DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NASA CR-
147521

(NASA-CR-147521) AN ITERATIVE PROCEDURE FOR OBTAINING MAXIMUM-LIKELIHOOD ESTIMATES OF THE PARAMETERS FOR A MIXTURE OF NORMAL DISTRIBUTIONS, ADDENDUM (Houston Univ.)
11 p HC $3.50 CSCL 12A G3/65

N76-20914

Unclas
21495

ADDENDUM TO "AN ITER PROC FOR
OBTAINING MAX-LIKELIHOOD ESTS OF
THE PARAM FOR A MIX OF NOR DISTS
B. C. PETERS, JR & H. F. WALKER
REPORT #47 SEPTEMBER 1975

APR 1976
RECEIVED
NASA STI FACILITY
INPUT BRANCH

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Addendum to "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

Report #47
September, 1975
NAS 9-12777

Addendum to "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.

In this report, we discuss new results and insights concerning an iterative procedure introduced in [1] for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. For any questions concerning notation, definitions, etc., the reader is referred to that report.

The iterative procedure in question is the following: Beginning with some starting value $\begin{pmatrix} \bar{\alpha}^{(1)} \\ \bar{\mu}^{(1)} \\ \bar{\Sigma}^{(1)} \end{pmatrix}$ in the space $\mathfrak{A} \oplus \mathfrak{M} \oplus \mathfrak{S}$ introduced in [1], define successive iterates inductively by the relationship

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

given in [1]. It is shown in [1] that, with probability approaching 1 as the sample size N approaches infinity, this procedure converges locally to the consistent maximum-likelihood estimate whenever ϵ is sufficiently small. (In particular, $\epsilon < \frac{4}{m(n+1)(n+2)}$ guarantees the local convergence of this procedure in probability.)

In this report, we prove that, in probability, the procedure (*) converges locally to the consistent maximum-likelihood estimate whenever $0 < \epsilon < 2$. We also show that the ϵ which yields optimal local convergence rates lies between 1 and 2. In fact, the optimal ϵ is near 1, if the component populations are widely separated, and near 2 if the component populations have nearly identical means and covariance matrices.

1. Local Convergence.

As in [1], we say that Φ_{ϵ} is locally contractive (in a norm $\| \;\|$ on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$) near $\begin{pmatrix} \bar{\alpha} \\ \bar{\mu} \\ \bar{\Sigma} \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$ if there is a number λ, $0 \le \lambda < 1$ such that

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

whenever $\begin{pmatrix}\bar{\alpha}'\\ \bar{\mu}'\\ \bar{\Sigma}'\end{pmatrix}$ lies sufficiently near $\begin{pmatrix}\bar{\alpha}\\ \bar{\mu}\\ \bar{\Sigma}\end{pmatrix}$. Our result is the following.

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

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

Proof: As observed in [1], the theorem will be proved if it can be shown that, for $0 < \epsilon < 2$, $E(\nabla \Phi_\epsilon(\bar{\alpha}^\circ, \bar{\mu}^\circ, \bar{\Sigma}^\circ))$ has operator norm less than 1 with respect to some vector norm on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$. (Throughout this note, the superscript "$\circ$" indicates that the superscripted parameters are the true parameters of the mixture density.) For $i=1,\ldots,m$, let $\langle\ ,\ \rangle_i'$ and $\langle\ ,\ \rangle_i''$ be the inner products on R^n and the space of real, symmetric $n \times n$ matrices introduced in [1], i.e., let

$$\langle v,w\rangle_i' = v^T(\alpha_i^\circ \Sigma_i^{\circ\,-1})w \quad \text{for } v,w \in R^n,$$

$$\langle A,B\rangle_i'' = \mathrm{tr}\{A(\tfrac{\alpha_i^\circ}{2}\Sigma_i^{\circ\,-1})B^T\} \quad \text{for real, symmetric } n \times n \ A \text{ and } B.$$

These inner products, together with scalar multiplication on R^1, induce an inner product $\langle\ ,\ \rangle$ on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$. Now $E(\nabla \Phi_\epsilon(\bar{\alpha}^\circ, \bar{\mu}^\circ, \bar{\Sigma}^\circ)) = I - \epsilon QR$, where

