NASA CR-
CR-147857

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

(NASA-CR-147857) THE NUMERICAL EVALUATION OF MAXIMUM-LIKELIHOOD ESTIMATES OF THE PARAMETERS FOR A MIXTURE OF NORMAL DISTRIBUTIONS FROM PARTIALLY IDENTIFIED SAMPLES (Houston Univ.) 17 p HC $3.50 N76-30891 Unclas G3/65 01742

THE NUMERICAL EVALUATION OF MAX LIKELIHOOD ESTIMATES OF THE PARAMETERS FOR A MIXTURE OF NORMAL DISTRIBUTIONS FROM PARTIALLY IDENTIFIED SAMPLES

HOMER F. WALKER

REPORT #54 JUNE, 1976

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-15000

HOUSTON, TEXAS 77004

The Numerical Evaluation of Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions from Partially Identified Samples

by

Homer F. Walker

Department of Mathematics, University of Houston

Houston, Texas 77004

June, 1976

Report #54

The Numerical Evaluation of Maximum-Likelihood
Estimates of the Parameters for a Mixture of Normal Distributions
from Partially Identified Samples

by

Homer F. Walker
Department of Mathematics, University of Houston
Houston, Texas 77004

1. Introduction.

Let $\pi_1,\ldots,\pi_m$ be populations whose multivariate observations in $\mathbb{R}^n$ are distributed with respective normal density functions

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

If π_o is a given mixture of members of these populations, then observations on π_o are distributed in $\mathbb{R}^n$ with density function

$$p(x) = \sum_{i=1}^m \alpha_i^o p_i(x)$$

for an appropriate set of proportions $\{\alpha_i^o\}_{i=1,-,m}$. These proportions necessarily satisfy $\sum_{i=1}^m \alpha_i^o = 1$ and $\alpha_i^o \geq 0$, $i = 1,-,m$. In this note, we also assume that each α_i^o is strictly positive.

We address here the problem of numerically approximating the maximum-likelihood estimates of the parameters $\{\alpha_i^o,\mu_i^o,\Sigma_i^o\}_{i=1,\ldots,m}$ determined by samples of two types. Samples of both types consist of sets $\{x_{ik}\}_{k=1,\ldots,N_i}$

of independent observations on π_i, $i = 0,\ldots,m$. (The sets $\{x_{ik}\}_{k=1,\ldots,N_i}$, $i = 1,\ldots,m$, comprise the identified observations of such samples, and such samples are said to be partially identified.) We distinguish samples of the two types according to whether the numbers N_i of identified observations contain information about the proportions α_i^o, $i = 1,\ldots,m$. If the numbers of identified observations contain no information about the proportions, then the sample is of the first type; otherwise, the sample is of the second type. The following are examples of how samples of the first and second types, respectively, might be obtained:

(1) For $i = 0,\ldots m$, numbers N_i are arbitrarily choosen and independent observations $\{x_{ik}\}_{k=1,-,N_i}$ are obtained from π_i.

(2) A number K_o of observations are obtained from π_o. For some $N_o < K_o$, N_o of these observations are left unidentified, while the remaining $K_o - N_o$ observations are identified. For $i = 1,\ldots,m$, a subset $\{x_{ik}\}_{k=1,\ldots,N_i}$ of the identified observations is determined whose member observations come from π_i.

In the following, we consider likelihood equations determined by the two types of samples which are necessary conditions for a maximum-likelihood estimate. These equations, which were derived by Coberly [1], suggest certain successive-approximations iterative procedures for obtaining maximum-likelihood estimates. These procedures, which are generalized steepest ascent (deflected gradient) procedures, contain those of Hosmer [2] as a special case. Using arguments that parallel those of [3], we show that, with probability 1 as

N_o approaches infinity (regardless of the relative sizes of N_o and N_i, $i = 1,\ldots,m$), these procedures converge locally to the strongly consistent maximum-likelihood estimates* 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.

