

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-

147738

(NASA-CR-147738) THE NUMERICAL EVALUATION
OF THE MAXIMUM-LIKELIHOOD ESTIMATE OF A
SUBSET OF MIXTURE PROPORTIONS (Houston
Univ.) 14 p HC \$3.50

N76-26928

CSSL 12A

G3/65

Unclas
42111

THE NUMERICAL EVALUATION OF THE
MAXIMUM-LIKELIHOOD ESTIMATE OF A
SUBSET OF MIXTURE PROPORTIONS,
BY B.C.PETERS, JR & H. F. WALKER
FEBRUARY 1976 REPORT #50

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



3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

*The Numerical Evaluation of the Maximum-Likelihood
Estimate of a Subset of Mixture Proportions, ■*

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

February, 1976

Report #50

ABSTRACT

THE NUMERICAL EVALUATION OF THE MAXIMUM- LIKELIHOOD ESTIMATE OF A SUBSET OF MIXTURE PROPORTIONS

In this note, we give necessary and sufficient conditions for a maximum-likelihood estimate of a subset of mixture proportions. From these conditions, we derive likelihood equations satisfied by the maximum-likelihood estimate and discuss a successive-approximations procedure suggested by these equations for numerically evaluating the maximum-likelihood estimate. It is shown that, with probability 1 for large samples, this procedure converges locally to the maximum-likelihood estimate whenever a certain step-size lies between 0 and 2. Furthermore, optimal rates of local convergence are obtained for a step-size which is bounded below by a number between 1 and 2.

The Numerical Evaluation of the Maximum-Likelihood
Estimate of a Subset of Mixture Proportions

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 is a convex combination of density functions p_0, p_1, \dots, p_m on \mathbb{R}^n . In particular, suppose that the density function of x is $p(x, \bar{\alpha})$, a member of the parametric family of density functions

$$p(x, \bar{\alpha}) = \sum_{i=1}^m \alpha_i p_i(x) + (1 - \beta) p_0(x)$$

for $x \in \mathbb{R}^n$, where $\bar{\alpha} = (\alpha_1, \dots, \alpha_m)^T \in \mathbb{R}^m$ and β satisfy the following con-

*This author was supported in part by NASA under Contract JSC-NAS-9-12777.

straints: (i) $0 \leq \beta \leq 1$ and $0 \leq \alpha_i \leq 1$ for $i = 1, \dots, m$; (ii) $\sum_{i=1}^m \alpha_i = \beta$. In this note, we assume that β and the density functions p_0, \dots, p_m are known, and we address the problem of numerically estimating $\bar{\alpha}^0$, the vector of unknown mixture proportions, on the basis of a given sample $\{x_k\}_{k=1, \dots, N}$ of independent observations on x .

To be more specific, we define a maximum-likelihood estimate of $\bar{\alpha}^0$, based on the given sample, to be a choice of $\bar{\alpha}$ which satisfies the constraints (i) and (ii) above and which maximizes the log-likelihood function

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

(We assume throughout this report that $p(x_k, \bar{\alpha}) \neq 0$ for $k = 1, \dots, N$ and for all $\bar{\alpha}$ satisfying the given constraints.) In the following, we derive necessary and sufficient conditions for $\bar{\alpha}$ to be a maximum-likelihood estimate of $\bar{\alpha}^0$, and we discuss a particular iterative procedure based on these conditions for the numerical evaluation of a maximum-likelihood estimate.

The results given here generalize those of [2], in which a restricted iterative procedure is considered in the special case $\beta = 1$. We also remark that our results apply to the problem of numerically evaluating a maximum-likelihood estimate of a proper subset $\{\alpha_i^0\}_{i=1, \dots, m}$ of mixture proportions in a density $p = \sum_{i=1}^s \alpha_i^0 p_i$ when the remaining $s-m$ proportions are known. Indeed, this problem is seen to be of the type considered here by taking $\beta = 1 - \sum_{i=m+1}^s \alpha_i^0$ and $p_0 = \frac{1}{1-\beta} \sum_{i=m+1}^s \alpha_i^0 p_i$.

