General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

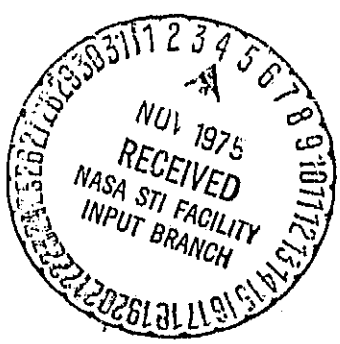
- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NASA CR-144463

AN ITER PROC FOR OBTAINING MAX LIKELIHOOD ESTIMATES OF THE PARAM FOR A MIXTURE OF NORMAL DISTS
BY B. PETERS,JR.,& H. WALKER
REPORT #43 JULY, 1975



PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-12777

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions

July, 1975

by

B. Charles Peters, Jr.

NASA/National Research Council Research Associate

Earth Observations Division, Johnson Space Center

and

Homer F. Walker

Department of Mathematics, University of Houston

Report 43

NAS-9-12777

An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions

by

B. Charles Peters, Jr.
NASA/National Research Council Research Associate
Earth Observations Division, Johnson Space Center

and

Homer F. Walker
Department of Mathematics, University of Houston
Houston, Texas

1. Introduction.

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

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

where

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

and

$$p_i(x) = \frac{1}{(2\pi)^{n/2}|\Sigma_i^o|^{1/2}} e^{-1/2(x-\mu_i^o)^T \Sigma_i^{o\,-1}(x-\mu_i^o)}.$$

If $\{x_k\}_{k=1,\ldots,N} \subseteq \mathbb{R}^n$ is an independent sample of observations on x, then a maximum-likelihood estimate of the parameters $\{\alpha_i^o, \mu_i^o, \Sigma_i^o\}_{i=1,\ldots,m}$ is a choice of parameters $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$ 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 $\{\alpha_i^o, \mu_i^o, \Sigma_i^o\}_{i=1,\ldots,m}$ replaced by the estimate $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$. (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.)

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 conditions for a maximum-likelihood estimate:

$$\left.\begin{aligned}
\text{(1.a)} \quad & \alpha_i = \frac{\alpha_i}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} \\
\text{(1.b)} \quad & \mu_i = \left\{\frac{1}{N} \sum_{k=1}^{N} x_k \frac{p_i(x_k)}{p(x_k)}\right\} \Big/ \left\{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right\} \\
\text{(1.c)} \quad & \Sigma_i = \left\{\frac{1}{N} \sum_{k=1}^{N} (x_k-\mu_i)(x_k-\mu_i)^T \frac{p_i(x_k)}{p(x_k)}\right\} \Big/ \left\{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right\}
\end{aligned}\right\} i=1,\ldots,m$$

These are known as the likelihood equations. As observed by Cramér [2], Huzurbazar [7], Wald [11], Chanda [1], and others, there is, loosely speaking, a unique solution of the likelihood equations which tends in probability to the true parameters as the sample size N approaches infinity. Furthermore, this solution is a maximum-likelihood estimate, indeed, the unique consistent maximum-likelihood estimate. (Strictly speaking, given any sufficiently small neighborhood of the true parameters, there is, with probability tending to 1 as N approaches infinity, a unique solution of the likelihood equations in that neighborhood, and this solution is a maximum-likelihood estimate. For completeness, we present a brief proof of this result in an appendix.) This note is addressed to the problem of determining this 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 (1.a), (1.b), and (1.c). 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 [8] and Wolfe [13] as well as the authors given below.

Empirical studies of Day [3], Duda and Hart [4], and Hasselblad [5] 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 [4]), which are not the consistent maximum-likelihood estimate if care is not taken in the choice of starting values. No theoretical evidence of convergence is given in [3], [4], or [5].

Peters and Coberly [10] have proved that, if all of the parameters μ_i and Σ_i are held fixed, then the iterative procedure suggested by the equation (1.a) alone converges locally to a maximum-likelihood estimate of the parameters α_i, $i=1,\ldots,m$. (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. Walker [12] has shown that, if all the parameters α_i and Σ_i are held fixed, then the iterative procedure suggested by the equation (1.b) converges locally to a maximum-likelihood estimate of the means μ_i, $i=1,\ldots,m$, provided that either $m = 2$ or the component normal densities in p are "widely separated" in a certain sense.