2. Samples of the first type.

We first assume that numbers $\{N_i\}_{i=0,\ldots,m}$ are given and that, for $i = 0,\ldots,m$, N_i independent observations $\{x_{ik}\}_{k=1,\ldots,N_i}$ are drawn on π_i. The log-likelihood function for a sample of this type is

$$L_1(\Theta) = \sum_{i=1}^{m} \sum_{k=1}^{N_i} \log p_i(x_{ik}) + \sum_{k=1}^{N_0} \log p(x_{ok}) \ .$$

In this expression, the parameter vector Θ (with components α_i, μ_i, Σ_i, $i = 1,\ldots,m$) belongs to the vector space $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$ defined in [3], and the density functions on the right-hand side are evaluated with the true parameter vector Θ^o (with components α_i^o, μ_i^o, Σ_i^o, $i = 1,\ldots,m$) replaced by Θ.

---

*As in [3], one can show that, given any sufficiently small neighborhood of the true parameters, there is, with probability 1 as N_o approaches infinity (regardless of the relative sizes of N_o and N_i, $i = 1,\ldots,m$), a unique solution of the likelihood equations for either type of sample in that neighborhood, and this solution is a maximum-likelihood estimate.

Differentiating $L_1(\Theta)$ and setting its partial derivatives to zero gives the likelihood equations

(1.a) $\alpha_i = A_i(\Theta) \equiv \frac{\alpha_i}{N_o} \sum_{k=1}^{N_o} \frac{p_i(x_{ok})}{p(x_{ok})}$

(1.b) $\mu_i = M_i(\Theta) \equiv \{ \sum_{k=1}^{N_i} x_{ik} + \sum_{k=1}^{N_o} x_{ok} \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \} \Big/ \{ N_i + \sum_{k=1}^{N_o} \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \}$

(1.c) $\Sigma_i = S_i(\Theta) \equiv \{ \sum_{k=1}^{N_i} (x_{ik}-\mu_i)(x_{ik}-\mu_i)^T + \sum_{k=1}^{N_o} (x_{ok}-\mu_i)(x_{ok}-\mu_i)^T \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \} \Big/$

$$\{ N_i + \sum_{k=1}^{N_o} \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \}$$

for $i = 1,\ldots,m$.

We set

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

and define an operator Φ_ϵ on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$ by

$$\Phi_\epsilon(\Theta) = (1 - \epsilon)\Theta + \epsilon \begin{pmatrix} A(\Theta) \\ M(\Theta) \\ S(\Theta) \end{pmatrix} .$$

Clearly, for any non-zero ϵ, the likelihood equations are satisfied by a vector $\Theta \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$ if and only if $\Theta = \Phi_\epsilon(\Theta)$.

We consider the following iterative procedure: Beginning with some starting value $\Theta^{(1)}$, define successive iterates inductively by

(2) $$\Theta^{(j+1)} = \Phi_\epsilon(\Theta^{(j)})$$

for $j = 1, 2, 3, \ldots$. Our local convergence result for this iterative procedure, as stated in the introduction, follows immediately from the theorem below.

Theorem 1: With probability 1 as N_0 approaches infinity, Φ_ϵ is a locally contractive operator (in some norm on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$) near the strongly consistent maximum-likelihood estimate whenever $0 < \epsilon < 2$.

In saying that Φ_ϵ is a locally contractive operator near a point $\Theta \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$, 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

$$||\Phi_\epsilon(\Theta') - \Theta|| \leq \lambda||\Theta' - \Theta||$$

whenever Θ' lies sufficiently near Θ.

Proof of Theorem 1: Let

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

be the strongly consistent maximum-likelihood estimate. We assume that

$\alpha_i \neq 0$, $i = 1,\ldots,m$. (As N_o approaches infinity, the probability is 1 that this is the case.) As in [3], it suffices to show that, with probability 1, $\nabla\Phi_\epsilon(\Theta)$ converges to an operator which has operator norm less than 1 with respect to a suitable vector norm on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$.

