The Annals of Statistics
1983, Vol. 11, No. 1, 86–94

THE GEOMETRY OF MIXTURE LIKELIHOODS: A GENERAL THEORY[1]

BY BRUCE G. LINDSAY

The Pennsylvania State University

In this paper certain fundamental properties of the maximum likelihood estimator of a mixing distribution are shown to be geometric properties of the likelihood set. The existence, support size, likelihood equations, and uniqueness of the estimator are revealed to be directly related to the properties of the convex hull of the likelihood set and the support hyperplanes of that hull. It is shown using geometric techniques that the estimator exists under quite general conditions, with a support size no larger than the number of distinct observations. Analysis of the convex dual of the likelihood set leads to a dual maximization problem. A convergent algorithm is described. The defining equations for the estimator are compared with the usual parametric likelihood equations for finite mixtures. Sufficient conditions for uniqueness are given. Part II will deal with a special theory for exponential family mixtures.

1. Introduction. Although the principal subject of this paper is the existence and uniqueness of the maximum likelihood estimator of a mixing distribution, the results arise from a study of the convex geometry of the likelihood when it is viewed as a curve in Euclidean space. Statistical problems such as existence, discreteness, support size characterization, and uniqueness are transformed into geometrical problems concerning support hyperplanes of the convex hull of the likelihood curve. This approach throws new light on what were considered to be difficult problems. The sequel to this paper, which considers mixtures of exponential family densities, will draw, in addition, on results from differential geometry and the study of totally positive kernels (Karlin, 1968).

The problem may be formulated as follows. Let $\{f_\theta\colon \theta \in \Omega\}$ be a parametric family of densities with respect to some sigma-finite measure. Let the parameter space Ω have a sigma field of measurable sets which contains all atomic sets $\{\theta\}$. Let $\mathscr{M}$ be the class of all probability measures on Ω. Define the function

$$f_Q(x) = \int f_\theta(x)\, dQ(\theta), \qquad Q \in \mathscr{M},$$

to be the *mixture density* corresponding to *mixing distribution* Q. Since the densities f_θ correspond to the atomic mixing distributions $\delta(\theta)$, which assign probability one to any set containing θ, they will be called the *atomic densities*. A finite discrete mixing distribution with support size J will be expressed as $Q = \sum \pi_j \delta(\theta_j)$, where it will be understood that j runs from 1 to J and that

(1.1) (a) the θ_j are distinct elements of Ω, (b) $\pi_j > 0$ for all j, (c) $\sum \pi_j = 1$.

Given a random sample $X_1, \cdots, X_n$ from the mixture density f_Q, the objective will be to estimate the mixing distribution Q by $\hat{Q}_n$, a maximizer of the likelihood

$$L(Q) = \textstyle\prod_{i=1}^n f_Q(x_i).$$

(The modifications necessary when f depends on i will be discussed in Section 2.) Of course, the mixture problem is poorly specified unless one can identify the mixing

Received February 1981; revised July 1982.

[1] This research was partially supported by The National Science Foundation under Grant MCS-8003081

AMS 1970 *subject classifications.* Primary 62A10, 62G05; secondary 52A40.

Key words and phrases. Mixture, maximum likelihood, geometry.

distribution Q from the distribution of the random variable X under the density function f_Q. It will be seen later, however, that there are identifiable functions of Q which will always be uniquely estimated by the method of maximum likelihood. Pertinent references for identifiability are Teicher (1963), Barndorff-Nielsen (1965), and Chandra (1977).

Under a fairly general set of assumptions, including identifiability, Kiefer and Wolfowitz (1956) identified consistency properties of the maximum likelihood estimator $\hat{Q}_n$. The first extensive examination of the structure of the estimator was by Simar (1976), who dealt exclusively with the case where the atomic densities are Poisson with mean θ. In this case, he showed consistency, and provided an algorithm for computation. Laird (1978) discussed the case of arbitrary atomic densities and gave firm conclusions about discreteness under regularity conditions on these densities. Lindsay (1981) presented some general structural properties, including an analogy to the usual likelihood equation and some moment properties of the estimator. Hill, *et. al.* (1980) consider a special case in where there are only a countable number of distinct θ in Ω. Jewell (1982) discusses mixtures of exponential densities, giving results analogous to Simar's.