In the following, we present a general iterative procedure for determining the consistent maximum-likelihood estimate, of which the above procedure is a special case. Indeed, our procedure is in some ways like a steepest-ascent method, and the above procedure is obtained when a certain "step-size" is taken to be 1. We show that, if the "step-size" is sufficiently small, then with probability approaching 1 as the sample size approaches infinity, this procedure converges locally to the consistent maximum-likelihood estimate. This scheme is as easily implemented in general as in the above special case, and it appears to hold considerable promise as an effective tool for obtaining consistent maximum-likelihood estimates in many situations of practical interest.

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, $1 \le i \le m$, 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 $\mathcal{A}, \mathcal{M}$, and $\mathcal{S}$ the respective m-fold direct sums of these spaces with themselves, and we represent elements of $\mathcal{A}, \mathcal{M}$, and $\mathcal{S}$ as columns

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

It will be convenient to represent elements of the direct sum $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$ as either

$$\begin{pmatrix} \overline{\alpha} \\ \overline{\mu} \\ \overline{\Sigma} \end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \\ \mu_1 \\ \vdots \\ \mu_m \\ \Sigma_1 \\ \vdots \\ \Sigma_m \end{pmatrix}$$

If, for $i = 1, \ldots, m$, we denote

$$A_i(\overline{\alpha}; \overline{\mu}, \overline{\Sigma}) = \frac{\alpha_i}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)},$$

$$M_i(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \frac{1}{N}\sum_{k=1}^{N} x_k \frac{p_i(x_k)}{p(x_k)} \Big/ \frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} ,$$

$$S_i(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \frac{1}{N}\sum_{k=1}^{N} (x_k-\mu_i)(x_k-\mu_i)^T \frac{p_i(x_k)}{p(x_k)} \Big/ \frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} ,$$

then the likelihood equations can be written as

$$(2) \qquad \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} = \begin{pmatrix} A(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ M(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ S(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \end{pmatrix} ,$$

where

$$A(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \begin{pmatrix} A_1(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ \vdots \\ A_m(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \end{pmatrix}, \quad M(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \begin{pmatrix} M_1(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ \vdots \\ M_m(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \end{pmatrix}, \quad S(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \begin{pmatrix} S_1(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ \vdots \\ S_m(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \end{pmatrix}$$

One can write (2) equivalently as

$$(3) \qquad \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} = \Phi_\epsilon(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \equiv (1-\epsilon) \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} + \epsilon \begin{pmatrix} A(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ M(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ S(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \end{pmatrix}$$

for any value of ϵ. (Of course, (3) becomes (2) when $\epsilon = 1$.) The following iterative procedure is suggested by (3) for obtaining a solution of the likelihood equations: Beginning with some starting value $\begin{pmatrix} \bar{\alpha}^{(1)} \\ \bar{\mu}^{(1)} \\ \bar{\Sigma}^{(1)} \end{pmatrix}$, define successive iterates inductively by

$$(4) \qquad \begin{pmatrix} \bar{\alpha}^{(k+1)} \\ \bar{\mu}^{(k+1)} \\ \bar{\Sigma}^{(k+1)} \end{pmatrix} = \Phi_\epsilon \left(\bar{\alpha}^{(k)}, \bar{\mu}^{(k)}, \bar{\Sigma}^{(k)}\right)$$

for $k = 1,2,3,\ldots$. This procedure becomes the procedure given in the introduction when $\epsilon = 1$.

In the next section, we show that if ϵ is a sufficiently small positive number, then, with probability approaching 1 as the sample size N approaches infinity, this procedure converges locally to the consistent maximum-likelihood estimate. This is done by showing that, with probability approaching 1 as N approaches infinity, the operator Φ_ϵ is <u>locally contractive</u> (in a suitable vector norm) near that estimate, provided ϵ is a sufficiently small positive number. In saying that Φ_ϵ is locally contractive near a point $\begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$, we mean that there is a vector norm $\| \ \|$ on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$ and a number λ, $0 \leq \lambda < 1$ such that

$$(5) \qquad \left\| \Phi_\epsilon(\bar{\alpha}', \bar{\mu}', \bar{\Sigma}') - \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} \right\| \leq \lambda \left\| \begin{pmatrix} \bar{\alpha}' \\ \bar{\mu}' \\ \bar{\Sigma}' \end{pmatrix} - \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} \right\|$$

whenever $\begin{pmatrix} \bar{\alpha}' \\ \bar{\mu}' \\ \bar{\Sigma}' \end{pmatrix}$ lies sufficiently near $\begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix}$

3. <u>The local contractibility and convergence results</u>.