Now

$$\nabla\Phi_\epsilon(\Theta) = (1-\epsilon)I + \epsilon\,\nabla\begin{pmatrix} A(\Theta) \\ M(\Theta) \\ S(\Theta) \end{pmatrix} ,$$

and we write

$$\nabla\begin{pmatrix} A \\ M \\ S \end{pmatrix} = \begin{pmatrix} \nabla_{\overline{\alpha}}A & \nabla_{\overline{\mu}}A & \nabla_{\overline{\Sigma}}A \\ \nabla_{\overline{\alpha}}M & \nabla_{\overline{\mu}}M & \nabla_{\overline{\Sigma}}M \\ \nabla_{\overline{\alpha}}S & \nabla_{\overline{\mu}}S & \nabla_{\overline{\Sigma}}S \end{pmatrix} .$$

Define inner products $\langle\ ,\ \rangle_i'$ on $\mathcal{M}$, $\langle\ ,\ \rangle_i''$ on $\mathcal{S}$, and $\langle\ ,\ \rangle$ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ as in [3]. Setting

$$\beta_i(x) = \frac{p_i(x)}{p(x)},\ \gamma_i(x) = (x-\mu_i),\ \delta_i(x) = [\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T - I], K_i = N_i + \alpha_i N_o$$

for $i = 1,\ldots,m$, one calculates

$$\nabla_{\overline{\alpha}}A(\Theta) = I - (\text{diag}\ \alpha_i)\ \frac{1}{N_o}\sum_1^{N_o}\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}^T$$

$$\nabla_{\overline{\mu}}A(\Theta) = -(\text{diag}\ \alpha_i)\ \frac{1}{N_o}\sum_1^{N_o}\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}\begin{pmatrix} \langle\beta_1\gamma_1,\cdot\rangle_1' \\ \vdots \\ \langle\beta_m\gamma_m,\cdot\rangle_m' \end{pmatrix}^T$$

$$\nabla_{\overline{\Sigma}}A(\Theta) = -(\text{diag}\ \alpha_i)\ \frac{1}{N_o}\sum_1^{N_o}\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}\begin{pmatrix} \langle\beta_1\delta_1,\cdot\rangle_1'' \\ \vdots \\ \langle\beta_m\delta_m,\cdot\rangle_m'' \end{pmatrix}^T$$

$$\nabla_{\underline{\alpha}} M(\Theta) = (\text{diag}\ \frac{1}{K_i}\ \sum_1^{N_o} \beta_i\gamma_i) - (\text{diag}\ \frac{\alpha_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix}\beta_1\gamma_1\\ \vdots \\ \beta_m\gamma_m\end{pmatrix} \begin{pmatrix}\beta_1\\ \vdots\\ \beta_m\end{pmatrix}^T \right\}$$

$$\nabla_{\underline{\mu}} M(\Theta) = (\text{diag}\ \frac{\alpha_i}{K_i}\ \sum_1^{N_o} \gamma_i\gamma_i^T\Sigma_i^{-1}\beta_i) - (\text{diag}\ \frac{\alpha_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix}\beta_1\gamma_1\\ \vdots \\ \beta_m\gamma_m\end{pmatrix} \begin{pmatrix}\langle\beta_1\gamma_1,\cdot\rangle_1'\\ \vdots\\ \langle\beta_m\gamma_m,\cdot\rangle_m'\end{pmatrix}^T \right\}$$

$$\nabla_{\underline{\Sigma}} M(\Theta) = (\text{diag}\ \frac{1}{K_i}\ \sum_1^{N_o} \beta_i\gamma_i\langle\delta_i,\cdot\rangle_i'') - (\text{diag}\ \frac{\alpha_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix}\beta_1\gamma_1\\ \vdots \\ \beta_m\gamma_m\end{pmatrix} \begin{pmatrix}\langle\beta_1\delta_1,\cdot\rangle_1''\\ \vdots\\ \langle\beta_m\delta_m,\cdot\rangle_m''\end{pmatrix}^T \right\}$$