2. Geometry and the likelihood. The geometric perspective is now introduced and the main results summarized. Suppose the observation vector $(x_1, \cdots, x_n)$ has K distinct data points $y_1, y_2, \cdots, y_K$. Let n_k be the number of x's which equal y_k. Define the *atomic* and *mixture likelihood vectors* to be $\mathbf{f}_\theta = (f_\theta(y_1), \cdots, f_\theta(y_K))$ and $f_Q = (f_Q(y_1), \cdots, f_0(y_K))$ respectively. The *likelihood curve* is the function from Ω to $\mathbb{R}$ defined by $\theta \to \mathbf{f}_\theta$. The trace of this curve, given by $\Gamma = \{\mathbf{f}_\theta : \theta \in \Omega\}$, represents all possible fitted values of the atomic likelihood vector. The key to the geometric approach is that an arbitrary convex combination of elements of Γ can be written as $\sum \pi_j \mathbf{f}_{\theta_j}$, with restrictions as in (1.1), and so is equal to the mixture likelihood vector $\mathbf{f}_Q$, where $Q = \sum \pi_j \delta(\theta_j)$.

If the observations $X_1, X_2, \cdots, X_n$ are independent but X_i has a density $f_{i\theta}(x)$ which depends on i in a known fashion, as would be the case if there were covariates or varying sample sizes, then geometric analysis of the likelihood vector $\mathbf{f}_\theta \equiv (f_{1\theta}(x_1), \cdots, f_{n\theta}(x_n))$ will yield conclusions as stated hereafter, but with n substituted for K in the theorems.

This interpretation of the likelihood allows one to draw on results from convex geometry such as to be found in Roberts and Varberg (1973), hereafter referenced as R&V. The convex hull of the set Γ, written conv(Γ), is the intersection of all convex sets containing Γ; it is itself a convex set. It is a fundamental result of convex geometry (R&V, page 76) that conv(Γ) is precisely the set of convex combinations of elements of Γ and so conv(Γ) $= \{\mathbf{f}_Q : Q \in \mathcal{M}, Q$ has finite support$\}$. If Γ is compact, as will be generally assumed, we have the stronger result that conv(Γ) $= \{f_Q : Q \in \mathcal{M}\}$ (Phelps, 1966, Proposition 1.2), under the measurability of the map $\theta \to \mathbf{f}_\theta$. As such the maximization problem can now be reformulated. Maximizing $L(Q)$ over $Q \in \mathcal{M}$ may be accomplished by maximizing the concave functional $\phi(\mathbf{f}) = \sum n_k \log f_k$ over $\mathbf{f}$ in the K-dimensional set conv(Γ).

If $\hat{\mathbf{f}} \in$ conv(Γ) maximizes $\phi(\mathbf{p})$, then there will exist $\hat{Q} \in \mathcal{M}$ such that $\mathbf{f}_{\hat{Q}} = \hat{\mathbf{f}}$ and $\hat{Q}$ maximizes $L(Q)$. Our approach will be to focus mainly on the geometric problem of determining $\hat{\mathbf{f}}$. This brings the general mixture theory close to those results from the theory of optimal design which arise from maximizing a concave functional such as log determinant on the convex set of possible design matrices. Many of the key references for this paper are thus from the design literature.

The following two examples illustrate the geometric approach and demonstrate the broad scope of the mixture model.

EXAMPLE 1. Define the atomic densities on $\mathbb{R}$ by

$$f_\theta(x) = \begin{cases} 1 & \text{if } \theta = x, \\ 0 & \text{otherwise.} \end{cases}$$

These functions are densities with respect to counting measure on $\mathbb{R}$. Although counting measure is not sigma-finite, if one restricts attention to those mixing distributions P which are discrete, then the mixture function f_P will also be a density with respect to counting

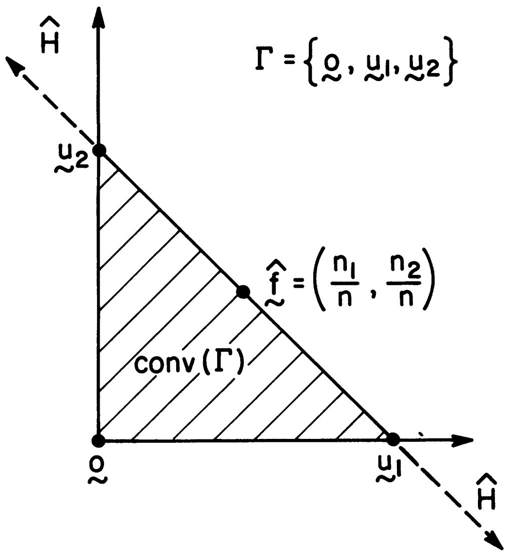
$\hat{H}$
$\Gamma = \{\underset{\sim}{0}, \underset{\sim}{u}_1, \underset{\sim}{u}_2\}$
$\underset{\sim}{u}_2$
$\hat{\underset{\sim}{f}} = \left(\frac{n_1}{n}, \frac{n_2}{n}\right)$
conv(Γ)
$\underset{\sim}{0}$
$\underset{\sim}{u}_1$
$\hat{H}$

FIG. 1. *In Example* 1 *with* $K = 2$ conv(Γ) *is the closed triangular region generated by extreme points* $\{\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2\}$.