We now establish the following

Theorem. With probability approaching 1 as N approaches infinity, Φ_ϵ is a locally contractive operator (in some norm on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$) near the consistent maximum-likelihood estimate whenever ϵ is a sufficiently small positive number.

Our main result is an immediate consequence of this theorem, which we state as a

Corollary. With probability approaching 1 as N approaches infinity, the iterative procedure (3) converges locally to the consistent maximum-likelihood estimate whenever ϵ is a sufficiently small positive number.

Throughout the proof of the theorem, 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 [9].

Proof of the theorem: Let $\begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix}$ be the consistent maximum-likelihood estimate. We assume that $\alpha_i \neq 0$, $i = 1,\ldots,m$. (As N tends to infinity, the probability approaches 1 that this is the case.) It must be shown that, with probability approaching 1 as N approaches infinity, an inequality of the form (5) holds whenever ϵ is a sufficiently small positive number.

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

$$\Phi_\epsilon(\bar{\alpha}',\bar{\mu}',\bar{\Sigma}') - \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} = \nabla \Phi_\epsilon(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \left[\begin{pmatrix} \bar{\alpha}' \\ \bar{\mu}' \\ \bar{\Sigma}' \end{pmatrix} - \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} \right] +$$

$$+ 0\left(\left\| \begin{pmatrix} \bar{\alpha}' \\ \bar{\mu}' \\ \bar{\Sigma}' \end{pmatrix} - \begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} \right\|^2 \right)$$

Consequently, the theorem will be proved if it can be shown that, for small positive ϵ, $\nabla \Phi_\epsilon(\bar{\alpha},\bar{\mu},\bar{\Sigma})$ converges in probability to an operator which has norm less than 1 with respect to a suitable norm on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$.

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

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

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

The entries of the above matrix can themselves be represented as matrices of Fréchet derivatives. For $i = 1,\ldots,m$, we introduce inner products $\langle x,y\rangle_i' = x^T(\alpha_i \Sigma_i^{-1})y$ on $\mathbb{R}^n$ and $\langle A,B\rangle_i'' = \mathrm{tr}\{A(\frac{\alpha_i}{2}\Sigma_i^{-1})B^T\}$ on the space of real, symmetric $n \times n$ matrices. After a straightforward but extremely tedious calculation, one obtains with the aid of equations (1) that

$$\nabla_{\bar{\alpha}} A(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = I - (\mathrm{diag}\ \alpha_i)\{\frac{1}{N}\sum_{k=1}^{N} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)} \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)} \end{pmatrix} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)} \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)} \end{pmatrix}^T \}$$

$$\nabla_{\bar{\mu}} A(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = -(\mathrm{diag}\ \alpha_i)\{\frac{1}{N}\sum_{k=1}^{N} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)} \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)} \end{pmatrix} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)} \langle x_k-\mu_1,\cdot\rangle_1' \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)} \langle x_k-\mu_m,\cdot\rangle_m' \end{pmatrix}^T \}.$$

$$\nabla_{\Sigma}\Lambda(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = -(\operatorname{diag}\alpha_i)\left\{\frac{1}{N}\sum_{k=1}^{N}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}\end{pmatrix}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\left\langle\Sigma_1^{-1}(x_k-\mu_1)(x_k-\mu_1)^T - I,\cdot\right\rangle_1'' \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}\left\langle\Sigma_m^{-1}(x_k-\mu_m)(x_k-\mu_m)^T - I,\cdot\right\rangle_m''\end{pmatrix}^T\right\}$$

$$\nabla_{\bar{\alpha}}M(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = -\left\{\frac{1}{N}\sum_{k=1}^{N}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}(x_k-\mu_1)\\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}(x_k-\mu_m)\end{pmatrix}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}\end{pmatrix}^T\right\}$$