2. The likelihood equations.

One easily verifies that the log-likelihood function L is a concave function of $\bar{\alpha}$ on the constraint set, i.e., the set of elements of \mathbb{R}^m satisfying the constraints (i) and (ii) given in the introduction. It follows that a necessary and sufficient condition for $\bar{\alpha}$ to be a maximum-likelihood estimate of $\bar{\alpha}^0$ is that $\nabla L(\bar{\alpha})(\bar{\alpha}' - \bar{\alpha}) \leq 0$ for all $\bar{\alpha}'$ in the constraint set, where $\nabla L(\bar{\alpha}) = (\frac{\partial L}{\partial \alpha_1}(\bar{\alpha}), \dots, \frac{\partial L}{\partial \alpha_m}(\bar{\alpha}))$. Since this inequality holds if and only if it holds whenever $\bar{\alpha}'$ is an extreme point of the constraint set, one concludes that $\bar{\alpha}$ is a maximum-likelihood estimate if and only if, for $i = 1, \dots, m$, $\beta \frac{\partial L}{\partial \alpha_i}(\bar{\alpha}) \leq \nabla L(\bar{\alpha})\bar{\alpha}$, with equality if $\alpha_i > 0$. We reformulate this result as the following necessary and sufficient condition for $\bar{\alpha}$ to be a maximum-likelihood estimate of $\bar{\alpha}^0$: For $i = 1, \dots, m$,

$$(1) \quad \beta \sum_{k=1}^N \frac{p_i(x_k)}{p(x_k, \bar{\alpha})} \leq \sum_{j=1}^m \sum_{k=1}^N \frac{\alpha_j p_j(x_k)}{p(x_k, \bar{\alpha})},$$

with equality if $\alpha_i > 0$.

Multiplying both sides of (1) by α_i and rearranging gives the following necessary condition for $\bar{\alpha}$ to be a maximum-likelihood estimate:

$$(2) \quad \alpha_i = A_i(\bar{\alpha}) \equiv \beta \sum_{k=1}^N \frac{\alpha_i p_i(x_k)}{p(x_k, \bar{\alpha})} \Bigg/ \sum_{j=1}^m \sum_{k=1}^N \frac{\alpha_j p_j(x_k)}{p(x_k, \bar{\alpha})}$$

for $i = 1, \dots, m$. This condition is not sufficient in general for $\bar{\alpha}$ to be a maximum-likelihood estimate. Indeed, this condition is satisfied by each extreme point of the constraint set. However, this condition is sufficient as well as necessary for $\bar{\alpha}$ to be a maximum-likelihood estimate which lies in the interior of the constraint set, i.e., the components of which satisfy $\alpha_i > 0$ for $i=1, \dots, m$. We refer to the equations (2) as the likelihood equations.

3. The iterative procedure.

We now define an iterative procedure based on the likelihood equations and discuss its applicability to the problem of numerically evaluating a maximum-likelihood estimate of $\bar{\alpha}^0$. Setting $A(\bar{\alpha}) = (A_1(\bar{\alpha}), \dots, A_m(\bar{\alpha}))^T$, we write the likelihood equations as

$$(3) \quad \bar{\alpha} = A(\bar{\alpha}).$$

Equivalent to (3) is the equation

$$(4) \quad \bar{\alpha} = \phi_\epsilon(\bar{\alpha}) \equiv (1 - \epsilon)\bar{\alpha} + \epsilon A(\bar{\alpha})$$

for any number ϵ . (Of course, (4) becomes (3) when $\epsilon = 1$.) Note that the continuous nonlinear operator A maps the constraint set into itself. For any ϵ and any $\bar{\alpha}$ in the constraint set, the components of $\phi_\epsilon(\bar{\alpha})$ sum to 1; however, the components of $\phi_\epsilon(\bar{\alpha})$ are guaranteed to be non-negative for all $\bar{\alpha}$ in the constraint set only if $0 \leq \epsilon \leq 1$.

The iterative procedure suggested by (4) is the following: Beginning with some starting value $\bar{\alpha}^{(1)}$ in the constraint set, define successive iterates inductively by

$$(5) \quad \bar{\alpha}^{(j+1)} = \phi_\epsilon(\bar{\alpha}^{(j)})$$

for $j = 1, 2, \dots$. We observe that if the sequence of iterates defined by (5) converges, then its limit is a fixed point of ϕ_ϵ and, hence, of A . Our first theorem gives sufficient conditions for such a limit to be a maximum-likelihood estimate. The proof of the theorem is virtually the same as that of the

corresponding theorem in [2], and we omit it.

THEOREM 1: Suppose that $\bar{\alpha}^{(1)}$ lies in the interior of the constraint set and that $0 < \epsilon \leq 1$. If the sequence of iterates defined by (5) converges, then its limit is a maximum-likelihood estimate of $\bar{\alpha}^0$.