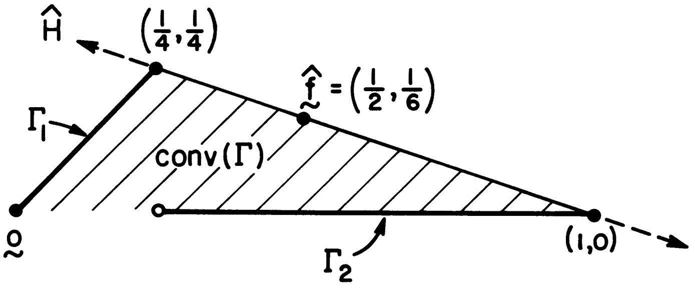
$\hat{H}$
$\left(\frac{1}{4}, \frac{1}{4}\right)$
$\hat{\underset{\sim}{f}} = \left(\frac{1}{2}, \frac{1}{6}\right)$
Γ_1
conv(Γ)
$\underset{\sim}{0}$
Γ_2
(1,0)

FIG. 2. *In Example* 2 *with* $K = 2$ *the set* Γ *is the union of* $\mathbf{0} = \{\mathbf{f}_\theta: \theta \in (0, y_1)\}$, $\Gamma_1 = \{\mathbf{f}_\theta: \theta \in [y_2, \infty)\}$ *and* $\Gamma_2 = \{\mathbf{f}_\theta: \theta \in [y_1, y_2)\}$. *Again* conv(Γ) *is a closed triangular region.*

measure, and the mixture density will be a density for P itself. The likelihood curve Γ consists of the origin $\mathbf{0}$ plus the unit vectors $\mathbf{u}_1 = (1, 0, \cdots, 0)$, $\mathbf{u}_2 = (0, 1, 0 \cdots 0) \cdots$, $\mathbf{u}_K = (0, 0, \cdots, 1)$, where the point $\mathbf{u}_k$ corresponds to the atomic likelihood vector evaluated at $\theta = y_k$. The convex hull of Γ is the K-dimensional solid figure with points of Γ as vertices. The maximum likelihood estimator for the mixing distribution is well known, being the empirical distribution function. It puts mass n_k/n at y_k, for $k = 1, \cdots, K$. In Figure 1, the sets Γ and conv(Γ) are illustrated for the case $K = 2$, along with the corresponding fitted likelihood vector $\hat{\mathbf{f}}$ and the supporting hyperplane $\hat{H}$ which contains $\hat{\mathbf{f}}$.

EXAMPLE 2. Suppose that the atomic densities are uniform on $(0, \theta]$: $f_\theta(x) = 1/\theta$ if $x \in (0, \theta]$. Let the parameter space Ω be $(0, \infty)$, with the Borel measurable sets. The family of mixture densities is now equivalent to the family of all nonincreasing density functions of $(0, \infty)$. The set Γ consists of K half-open line segments in K-space, plus the origin $\mathbf{0}$, which correspond respectively to values of θ in the intervals $(0, y_1)$, $[y_1, y_2)$, $\cdots$, $[y_K, \infty)$. In Figure 2, the sets Γ and conv(Γ) are illustrated for $y_1 = 1$, $y_2 = 4$, $n_1 = 1$, $n_2 = 1$. Once again the maximum likelihood estimator $\hat{Q}$ is known. It is a discrete measure with support on a subset of $\{y_1, \cdots, y_K\}$. The weights are such that the resulting estimated distribution function $F_{\hat{Q}}$ is the least concave majorant to the empirical distribution function of the sample. See, for example, Barlow, *et al* (1972, page 223).

3. Existence and discreteness. A famous theorem of Caratheodory (R&V, page 76) ensures that each element $\mathbf{f}$ of conv (Γ) has a representation as $\mathbf{f}_Q$, where Q has no more than $K + 1$ points of support. The following theorem presents results concerning the existence and support size of the estimator $\hat{\mathbf{f}}$.

THEOREM 3.1. *If* Γ *is closed and bounded (compact), then there exists a unique vector* $\hat{\mathbf{f}}$ *on the boundary of* $\operatorname{conv}(\Gamma)$ *which maximizes the log likelihood* $\phi(\cdot)$ *on that set. The point* $\hat{\mathbf{f}}$ *can be expressed as* $\mathbf{f}_Q$, *where* Q *has* K *or fewer points of support.*

This theorem is very similar to theorems from design. The proof is elementary: see Silvey (1980, page 72). In regard to the necessity of the conditions, we note that if Γ is unbounded in the positive orthant, then no maximum to $\phi(\cdot)$ exists. If Γ is not closed, the theorem can be applied to the closure; one must then determine if the limit points show up in the maximizing mixture and, if so, how to interpret them. This theorem settles an issue raised by Laird (1978) as to conditions necessitating the existence of maximizing mixtures with support size no greater than the sample size.