$$\nabla_{\bar{\mu}}M(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = I \otimes \left\{\frac{1}{N}\sum_{k=1}^{N}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}(x_k-\mu_1)\\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}(x_k-\mu_m)\end{pmatrix}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\left\langle x_k-\mu_1,\cdot\right\rangle_1' \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}\left\langle x_k-\mu_m,\cdot\right\rangle_m'\end{pmatrix}^T\right\}$$

$$\nabla_{\bar{\Sigma}}M(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = (\operatorname{diag}\frac{1}{\alpha_i N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)}(x_k-\mu_i)\left\langle\Sigma_i^{-1}(x_k-\mu_i)(x_k-\mu_i)^T - I,\cdot\right\rangle_i'' -$$

$$-\left\{\frac{1}{N}\sum_{k=1}^{N}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}(x_k-\mu_1)\\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}(x_k-\mu_m)\end{pmatrix}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\left\langle\Sigma_1^{-1}(x_k-\mu_1)(x_k-\mu_1)^T - I,\cdot\right\rangle_1'' \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}\left\langle\Sigma_m^{-1}(x_k-\mu_m)(x_k-\mu_m)^T - I,\cdot\right\rangle_m''\end{pmatrix}^T\right\}$$

$$\nabla_{\bar{\alpha}} S(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = -(\operatorname{diag}\Sigma_i)\{\tfrac{1}{N}\sum_{k=1}^{N}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}[\Sigma_1^{-1}(x_k-\mu_1)(x_k-\mu_1)^T-I]\\ \vdots\\ \frac{p_m(x_k)}{p(x_k)}[\Sigma_m^{-1}(x_k-\mu_m)(x_k-\mu_m)^T-I]\end{pmatrix}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\\ \vdots\\ \frac{p_m(x_k)}{p(x_k)}\end{pmatrix}^T\}$$

$$\nabla_{\bar{\mu}} S(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = -(\operatorname{diag}\Sigma_i)\{\tfrac{1}{N}\sum_{k=1}^{N}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}[\Sigma_1^{-1}(x_k-\mu_1)(x_k-\mu_1)^T-I]\\ \vdots\\ \frac{p_m(x_k)}{p(x_k)}[\Sigma_m^{-1}(x_k-\mu_m)(x_k-\mu_m)^T-I]\end{pmatrix}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\langle x_k-\mu_1,\cdot\rangle_1'\\ \vdots\\ \frac{p_m(x_k)}{p(x_k)}\langle x_k-\mu_m,\cdot\rangle_m'\end{pmatrix}^T\}$$

$$\nabla_{\bar{\Sigma}} S(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = (\operatorname{diag}\Sigma_i\,\frac{1}{\alpha_i N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)}[\Sigma_i^{-1}(x_k-\mu_i)(x_k-\mu_i)^T-I]\langle\Sigma_i^{-1}(x_k-\mu_i)(x_k-\mu_i)^T-I,\cdot\rangle_i'') -$$

$$-(\operatorname{diag}\Sigma_i)\{\tfrac{1}{N}\sum_{k=1}^{N}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}[\Sigma_1^{-1}(x_k-\mu_1)(x_k-\mu_1)^T-I]\\ \vdots\\ \frac{p_m(x_k)}{p(x_k)}[\Sigma_m^{-1}(x_k-\mu_m)(x_k-\mu_m)^T-I]\end{pmatrix}\begin{pmatrix}\frac{p_1(x_k)}{p(x_k)}\langle\Sigma_1^{-1}(x_k-\mu_1)(x_k-\mu_1)^T-I,\cdot\rangle_1''\\ \vdots\\ \frac{p_m(x_k)}{p(x_k)}\langle\Sigma_m^{-1}(x_k-\mu_m)(x_k-\mu_m)^T-I,\cdot\rangle_m''\end{pmatrix}^T\}.$$