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

and

$$\int_{R_n} \begin{pmatrix} \frac{p_1}{p} \\ \vdots \\ \frac{p_m}{p} \\ \frac{p_1}{p}(x-\mu_1^\circ) \\ \vdots \\ \frac{p_m}{p}(x-\mu_m^\circ) \\ \frac{p_1}{p}[\Sigma_1^{\circ -1}(x-\mu_1^\circ)(x-\mu_1^\circ)^T-I] \\ \vdots \\ \frac{p_m}{p}[\Sigma_m^{\circ -1}(x-\mu_m^\circ)(x-\mu_m^\circ)^T-I] \end{pmatrix} \left\langle \begin{pmatrix} \frac{p_1}{p} \\ \vdots \\ \frac{p_m}{p} \\ \frac{p_1}{p}(x-\mu_1^\circ) \\ \vdots \\ \frac{p_m}{p}(x-\mu_m^\circ) \\ \frac{p_1}{p}[\Sigma_1^{\circ -1}(x-\mu_1^\circ)(x-\mu_1^\circ)^T-I] \\ \vdots \\ \frac{p_m}{p}[\Sigma_m^{\circ -1}(x-\mu_m^\circ)(x-\mu_m^\circ)^T-I] \end{pmatrix}, \right\rangle p\, dx$$

One sees that the theorem will be proved if it can be shown that, with respect to some vector norm on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$, the operator norm of QR is no greater than 1. Since QR is positive definite and symmetric with respect to the inner product $\langle\ ,Q^{-1}\ \rangle$, it follows that the theorem will be proved if it can be shown that $\langle V,Q^{-1}[QR]V\rangle = \langle V,RV\rangle \le \langle V,Q^{-1}V\rangle$ for $V \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$.

For

$$V = \begin{pmatrix} y_1 \\ \vdots \\ y_m \\ v_1 \\ \vdots \\ v_m \\ A_1 \\ \vdots \\ A_m \end{pmatrix} \in \mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S},$$

one has

$$<V,RV> = \int_{R^n} \Big(\sum_{i=1}^{m} y_i \frac{p_i}{p} + \sum_{i=1}^{m} v_i^T(\alpha_i^{\circ}\Sigma_i^{\circ -1})\frac{p_i}{p}(x-\mu_i^{\circ}) +$$

$$\sum_{i=1}^{m} \mathrm{tr}\{A_i(\frac{\alpha_i^{\circ}}{2}\Sigma_i^{\circ -1})\frac{p_i}{p}[\Sigma_i^{\circ -1}(x-\mu_i^{\circ})(x-\mu_i^{\circ})^T - I]^T\}\Big)^2 p\ dx$$

$$= \int_{R^n} \Big(\sum_{i=1}^{m} [\alpha_i^{\circ -1} y_i + v_i^T\Sigma_i^{\circ -1}(x-\mu_i^{\circ}) + \mathrm{tr}\{A_i(\frac{1}{2}\Sigma_i^{\circ -1})[\Sigma_i^{\circ -1}(x-\mu_i^{\circ})(x-\mu_i^{\circ})^T - I]^T\}]\frac{\alpha_i^{\circ}p_i}{p}\Big)^2 p\ dx$$

$$\leq \int_{R^n} \Big(\sum_{i=1}^{m} [\alpha_i^{\circ -1} y_i + v_i^T\Sigma_i^{\circ -1}(x-\mu_i^{\circ}) + \mathrm{tr}\{A_i(\frac{1}{2}\Sigma_i^{\circ -1})[\Sigma_i^{\circ -1}(x-\mu_i^{\circ})(x-\mu_i^{\circ})^T - I]\}]^2 \frac{\alpha_i^{\circ}p_i}{p}\Big) p\ dx$$

by Schwarz's inequality. 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,

$$<V,RV> \leq \int_{R^n} \Big(\sum_{i=1}^{m} [\alpha_i^{\circ -2} y_i^2 + (v_i\Sigma_i^{\circ -1}(x-\mu_i^{\circ}))^2 + (\mathrm{tr}\{A_i(\frac{1}{2}\Sigma_i^{\circ -1})[\Sigma_i^{\circ -1}(x-\mu_i^{\circ})(x-\mu_i^{\circ})^T - I]^T\})]^2 \alpha_i^{\circ} p_i\Big) dx$$