$$\nabla_{\underline{\alpha}} S(\Theta) = (\text{diag}\ \frac{\Sigma_i}{K_i}\ \sum_1^{N_o} \beta_i\delta_i) - (\text{diag}\ \frac{\alpha_i\Sigma_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix}\beta_1\delta_1\\ \vdots \\ \beta_m\delta_m\end{pmatrix} \begin{pmatrix}\beta_1\\ \vdots\\ \beta_m\end{pmatrix}^T \right\}$$

$$\nabla_{\underline{\mu}} S(\Theta) = (\text{diag}\ \frac{1}{K_i}\ \{-\sum_1^{N_i}[(\cdot)\gamma_i^T+\gamma_i(\cdot)^T] - \alpha_i\sum_1^{N_o}[(\cdot)\gamma_i^T+\gamma_i(\cdot)^T]\beta_i+\Sigma_i\sum_1^{N_o}\delta_i\langle\beta_i\gamma_i,\cdot\rangle_i'\}) -$$

$$- (\text{diag}\ \frac{\alpha_i\Sigma_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix}\beta_1\delta_1\\ \vdots \\ \beta_m\delta_m\end{pmatrix} \begin{pmatrix}\langle\beta_1\gamma_1,\cdot\rangle_1'\\ \vdots\\ \langle\beta_m\gamma_m,\cdot\rangle_m'\end{pmatrix}^T \right\}$$

$$\nabla_{\underline{\Sigma}} S(\Theta) = (\text{diag}\ \frac{\Sigma_i}{K_i}\ \sum_1^{N_o} \beta_i\delta_i\langle\delta_i,\cdot\rangle_i'') - (\text{diag}\ \frac{\alpha_i\Sigma_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix}\beta_1\delta_1\\ \vdots \\ \beta_m\delta_m\end{pmatrix} \begin{pmatrix}\langle\beta_1\delta_1,\cdot\rangle\\ \vdots\\ \langle\beta_m\delta_m,\cdot\rangle\end{pmatrix}^T \right\}$$

Here, the arguments of β_i, γ_i and δ_i can be determined from the indices of summation, e.g.,

$$\sum_1^{N_o} \beta_i\gamma_i = \sum_{k=1}^{N_o} \beta_i(x_{ok})\gamma_i(x_{ok}) .$$

Setting

$$V = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \beta_1\gamma_1 \\ \vdots \\ \beta_m\gamma_m \\ \beta_1\delta_1 \\ \vdots \\ \beta_m\delta_m \end{pmatrix}$$

one obtains at Θ

$$\nabla \begin{pmatrix} A \\ M \\ S \end{pmatrix} = \begin{pmatrix} I & 0 & 0 \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix} - \begin{pmatrix} (\mathrm{diag}\, \frac{\alpha_i}{N_O}) & 0 & 0 \\ 0 & (\mathrm{diag}\, \frac{\alpha_i}{K_i}) & 0 \\ 0 & 0 & (\mathrm{diag}\, \frac{\alpha_i \Sigma_i}{K_i}) \end{pmatrix} \{\sum_1^{N_O} V(x_{ok}) \cdot V(x_{ok}), \cdot\} ,$$

where

$$B_{21} = (\mathrm{diag}\, \frac{1}{K_i} \sum_1^{N_O} \beta_i \gamma_i)$$

$$B_{22} = (\mathrm{diag}\, \frac{\alpha_i}{K_i} \sum_1^{N_O} \gamma_i \gamma_i^T \Sigma_i^{-1} \beta_i)$$

$$B_{23} = (\mathrm{diag}\, \frac{1}{K_i} \sum_1^{N_O} \beta_i \gamma_i <\delta_i, \cdot>''_i)$$

$$B_{31} = (\mathrm{diag}\, \frac{\Sigma_i}{K_i} \sum_1^{N_O} \beta_i \delta_i)$$