In order to give sufficient conditions for the convergence of the iterates defined by (5), we need to make further assumptions concerning the density functions p_0, \dots, p_m . Henceforth, we assume that they are linearly independent, i.e., that any linear combination $\sum_{i=0}^m c_i p_i$, with $\sum_{i=0}^m c_i^2 \neq 0$, does not vanish identically on \mathbb{R}^n . This insures that, with probability 1, there exists a unique maximum-likelihood estimate for large N which converges to $\bar{\alpha}^0$ as N approaches infinity. (See, for example, Appendix 1 of [3].) Our aim is to establish the following result.

THEOREM 2: Suppose that $\bar{\alpha}^0$ lies in the interior of the constraint set and that $0 < \epsilon < 2$. Then with probability 1 as N approaches infinity, ϕ_ϵ is a local contraction on the constraint set near $\bar{\alpha}$, the (unique) maximum-likelihood estimate of $\bar{\alpha}^0$. If the density functions p_0, \dots, p_m are analytic as well as linearly independent, then ϕ_ϵ is a local contraction on the constraint set near $\bar{\alpha}$ with probability 1 whenever $\bar{\alpha}$ lies in the interior of the constraint set and $N \geq m$.

In saying that ϕ_ϵ is a local contraction on the constraint set near $\bar{\alpha}$, we mean that there exists a norm $|| \cdot ||$ on \mathbb{R}^m and a constant λ , $0 \leq \lambda < 1$, such that

$$(6) \quad ||\phi_\epsilon(\bar{\alpha}') - \alpha|| \leq \lambda ||\bar{\alpha}' - \alpha||$$

for all $\bar{\alpha}'$ in the constraint set which lie sufficiently near $\bar{\alpha}$. Our sufficient conditions for the convergence of the iterates defined by (5) are stated in the corollary below, which is an immediate consequence of Theorem 2 and the inequality (6).

COROLLARY: Suppose that $\bar{\alpha}^0$ lies in the interior of the constraint set and that $0 < \epsilon < 2$. Then with probability 1 as N approaches infinity, the iterates defined by (5) converge to $\bar{\alpha}$, the (unique) maximum-likelihood estimate of $\bar{\alpha}^0$, whenever $\bar{\alpha}^{(1)}$ lies sufficiently near $\bar{\alpha}$. If the density functions p_0, \dots, p_m are analytic as well as linearly independent, then, with probability 1 whenever $\bar{\alpha}$ lies in the interior of the constraint set and $N \geq m$, the iterates converge to $\bar{\alpha}$ whenever $\bar{\alpha}^{(1)}$ lies sufficiently near $\bar{\alpha}$.

Proof of Theorem 2: In proving the first statement of the theorem, it may be assumed that the (unique) maximum-likelihood estimate $\bar{\alpha}$ lies in the interior of the constraint set. (By the remarks preceding the theorem, the probability is 1 that this occurs for large N .) Assuming $0 < \epsilon < 2$, we must show that, with probability 1 as N approaches infinity, an inequality of the form (6) holds.

For any norm on \mathbb{R}^m , one can write

$$\phi_\epsilon(\bar{\alpha}') - \bar{\alpha} = \nabla \phi_\epsilon(\bar{\alpha})[\bar{\alpha}' - \bar{\alpha}] + o(\|\bar{\alpha}' - \bar{\alpha}\|^2).$$

In this expression, $\nabla \phi_\epsilon(\bar{\alpha})$ denotes the $m \times m$ matrix whose ij^{th} entry is the i^{th} component of $\frac{\partial}{\partial \alpha_j} \phi_\epsilon(\bar{\alpha})$. It follows that the first statement of the theorem will be proved if it can be shown that, with probability 1 as N

approaches infinity, there exists a norm $|| \cdot ||$ on \mathbb{R}^m and a number λ , $0 \leq \lambda < 1$, for which an inequality of the form

$$(7) \quad ||\nabla\phi_\epsilon(\bar{\alpha})\bar{\gamma}'|| \leq \lambda ||\bar{\gamma}'||$$

holds for all $\bar{\gamma}'$ in the subspace

$$\mathcal{E} = \{\bar{\gamma}' = (\gamma_1, \dots, \gamma_m)^T : \sum_{i=1}^m \gamma_i = 0\} \subseteq \mathbb{R}^m.$$