Now

(1) $$\int_{R^n} \alpha_i^{o-1} y_i^2 p_i dx = \alpha_i^{o-1} y_i^2$$

(2) $$\int_{R^n} (v_i^T \Sigma_i^{o-1}(x-\mu_i^o))^2 \alpha_i^o p_i dx = \int_{R^n} v_i^T \Sigma_i^{o-1}(x-\mu_i^o)(x-\mu_i^o)^T \Sigma_i^{o-1} v_i \alpha_i^o p_i dx$$

$$= \langle v_i, v_i \rangle_i'$$

(3) $$\int_{R^n} (\mathrm{tr}\{A_i(\tfrac{1}{2}\Sigma_i^{o-1})[\Sigma_i^{o-1}(x-\mu_i^o)(x-\mu_i^o)^T - I]^T\})^2 \alpha_i^o p_i dx = \langle A_i, \Sigma_i^{o-1} A_i \rangle_i''$$

(A proof of (3) follows below.) From (1), (2), and (3), one concludes that

$$\langle V, RV \rangle \le \sum_{i=1}^{m} \alpha_i^{o-1} y_i^2 + \langle v_i, v_i \rangle_i' + \langle A_i, \Sigma_i^{o-1} A_i \rangle_i'' = \langle V, Q^{-1} V \rangle .$$

This completes the proof of the theorem.

<u>Proof of (3)</u>: Setting $y = \Sigma_i^{o-1/2}(x-\mu_i^o)$ and

$$I = \int_{R^n} (\mathrm{tr}\{A_i(\tfrac{1}{2}\Sigma_i^{o-1})[\Sigma_i^{o-1}(x-\mu_i^o)(x-\mu_i^o)^T - I]^T\})^2 \alpha_i^o p_i dx,$$

one obtains

$$I = \frac{\alpha_i^o}{4} \int_{R^n} (\mathrm{tr}\{A_i[\Sigma_i^{o-1/2} y y^T \Sigma_i^{o-1/2} - \Sigma_i^{o-1}]^T\})^2 p_o dy,$$

where $p_o \sim N(0, I)$. Denoting $\Sigma_i^{o-1/2} A_i \Sigma_i^{o-1/2} = B = (b_{jk})$, one then derives

$$I = \frac{\alpha_i^o}{4} \int_{R^n} (\mathrm{tr}\{B[yy^T - I]\})^2 p_o dy$$

$$= \frac{\alpha_i^o}{4} \int_{R^n} [(\mathrm{tr}\{Byy^T\})^2 - 2\mathrm{tr}\{B\}\mathrm{tr}\{Byy^T\} + (\mathrm{tr}\{B\})^2] p_o dy$$

$$= \frac{\alpha_i^o}{4}\{\sum_{j,k,p,q} \beta_{jk}\beta_{pq} \int_{R^n} y_k y_j y_q y_p p_o dy - 2\ \mathrm{tr}\{B\} \sum_{j,k} \beta_{jk} \int_{R^n} y_k y_j p_o dy + (\mathrm{tr}\{B\})^2\}$$

$$= \frac{\alpha_i^o}{4}\{\sum_k \sum_{p \neq k} \beta_{kk}\beta_{pp} + \sum_k \sum_{j \neq k} \beta_{jk}\beta_{jk} + \sum_k \sum_{j \neq k} \beta_{jk}\beta_{kj} + 3\sum_k \beta_{kk}^2 - 2(\mathrm{tr}\{B\})^2 + (\mathrm{tr}\{B\})^2\}$$

$$= \frac{\alpha_i^o}{2} \mathrm{tr}\{B^2\} = \frac{\alpha_i^o}{2} \mathrm{tr}\{\Sigma_i^{o-1/2} A_i \Sigma_i^{o-1} A_i \Sigma_i^{o-1/2}\} = \mathrm{tr}\{A_i (\frac{\alpha_i^o}{2} \Sigma_i^{o-1})(\Sigma_i^{o-1} A_i)^T\}$$

$$= \langle A_i, \Sigma_i^{o-1} A_i \rangle_i''$$

3. The optimal ϵ.

From the proof of the theorem, one sees that, asymptotically as N approaches infinity, the value of ϵ which yields optimal local convergence rates is that which minimizes the spectral radius of $E(\nabla\Phi_\epsilon(\bar{\alpha}^o, \bar{\mu}^o, \bar{\Sigma}^o))$. (Indeed, $E(\nabla\Phi_\epsilon(\bar{\alpha}^o, \bar{\mu}^o, \bar{\Sigma}^o)) = I - \epsilon\ QR$ is symmetric with respect to the inner product $\langle\ , Q^{-1}\ \rangle$; hence, its operator norm with respect to this inner product is equal to its spectral radius.) Letting ρ and τ denote, respectively, the largest and smallest eigenvalues of QR, one verifies that the spectral radius of $E(\nabla\Phi_\epsilon(\bar{\alpha}^o, \bar{\mu}^o, \bar{\Sigma}^o))$ is minimized when $1 - \epsilon\tau = \epsilon\rho - 1$, i.e., when $\epsilon = \frac{2}{\rho + \tau}$. Now $\rho = 1$ always, for it follows from the proof of the theorem that ρ is never greater than 1, and

