SIAM J. APPL. MATH.
Vol. 35, No. 3, November 1978

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

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

Abstract. In this note, we give necessary and sufficient conditions for a maximum-likelihood estimate of a subset of the proportions in a mixture of specified distributions. 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.

1. Introduction. Let x be an n-dimensional random variable whose density function is a convex combination of density functions $p_0, p_1, \ldots, p_m$ on $\mathbb{R}^n$. In particular, suppose that the density function of x is $p(x, \bar{\alpha}^0)$, 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 $\bar{\alpha} = (\alpha_1, \ldots, \alpha_m)^T \in \mathbb{R}^m$ and β satisfy the following constraints: (i) $0 \leqq \beta \leqq 1$ and $0 \leqq \alpha_i \leqq 1$ for $i = 1, \cdots, m$; (ii) $\sum_{i=1}^m \alpha_i = \beta$. In this note, we assume that β and the density functions $p_0, \cdots, p_m$ 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 $\{x_k\}_{k=1,\cdots,N}$ 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 $k = 1, \cdots, N$ and for all $\bar{\alpha}$ satisfying the given constraints.) Taking advantage of the fact that the log-likelihood function is concave, we derive necessary and sufficient conditions for $\bar{\alpha}$ to be a maximum-likelihood estimate of $\bar{\alpha}^0$. These conditions, in turn, lead naturally to a particular successive approximations procedure 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 $\{\alpha_i^0\}_{i=1,\cdots,m}$ of mixture proportions in a density $p = \sum_{i=1}^s \alpha_i^0 p_i$ when the

* Received by the editors June 22, 1976.

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

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

remaining $s-m$ proportions are known. Indeed, this problem is seen to be of the type considered here by taking

$$\beta = 1 - \sum_{i=m+1}^{s} \alpha_i^0 \quad \text{and} \quad p_0 = \frac{1}{1-\beta} \sum_{i=m+1}^{s} \alpha_i^0 p_i.$$

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 $\nabla L(\bar{\alpha})(\bar{\alpha}' - \bar{\alpha}) \leqq 0$ for all $\bar{\alpha}'$ in the constraint set, where $\nabla L(\bar{\alpha}) = ((\partial L/\partial \alpha_1)(\bar{\alpha}), \cdots, (\partial L/\partial \alpha_m)(\bar{\alpha}))$. 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 $i = 1, \cdots, m$, $\beta(\partial L/\partial \alpha_i)(\bar{\alpha}) \leqq \nabla L(\bar{\alpha})\bar{\alpha}$, 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 $i = 1, \cdots, m$,

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

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:

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

for $i = 1, \cdots, m$. 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 $i = 1, \cdots, m$. We refer to the equations (2) as the *likelihood equations.*

For the case $\beta = 1$, there is an interpretation of equation (2) which has considerable heuristic appeal. For $\beta = 1$, equation (2) becomes

$$\alpha_i = \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha_i p_i(x_k)}{p(x_k, \bar{\alpha})}. \tag{2a}$$

If the density functions $p_1, \cdots, p_m$ are interpreted as conditional densities $p_j(x) = p(x|S_j)$ for some partition $\{S_1, \cdots, S_m\}$ of the underlying probability space into measureable sets, and the coefficients α_j are interpreted as a priori probabilities $\alpha_j = P(S_j)$, then equation (2a) states that the estimate of $P(S_j)$ is equal to the sample average of the estimates of the conditional probabilities $P(S_j|x_k)$ of S_j given the observation x_k.

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 $A(\bar{\alpha}) = (A_1(\bar{\alpha}), \cdots, A_m(\bar{\alpha}))^T$, we write the likelihood equation as

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

Equivalent to (3) is the equation

$$\bar{\alpha} = \Phi_\varepsilon(\bar{\alpha}) \equiv (1-\varepsilon)\bar{\alpha} + \varepsilon A(\bar{\alpha}) \tag{4}$$

for any number ε. (Of course, (4) becomes (3) when $\varepsilon = 1$.) 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 $\Phi_\varepsilon(\bar{\alpha})$ sum to 1; however, the components of $\Phi_\varepsilon(\bar{\alpha})$ are guaranteed to be nonnegative for all $\bar{\alpha}$ in the constraint set only if $0 \leqq \varepsilon \leqq 1$.