$$B_{32} = (\mathrm{diag}\, \frac{1}{K_i} \{- \sum_1^{N_i} [(\cdot)\gamma_i^T + \gamma_i (\cdot)^T] - \alpha_i \sum_1^{N_O} [(\cdot)\gamma_i^T + \gamma_i (\cdot)^T] \beta_i + \Sigma_i \sum_1^{N_O} \delta_i <\beta_i \gamma_i, \cdot>'_i \})$$

$$B_{33} = (\mathrm{diag}\, \frac{\Sigma_i}{K_i} \sum_1^{N_O} \beta_i \delta_i <\delta_i, \cdot>''_i) .$$

We have assumed that Θ is the strongly consistent maximum-likelihood estimate. Then, regardless of the relative sizes of N_i and N_o, one can show as in [3] that, with probability 1, $\{\nabla\Phi_\epsilon(\Theta) - E(\nabla\Phi_\epsilon(\Theta^o))\}$ converges to zero as N_o approaches infinity. Now

$$E\left(\nabla \begin{pmatrix} A(\Theta^o) \\ M(\Theta^o) \\ S(\Theta^o) \end{pmatrix}\right) = \begin{pmatrix} I & 0 & 0 \\ 0 & (\mathrm{diag}\, \frac{\alpha_i^o N_o}{K_i}\, I) & 0 \\ 0 & 0 & (\mathrm{diag}\, \frac{\alpha_i N_o}{K_i}\, I) \end{pmatrix} - \ldots$$

$$- \begin{pmatrix} (\mathrm{diag}\, \alpha_i^o) & 0 & 0 \\ 0 & (\mathrm{diag}\, \frac{\alpha_i^o N_o}{K_i}\, I) & 0 \\ 0 & 0 & (\mathrm{diag}\, \frac{\alpha_i^o N_o}{K_i}\, \Sigma_i^o) \end{pmatrix} \{ \int_{\mathbb{R}^n} V(x) \langle V(x), \cdot \rangle p(x)dx \}$$

$$= B(I - QR),$$

where

$$B = \begin{matrix} I & 0 & 0 \\ 0 & (\mathrm{diag}\, \frac{\alpha_i^o N_o}{K_i}\, I) & 0 \\ 0 & 0 & (\mathrm{diag}\, \frac{\alpha_i^o N_o}{K_i}\, I) \end{matrix}$$

$$Q = \begin{matrix} (\mathrm{diag}\, \alpha_i^o) & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & (\mathrm{diag}\, \Sigma_i^o) \end{matrix}$$

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

It was shown in [3] that QR is positive-definite and symmetric with operator norm less than 1 with respect to the inner product $\langle \cdot, Q^{-1} \cdot \rangle$ on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$. It follows that $I-QR$ is positive-definite and symmetric with norm less than 1 with respect to $\langle \cdot, Q^{-1} \cdot \rangle$. Since B and Q commute, $\langle \cdot, Q^{-1}B^{-1} \cdot \rangle$ is an inner product on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$, and one sees that $\langle W, Q^{-1}W \rangle \le \langle W, Q^{-1}B^{-1}W \rangle$ for $W \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$. Consequently, $B(I-QR)$ is positive-definite and symmetric with norm less than 1 with respect to the inner product $\langle \cdot, Q^{-1}B^{-1} \cdot \rangle$. One concludes that

$$E(\nabla\Phi_{\epsilon}(\Theta^{o})) = (1 - \epsilon)I + \epsilon\, E(V \begin{pmatrix} A(\Theta^{o}) \\ M(\Theta^{o}) \\ S(\Theta^{o}) \end{pmatrix})$$

has norm less than 1 with respect to $\langle \cdot, Q^{-1}B^{-1} \cdot \rangle$ whenever $0 < \epsilon < 2$. This completes the proof of the theorem.