We note that while the uniqueness of $\hat{\mathbf{f}}$ will guarantee that some functions of Q are uniquely estimated, there may be more than one mixture Q which satisfies $\mathbf{f}_Q = \hat{\mathbf{f}}$. The uniqueness of $\hat{Q}$ is discussed in Section 8.

4. Differentiation of the log likelihood. In this section the maximization problem is recast by means of differentiation into a more useful form. In particular, the function $D(\theta; Q)$ defined below (4.1) will be repeatedly useful for determining properties of the estimator.

Viewed as a function of K variables, the function $\phi(\mathbf{p}) = \sum n_k \log p_k$ is differentiable in the positive orthant, with the directional derivative of ϕ at f_{Q_0} towards f_{Q_1} being

$$\Phi(\mathbf{f}_{Q_1}; \mathbf{f}_{Q_0}) = \lim \varepsilon^{-1}\{\phi((1-\varepsilon)\mathbf{f}_{Q_0} + \varepsilon\mathbf{f}_{Q_1}) - \phi(f_{Q_0})\}$$

$$= \sum n_k \{f_{Q_1}(y_k) - f_{Q_0}(y_k)\}/f_{Q_0(y_k)}.$$

We write

$$D(\theta; Q) = \Phi(\mathbf{f}_\theta; \mathbf{f}_Q) = \sum n_k \left\{\frac{f_\theta(y_k)}{f_Q(y_k)} - 1\right\}, \tag{4.1}$$

and note that $\int D(\theta; Q_0)\, dQ_1(\theta) = \Phi(\mathbf{f}_{Q_1}; \mathbf{f}_{Q_0})$.

By analogy to the equivalence theorems of design optimality (see Whittle, 1973, Theorem 1), we have the following fundamental theorem of mixture estimation.

THEOREM 4.1. A. *The measure* $\hat{Q}$ *which maximizes* $\log L(Q)$ *can be equivalently characterized by three conditions*: (a) $\hat{Q}$ *maximizes* $L(Q)$, (b) $\hat{Q}$ *minimizes* $\sup_\theta D(\theta; Q)$, (c) $\sup D(\theta, \hat{Q}) = 0$.

B. *The point* $(\hat{\mathbf{f}}, \hat{\mathbf{f}})$ *is a saddle point of* Φ, *in the sense that*

$$\Phi(\mathbf{f}_{Q_0}, \hat{\mathbf{f}}) \le 0 = \Phi(\hat{\mathbf{f}}, \hat{\mathbf{f}}) \le \Phi(\hat{\mathbf{f}}, \mathbf{f}_{Q_1})$$

for $Q_0, Q_1 \in \mathcal{M}$.

C. *The support of* $\hat{Q}$ *is contained in the set of* θ *for which* $D(\theta, \hat{Q}) = 0$.

This theorem may be given the following geometric interpretation. For $\mathbf{f}_Q$ in $\operatorname{conv}(\Gamma)$, let $\mathbf{f}_Q^*$ be the vector $(n_1/f_Q(y_1), \cdots, n_K/f_Q(y_K))$. With this notation $D(\theta, Q) = \langle \mathbf{f}_Q^*, \mathbf{f}_\theta \rangle - n$. The equation $\langle \mathbf{f}_Q^*, \mathbf{z} \rangle = n$ defines a hyperplane of points in $\mathbb{R}^K$; the measure Q is maximal if and only if $\{z: \langle f_{\hat{Q}}^*, \mathbf{z} \rangle = n\}$ is a support hyperplane of the set $\operatorname{conv}(\Gamma)$; moreover, since $\langle \mathbf{f}_Q, \mathbf{f}_\theta \rangle = n$ for all θ in the support of $\hat{Q}$, the supporting vectors $\mathbf{f}_\theta$ lie in that hyperplane.

5. Inequalities; duality. In this section several inequalities are derived which will be needed in the discussion of algorithms in Section 6. Along the way a dual maximization problem is presented. The setting is as follows: suppose at stage m of a maximization algorithm we have obtained estimator Q, or equivalently, $\mathbf{f}_Q$. Given the maximal gradient $\delta = \sup\{D(\theta; Q) : \theta \in \Omega\}$ and the value $\log L(Q)$, what can be said about the residual $\Delta = \log L(\hat{Q}) - \log L(Q)$? We derive the best possible upper and lower bounds for Δ given knowledge only of δ.

We first state and prove a duality result for the mixture problem whose design counterpart can be found in Silvey and Titterington (1973, 1974) and Pukelsheim (1980). The method of proof was also used in a duality result of Böhning and Hoffmann (1981). There are two equivalent forms for the dual problem, which we call Problems 1 and 2.

PROBLEM 1. Minimize $\phi(\mathbf{p})$ subject to $\mathbf{p} \in (\mathbb{R}^+)^K$ and $\Phi(f_\theta; \mathbf{p}) \le 0$ for all $\theta \in \Omega$.

