_i(\Theta)=\frac{1}{N}\sum_{k=1}^{N}x_k\frac{p_i(x_k)}{p(x_k)}\Big/\left\{\frac{1}{N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)}\right\},$$

$$S_i(\Theta)=\frac{1}{N}\sum_{k=1}^{N}(x_k-\mu_i)(x_k-\mu_i)^T\frac{p_i(x_k)}{p(x_k)}\Big/\left\{\frac{1}{N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)}\right\},$$

then the likelihood equations can be written as

$$\Theta=\begin{pmatrix}A(\Theta)\\ M(\Theta)\\ S(\Theta)\end{pmatrix}, \tag{2}$$

where

$$A(\Theta)=\begin{pmatrix}A_1(\Theta)\\ \vdots\\ A_m(\Theta)\end{pmatrix},\quad M(\Theta)=\begin{pmatrix}M_1(\Theta)\\ \vdots\\ M_m(\Theta)\end{pmatrix},\quad S(\Theta)=\begin{pmatrix}S_1(\Theta)\\ \vdots\\ S_m(\Theta)\end{pmatrix}.$$

One can write (2) more generally as

$$\Theta=\Phi_\varepsilon(\Theta)\equiv(1-\varepsilon)\Theta+\varepsilon\begin{pmatrix}A(\Theta)\\ M(\Theta)\\ S(\Theta)\end{pmatrix} \tag{3}$$

for any value of ε. Of course, (3) becomes (2) when $\varepsilon=1$.

The following iterative procedure is suggested by (3) for obtaining a solution of the likelihood equations: Beginning with some starting value $\Theta^{(1)}$, define successive iterates inductively by

$$\Theta^{(k+1)}=\Phi_\varepsilon(\Theta^{(k)}) \tag{4}$$

for $k=1,2,3,\cdots$. This is the general iterative procedure with which this note is concerned. Clearly, this procedure becomes the procedure given in the Introduction when $\varepsilon=1$.

In the next section, we show that if $0<\varepsilon<2$, then, with probability 1 as N approach infinity, this procedure converges locally to the strongly consistent maximum-likelihood estimate. This is done by showing that, with probability 1 as N approaches infinity, the operator Φ_ε is *locally contractive* (in a suitable vector norm) near that estimate, provided $0<\varepsilon<2$. In saying that Φ_ε is locally contractive near a point $\Theta\in\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$, we mean that there is a vector norm $\|\cdot\|$ on $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$, and a number λ, $0\leqq\lambda<1$, such that

$$\|\Phi_\varepsilon(\Theta')-\Theta\|\leqq\lambda\|\Theta'-\Theta\|, \tag{5}$$

whenever Θ' lies sufficiently near Θ.

3. The local contractibility and convergence results. We now establish the following

THEOREM. *With probability* 1 *as N approaches infinity,* Φ_ε *is a locally contractive operator* (*in some norm on* $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$) *near the strongly consistent maximum-likelihood estimate whenever* $0<\varepsilon<2$.

Our main result, given by the following corollary, is an immediate consequence of this theorem.

COROLLARY. *With probability* 1 *as N approaches infinity, the iterative procedure* (4) *converges locally to the strongly consistent maximum-likelihood estimate whenever* $0<\varepsilon<2$.

Throughout the remainder of this paper, the symbol "∇" denotes the Fréchet derivative of a vector-valued function of a vector variable. When ambiguity exists, the specific vector variable of differentiation appears as a subscript of this symbol. For questions concerning the definition and properties of Fréchet derivatives, see Luenberger [11].

Proof of the theorem. Let

$$\Theta=\begin{pmatrix}\bar{\alpha}\\ \bar{\mu}\\ \bar{\Sigma}\end{pmatrix}=\begin{pmatrix}\alpha_1\\ \cdot\\ \cdot\\ \cdot\\ \alpha_m\\ \mu_1\\ \cdot\\ \cdot\\ \cdot\\ \mu_m\\ \Sigma_1\\ \cdot\\ \cdot\\ \cdot\\ \Sigma_m\end{pmatrix}$$

be the strongly consistent maximum-likelihood estimate. We assume that $\alpha_i\neq 0$, $i=1,\cdots,m$. (As N tends to infinity, the probability is 1 that this is the case.) It must be shown that, with probability 1 as N approaches infinity, an inequality of the form (5) holds whenever $0<\varepsilon<2$.

For any norm on $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$, one can write

$$\Phi_\varepsilon(\Theta')-\Theta=\nabla\Phi_\varepsilon(\Theta)[\Theta'-\Theta]+O(\|\Theta'-\Theta\|^2).$$

Consequently, the theorem will be proved if it can be shown that, for $0<\varepsilon<2$, $\nabla\Phi_\varepsilon(\Theta)$ converges with probability 1 to an operator which has operator norm less than 1 with respect to a suitable vector norm on $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$.

One can write $\nabla\Phi_\varepsilon$ as $(1-\varepsilon)I$ plus a matrix of Fréchet derivatives:

$$\nabla\Phi_\varepsilon=(1-\varepsilon)I+\varepsilon\begin{pmatrix}\nabla_{\bar{\alpha}}A & \nabla_{\bar{\mu}}A & \nabla_{\bar{\Sigma}}A\\ \nabla_{\bar{\alpha}}M & \nabla_{\bar{\mu}}M & \nabla_{\bar{\Sigma}}M\\ \nabla_{\bar{\alpha}}S & \nabla_{\bar{\mu}}S & \nabla_{\bar{\Sigma}}S\end{pmatrix}.$$

This is consistent with our representation of elements of $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$ as columns.

The entries of the above matrix can themselves be represented as matrices of Fréchet derivatives. For $i=1,\cdots,m$, we introduce inner products $\langle x,y\rangle_i'=x^T(\alpha_i\Sigma_i^{-1})y$ on $\mathbb{R}^n$ and $\langle A,B\rangle_i''=\operatorname{tr}\{A((\alpha_i/2)\Sigma_i^{-1})B^T\}$ on the space of real, symmetric $n\times n$ matrices, and we define $\beta_i(x)=p_i(x)/p(x)$, $\gamma_i(x)=(x-\mu_i)$, and $\delta_i(x)=[\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T-I]$. By the notation $\langle x,\cdot\rangle_i'$ we mean the operator which when evaluated at $y\in\mathbb{R}^n$, is $\langle x,y\rangle_i'$. Similarly, the notation $\langle A,\cdot\rangle_i''$ means the operator which when evaluated at a real, symmetric $n\times n$ matrix B is $\langle A,B\rangle_i''$.