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

$$\bar{\alpha}^{(j+1)} = \Phi_\varepsilon(\bar{\alpha}^{(j)}) \tag{5}$$

for $j = 1, 2, \cdots$. 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 $0 < \varepsilon \leqq 1$. 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 $p_0, \cdots, p_m$. 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 yanish 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 $0 < \varepsilon < 2$. 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 $p_0, \cdots, p_m$ 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 $N \geqq m$.*

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, $0 \leqq \lambda < 1$, such that

$$\|\Phi_\varepsilon(\bar{\alpha}') - \bar{\alpha}\| \leqq \lambda \|\bar{\alpha}' - \bar{\alpha}\| \tag{6}$$

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 $0 < \varepsilon < 2$. 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 $p_0, \cdots, p_m$ are analytic as well as linearly independent, then, with probability* 1 *whenever $\bar{\alpha}$ lies in the interior of the constraint set and $N \geqq m$, 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 $0<\varepsilon<2$, 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

$$\Phi_\varepsilon(\bar{\alpha}')-\bar{\alpha}=\nabla\Phi_\varepsilon(\bar{\alpha})[\bar{\alpha}'-\bar{\alpha}]+O(\|\bar{\alpha}'-\bar{\alpha}\|^2).$$

In this expression, $\nabla\Phi_\varepsilon(\bar{\alpha})$ denotes the $m\times m$ matrix whose ijth entry is the ith component of $(\partial/\partial\alpha_j)\Phi_\varepsilon(\bar{\alpha})$. 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 exist a norm $\|\cdot\|$ on $\mathbb{R}^m$ and a number λ, $0\leqq\lambda<1$, for which an inequality of the form

$$(7)\qquad \|\nabla\Phi_\varepsilon(\bar{\alpha})\bar{\gamma}\|\leqq\lambda\|\bar{\gamma}\|$$

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

$$\mathscr{E}=\left\{\bar{\gamma}=(\gamma_1,\cdots,\gamma_m)^T:\sum_{i=1}^m\gamma_i=0\right\}\subseteq\mathbb{R}^m.$$

Using the fact that $\bar{\alpha}$ satisfies the likelihood equations (2), one verifies that $\nabla\Phi_\varepsilon(\bar{\alpha})=I-\varepsilon Q$. Here, Q is defined by

$$Q=\frac{1}{b(\bar{\alpha})}D\sum_{k=1}^N[\beta\bar{\delta}(x_k,\bar{\alpha})+(1-\beta)\delta_0(x_k,\bar{\alpha})\bar{e}]\bar{\delta}(x_k,\bar{\alpha})^T,$$

where $\bar{e}=(1,\cdots,1)^T$, $\delta_i(x,\bar{\alpha})=p_i(x)/p(x,\bar{\alpha})$ for $i=0,\cdots,m$, $\bar{\delta}(x,\bar{\alpha})=(\delta_1(x,\bar{\alpha}),\cdots,\delta_m(x,\bar{\alpha}))^T$, $b(\bar{\alpha})=\sum_{k=1}^N\bar{\alpha}^T\bar{\delta}(x_k,\bar{\alpha})$, and D is a diagonal matrix (d_{ij}) with $d_{ii}=\alpha_i$ for $i=1,\cdots,m$. One verifies without difficulty that $\mathscr{E}$ is invariant under Q and, hence, under $\nabla\Phi_\varepsilon(\bar{\alpha})$. 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 $\mathscr{E}$ with respect to which the operator norm of $\nabla\Phi_\varepsilon(\bar{\alpha})$ is less than 1.

Define an inner product $\langle\cdot,\cdot\rangle$ on $\mathscr{E}$ by $\langle\bar{\gamma},\bar{\gamma}'\rangle=\bar{\gamma}^TD^{-1}\bar{\gamma}'$ for $\bar{\gamma}$ and $\bar{\gamma}'$ in $\mathscr{E}$. It is easily shown that, with respect to this inner product, Q is symmetric (in fact, positive semidefinite) on $\mathscr{E}$. Indeed, for $\bar{\gamma}$ and $\bar{\gamma}'$ in $\mathscr{E}$, the fact that $\bar{e}^T\bar{\gamma}=\bar{e}^T\bar{\gamma}'=0$ yields

$$\begin{aligned}\langle\bar{\gamma},Q\bar{\gamma}'\rangle&=\bar{\gamma}^T\left\{\frac{1}{b(\bar{\alpha})}\sum_{k=1}^N\beta\bar{\delta}(x_k,\bar{\alpha})\bar{\delta}(x_k,\bar{\alpha})^T\right\}\bar{\gamma}'\\&=\left\{\frac{1}{b(\bar{\alpha})}\sum_{k=1}^N\beta\bar{\delta}(x_k,\bar{\alpha})\bar{\delta}(x_k,\bar{\alpha})^T\bar{\gamma}\right\}^T\bar{\gamma}'=\langle Q\bar{\gamma},\bar{\gamma}'\rangle.\end{aligned}$$