PROBLEM 2. Maximize $\phi(\mathbf{w})$ subject to $\mathbf{w} \in (\mathbb{R}^+)^K$ and $\langle \mathbf{f}_\theta, \mathbf{w}\rangle \le n$ for all $\theta \in \Omega$.

THEOREM 5.1. *Suppose that Γ is compact and that the mixture maximum likelihood estimator is $\hat{\mathbf{f}}$. Then $\hat{\mathbf{f}}$ is the solution to Problem* 1; $\hat{\mathbf{w}} = \hat{\mathbf{f}}^*$ *is the solution to Problem* 2.

PROOF. Problems 1 and 2 are shown equivalent by the change of variable $(p_1, \cdots, p_K) = (n_1/w_1, \cdots, n_K/w_K)$. To solve Problem 1, note that ϕ's concavity implies

$$\phi(\mathbf{p}) + \Phi(\hat{\mathbf{f}}; \mathbf{p}) \ge \phi(\hat{\mathbf{f}}).$$

For feasible values of $\mathbf{p}$ the second summand is nonpositive, so $\phi(\mathbf{p}) \ge \phi(\hat{\mathbf{f}})$. But $\hat{\mathbf{f}}$ is feasible, so it is the solution. □

If the set of extreme points of Γ in conv(Γ) is $\{f_{\theta_1}, \cdots, f_{\theta_m}\}$, then Problem 2 is equivalent to the maximization of concave function $\phi(\mathbf{w})$ subject to finitely many linear constraints $\langle \mathbf{w}, \mathbf{f}_{\theta_i}\rangle \le n$ and so could be solved using convex programming techniques. Example 1 of Section 1 has a particularly simple formulation as Problem 2: maximize $\sum n_i \log w_i$ subject to $0 \le w_i \le n$. Solution: $\hat{w}_i = n$ implies $\hat{f}_Q(y_i) = n_i/n$.

COROLLARY 5.2. *If Γ is compact, then*

$$\Delta \le n \log(1 + \delta/n). \tag{5.1}$$

This bound is the best possible given knowledge of δ only.

PROOF. If $\sup_\theta D(\theta; Q) = \delta$, then for $\mathbf{f}_Q^* = (n_1/f_Q(y_1), \cdots n_K/f_Q(y_K))$ we have

$$\langle \mathbf{f}_Q^*, \mathbf{f}_\theta\rangle \le n + \delta \quad \text{for all} \quad \theta \in \Omega$$

so

$$\left\langle \frac{n}{n+\delta}\mathbf{f}_Q^*, \mathbf{f}_\theta \right\rangle \le n \quad \text{for all} \quad \theta \in \Omega.$$

Thus from Problem 2 we have $\phi((n/(n + \delta))\mathbf{f}_Q^*) \le \phi(\hat{\mathbf{w}})$ which gives result (5.1). This last inequality is an equality if $\hat{\mathbf{f}} = (n + \delta)n^{-1}\mathbf{f}_Q$. □

Next, geometric methods will be used to generate a lower bound for Δ similar to a design inequality given by Wynn (1970, equation 4.5; see also Fedorov, 1972, Theorem 2.5.2). Suppose that $D(\theta; Q) = \alpha > 0$. Then $\mathbf{f}_\theta$ is in the hyperplane $H = \{\mathbf{z}: \langle \mathbf{z}, \mathbf{f}_Q^*\rangle = n + \alpha\}$. Since it is also in the nonnegative orthant, it has a convex representation as $\sum \pi_k \mathbf{p}_k$ where $\mathbf{p}_k$ are the extreme points of the convex set C generated by intersecting H with the nonnegative orthant:

$$\mathbf{p}_1 = \left(\frac{n+\alpha}{n_1} f_Q(y_1), 0, 0, \cdots, 0\right), \qquad \mathbf{p}_2 = \left(0, \frac{n+\alpha}{n_2} f_Q(y_2), 0, \cdots, 0\right), \cdots$$

$$\vdots$$

$$\mathbf{p}_K = \left(0, 0, \cdots, \frac{n+\alpha}{n_k} f_Q(y_K)\right).$$

For every $0 \le \varepsilon < 1$ we have the system of inequalities

$$\inf_{\mathbf{p}\in C}\phi((1-\varepsilon)\mathbf{f}_Q + \varepsilon\mathbf{p}) \le \phi((1-\varepsilon)\mathbf{f}_Q + \varepsilon\mathbf{f}_\theta) \le \phi(\hat{\mathbf{f}}),$$

and so

$$\sup_\varepsilon \inf_{\mathbf{p}\in C}\phi((1-\varepsilon)\mathbf{f}_Q + \varepsilon\mathbf{p}) \le \phi(\hat{\mathbf{f}}). \tag{5.2}$$

For fixed $0 < \varepsilon < 1$ the function $\phi((1-\varepsilon)\mathbf{f}_Q + \varepsilon\mathbf{p})$ is concave and continuous as a function of $\mathbf{p} \in C$, so it will attain its global minimum somewhere on the extreme points of C, namely, $\{\mathbf{p}_1, \cdots, \mathbf{p}_K\}$ (R&V, page 124). The maximizing value of ε may be solved for explicitly for each $\mathbf{p}_k$, yielding $\hat{\varepsilon}_k = \alpha n_k / \{n(n + \alpha - n_k)\}$.