After a straightforward but extremely tedious calculation, one obtains with the aid of equations (1) that at the maximum likelihood estimate

$$\nabla_{\bar{\alpha}} A(\Theta) = I - (\operatorname{diag} \alpha_i) \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k) \\ \vdots \\ \beta_m(x_k) \end{pmatrix} \begin{pmatrix} \beta_1(x_k) \\ \vdots \\ \beta_m(x_k) \end{pmatrix}^T \right\},$$

$$\nabla_{\bar{\mu}} A(\Theta) = -(\operatorname{diag} \alpha_i) \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k) \\ \vdots \\ \beta_m(x_k) \end{pmatrix} \begin{pmatrix} \langle \beta_1(x_k)\gamma_1(x_k), \cdot \rangle_1' \\ \vdots \\ \langle \beta_m(x_k)\gamma_m(x_k), \cdot \rangle_m' \end{pmatrix}^T \right\},$$

$$\nabla_{\bar{\Sigma}} A(\Theta) = -(\operatorname{diag} \alpha_i) \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k) \\ \vdots \\ \beta_m(x_k) \end{pmatrix} \begin{pmatrix} \langle \beta_1(x_k)\delta_1(x_k), \cdot \rangle_1'' \\ \vdots \\ \langle \beta_m(x_k)\delta_m(x_k), \cdot \rangle_m'' \end{pmatrix}^T \right\},$$

$$\nabla_{\bar{\alpha}} M(\Theta) = -\left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k)\gamma_1(x_k) \\ \vdots \\ \beta_m(x_k)\gamma_m(x_k) \end{pmatrix} \begin{pmatrix} \beta_1(x_k) \\ \vdots \\ \beta_m(x_k) \end{pmatrix}^T \right\},$$

$$\nabla_{\bar{\mu}} M(\Theta) = I - \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k)\gamma_1(x_k) \\ \vdots \\ \beta_m(x_k)\gamma_m(x_k) \end{pmatrix} \begin{pmatrix} \langle \beta_1(x_k)\gamma_1(x_k), \cdot \rangle_1' \\ \vdots \\ \langle \beta_m(x_k)\gamma_m(x_k), \cdot \rangle_m' \end{pmatrix}^T \right\},$$

$$\begin{aligned} \nabla_{\bar{\Sigma}} M(\Theta) = {} & \left(\operatorname{diag} \frac{1}{\alpha_i N} \sum_{k=1}^{N} \beta_i(x_k)\gamma_i(x_k) \langle \delta_i(x_k), \cdot \rangle_i'' \right) \\ & - \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k)\gamma_1(x_k) \\ \vdots \\ \beta_m(x_k)\gamma_m(x_k) \end{pmatrix} \begin{pmatrix} \langle \beta_1(x_k)\delta_1(x_k), \cdot \rangle_1'' \\ \vdots \\ \langle \beta_m(x_k)\delta_m(x_k), \cdot \rangle_m'' \end{pmatrix}^T \right\}, \end{aligned}$$

$$\nabla_{\bar{\alpha}} S(\Theta) = -(\operatorname{diag} \Sigma_i) \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k)\delta_1(x_k) \\ \vdots \\ \beta_m(x_k)\delta_m(x_k) \end{pmatrix} \begin{pmatrix} \beta_1(x_k) \\ \vdots \\ \beta_m(x_k) \end{pmatrix}^T \right\},$$

$$\begin{aligned} \nabla_{\bar{\mu}} S(\Theta) = {} & \left(\operatorname{diag} \Sigma_i \frac{1}{\alpha_i N} \sum_{k=1}^{N} \beta_i(x_k)\delta_i(x_k) \langle \gamma_i(x_k), \cdot \rangle_i' \right) \\ & -(\operatorname{diag} \Sigma_i) \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k)\delta_1(x_k) \\ \vdots \\ \beta_m(x_k)\delta_m(x_k) \end{pmatrix} \begin{pmatrix} \langle \beta_1(x_k)\gamma_1(x_k), \cdot \rangle_1' \\ \vdots \\ \langle \beta_m(x_k)\gamma_m(x_k), \cdot \rangle_m' \end{pmatrix}^T \right\}, \end{aligned}$$

$$\nabla_{\bar{\bar{\Sigma}}} S(\Theta) = \left(\operatorname{diag} \Sigma_i \frac{1}{\alpha_i N} \sum_{k=1}^{N} \beta_i(x_k)\delta_i(x_k)\langle \delta_i(x_k), \cdot\rangle_i''\right)$$

$$-(\operatorname{diag} \Sigma_i) \left\{ \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} \beta_1(x_k)\delta_1(x_k) \\ \vdots \\ \beta_m(x_k)\delta_m(x_k) \end{pmatrix} \begin{pmatrix} \langle \beta_1(x_k)\delta_1(x_k), \cdot\rangle_1'' \\ \vdots \\ \langle \beta_m(x_k)\delta_m(x_k), \cdot\rangle_m'' \end{pmatrix}^T \right\}.$$

The inner products $\langle \cdot, \cdot\rangle_i'$ and $\langle \cdot, \cdot\rangle_i''$, together with scalar multiplication on $\mathbb{R}^1$, induce an inner product $\langle \cdot, \cdot\rangle$ and $\mathscr{A} \oplus \mathscr{M} \oplus \mathscr{S}$. Setting

$$V(x) = \begin{pmatrix} \beta_1(x) \\ \vdots \\ \beta_m(x) \\ \beta_1(x)\gamma_1(x) \\ \vdots \\ \beta_m(x)\gamma_m(x) \\ \beta_1(x)\delta_1(x) \\ \vdots \\ \beta_m(x)\delta_m(x) \end{pmatrix} \in \mathscr{A} \oplus \mathscr{M} \oplus \mathscr{S},$$

$$B_{23} = \left(\operatorname{diag} \frac{1}{\alpha_i N} \sum_{k=1}^{N} \beta_i(x_k)\gamma_i(x_k)\langle \delta_i(x_k), \cdot\rangle_i''\right),$$

$$B_{32} = \left(\operatorname{diag} \Sigma_i \frac{1}{\alpha_i N} \sum_{k=1}^{N} \beta_i(x_k)\delta_i(x_k)\langle \gamma_i(x_k), \cdot\rangle_i'\right),$$

$$B_{33} = \left(\operatorname{diag} \Sigma_i \frac{1}{\alpha_i N} \sum_{k=1}^{N} \beta_i(x_k)\delta_i(x_k)\langle \delta_i(x_k), \cdot\rangle_i''\right),$$

one obtains

$$\nabla\Phi_\varepsilon(\Theta) = \begin{pmatrix} I & 0 & 0 \\ 0 & I & \varepsilon B_{23} \\ 0 & \varepsilon B_{32} & (1-\varepsilon)I + \varepsilon B_{33} \end{pmatrix}$$

$$-\varepsilon \begin{pmatrix} (\operatorname{diag} \alpha_i) & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & (\operatorname{diag} \Sigma_i) \end{pmatrix} \left\{ \frac{1}{N} \sum_{k=1}^{N} V(x_k)\langle V(x_k), \cdot\rangle \right\}.$$