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 $\langle\cdot,\cdot\rangle$ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$. Setting

$$V(x) = \begin{pmatrix} \dfrac{p_1(x)}{p(x)} & & \\ \vdots & & \\ \dfrac{p_m(x)}{p(x)} & & \\ \dfrac{\Gamma_1(x)}{p(x)}(x-\mu_1) & & \\ \vdots & & \\ \dfrac{p_m(x)}{p(x)}(x-\mu_m) & & \\ \dfrac{p_1(x)}{p(x)}[\Sigma_1^{-1}(x-\mu_i)(x-\mu_1)^T - I] & & \\ \vdots & & \\ \dfrac{p_m(x)}{p(x)}[\Sigma_m^{-1}(x-\mu_m)(x-\mu_m)^T - I] & & \end{pmatrix} \in \mathcal{K} \oplus \mathcal{M} \oplus \mathcal{S},$$

one obtains

$$\mathcal{V}\Phi_\epsilon(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \begin{pmatrix} I & 0 & 0 \\ 0 & I & \epsilon(\operatorname{diag}\frac{1}{\alpha_i N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)}(x_k-\mu_i)\langle\Sigma_i^{-1}(x_k-\mu_i)(x_k-\mu_i)^T-I,\cdot\rangle_i'') \\ 0 & 0 & (1-\epsilon)I + \epsilon(\operatorname{diag}\Sigma_i\frac{1}{\alpha_i N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)}[\Sigma_i^{-1}(x_k-\mu_i)(x_k-\mu_i)^T-I]\cdot \\ & & \cdot\langle\Sigma_i^{-1}(x_k-\mu_i)(x_k-\mu_i)^T - I,\cdot\rangle_i'') \end{pmatrix}$$

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

We have assumed that the solution $\begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix}$ of the likelihood equations is consistent. Denoting the true parameters by $\begin{pmatrix} \bar{\alpha}^o \\ \bar{\mu}^o \\ \bar{\Sigma}^o \end{pmatrix}$, one verifies without difficulty that $\nabla\Phi_\epsilon(\bar{\alpha},\bar{\mu},\bar{\Sigma})$ converges in probability to $E(\nabla\Phi_\epsilon(\bar{\alpha}^o,\bar{\mu}^o,\bar{\Sigma}^o))$ as N approaches infinity. A straightforward calculation yields

$$E(\nabla\Phi_\epsilon(\bar{\alpha}^o,\bar{\mu}^o,\bar{\Sigma}^o)) = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} -$$

$$\cdot \epsilon \begin{pmatrix} (\text{diag } \alpha_i^o) & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & (\text{diag } \Sigma_i^o) \end{pmatrix} \{ \int_{\mathbb{R}^n} V^o(x) \langle V^o(x), \cdot \rangle p^o(x)\,dx \}.$$

(In this expression, the superscript "o" on V and p indicates that the true parameters are used in evaluating these functions.) Thus $E(\nabla\Phi_\epsilon(\bar{\alpha}^o,\bar{\mu}^o,\bar{\Sigma}^o))$ is an operator on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$ of the form $I - \epsilon QR$, where Q and R are positive-definite and symmetric 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 $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$, it must be the case that, for small positive ϵ, the operator norm of $E(\nabla\Phi_\epsilon(\bar{\alpha}^o,\bar{\mu}^o,\bar{\Sigma}^o))$, with respect to the inner product $\langle \cdot, Q^{-1} \cdot \rangle$, is less than 1. So, for small positive ϵ, $\nabla\Phi_\epsilon(\bar{\alpha},\bar{\mu},\bar{\Sigma})$ converges in probability to an operator having norm less than 1 with respect to the inner product $\langle \cdot, Q^{-1} \cdot \rangle$ on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$. This completes the proof of the theorem.

We remark that, in order for the conclusion of the theorem to hold, it is sufficient to take ϵ less than $\frac{4}{m(n+1)(n+2)}$. Indeed, it is observed in the proof of the theorem that $E(\nabla\Phi_\epsilon(\bar{\alpha}^o,\bar{\mu}^o,\bar{\Sigma}^o)) = I - \epsilon QR$, where QR is positive-definite and symmetric with respect to a certain inner product and, hence, has positive eigenvalues. Denoting the spectral radius of QR by $\rho(QR)$, one then verifies that $E(\nabla\Phi_\epsilon(\bar{\alpha}^o,\bar{\mu}^o,\bar{\Sigma}^o))$ has operator norm less than 1, with respect to some vector norm, whenever ϵ is less than $\frac{2}{\rho(QR)}$. (See [6].) Now