We remark that, reasoning as in [3], one may determine a particular value of ϵ (the "optimal ϵ") which yields, with probability 1 as N_o approaches infinity, the fastest asymptotic uniform rates of local convergence of the iterative procedure (2) near Θ. This optimal ϵ is given by

$$\epsilon = \frac{2}{2 - (\tau+\rho)}$$

where ρ and τ are, respectively the largest and smallest eigenvalues of $B(I-QR)$ regarded as an operator on $\mathcal{E} \oplus \mathcal{M} \oplus \mathcal{S}$ ($\mathcal{E}$ is the subspace of $\mathcal{A}$ whose components sum to zero.) Since ρ and τ lie between zero and 1, one sees that the optimal ϵ is always greater than 1. If the component populations are "widely separated," then ρ and τ are near zero and, hence, the optimal ϵ is near 1. If two or more of the component populations are nearly indistinguishable and if N_o is large relative to the N_i's, then τ is near zero, and the optimal ϵ cannot be much smaller than 2.

3. Samples of the second type.

We now assume that K_o observations are obtained from the mixture population π_o, and that, for some $N_o < K_o$, N_o of these observations are left unidentified, while the remaining $K_o - N_o$ observations are identified. For $i = 1,\ldots,m$, let $\{x_{ik}\}_{k=1,\ldots,N_i}$ denote the subset of the identified observations which come from π_i, and let $\{x_{ok}\}_{k=1,\ldots,N_o}$ be the set of unidentified observations from π_o. The log-likelihood function for this sample is

$$L_2(\Theta) = \log\left\{\frac{(\sum_{i=1}^m N_i)!}{N_1!\ldots N_m!}\,\alpha_1^{N_1}\ldots\alpha_m^{N_m}\right\} + \sum_{i=1}^m \sum_{k=1}^{N_i} \log p_i(x_{ik}) + \sum_{k=1}^{N_o} \log p(x_{ok})$$

$$= \log\left\{\frac{(\sum_{i=1}^m N_i)!}{N_1!\ldots N_m!}\right\} + \sum_{i=1}^m \sum_{k=1}^{N_i} \log[\alpha_i p_i(x_{ik})] + \sum_{k=1}^{N_o} \log p(x_{ok}) .$$

Differentiating L_2 and setting its partial derivatives to zero gives the likelihood equations

$$\alpha_i = \tilde{A}_i(\Theta) \equiv \frac{N_i}{K_o} + \frac{\alpha_i}{K_o}\sum_{k=1}^{N_o}\frac{p_i(x_{ok})}{p(x_{ok})} \tag{3.a}$$

$$\mu_i = M_i(\Theta) \tag{3.b}$$

$$\Sigma_i = S_i(\Theta) \tag{3.c}$$

for $i = 1,\ldots,m$.

We set

$$\tilde{A}(\Theta) = \begin{pmatrix} \tilde{A}_1(\Theta) \\ \vdots \\ \tilde{A}_m(\Theta) \end{pmatrix}$$

and define an operator $\tilde{\Phi}_\epsilon$ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ by

$$\tilde{\Phi}_\epsilon(\Theta) = (1-\epsilon)\Theta + \epsilon \begin{pmatrix} A(\Theta) \\ M(\Theta) \\ S(\Theta) \end{pmatrix} \quad .$$

Our iterative procedure is the following: Beginning with some starting value $\Theta^{(1)}$, define successive iterates inductively by

$$\Theta^{(j+1)} = \tilde{\Phi}_\epsilon(\Theta^{(j)}) \tag{4}$$

for $j = 1,2,3,\ldots$. As before, the desired local convergence result for this iterative procedure follows from the theorem below.

Theorem 2: With probability 1 as N_0 approaches infinity, $\tilde{\Phi}_\epsilon$ is a locally contractive operator (in some norm on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$) near the strongly consistent maximum-likelihood estimate whenever $0 < \epsilon < 2$.