Denoting the vector of true parameters by Θ^0, one verifies without difficulty that $\nabla\Phi_\varepsilon$ is of the form

$$\nabla\Phi_\varepsilon(\Theta) = \frac{1}{N} \sum_{k=1}^{N} F(x_k, \Theta),$$

where the operator $F(x, \Theta)$ not only has finite expectation (in norm) at Θ^0 but also has a Fréchet derivative with respect to Θ for which the following holds: If $\|\cdot\|$ is any operator norm on $\nabla_\Theta F$, then there exists a real-valued function f on $\mathbb{R}^n$ such that

$$\int_{\mathbb{R}^n} f(x)p(x)\,dx<\infty$$

at Θ^0 and such that $\|\nabla_\Theta F(x,\Theta)\|\leqq f(x)$ for all $x\in\mathbb{R}^n$ and all Θ in a sufficiently small neighborhood of Θ^0. Since the solution Θ of the likelihood equations is strongly consistent, it follows from the strong law of large numbers (see Loève [10]) that $\nabla\Phi_\varepsilon(\Theta)$ converges with probability 1 to $E(\nabla\Phi_\varepsilon(\Theta^0))$ as N approaches infinity.

To complete the proof of the theorem, it must be shown that $E(\nabla\Phi_\varepsilon(\Theta^0))$ has operator norm less than 1 with respect to some vector norm on $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$ whenever $0<\varepsilon<2$. A straightforward calculation yields

$$E(\nabla\Phi_\varepsilon(\Theta^0))=\begin{pmatrix} I & 0 & 0\\ 0 & I & 0\\ 0 & 0 & I\end{pmatrix}-\varepsilon\begin{pmatrix}(\operatorname{diag}\alpha_i^0) & 0 & 0\\ 0 & I & 0\\ 0 & 0 & (\operatorname{diag}\Sigma_i^0)\end{pmatrix}\left\{\int_{\mathbb{R}^n} V(x)\langle V(x),\cdot\rangle p(x)\,dx\right\},$$

where α_i^0, μ_i^0, and Σ_i^0, $i=1,\cdots,m$, are the components of Θ^0. Thus $E(\nabla\Phi_\varepsilon(\Theta^0))$ is an operator on $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$ of the form $I-\varepsilon QR$, where

$$Q=\begin{pmatrix}(\operatorname{diag}\alpha_i^0) & 0 & 0\\ 0 & I & 0\\ 0 & 0 & (\operatorname{diag}\Sigma_i^0)\end{pmatrix}$$

and

$$R=\int_{\mathbb{R}^n} V(x)\langle V(x),\cdot\rangle p(x)\,dx$$

are positive-definite and symmetric[1] with respect to the inner product $\langle\cdot,\cdot\rangle$. Since QR is positive-definite and symmetric with respect to the inner product $\langle\cdot,Q^{-1}\cdot\rangle$ on $\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$, it suffices to show that

$$\langle W,RW\rangle=\langle W,Q^{-1}[QR]W\rangle\leqq\langle W,Q^{-1}W\rangle,$$

for all $W\in\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S}$. Indeed, it follows from this inequality that, with respect to the inner product $\langle\cdot,Q^{-1}\rangle$, the operator norm of QR is no greater than 1 and, hence, the operator norm of $E(\nabla\Phi_\varepsilon(\Theta^0))$ is less than 1 whenever $0<\varepsilon<2$.

For

$$W=\begin{pmatrix} y_1\\ \vdots\\ y_m\\ v_1\\ \vdots\\ v_m\\ B_1\\ \vdots\\ B_m\end{pmatrix}\in\mathscr{A}\oplus\mathscr{M}\oplus\mathscr{S},$$

[1] An operator T is *symmetric* with respect to an inner product $(\cdot,\cdot)$ if $(x,Ty)=(Tx,y)$ and *positive definite* if it is symmetric and $(x,Tx)>0$ for $x\neq 0$.

one has

$$\langle W, RW\rangle = \int_{\mathbb{R}^n} \Bigg[\sum_{i=1}^{m} y_i\beta_i(x) + \sum_{i=1}^{m} v_i^T(\alpha_i^0\Sigma_i^{0-1})\beta_i(x)\gamma_i(x) + \sum_{i=1}^{m} \operatorname{tr}\left\{B_i\left(\frac{\alpha_i^0}{2}\Sigma_i^{0-1}\right)\beta_i(x)\delta_i(x)^T\right\}\Bigg]^2 p(x)\,dx$$

$$= \int_{\mathbb{R}^n} \left[\sum_{i=1}^{m} (\alpha_i^{0-1}y_i + v_i^T\Sigma_i^{0-1}\gamma_i(x) + \operatorname{tr}\{B_i(\tfrac{1}{2}\Sigma_i^{0-1})\delta_i(x)^T\})\alpha_i^0\beta_i(x)\right]^2 p(x)\,dx$$

$$\leqq \int_{\mathbb{R}^n} \left[\sum_{i=1}^{m} (\alpha_i^{0-1}y_i + v_i^T\Sigma_i^{0-1}\gamma_i(x) + \operatorname{tr}\{B_i(\tfrac{1}{2}\Sigma_i^{0-1})\delta_i(x)^T\})^2\alpha_i^0\beta_i(x)\right] p(x)\,dx.$$

The inequality is a consequence of the following corollary of Schwarz's inequality: If $\eta_i \geqq 0$ for $i = 1, \cdots, m$ and if $\sum_{i=1}^m \eta_i = 1$, then $|\sum_{i=1}^m \xi_i\eta_i|^2 \leqq \sum_{i=1}^m \xi_i^2\eta_i$ for all $\{\xi\}_{i=1,\cdots,m}$. If the squared expressions in the last sum above are written out in full, one sees that the integrals of the cross terms in these expressions vanish. Consequently

$$\langle W, RW\rangle \leqq \int_{\mathbb{R}^n} \sum_{i=1}^{m} [\alpha_i^{0-2}y_i^2 + (v_i^T\Sigma^{0-1}\gamma_i(x))^2 + (\operatorname{tr}\{B_i(\tfrac{1}{2}\Sigma_i^{0-1})\delta_i(x)^T\})^2]\alpha_i^0 p_i(x)\,dx.$$