For this minimizing value the log likelihood can be written as

$$\phi((1-\hat{\varepsilon}_k)\mathbf{f}_Q + \hat{\varepsilon}_k\mathbf{p}_k) = \phi(\mathbf{f}_Q) + \beta_k \tag{5.3}$$

where

$$\beta_k = \log\left\{\left(1 + \frac{\alpha}{n}\right)^n \bigg/ \left(1 + \frac{\alpha}{n - n_k}\right)^{n-n_k}\right\}.$$

Let $h(x) = x \log(1 + \alpha/x)$. It is easily shown that h is concave and increasing in x, and so $\beta_k = h(n) - h(n - n_k) > 0$.

Since the above arguments hold for every $0 < \alpha \le \delta$, (5.2) and (5.3) imply the following theorem.

THEOREM 5.3. *Suppose that* Γ *is closed and bounded. If* Q *is a mixing distribution, let* $\delta = \sup\{D(\mathbf{f}_\theta; \mathbf{f}_Q): \theta \in \Omega\}$, $\Delta = \log L(\hat{Q}) - \log L(Q)$, *and* $n^* = \min_k\{n_k\}$. *Then*

$$\Delta \ge n \log\left(1 + \frac{\delta}{n}\right) - (n - n^*)\log\left(1 + \frac{\delta}{n - n^*}\right) \tag{5.4}$$

REMARK. The lower bound (5.4) will be met if $\sup\{\phi((1-\varepsilon)\mathbf{f}_Q + \varepsilon\mathbf{f}_\theta): 0 \le \notin \le 1, \theta \in \Omega\}$ occurs at $\mathbf{f}_\theta = \mathbf{p}_k$, with $n_k = n^*$, and $\hat{\mathbf{f}} = (1 - \hat{\varepsilon})\mathbf{f}_Q + \hat{\varepsilon}\,\mathbf{p}_k$.

6. Algorithms. Finding the estimator $\hat{\mathbf{f}}$ generally requires an iterative algorithm. Herein we first define a somewhat inefficient vertex direction method (VDM) whose simple proof of convergence will provide proof of the convergence of the more complicated procedures to be discussed later. At the mth step let Q_m be the current estimator; also let $\Delta_m = \log L(\hat{Q}) - \log L(Q_m)$. Procedure: find θ_m^* to maximize $D(\theta; Q_m)$; then find $\hat{\varepsilon}_m$ to maximize $\phi(\mathbf{f}_{Q_m}(1-\varepsilon) + \varepsilon\, \mathbf{f}_{\theta_m^*})$; then set $Q_{m+1} = (1 - \hat{\varepsilon}_m)Q_m + \hat{\varepsilon}_m\delta(\theta_m^*)$.

THEOREM 6.1. *If* Γ *is compact, then the above described* VDM *algorithm converges; that is,*

$$\lim_{m\to\infty}\mathbf{f}_{Q_m} = \hat{\mathbf{f}}.$$

PROOF. The proof follows Fedorov (1972, Theorem 2.5.3), with his inequality (2.5.20) being matched by inequality (5.2) of this paper. See also Wynn (1970). □

An analysis of the rate of convergence demonstrates that the VDM can slow down (i.e., $\Delta_m - \Delta_{m+1}$ proportional to Δ_m^2) as it nears the maximum if the support points of the solution $\hat{\mathbf{f}}$ are in the extreme rays of the positive orthant; in those cases the net increment $\Delta_m - \Delta_{m+1}$ will be near the minimal value in (5.4), while the residual likelihood Δ_m is near the upper bound (5.1). Various modifications have been proposed to the VDM in the design literature; see, for example, Wu (1978) and Böhning (1981).

We now consider several methods from the mixture literature. The iterative method of Simar (1976), given for the Poisson case but more generally applicable, involves finding the maxima θ^* over θ of $D(\theta; Q_m)$, then forming new estimator Q_{m+1} by maximizing weights π over the set of support points $\theta^* \cup$ support $\{Q_m\}$. The convergence theorem 6.1 now gives proof of the convergence of Simar's method, as each step of his algorithm has a greater increase in likelihood than the VDM. Böhning (1982) has a proof of convergence for the Poisson case only, along with a discussion of possible modifications for improved convergence. Silvey and Titterington (1973) prove the design version of this for D-optimality.