$$\begin{pmatrix} \alpha_1^o \\ \cdot \\ \cdot \\ \cdot \\ \alpha_m^o \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{S}$$

is always an eigenvector of QR with eigenvalue 1. Thus optimal convergence rates are obtained when $\epsilon = \frac{2}{1+\tau}$, where τ lies between 0 and 1. In particular, the best choice of ϵ lies between 1 and 2.

Suppose that the component populations in the mixture are "widely separated" in the sense that each pair (μ_i^o, Σ_i^o) differs greatly from every other such pair. Then

$$\left(\frac{\alpha_i^o p_i(x)}{p(x)}\right)\left(\frac{\alpha_j^o p_j(x)}{p(x)}\right) \approx \delta_{ij} \quad \text{for} \quad x \in R^n \quad \text{and} \quad i,j = 1,\dots,m,$$

and one verifies that $QR \approx I$. Consequently, optimal convergence rates are obtained for an ϵ near 1 and, for the optimal ϵ, $E(\nabla \Phi_\epsilon(\bar{\alpha}^\circ, \bar{\mu}^\circ, \bar{\Sigma}^\circ)) = I - \epsilon\, QR \approx 0$. Thus for mixtures whose component populations are "widely separated", optimal convergence rates are obtained for an ϵ near 1, and rapid first-order convergence can be expected for this ϵ.

Now suppose that the component populations in the mixture are such that each pair $(\mu_i^\circ, \Sigma_i^\circ)$ differs little from every other such pair. Then

$p(x) \approx p_i(x)$ and $\frac{p_i(x)}{p(x)} \approx 1$ for $x \in R^n$ and $i = 1,\ldots,m$, and one verifies that the smallest eigenvalue of QR is near zero. It follows that optimal convergence rates are obtained for an ϵ near 2. In this case, the spectral radius of $E(\nabla \Phi_\epsilon(\bar{\alpha}^o,\bar{\mu}^o,\bar{\Sigma}^o))$ is near 1, even for the optimal value of ϵ; hence, slow first-order convergence is to be expected.

We conclude by observing that the major practical implication of this note is that the iterative procedure under consideration converges whenever the step-size ϵ lies in an interval which is completely independent of the particular mixture problem at hand. It is readily ascertained that this cannot be said for the regular steepest descent procedure

$$\alpha_i^{(q+1)} = \alpha_i^{(q)} + \epsilon\left[\frac{1}{N}\sum_{k=1}^{N}\frac{p_i(x_k)}{p(x_k)} - \frac{1}{mN}\sum_{j=1}^{m}\sum_{k=1}^{N}\frac{p_j(x_k)}{p(x_k)}\right]$$

$$\mu_i^{(q+1)} = \mu_i^{(q)} + \epsilon\left[\frac{1}{N}\sum_{k=1}^{N}\frac{\alpha_i^{(q)}p_i(x_k)}{p(x_k)}\Sigma_i^{(q)-1}(x_k-\mu_i^{(q)})\right]$$

$$\Sigma_i^{(q+1)} = \Sigma_i^{(q)} + \epsilon\left[\frac{1}{2N}\sum_{k=1}^{N}\frac{\alpha_i^{(q)}p_i(x_k)}{p(x_k)}\left[-\Sigma_i^{(q)-1} + \Sigma_i^{(q)-1}(x_k-\mu_i^{(q)})(x_k-\mu_i^{(q)})^T\Sigma_i^{(q)-1}\right]\right].$$

Thus the procedure considered here offers considerable practical advantages over the steepest descent procedure, even though it is itself a generalized steepest descent (deflected gradient) procedure.

REFERENCE

1. B.C. Peters and H.F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions, "Report #43, NASA Contract NAS-9-12777, University of Houston, Department of Mathematics.