Now

$$\int_{\mathbb{R}^n} \alpha_i^{0-1}y_i^2 p_i(x)\,dx = \alpha_i^{0-1}y_i^2, \tag{6a}$$

$$\int_{\mathbb{R}^n} (v_i^T\Sigma_i^{0-1}\gamma_i(x))^2\alpha_i^0 p_i(x)\,dx = \int_{\mathbb{R}^n} v_i^T\Sigma_i^{0-1}(x-\mu_i^0)(x-\mu_i^0)^T\Sigma_i^{0-1}v_i\alpha_i^0 p_i(x)\,dx = \langle v_i, v_i\rangle_i', \tag{6b}$$

$$\int_{\mathbb{R}^n} (\operatorname{tr}\{B_i(\tfrac{1}{2}\Sigma_i^{0-1})\delta_i(x)^T\})^2\alpha_i^0 p_i(x)\,dx = \langle B_i, \Sigma_i^{0-1}B_i\rangle_i''. \tag{6c}$$

(A proof of (6c) follows below.) From (6a), (6b), and (6c), one concludes that

$$\langle W, RW\rangle \leqq \sum_{i=1}^{m} \alpha_i^{0-1}y_i^2 + \sum_{i=1}^{m} \langle v_i, v_i\rangle_i' + \sum_{i=1}^{m} \langle B_i, \Sigma_i^{0-1}B_i\rangle_i'' = \langle W, Q^{-1}W\rangle.$$

This completes the proof of the theorem.

Proof of (6c). Setting $y = \Sigma_i^{0-1/2}(x-\mu_i^0)$ and

$$C = \int_{\mathbb{R}^n} (\operatorname{tr}\{B_i(\tfrac{1}{2}\Sigma_i^{0-1})\delta_i(x)^T\})^2\alpha_i^0 p_i(x)\,dx,$$

one verifies that

$$C = \frac{\alpha_i^0}{4}\int_{\mathbb{R}^n} (\operatorname{tr}\{B_i[\Sigma_i^{0-1/2}yy^T\Sigma_i^{0-1/2} - \Sigma_i^{0-1})^T\})^2 p_0(y)\,dy,$$

where $p_0 \sim N(0, I)$. Denoting $\Sigma_i^{0-1/2} B_i \Sigma_i^{0-1/2} = D = (d_{jk})$, one then derives

$$\begin{aligned}
C &= \frac{\alpha_i^0}{4} \int_{\mathbb{R}^n} (\operatorname{tr}\{D[yy^T - I]\})^2 p_0(y)\, dy \\
&= \frac{\alpha_i}{4} \int_{\mathbb{R}^n} [(\operatorname{tr}\{Dyy^T\})^2 - 2\operatorname{tr}\{D\} \operatorname{tr}\{Dyy^T\} + (\operatorname{tr}\{D\})^2] p_0(y)\, dy \\
&= \frac{\alpha_i^0}{4} \left\{ \sum_{j,k,p,q} d_{jk} d_{pq} \int_{\mathbb{R}^n} y_k y_j y_q y_p p_0(y)\, dy - 2(\operatorname{tr}\{D\})^2 + (\operatorname{tr}\{D\})^2 \right\} \\
&= \frac{\alpha_i^0}{4} \left\{ \sum_k \sum_{p \neq k} d_{kk}\, d_{pp} + \sum_k \sum_{j \neq k} d_{jk}\, d_{jk} + \sum_k \sum_{j \neq k} d_{jk}\, d_{kj} + 3 \sum_k d_{kk}^2 - (\operatorname{tr}\{D\})^2 \right\} \\
&= \frac{\alpha_i^0}{2} \operatorname{tr}\{D^2\} = \frac{\alpha_i^0}{2} \operatorname{tr}\{\Sigma_i^{0-1/2} B_i \Sigma_i^{0-1/2} B_i \Sigma_i^{0-1/2}\} = \operatorname{tr}\left\{ B_i \left(\frac{\alpha_i}{2} \Sigma_i^{0-1} \right) (\Sigma_i^{0-1} B_i)^T \right\} \\
&= \langle B_i, \Sigma^{0-1} B_i \rangle_i''.
\end{aligned}$$

4. The optimal ε. The results just obtained state that, with probability 1 as N approaches infinity, the iterative procedure (4) converges locally to the strongly consistent maximum-likelihood estimate Θ whenever $0 < \varepsilon < 2$. In this section we observe that there exists a particular value of ε, referred to as "the optimal ε," which yields, with probability 1, the fastest asymptotic uniform rate of local convergence of (4) near Θ. We derive a lower bound between 1 and 2 on the optimal ε and relate it to the separation of the component populations in the mixture.

From the proof of the theorem, one sees that the optimal ε is that which minimizes the spectral radius of the operator $E(\nabla\Phi_\varepsilon(\Theta^0))$ restricted to the space $\mathscr{C} \oplus \mathscr{M} \oplus \mathscr{S}$, where $\mathscr{C}$ is the subspace of $\mathscr{A}$ whose components sum to zero. Indeed, the restricted operator $E(\nabla\Phi_\varepsilon(\Theta)) = I - \varepsilon QR$ is symmetric on $\mathscr{C} \oplus \mathscr{M} \oplus \mathscr{S}$ with respect to the inner product $\langle \cdot, Q^{-1} \cdot \rangle$. Consequently, its operator norm with respect to this inner product is equal to its spectral radius and, hence, minimal. We observe that the restriction of QR to $\mathscr{C} \oplus \mathscr{M} \oplus \mathscr{S}$ is positive-definite and symmetric with respect to the inner product $\langle \cdot, Q^{-1} \cdot \rangle$. Letting ρ and τ denote, respectively, the largest and smallest eigenvalues of this restriction of QR, one verifies that the spectral radius of $E(\nabla\Phi_\varepsilon(\Theta^0))$, restricted to $\mathscr{C} \oplus \mathscr{M} \oplus \mathscr{S}$, is minimized when $1 - \varepsilon\tau = \varepsilon\rho - 1$, i.e., when $\varepsilon = 2/(\rho + \tau)$.

It follows from the proof of the theorem that ρ is never greater than 1. Thus the optimal ε is bounded below by $2/1 + \tau$, where τ lies between 0 and 1. In particular, this lower bound on the optimal ε lies between 1 and 2. We have been unable to determine ρ more precisely in general. It should be noted that, if ρ is strictly less than $\frac{1}{2}$, then the optimal ε is actually greater than 2, even though the theorem just proved fails to guarantee the local convergence of (4) for such values of ε.