Proof of Theorem 2: If Θ is the strongly consistent maximum-likelihood estimate, then, as before, it suffices to show that, with probability 1, $\nabla\tilde{\Phi}_\epsilon(\Theta)$ converges as N_0 approaches infinity to an operator which has operator norm less than 1 with respect to some vector norm on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$. Proceeding as before, one sees that

$$\nabla_{\overline{\alpha}}\widetilde{A}(\Theta) = (\text{diag}\,(1 - \frac{N_i}{\alpha_i K_o})) - (\text{diag}\,\frac{\alpha_i}{K_o}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}^T \right\}$$

$$\nabla_{\overline{\mu}}\widetilde{A}(\Theta) = -(\text{diag}\,\frac{\alpha_i}{K_o}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \begin{pmatrix} \langle\beta_1\gamma_1, \cdot\rangle'_1 \\ \vdots \\ \langle\beta_m\gamma_m, \cdot\rangle'_m \end{pmatrix}^T \right\}$$

$$\nabla_{\overline{\Sigma}}\widetilde{A}(\Theta) = -(\text{diag}\,\frac{\alpha_i}{K_o}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \begin{pmatrix} \langle\beta_1\delta_1, \cdot\rangle''_1 \\ \vdots \\ \langle\beta_m\delta_m, \cdot\rangle''_m \end{pmatrix}^T \right\}$$

The remaining Fréchet derivatives, i.e., the derivatives at Θ of M and S with respect to $\overline{\alpha}$, $\overline{\mu}$, and $\overline{\Sigma}$, are unchanged, except that K_i must be replaced by $\alpha_i K_o$ wherever it appears.

One obtains at Θ

$$(4) \qquad \nabla \begin{pmatrix} \widetilde{A} \\ M \\ S \end{pmatrix} = \begin{pmatrix} (\text{diag}(1 - \frac{N_i}{\alpha_i K_o})) & 0 & 0 \\ \widetilde{B}_{21} & \widetilde{B}_{22} & \widetilde{B}_{23} \\ \widetilde{B}_{31} & \widetilde{B}_{32} & \widetilde{B}_{33} \end{pmatrix} -$$

$$\begin{pmatrix} (\text{diag}\,\frac{\alpha_i}{K_o}) & 0 & 0 \\ 0 & \frac{1}{K_o} I & 0 \\ 0 & 0 & (\text{diag}\,\frac{\Sigma_i}{K_o}) \end{pmatrix} \left\{ \sum_{k=1}^{N_o} V(x_{ok}) \langle V(x_{o,k}), \cdot \rangle \right\}$$

In this expression, each $\widetilde{B}_{jk}$ is the same as the corresponding B_{jk} defined previously, except that each K_i in the latter is replaced by $\alpha_i K_o$ in the former. One verifies that, with probability 1 as N_o approaches infinity, (4) has the same limit as $\tilde{B}(I-QR)$, where Q and R are as before and $\tilde{B} = \frac{N_o}{K_o} I$. Repeating our earlier reasoning, one verifies that $\tilde{B}(I-QR)$ is positive-definite and symmetric with norm less than 1 with respect to the inner product $\langle \cdot, Q^{-1}\tilde{B}^{-1} \cdot \rangle$. Hence

$$\nabla\tilde{\Phi}_\epsilon(\Theta) = (1 - \epsilon) + \epsilon\nabla \begin{pmatrix} \tilde{A}(\Theta) \\ M(\Theta) \\ S(\Theta) \end{pmatrix}$$

converges to an operator which has norm less than 1 with respect to $\langle \cdot, Q^{-1}\tilde{B}^{-1} \cdot \rangle$ whenever $0 < \epsilon < 2$. This completes the proof of the theorem.

The remarks concerning the "optimal ϵ" at the conclusion of the preceding section are valid here verbatim.

BIBLIOGRAPHY

1. W. H. Coberly, private communication.
2. D. W. Hosmer, Jr., "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), pp. 761-770.
3. B. C. Peters, Jr., and H. F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions," Report #51, NASA contract NAS-9-12777, University of Houston, Department of Mathematics.