It has also been suggested (Laird, 1978; Jewell, 1982) that one use the EM algorithm (Dempster, Laird, and Rubin, 1977). This algorithm is directly applicable only to the constrained problem of maximization over discrete mixtures Q with a fixed number of support points. (See also Section 7). It can be adapted to the unconstrained problem by maximizing repeatedly with an increasing number of support points until the estimator stops changing. Although multiple initial values are generally taken to avoid termination at a local maximum, verification that $D(\theta; \hat{Q}) \le 0$ for all θ would seem both useful and necessary to ensure a global maximum has been obtained. Alternation of the EM algorithm with VDM would guarantee convergence, as the likelihood cannot decrease on the EM portion.

Silvey, Titterington and Torsney (1978) use a modified version of the EM algorithm for a design problem where points of support are fixed and weights are to be estimated. Böhning and Hoffmann (1981) have a discussion of other iteration procedures for this same problem.

7. Restricted support size estimators. In this section, the preceding theory of the mixture maximum likelihood estimator is linked to the more familiar problem of estimating a discrete mixing distribution when the support size is fixed in advance. Everitt and Hand (1981, page 8) have a general discussion of the latter problem. The two theories are linked through the differential $D(\theta; Q)$.

Theorem 4.1 indicates that the support points of the global maximum likelihood estimator $\hat{Q}$ are local maxima of $D(\theta; \hat{Q})$ and so under assumptions of differentiability in θ of D the measure $\hat{Q}$ satisfies the following at each interior support point θ^*:

$$D(\theta^*; \hat{Q}) = 0, \qquad D'(\theta^*, \hat{Q}) = 0, \qquad D''(\theta^*, \hat{Q}) \le 0. \tag{7.1}$$

On the other hand, a mixture Q which is a local maximum to $L(Q)$ within the $(2m-1)$-dimensional parameter space (π, θ), m fixed, satisfies the following:

THEOREM 7.1. *If $\tilde{Q}$ is an m-point maximum likelihood estimator then at each support point θ^* of $\tilde{Q}$ one has*

(a) $D(\theta^*; \tilde{Q}) = 0$; (b) *if $f_\theta(y)$ has a continuous first derivative in a neighborhood of θ^* for each data point y, then $D'(\theta^*; \tilde{Q}) = 0$; and* (c) *if $f_\theta(y)$ has a continuous second derivative in a neighborhood of θ^* for data point y, then*

$$D''(\theta^*; \tilde{Q}) - \textstyle\sum_{k=1}^{K} n_k \left\{\dfrac{f_{\theta^*}'(y_k)}{f_{\tilde{Q}}(y_k)}\right\}^2 \le 0.$$

This lemma can be proved by using the parametric likelihood equations for the problem. In Part II of this paper we will consider the number and location of the zeroes of functions

$D'(\theta; Q)$ in the exponential family. Part (b) of the lemma ties this investigation to the restricted support size problem. Part (c) indicates that a key distinction between the global maximum and restricted estimators is that the support points of a restricted estimator $\tilde{Q}$ may show up as *minima* to $D(\theta; \tilde{Q})$. Since the EM algorithm terminates at restricted maxima, a simple first order check for global maximality in this algorithm is to verify that $D''(\theta^*; \tilde{Q}) \le 0$ at the support points θ^* of the measure $\tilde{Q}$.

8. Uniqueness of the estimator. The next objective is to characterize conditions under which the estimator of the mixing distribution is unique. That is, when does there exist a unique $\hat{Q}$ with $\hat{\mathbf{f}} = f_{\hat{Q}}$? The point $\hat{\mathbf{f}}$ sits in at least one support hyperplane of the mixture likelihood set, namely $\{\mathbf{z}\colon \langle \hat{\mathbf{f}}^*, \mathbf{z}\rangle = n\}$. We first identify a property of this hyperplane which gives uniqueness.

THEOREM 8.1. *Assume* Γ *is compact. The point* $\hat{\mathbf{f}}$ *in the boundary of* $\operatorname{conv}(\Gamma)$ *has a unique convex representation in terms of elements of* Γ *if* $\hat{\mathbf{f}}$ *lies in a support hyperplane* H *of* $\operatorname{conv}(\Gamma)$ *which intersects* Γ *in a set of affinely independent vectors.* (*These must be fewer than* K *in number.*)

PROOF. If $\hat{\mathbf{f}}$ is in the support hyperplane $H = \{\mathbf{z}\colon \langle \mathbf{w}, \mathbf{z}\rangle = n\}$, then all possible support points of $\hat{\mathbf{f}}$ under any possible representation Q lie in H, and hence in $H \cap \Gamma$. If $H \cap \Gamma$ consists of affinely independent vectors, then there exist at most K of them; say they are $\{\mathbf{f}_{\theta_1}, \cdots, \mathbf{f}_{\theta_J}\}$. The point $\hat{\mathbf{f}}$ lies in the convex hull of Γ, and hence the convex hull of $H \cap \Gamma$, so there exists a convex representation as $\sum \pi_j \hat{\mathbf{f}}_{\theta_j} = \hat{\mathbf{f}}$. The weights $\pi_1, \cdots, \pi_J$, with $\sum \pi_j = 1$, are unique by affine independence.