Suppose that the component populations in the mixture are "widely separated" in the sense that each pair (μ_i^0, Σ_i^0) differs greatly from every other such pair. Then, for $i, j = 1, \cdots, m$,

$$\frac{\alpha_i^0 p_i(x)}{p(x)} \frac{\alpha_j^0 p_j(x)}{p(x)} \approx 0 \quad \text{for } x \in \mathbb{R}^n, \text{ whenever } i \neq j.$$

One sees that $QR \approx I$ and, hence, ρ and τ must both lie near 1. Consequently, fastest asymptotic local convergence rates are obtained for ε near 1, and, for the optimal ε, $E(\nabla\Phi_\varepsilon(\Theta^0)) = I - \varepsilon QR \approx 0$. Thus for mixtures whose component populations are "widely separated," the optimal ε is only slightly greater than 1, and rapid first-order local convergence of the iterative procedure (4) to Θ can be expected asymptotically for this ε.

Now suppose that the component populations in the mixture are such that at least two pairs (μ_i^0, Σ_i^0) and (μ_j^0, Σ_j^0), $i \neq j$, are nearly identical. Then $\beta_i(x) \approx \beta_j(x)$, $\beta_i(x)\gamma_i(x) \approx \beta_j(x)\gamma_j(x)$ and $\beta_i(x)\delta_i(x) \approx \beta_j(x)\delta_j(x)$, and it follows that R is nearly singular and, hence, that τ is near zero. One concludes that the optimal ε cannot be much smaller than 2. In fact if ρ is near 1, as is the case when all pairs (μ_i^0, Σ_i^0) are nearly identical, then the optimal ε must lie near 2. Furthermore, the spectral radius of $E(\nabla\Phi_\varepsilon(\Theta))$ is near 1, even for the optimal ε; therefore, slow first-order convergence can be expected asymptotically in this case.

5. Maximum-likelihood estimates of the a priori probabilities and the means. It happens that, if the covariance matrices Σ_i, $i = 1, \cdots, m$, are held fixed, then, under certain conditions, an appropriately restricted version of the iterative procedure (4) converges locally with probability 1 to a maximum-likelihood estimate of the parameters α_i^0 and μ_i^0, $i = 1, \cdots, m$, whenever the number of observations in the sample reaches a certain finite size. To be more specific, we introduce the following notation: For $\tilde{\Theta} = \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M}$ and $\bar{\Sigma} \in \mathcal{S}$, denote

$$\Theta = \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$$

by $(\tilde{\Theta}, \bar{\Sigma})$. Then, for given $\bar{\Sigma}$, the likelihood equations for the parameters α_i and μ_i, $i = 1, \cdots, m$ can be written as

$$\tilde{\Theta} = \begin{pmatrix} A(\tilde{\Theta}, \bar{\Sigma}) \\ M(\tilde{\Theta}, \bar{\Sigma}) \end{pmatrix},$$

or, more generally as

$$\tilde{\Theta} = \tilde{\Phi}_\varepsilon(\tilde{\Theta}, \bar{\Sigma}) \equiv (1-\varepsilon)\tilde{\Theta} + \varepsilon \begin{pmatrix} A(\tilde{\Theta}, \bar{\Sigma}) \\ M(\tilde{\Theta}, \bar{\Sigma}) \end{pmatrix} \tag{7}$$

for any ε. The appropriate iterative procedure to consider is the following: Beginning with some starting value $\tilde{\Theta}^{(1)}$, define successive iterates inductively by

$$\tilde{\Theta}^{(k+1)} = \tilde{\Phi}_\varepsilon(\tilde{\Theta}^{(k)}, \bar{\Sigma}) \tag{8}$$

for $k = 1, 2, 3, \cdots$. Our result concerning this procedure is given by the theorem and its corollary below.

THEOREM. *If $N \geqq m(n+1)$ and if $(\tilde{\Theta}, \bar{\Sigma})$ is a solution of* (7) *which lies sufficiently near a solution of* (3), *then, with probability* 1, *$\tilde{\Phi}_\varepsilon$ is a locally contractive operator* (*in some norm on $\mathcal{A} \oplus \mathcal{M}$*) *near $\tilde{\Theta}$ whenever $0 < \varepsilon < 2$.*

COROLLARY. *If $N \geqq m(n+1)$ and if $(\tilde{\Theta}, \bar{\Sigma})$ is a solution of* (7) *which lies sufficiently near a solution of* (3), *then, with probability* 1, *the iterative procedure* (8) *converges locally to $\tilde{\Theta}$ whenever $0 < \varepsilon < 2$.*

Proof of the theorem. Suppose that $N \geqq m(n+1)$ and $0 < \varepsilon < 2$. As in the proof of the preceding theorem, it suffices to show that, with probability 1, $\nabla_{\tilde{\Theta}}\tilde{\Phi}_\varepsilon(\tilde{\Theta}, \bar{\Sigma})$ has operator norm less than 1 with respect to some vector norm on $\mathcal{A} \oplus \mathcal{M}$. Since $\nabla_{\tilde{\Theta}}\tilde{\Phi}_\varepsilon$ depends continuously on $\tilde{\Theta}$ and $\bar{\Sigma}$, this need only be shown when $(\tilde{\Theta}, \bar{\Sigma})$ is a solution of (3).

From the proof of the preceding theorem, one sees that if $(\tilde{\Theta}, \bar{\Sigma})$ is a solution of (3), then

$$\nabla_{\tilde{\Theta}} \tilde{\Phi}_\varepsilon(\tilde{\Theta}, \bar{\Sigma}) = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} (\operatorname{diag} \alpha_i) & 0 \\ 0 & I \end{pmatrix} \left\{ \frac{1}{N} \sum_{k=1}^{N} \tilde{V}(x_k) \langle \tilde{V}(x_k), \cdot \rangle \right\}$$

where

$$\tilde{V}(x) = \begin{pmatrix} \beta_1(x) \\ \vdots \\ \beta_m(x) \\ \beta_1(x)\gamma_1(x) \\ \vdots \\ \beta_m(x)\gamma_m(x) \end{pmatrix},$$

and the inner product $\langle \cdot, \cdot \rangle$ is now the inner product induced on $\mathscr{A} \oplus \mathscr{M}$ by scalar multiplication on $\mathbb{R}^1$ and the inner products $\langle \cdot, \cdot \rangle_i'$ on $\mathbb{R}^n$. As before, $\nabla_{\tilde{\Theta}} \tilde{\Phi}_\varepsilon(\tilde{\Theta}, \bar{\Sigma})$ is of the form $I - \varepsilon \tilde{Q} \tilde{R}$, where

$$\tilde{Q} = \begin{pmatrix} (\operatorname{diag} \alpha_i) & 0 \\ 0 & I \end{pmatrix},$$

and

$$\tilde{R} = \left\{ \frac{1}{N} \sum_{k=1}^{N} \tilde{V}(x_k) \langle \tilde{V}(x_k), \cdot \rangle \right\}.$$