$$\begin{aligned}
\rho(QR) &< \mathrm{tr}\{QR\} \\
&= \sum_{i=1}^m \alpha_i \int_{\mathbb{R}^n} \frac{p_i(x)^2}{p(x)}\,dx + \sum_{i=1}^m \mathrm{tr}\Big\{\int_{\mathbb{R}^n} \frac{p_i(x)^2}{p(x)}(x-\mu_i)\langle x-\mu_i,\cdot\rangle_i'\,dx\Big\} \\
&\quad + \sum_{i=1}^m \mathrm{tr}\Big\{\Sigma_i \int_{\mathbb{R}^n} \frac{p_i(x)^2}{p(x)}[\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T - I]\langle \Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T - I,\cdot\rangle_i''\,dx\Big\} \\
&< \sum_{i=1}^m \int_{\mathbb{R}^n} p_i(x)\,dx + \sum_{i=1}^m \int_{\mathbb{R}^n} (x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)p_i(x)\,dx \\
&\quad + \sum_{i=1}^m \int_{\mathbb{R}^n} \frac{1}{2}\,\mathrm{tr}\{(\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T - I)^2\}p_i(x)\,dx \\
&= m + mn + \frac{m}{2}(n^2+n) = \frac{m(n+1)(n+2)}{2}\,.
\end{aligned}$$

It follows that the conclusion of the theorem holds whenever $\epsilon < \frac{4}{m(n+1)(n+2)}$.

4. 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 [8] for obtaining solutions of the likelihood equations. It is our feeling

that the iterative procedure presented here offers considerable computational advantages over these procedures in many cases of practical interest.

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 $\frac{m(n+1)(n+2)}{2} - 1$. 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.

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) appears to have two particular advantages over the steepest-ascent method. First, the successive iterates defined by (4) automatically satisfy the requisite constraints on the parameters, i.e., the successive Σ_i's are, in probability, positive-definite and the successive α_i's are positive and sum to 1. Second, by the remarks following the proof of the theorem, one knows that, in probability, there is a value of ϵ, depending only on m and n, for which the procedure (4) converges locally to the consistent maximum-likelihood estimate. We doubt that there exists a step-size depending only on m and n which is similarly sufficient for the local convergence of the steepest-ascent procedure.

Appendix

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

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

$$\frac{\partial L}{\partial \theta_i} = 0, \quad i = 1,\ldots,\nu,$$

be satisfied, where θ_i is the $i^{\underline{th}}$ component of Θ. In the following, our objective is to show that if p satisfies certain conditions, then, given any sufficiently small neighborhood of Θ^o, there is, with probability approaching 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^o.

We assume that $p(x,\Theta)$ satisfies the following conditions of Chanda [1]:

(a) There is a neighborhood Ω of Θ^o such that for all $\Theta \in \Omega$, for almost

all $x \in \mathbb{R}^n$, and for $i,j,k=1,\ldots,\nu$, $\frac{\partial p}{\partial \theta_i}$, $\frac{\partial^2 p}{\partial \theta_i \partial \theta_j}$, and $\frac{\partial^3 p}{\partial \theta_i \partial \theta_j \partial \theta_k}$ exist and satisfy

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

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

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

(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^o.

Let $\mathcal{L}(\Theta) = \begin{pmatrix} \frac{1}{N} \frac{\partial L}{\partial \theta_1} \\ \vdots \\ \frac{1}{N} \frac{\partial L}{\partial \theta_\nu} \end{pmatrix}$.

It is immediately seen that $\mathcal{L}(\Theta) = 0$ if and only if the likelihood equations are satisfied, and that, by the weak law of large numbers, $\mathcal{L}(\Theta^o)$ converges in probability to zero. Furthermore, it follows from assumptions (a) and (b) above that there exists a neighborhood Ω^o of Θ^o (contained in Ω and, for convenience, convex) and a positive ϵ such that, with probability approaching 1 as N approaches infinity, $\nabla\mathcal{L}(\Theta) \le -\epsilon I$ for all $\Theta \in \Omega^o$. (The inequality is with respect to the usual ordering on symmetric matrices.) Denoting the spherical neighborhood of radius δ about Θ^o by Ω_δ, we establish the following

Lemma: With probability approaching 1 as N approaches infinity,

(i) $\mathcal{L}$ is one-to-one on Ω^o,

(ii) $\mathcal{L}(\Omega_\delta)$ contains the ball of radius $\epsilon\delta$ about $\mathcal{L}(\theta^o)$ whenever $\Omega_\delta \subseteq \Omega^o$.