Using the fact that $\bar{\alpha}$ satisfies the likelihood equations (2), one verifies that $\nabla\phi_\epsilon(\bar{\alpha}) = I - \epsilon Q$, where

$$Q = \frac{1}{\sum_{i=1}^m \sum_{k=1}^N \frac{\alpha_i p_i(x_k)}{p(x_k, \bar{\alpha})}} (\text{diag } \alpha_i) \left\{ \sum_{k=1}^N \beta \begin{pmatrix} \frac{p_1(x_k)}{p(x_k, \bar{\alpha})} \\ \vdots \\ \frac{p_m(x_k)}{p(x_k, \bar{\alpha})} \end{pmatrix} + (1-\beta) \frac{p_0(x_k)}{p(x_k, \bar{\alpha})} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k, \bar{\alpha})} \\ \vdots \\ \frac{p_m(x_k)}{p(x_k, \bar{\alpha})} \end{pmatrix}$$

It is easily shown that \mathcal{E} is invariant under Q and, hence, under $\nabla\phi_\epsilon(\bar{\alpha})$. To establish an inequality of the form (7), it suffices to show that, with probability 1 as N approaches infinity, there exists a norm on \mathcal{E} with respect to which the operator norm of $\nabla\phi_\epsilon(\bar{\alpha})$ is less than 1.

Define an inner product $\langle \cdot, \cdot \rangle$ on \mathcal{E} by $\langle \bar{\gamma}, \bar{\gamma}' \rangle = \bar{\gamma}^T (\text{diag } \alpha_i^{-1}) \bar{\gamma}'$ for $\bar{\gamma}$ and $\bar{\gamma}'$ in \mathcal{E} . With respect to this inner product, Q is symmetric (in fact, positive semi-definite) on \mathcal{E} . It follows that, with respect to the

norm on \mathcal{E} defined by this inner product, the operator norm of $V\phi_\epsilon(\bar{\alpha})$ on \mathcal{E} is less than 1 if the eigenvalues of Q which correspond to eigenvectors in \mathcal{E} lie in the interval $(0,1]$.

Now Q is column stochastic; hence, no eigenvalue of Q on \mathcal{E} is greater than 1. (See [1] for a discussion of column stochastic matrices.) To complete the proof of the first statement of the theorem, it need only be shown that, with probability 1 as N approaches infinity, Q is positive-definite on \mathcal{E} . Now, with probability 1, $\bar{\alpha}$ converges to $\bar{\alpha}^0$ as N approaches infinity. Using arguments analogous to those employed in [3], one verifies that, with probability 1, Q converges as N approaches infinity to

$$\frac{1}{\beta}(\text{diag } \alpha_1^0) \left\{ \int_{\mathbb{R}^n} \beta \begin{pmatrix} \frac{p_1(x)}{p(x, \bar{\alpha}^0)} \\ \vdots \\ \frac{p_m(x)}{p(x, \bar{\alpha}^0)} \end{pmatrix} + (1-\beta) \frac{p_0(x)}{p(x, \bar{\alpha}^0)} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\} \begin{pmatrix} \frac{p_1(x)}{p(x, \bar{\alpha}^0)} \\ \vdots \\ \frac{p_m(x)}{p(x, \bar{\alpha}^0)} \end{pmatrix}^T p(x, \bar{\alpha}^0) dx,$$

a positive-definite operator on \mathcal{E} . It follows that Q is positive-definite on \mathcal{E} with probability 1 as N approaches infinity, and the first statement of the theorem is proved.

To prove the second statement of the theorem, suppose that $N \geq m$, that $\bar{\alpha}$ lies in the interior of the constraint set, and that p_0, \dots, p_m are analytic as well as linearly independent. Repeating the above argument with only minor changes, one obtains the desired result by finally observing that, as a consequence of the lemma in Appendix 2 of [3], Q is positive-definite on \mathcal{E} with probability 1 whenever $N \geq m$. This completes the proof of the theorem.