We observe that $\tilde{Q}\tilde{R}$ is symmetric and positive semi-definite with respect to the inner product $\langle \cdot, \tilde{Q}^{-1} \cdot \rangle$. In fact, it is shown in Appendix B that, with probability 1, $\tilde{Q}\tilde{R}$ is positive-definite with respect to this inner product. Consequently, the theorem will be proved if it can be shown that

$$\langle \tilde{W}, \tilde{Q}^{-1}[\tilde{Q}\tilde{R}]\tilde{W} \rangle = \langle \tilde{W}, \tilde{R}\tilde{W} \rangle \leqq \langle \tilde{W}, \tilde{Q}^{-1}\tilde{W} \rangle$$

for all $\tilde{W} \in \mathscr{A} \oplus \mathscr{M}$.

For

$$\tilde{W} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \\ v_1 \\ \vdots \\ v_m \end{pmatrix} \in \mathscr{A} \oplus \mathscr{M},$$

one has

$$\begin{aligned} \langle \tilde{W}, \tilde{R}\tilde{W} \rangle &= \frac{1}{N} \sum_{k=1}^{N} \left[\sum_{i=1}^{m} y_i \beta_i(x_k) + \sum_{i=1}^{m} v_i^T (\alpha_i \Sigma_i^{-1}) \beta_i(x_k) \gamma_i(x_k) \right]^2 \\ &= \frac{1}{N} \sum_{k=1}^{N} \left[\sum_{i=1}^{m} (\alpha_i^{-1} y_i + v_i^T \Sigma_i^{-1} \gamma_i(x_k)) \alpha_i \beta_i(x_k) \right]^2 \\ &\leqq \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{m} [\alpha_i^{-1} y_i + v_i^T \Sigma_i^{-1} \gamma_i(x_k)]^2 \alpha_i \beta_i(x_k) \end{aligned}$$

by Schwarz's inequality. Since $(\tilde{\Theta}, \tilde{\Sigma})$ is a solution of (3), this easily yields

$$\langle \tilde{W}, \tilde{R}\tilde{W}\rangle \leqq \sum_{i=1}^{m} \alpha_i^{-1} y_i^2 + \sum_{i=1}^{m} v_i^T(\alpha_i\Sigma_i^{-1})v_i = \langle \tilde{W}, \tilde{Q}^{-1}\tilde{W}\rangle,$$

and the proof is complete.

If the conclusion of the theorem holds for some solution $(\tilde{\Theta}, \tilde{\Sigma})$ of (7), then, as in the preceding section, a particular value of ε can be determined which yields the fastest uniform rate of local convergence of (8) near $\tilde{\Theta}$. With respect to the inner product $\langle\cdot, \tilde{Q}^{-1}\cdot\rangle$, $\tilde{Q}\tilde{R}$ is positive-definite and symmetric on $\mathscr{E}\oplus\mathscr{M}$. Denoting the largest and smallest eigenvalues of the restriction of $\tilde{Q}\tilde{R}$ to $\mathscr{E}\oplus\mathscr{M}$ by ρ and τ, respectively, one sees that the optimal ε is again given by $\varepsilon = 2/(\rho+\tau)$. Since the restriction of $\tilde{Q}\tilde{R}$ has operator norm no greater than 1 with respect to the inner product $\langle\cdot, \tilde{Q}^{-1}\cdot\rangle$, ρ must be no greater than 1. Hence, $\varepsilon \geqq 1/(1+\tau)$, where τ lies between 0 and 1. Reasoning as before, one sees that the optimal ε lies near 1 if the component populations are "widely separated," and cannot be much less than 2 if two or more of the populations have nearly identical means and covariance matrices. In the former case, rapid first-order local convergence of (8) can be expected for the optimal ε. In the latter case, if ρ is near 1, then the optimal ε must be near 2, and slow first-order convergence of (8) can be expected, even for the optimal ε.

6. Concluding remarks. A number of numerical techniques for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions have been discussed in the literature. In addition to the usual steepest-ascent method for obtaining a local maximum of the log-likelihood function, we mention in particular Newton's method, the method of scoring, and the modifications of these procedures investigated by Kale [9] for obtaining solutions of the likelihood equations. It is our feeling that the iterative procedure (4) offers considerable computational advantages over these procedures in many cases of practical interest.

Even though the partial derivatives of the log-likelihood function are not appreciably more difficult to evaluate than the expressions used in defining the function Φ_ε, the procedure (4), which is a generalized steepest–ascent (deflected gradient) method appears to have two particular advantages over the usual steepest-ascent method. First, the major practical advantage of procedure (4) is that if ε is no greater than 1, then the constraints of the problem are automatically satisfied by the successive iterates for any feasible choice of the starting value $\Theta^{(1)}$; i.e., the successive Σ_i's are symmetric ad positive-definite and the successive α_i's are positive and sum to 1. With more conventional ascent procedures, special precautions must be taken to insure that the inequality constraints are not violated, at least in the initial stages of the iteration. Second, in the interval of step sizes $0<\varepsilon\leqq 1$ in which the constraints are preserved, there is one step-size, namely $\varepsilon = 1$, which is best in terms of the asymptotic rate of convergence, regardless of the particular mixture problem at hand. This suggests that the likelihood function is actually increased at each stage in procedure (4) with $\varepsilon = 1$, a conjecture which is supported by our experience and that of others [6], but which we have been unable to prove.

Although Newton's method and the method of scoring offer quadratic and near-quadratic convergence, respectively, for large sample sizes, they require at each iteration the inversion of a square matrix whose dimension is equal to the number of independent variables among the parameters, namely $(m(n+1)(n+2)/2)-1$. Thus these methods may be less efficient computationally than the iterative procedure (4) if m and n are large, even though they may yield a satisfactory approximate solution after fewer iterations. The modified versions of Newton's method and the method of scoring do not require the re-calculation of the inverse of a large matrix at each step. However, quadratic convergence is not achieved with these modified methods, and multiplication by a large matrix must still be carried out at each iteration.