Proof: We may assume that $\nabla\mathcal{L}(\theta) \leq -\epsilon I$ for all $\theta \in \Omega^o$, since the probability that this is the case tends to 1 as N approaches infinity. To prove (i), suppose that $\mathcal{L}(\theta^1) = \mathcal{L}(\theta^2)$ for θ^1 and θ^2 in Ω^o. Then

$$\begin{aligned} 0 &= (\theta^1 - \theta^2)^T[\mathcal{L}(\theta^1) - \mathcal{L}(\theta^2)] \\ &= (\theta^1 - \theta^2)^T\{\int_0^1 \nabla\mathcal{L}(\theta^2 + t[\theta^1 - \theta^2])dt\}(\theta^1 - \theta^2). \end{aligned}$$

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

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

$$\mathcal{L}(\theta^1) - \mathcal{L}(\theta^o) = \{\int_0^1 \nabla\mathcal{L}(\theta^o + t[\theta^1 - \theta^o])dt\}(\theta^1 - \theta^o).$$

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

$$\left\|\mathcal{L}(\theta^1) - \mathcal{L}(\theta^o)\right\| \geq \epsilon \left\|\theta^1 - \theta^o\right\| = \epsilon\,\delta,$$

where $\|\ \|$ denotes the usual Euclidean norm on $\mathbb{R}^\nu$. Since all boundary points of $\mathcal{L}(\Omega_\delta)$ are images under $\mathcal{L}$ 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^o which is contained in Ω^o, then one can find a δ for which $\Omega_\delta \subseteq \Omega^1 \subseteq \Omega^o$. By the lemma, the probability approaches 1 as N tends to infinity that $\mathcal{L}$ is one-to-one on Ω^1 and that $\mathcal{L}(\Omega_\delta)$ and, hence, $\mathcal{L}(\Omega^1)$ contain the ball of radius $\epsilon\delta$ about $\mathcal{L}(\Theta^o)$. Since $\mathcal{L}(\Theta^o)$ converges in probability to zero, one concludes that, with probability tending to 1 as N approaches infinity, there exists a unique $\Theta \in \Omega^1$ for which $\mathcal{L}(\Theta) = 0$. Since the probability also tends to 1 that $\nabla\mathcal{L}$ is negative-definite on Ω^1, this Θ is, with probability approaching 1, a maximum-likelihood estimate.

BIBLIOGRAPHY

1. K. C. Chanda, "A note on the consistency and maxima of the roots of the likelihood equations," Biometrika 41 (1954), pp. 56-61.

2. H. Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946.

3. N. E. Day, "Estimating the components of a mixture of normal distributions," Biometrika 56 (1969), pp. 463-474.

4. R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, Inc., New York, 1973.

5. V. Hasselblad, "Estimation of parameters for a mixture of normal distributions," Technometrics 8 (1966),pp. 431-446.

6. A. S. Householder, Theory of Matrices and Numerical Analysis, Blaisdell Publishing Co., New York, 1964.

7. V. S. Huzurbazar, "The likelihood equation, consistency and the maxima of the likelihood function," Ann. Eugen., Lond. 14 (1948), pp. 185-200.

8. B. K. Kale, " On the solution of the likelihood equations by iteration processes. The multiparametric case," Biometrika 49 (1962), pp. 479-486.

9. D. G. Luenberger, Optimization by Vector Space Methods, John Wiley and Sons, Inc., New York, 1969.

10. B. C. Peters and W. H. Coberly, "The numerical evaluation of the maximum-likelihood estimate of mixture proportions," to appear.

11. A. Wald, "Note on the consistency of the maximum-likelihood estimate," Ann. Math. Stat. 20 (1949), p. 595.

12. H. F. Walker, "On the numerical evaluation of the maximum-likelihood estimate of mixture means," to appear.

13. J. H. Wolfe, "Pattern clustering by multivariate mixture analysis," Multivariate Behavioral Research 5 (1970), pp. 329-350.