Similarly,

$$\langle\bar{\gamma},Q\bar{\gamma}\rangle=\frac{1}{b(\bar{\alpha})}\sum_{k=1}^N\beta[\bar{\delta}(x_k,\bar{\alpha})^T\bar{\gamma}]^2\geqq0.$$

From the symmetry of Q on $\mathscr{E}$ with respect to the inner product $\langle\cdot,\cdot\rangle$, it follows that, if ρ and τ denote the largest and smallest eigenvalues of Q corresponding to eigenvectors in $\mathscr{E}$, then the operator norm of $\nabla\Phi_\varepsilon(\bar{\alpha})$ on $\mathscr{E}$ with respect to this inner product is equal to the larger of $|1-\varepsilon\rho|$ and $|1-\varepsilon\tau|$. Thus, the first statement of the theorem will be proved if it can be shown that $\rho\leqq1$ and $0<\tau$. Now Q is a Markov matrix and it follows that $\rho\leqq1$. (See [1, pp. 265–270] for a discussion of Markov matrices.) Noting that, with probability 1, $\bar{\alpha}$ converges to $\bar{\alpha}^0$ as N approaches infinity, one can use arguments analogous to those employed in [3] to verify that, with probability 1, Q converges to

$$\frac{1}{\beta}D^0\int_{\mathbb{R}^n}[\beta\bar{\delta}(x,\bar{\alpha}^0)+(1-\beta)\delta_0(x,\bar{\alpha}^0)\bar{e}]\bar{\delta}(x,\bar{\alpha}^0)^T p(x,\bar{\alpha}^0)\,dx,$$

a positive definite operator on $\mathscr{E}$. Here $D^0=(d_{ij}^0)$ is a diagonal matrix with $d_{ii}^0=\alpha_i^0$ for $i=1,\cdots,m$. One concludes that, with probability 1 as N approaches infinity, Q is positive definite on $\mathscr{E}$ and, hence, that $0<\tau$.

To prove the second statement of the theorem, suppose that $N\geqq m$, that $\bar{\alpha}$ lies in the interior of the constraint set, and that $p_0,\cdots,p_m$ 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 $\mathscr{E}$ with probability 1 whenever $N\geqq m$. 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 $0<\varepsilon<2$. 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

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

for $i=1,\cdots,m$. While determining a proper step size for the usual steepest ascent procedure is not usually a serious problem, it can be a time-consuming nuisance.

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 $\nabla\Phi_\varepsilon(\bar{\alpha})$ is positive-definite on $\mathscr{E}$. (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 $\nabla\Phi_\varepsilon(\bar{\alpha})=I-\varepsilon Q$, regarded as an operator on $\mathscr{E}$. ($\nabla\Phi_\varepsilon(\bar{\alpha})$ is symmetric on $\mathscr{E}$ 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 $1-\varepsilon\tau=\varepsilon\rho-1$, i.e., $\varepsilon=2/(\rho+\tau)$, where ρ and τ are, respectively, the largest and smallest eigenvalues of the operator Q restricted to $\mathscr{E}$.

It is shown in the proof of Theorem 2 that ρ is never greater than 1. Thus the optimal ε is bounded below by $2/(1+\tau)$, where τ lies between 0 and 1. In particular, this lower bound on the optimal ε lies between 1 and 2. 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 Markov matrix Q always has 1 as an eigenvalue, the eigenvalue ρ of the restricted operator Q on $\mathscr{E}$ can be arbitrarily small (and, hence, the optimal ε can be arbitrarily large). Indeed, Q is nearly the zero operator on $\mathscr{E}$ 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 $k = 1, \cdots, N$. 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, $\nabla\Phi_\varepsilon(\bar{\alpha}) = I - \varepsilon Q \approx 0$. 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, nonzero indices i and j, $p_i(x_k) \approx p_j(x_k)$ for $k = 1, \cdots, N$. 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 $\nabla\Phi_\varepsilon(\bar{\alpha})$ on $\mathscr{E}$ 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.

From the above considerations, one concludes that $\varepsilon < 1$ always gives a suboptimal asymptotic rate of convergence. We also remark that experience indicates that care should be taken in choosing ε greater than 1, at least initially, since the constraints could then be violated for a poor choice of the starting value $\bar{\alpha}^{(1)}$. We feel that a rapid and reliable iterative procedure can be developed in which ε is initially chosen to be 1 and then, after a number of iterations, modified once to speed up convergence near the maximum-likelihood estimate. Such a procedure would be especially useful when the component populations are not well separated.

REFERENCES

[1] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[2] B. C. PETERS AND W. A. COBERLY, *The numerical evaluation of the maximum-likelihood estimate of mixture proportions*, Commun. Statist.—Theor. Meth., A5 (1976), no. 12, pp. 1127–1135.

[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*, SIAM J. Appl. Math., to appear.