Appendix A. We now give a brief proof of the existence and uniqueness of the strongly consistent maximum-likelihood estimate. For the sake of generality, this is done in a somewhat broader context than is necessary for this paper.

Let $p(x, \Theta^0)$ be a probability density function of a vector variable $x \in \mathbb{R}^n$ and a vector parameter $\Theta \in \mathbb{R}^\nu$. If $\{x_k\}_{k=1,\cdots,N}$ is an independent sample of observations on a random variable $x \in \mathbb{R}^n$ whose probability density function is $p(x, \Theta^0)$ for some $\Theta^0 \in \mathbb{R}^\nu$, then a maximum-likelihood estimate of Θ^0 is a choice of Θ which locally maximizes the log-likelihood function

$$L = \sum_{k=1}^{N} \log p(x_k, \Theta).$$

If p is a differentiable function of Θ, then a necessary condition for a maximum-likelihood estimate is that the likelihood equations

$$\frac{\partial L}{\partial \Theta_i} = 0, \qquad i = 1, \cdots, \nu,$$

be satisfied, where Θ_i is the ith component of Θ. In the following, our objective is to show that if ρ satisfies certain conditions, then, given any sufficiently small neighborhood of Θ^0, there is, with probability 1 as N approaches infinity, a unique solution of the likelihood equations in that neighborhood, and this solution is a maximum-likelihood estimate of Θ^0.

We assume that $p(x, \Theta)$ satisfies the following conditions of Chanda [2]: (a) There is a neighborhood Ω of Θ^0 such that for all $\Theta \in \Omega$, for almost all $x \in \mathbb{R}^n$, and for $i, j, k = 1, \cdots, \nu$, $\partial p/\partial \theta_i$, $\partial^2 p/\partial \theta_i \, \partial \theta_j$, and $\partial^3 p/\partial \theta_i \, \partial \theta_j \, \partial \theta_k$ exist and satisfy

$$\left|\frac{\partial p}{\partial \theta_i}\right| \leqq f_i(x), \qquad \left|\frac{\partial^2 p}{\partial \theta_i \, \partial \theta_j}\right| \leqq f_{ij}(x), \qquad \left|\frac{\partial^3 \log p}{\partial \theta_i \, \partial \theta_j \, \partial \theta_k}\right| \leqq f_{ijk}(x),$$

where f_i and f_{ij} are integrable and f_{ijk} satisfies

$$\int_{\mathbb{R}^n} f_{ijk}(x) p(x, \Theta^0) \, dx < \infty.$$

(b) The matrix

$$J(\theta) = \left(\int_{\mathbb{R}^n} \frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j} p \, dx \right)$$

is positive-definite at θ^0.

$$\text{Let } \mathscr{L}(\Theta) = \begin{pmatrix} \dfrac{1}{N} \dfrac{\partial L}{\partial \theta_1} \\ \\ \dfrac{1}{N} \dfrac{\partial L}{\partial \theta_\nu} \end{pmatrix}.$$

It is immediately seen that $\mathscr{L}(\Theta)=0$ if and only if the likelihood equations are satisfied, and that, by the strong law of large numbers [10], $\mathscr{L}(\Theta^0)$ converges with probability 1 to zero. Furthermore, it follows from assumptions (a) and (b) above that there exists a neighborhood Ω^0 of Θ^0 (contained in Ω and, for convenience, convex) and a positive ε such that, with probability 1 as N approaches infinity, $\nabla\mathscr{L}(\Theta)\leqq -\varepsilon I$ for all $\Theta\in\Omega^0$. (That is, $\nabla\mathscr{L}(\Theta)+\varepsilon I$ is negative-semidefinite.) Denoting the spherical neighborhood of radius δ about Θ^0 by Ω_δ, we establish the following

LEMMA. *With probability* 1 *as N approaches infinity,*

(i) *$\mathscr{L}$ is one-to-one on Ω^0,*

(ii) *$\mathscr{L}(\Omega_\delta)$ contains the ball of radius $\varepsilon\delta$ about $\mathscr{L}(\Theta^0)$ whenever $\Omega_\delta\subseteq\Omega^0$.*

Proof. We may assume that $\nabla\mathscr{L}(\Theta)\leqq -\varepsilon I$ for all $\Theta\in\Omega^0$, since the probability that this is the case is 1 as N approaches infinity. To prove (i), suppose that $\mathscr{L}(\Theta^1)=\mathscr{L}(\Theta^2)$ for Θ^1 and Θ^2 in Ω^0. Then

$$0=(\Theta^1-\Theta^2)[\mathscr{L}(\Theta^1)-\mathscr{L}(\Theta^2)]$$

$$=(\Theta^1-\Theta^2)^T\left\{\int_0^1 \nabla\mathscr{L}(\Theta^2+t[\Theta^1-\Theta^2])\,dt\right\}(\Theta^1-\Theta^2).$$

The negative-definiteness of $\nabla\mathscr{L}$ implies that $\Theta^1=\Theta^2$, and (i) is proved.

To prove (ii), suppose that $\Omega_\delta\subseteq\Omega^0$, and let Θ^1 be a boundary point of Ω_δ. Then

$$\mathscr{L}(\Theta^1)-\mathscr{L}(\Theta^0)=\left\{\int_0^1 \nabla\mathscr{L}(\Theta^0+t[\Theta^1-\Theta^0])\,dt\right\}(\Theta^1-\Theta^0).$$

After left-multiplying this equation by $(\Theta^1-\Theta^0)^T$, one verifies using Schwarz's inequality and the negative-definiteness of $\nabla\mathscr{L}$ that

$$\|\mathscr{L}(\Theta^1)-\mathscr{L}(\Theta^0)\|\geqq\varepsilon\|\Theta^1-\Theta^0\|=\varepsilon\delta,$$

where $\|\cdot\|$ denotes the usual Euclidean norm on $\mathbb{R}^\nu$. Since all boundary points of $\mathscr{L}(\Omega_\delta)$ are images under $\mathscr{L}$ of boundary points of Ω_δ, the proof of (ii) is complete.

The desired result of this appendix follows immediately from this lemma and the remarks preceding it. Indeed, if Ω^1 is any neighborhood of Θ^0 which is contained in Ω^0, then one can find a δ for which $\Omega_\delta\subseteq\Omega^1\subseteq\Omega^0$. By the lemma, the probability is 1 as N tends to infinity that $\mathscr{L}$ is one-to-one on Ω^1 and that $\mathscr{L}(\Omega_\delta)$ and, hence, $\mathscr{L}(\Omega^1)$ contain the ball of radius $\varepsilon\delta$ about $\mathscr{L}(\Theta^0)$. Since $\mathscr{L}(\Theta^0)$ converges with probability 1 to zero, one concludes that, with probability 1 as N approaches infinity, there exists a unique $\Theta\in\Omega^1$ for which $\mathscr{L}(\Theta)=0$. Since the probability also is 1 as N approaches infinity that $\nabla\mathscr{L}$ is negative-definite on Ω^1, this Θ is, with probability 1, a maximum-likelihood estimate.