4. The optimal ϵ .

The corollary of Theorem 2 may be summarized by saying that, if $\bar{\alpha}^0$ lies in the interior of the constraint set, then, with probability 1 for large samples, the iterates defined by (5) converge locally to the maximum-likelihood estimate $\bar{\alpha}$ whenever $0 < \epsilon < 2$. Thus the iterative procedure (5), which is a generalized steepest-ascent (deflected-gradient) method, has the particularly important property of converging locally to $\bar{\alpha}$ whenever the step-size ϵ lies in an interval which is completely independent of the particular mixture problem at hand. Furthermore, if ϵ is no greater than 1, then the successive iterates defined by (5) are guaranteed to remain in the constraint set. It is readily ascertained that these properties are not shared by the usual steepest-ascent procedure, given by

$$\alpha_i^{(q+1)} = \alpha_i^{(q)} + \epsilon \left[\frac{1}{N} \sum_{k=1}^N \frac{P_i(x_k)}{p(x_k, \bar{\alpha}^{(q)})} - \frac{1}{mN} \sum_{j=1}^m \sum_{k=1}^N \frac{P_j(x_k)}{p(x_k, \bar{\alpha}^{(q)})} \right]$$

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

We now observe that there exists a particular value of ϵ , referred to as "the optimal ϵ ," which yields, with probability 1 for large samples, the fastest uniform rate of local convergence of (5) near $\bar{\alpha}$. Indeed, suppose that $\bar{\alpha}$ is an interior point of the constraint set and that $\nabla \phi_\epsilon(\bar{\alpha})$ is positive-definite on \mathcal{E} . (Recall that, with probability 1, these assumptions are valid for large samples.) Then one sees from the proof of Theorem 2 that the optimal ϵ is the unique value of ϵ which minimizes the spectral radius of $\nabla \phi_\epsilon(\bar{\alpha}) = I - \epsilon Q$, regarded as an operator on \mathcal{E} . ($\nabla \phi_\epsilon(\bar{\alpha})$ is symmetric on \mathcal{E} with respect to the inner product $\langle \cdot, \cdot \rangle$ defined previously. Consequently, its operator norm with respect to this inner product is equal to its spectral

radius and, hence, minimal.) It is easily verified that the optimal ϵ is given by $1 - \epsilon\tau = \epsilon\rho - 1$, i.e., $\epsilon = \frac{2}{\rho + \tau}$, where ρ and τ are, respectively, the largest and smallest eigenvalues of the operator Q restricted to \mathcal{E} .

It follows from the proof of Theorem 2 that ρ is never greater than 1. Thus the optimal ϵ is bounded below by $\frac{2}{1 + \tau}$, where τ lies between 0 and 1. In particular, this lower bound on the optimal ϵ lies between 1 and 2. It should be noted that, if ρ is strictly less than 1, then the optimal ϵ is actually greater than 2, even though Theorem 2 fails to guarantee the local convergence of (5) for such values of ϵ . We also observe that, despite the fact that the column-stochastic matrix Q always has 1 as an eigenvalue, the eigenvalue ρ of the restricted operator Q on \mathcal{E} can be arbitrarily small (and, hence, the optimal ϵ can be arbitrarily large). Indeed, Q is nearly the zero operator on \mathcal{E} if the component populations in the mixture are nearly identical.

Suppose that the component populations in the mixture are "widely separated" in the sense that, for $i \neq j$,

$$\frac{p_i(x_k) p_j(x_k)}{p(x_k, \bar{\alpha})^2} \approx 0$$

for $k = 1, \dots, N$. Then $Q \approx I$ and, hence, ρ and τ must lie near 1. One concludes that, with probability 1 for large samples, the fastest uniform rate of local convergence of (5) is obtained for ϵ near 1, and for the optimal ϵ , $\nabla \phi_{\epsilon}(\bar{\alpha}) = I - \epsilon Q \approx 0$. 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 (5) to $\bar{\alpha}$ can be expected for this ϵ .

Now suppose that two or more of the component populations in the mixture are nearly identical in the sense that, for some pair of distinct, non-zero indices i and j , $p_i(x_k) \approx p_j(x_k)$ for $k = 1, \dots, N$. Then Q is nearly singular, and hence, τ is near zero. Consequently, the optimal ϵ cannot be much smaller than 2. We remark that, if ρ is near 1 in this case, then the optimal ϵ must lie near 2. Then the spectral radius of $\nabla\Phi_\epsilon(\bar{\alpha})$ on \mathcal{E} is near 1, even for the optimal ϵ , and it follows that slow first-order local convergence of (5) to $\bar{\alpha}$ can be expected in this case.

BIBLIOGRAPHY

1. R. Bellman, Introduction to Matrix Analysis, McGraw-Hill Book Company, New York, 1960.
2. B.C. Peters and W.H. Coberly, "The numerical evaluation of the maximum-likelihood estimate of mixture proportions," to appear.
3. B.C. Peters and H.F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions," to appear.