Appendix B. We now prove that the operator $\tilde{Q}\tilde{R}$ is positive-definite on $\mathscr{A}\oplus\mathscr{M}$ with probability 1 whenever $N\geqq m(n+1)$. Since

$$\tilde{Q}\tilde{R}=\begin{pmatrix}(\operatorname{diag}\alpha_i) & 0\\ 0 & I\end{pmatrix}\left\{\frac{1}{N}\sum_{k=1}^N \tilde{V}(x_k)\langle\tilde{V}(x_k),\cdot\rangle\right\},$$

it suffices to show that the vectors

$$\tilde{V}(x_k)=\begin{pmatrix} \dfrac{p_1(x_k)}{p(x_k)} \\ \vdots \\ \dfrac{p_m(x_k)}{p(x_k)} \\ \dfrac{p_1(x_k)}{p(x_k)}(x_k-\mu_1) \\ \vdots \\ \dfrac{p_m(x_k)}{p(x_k)}(x_k-\mu_m) \end{pmatrix}, \qquad k=1,\cdots,N,$$

span $\mathscr{A}\oplus\mathscr{M}$ with probability 1 whenever $N \geqq m(n+1)$. This follows from the more general result below.

LEMMA. *Let $\{x_k\}_{k=1,\cdots,N}$ be an independent sample of observations on a random variable x in $\mathbb{R}^s$ which is distributed with a probability density function p. If V is a real-analytic function from $\mathbb{R}^s$ to $\mathbb{R}^t$ whose component functions are linearly independent, then the vectors $V(x_k)$, $k=1,\cdots,N$, span $\mathbb{R}^t$ with probability* 1 *whenever $N \geqq t$.*

Proof. Denoting the jth component function of V by v_j, we define a real-analytic function v_j from $\mathbb{R}^s$ to $\mathbb{R}^j$ by

$$V_j(x)=\begin{pmatrix} v_1(x) \\ \vdots \\ v_j(x) \end{pmatrix}$$

for $j=1,\cdots,t$. Our proof of the lemma consists of showing inductively that, for $j=1,\cdots,t$, the set $\{V_j(x_k)\}_{k=1,\cdots,j}$ spans $\mathbb{R}^j$ with probability 1. We make the preliminary observation that, since the real-analytic functions v_j are assumed to be linearly independent, any nonzero linear combination of them vanishes only on a set of Lebesgue measure zero in $\mathbb{R}^s$.

From the observation above, $V_1(x_1)$ is nonzero with probability 1; hence $V_1(x_1)$ spans $\mathbb{R}^1$ with probability 1. Suppose now that, for some j, $1 \leqq j < t$, the set $\{V_j(x_k)\}_{k=1,\cdots,j}$ spans $\mathbb{R}^j$ with probability 1. Then, with probability 1, the set $\{V_{j+1}(x_k)\}_{k=1,\cdots,j+1}$ fails to span $\mathbb{R}^{j+1}$ if and only if

$$V_{j+1}(x_{j+1})=\sum_{k=1}^{j} c_k V_{j+1}(x_k) \tag{B.1}$$

for some set of constants $\{c_k\}_{k=1,\cdots,j}$. If (B.1) holds, the constants c_k are determined by

$$\begin{pmatrix} c_1 \\ \vdots \\ c_j \end{pmatrix} = \mathscr{V}_j^{-1} V_j(x_{j+1})$$

with probability 1, where $\mathscr{V}_j$ is the $j \times j$ matrix whose kth column is $V_j(x_k)$. Thus, with probability 1, (B.1) holds if and only if

$$[\mathcal{V}_j^{-1} V_j(x_{j+1})]^T \begin{pmatrix} v_{j+1}(x_1) \\ \vdots \\ v_{j+1}(x_j) \end{pmatrix} - v_{j+1}(x_{j+1}) = 0.$$

Now

$$[\mathcal{V}_j^{-1} V_j(x)]^T \begin{pmatrix} v_{j+1}(x_1) \\ \vdots \\ v_{j+1}(x_j) \end{pmatrix} - v_{j+1}(x)$$

is a nonzero linear combination of the functions $v_1, \cdots, v_{j+1}$ and, hence, vanishes only on a set of Lebesgue measure zero in $\mathbb{R}^s$. One concludes that $\{V_{j+1}(x_k)\}_{k=1,\cdots,j+1}$ fails to span $\mathbb{R}^{j+1}$ with probability zero. This completes the induction, and the lemma is proved.

REFERENCES

[1] J. Aitchison and S. D. Silvey, *Maximum likelihood estimation of parameters subject to restraints*, Ann. Math. Statis., 3 (1958), pp. 813–828.

[2] K. C. Chanda, *A note on the consistency and maxima of the roots of the likelihood equations*, Biometrika, 41 (1954), pp. 56–61.

[3] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946.

[4] N. E. Day, *Estimating the components of a mixture of normal distributions*, Biometrika, 56 (1969), pp. 463–474.

[5] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.

[6] V. Hasselblad, *Estimation of parameters for a mixture of normal distributions*, Technometrics, 8 (1966), pp. 431–446.

[7] A. S. Householder, *Theory of Matrices and Numerical Analysis*, Blaisdell, New York, 1964.

[8] V. S. Huzurbazar, *The likelihood equation, consistency and the maxima of the likelihood function*, Annals of Eugenics, Lond., 14 (1948), pp. 185–200.

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

[10] M. Loéve, *Probability Theory*, Van Nostrand, New York, 1963.

[11] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[12] B. C. Peters and W. A. Coberly, *The numerical evaluation of a the maximum-likelihood estimate of mixture proportions*, Commun. Statist.—Theor. Meth., A5 (1976), No. 12, pp. 1127–1135.

[13] A. Wald, *Note on the consistency of the maximum-likelihood estimate*, Ann. Math. Statist., 20 (1949), p. 595.

[14] J. H. Wolfe, *Pattern clustering by multivariate mixture analysis*, Multivariate Behaviorial Research, 5 (1970), pp. 329–350.

[15] S. Zacks, *The Theory of Statistical Inference*, John Wiley, New York, 1971.