Given an estimator $\hat{Q}$, one can thus verify that it is unique by finding all the solutions $\mathbf{f}_\theta$ to $\langle \hat{\mathbf{f}}^*, \mathbf{f}_\theta\rangle = n$ and checking if they form an affinely independent set. It is more difficult to verify in advance that a given family of densities will always generate unique estimators. One useful approach seems to be the following.

STEP 1. Ensure that there can be no more than K solutions to $\langle \hat{\mathbf{f}}^*, \mathbf{f}_\theta\rangle = n$ given the constraint $\langle \hat{\mathbf{f}}^*, \mathbf{f}_\theta\rangle \le n$.

STEP 2. Verify that every set of K or fewer vectors $\mathbf{f}_{\theta_1}, \cdots, \mathbf{f}_{\theta_J}$ are affinely independent.

This approach leads to uniqueness for the exponential family; however, only a few special cases of Step 1 will be handled in this paper; the general case is treated in part two. These special cases have a simple treatment because of the following result.

LEMMA 8.2. *An exponential polynomial of the form* $h(x) = \sum_k q_k(x)\exp(c_k z)$, *where* $q_k(x)$ *is a real ordinary polynomial of degree* a_k, *admits at most* $K - 1 + \sum a_k$ *zeroes, counting multiplicities* (*Polya and Szego*, 1925).

In order to bound the number of maxima (zeroes) to $\langle \hat{\mathbf{f}}^*, \mathbf{f}_\theta\rangle - n$ at K, we note that if $\sum \hat{w}_i f_\theta(y_i)$ is analytic in θ, then it suffices to show that $\sum \hat{w}_i f'_\theta(y_i) = D'(\theta; \hat{Q})$ has at most $2K - 1$ zeroes, as maxima must alternate with minima. For the exponential family with densities of the form $f_\theta(x) = \exp(\theta x - \kappa(\theta))$, we have

$$D'(\theta; \hat{Q}) = \sum \hat{w}_k\{y_k - \kappa'(\theta)\}\exp\{\theta y_k - \kappa(\theta)\}. \tag{8.2}$$

Thus if $\kappa'(\theta)$ is of the right form, Lemma 8.2 can be applied. In particular, the normal density with mean θ and fixed variance 1 satisfies $\kappa'(\theta) = \theta$, so (8.2) has at most $2K - 1$ zeroes. The gamma density with scale parameter $1/\theta$ and known shape parameter α has $\kappa'(\theta) = \alpha/\theta$, so the same bound applies. (Jewell, 1981, has a slightly different proof.) The Poisson density has $\kappa(\theta) = e^\theta$, so that the maximal number of zeroes to (8.2) equals the number of distinct elements in

$$\{y_1, \cdots, y_K, y_1 + 1, \cdots, y_K + 1\}$$

minus one. Thus for the observed set {1, 2, 3, 7, 8, 9} there are at most 8 zeroes to (8.2) and hence at most 4 points of support in the interior of the natural parameter space.

Moving to Step 2, we note that it suffices that the vectors $\mathbf{f}_{\theta_1}, \cdots, \mathbf{f}_{\theta_K}$ be linearly independent, which is equivalent to

$$\det[f_{\theta_1}, \cdots, f_{\theta_K}] \neq 0$$

where $[\mathbf{f}_{\theta_1}, \cdots, \mathbf{f}_{\theta_K}]$ is the $K \times K$ matrix with kth column $\mathbf{f}_{\theta_k}$. This is true for the exponential family densities because they are generated by a totally positive kernel; see Karlin (1968, pages 18-20, 117-120) for other examples which satisfy Step 2 for this reason.

Acknowledgments. I wish to express my thanks to the referee who pointed out the link to optimal design theory and the key references therein.

REFERENCES

BARLOW, R. E., et al. (1972). *Statistical Inference Under Order Restrictions.* Wiley, New York.

BARNDORFF-NEILSON, O. (1965). Identifiability of mixtures of exponential families. *J. Math. Anal. Appl.* **12** 115-121.

BÖHNING, D. (1981). An exchange algorithm for optimal design problems (abstract). *International Symposium on Semi-Infinite Programming and Applications, Book of Abstracts* (edited by S. Zlobec).

BÖHNING, D. (1982). Convergence of Simar's Algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.* **10** 1006-1008.

BÖHNING, D. and HOFFMANN, K. M. (1982). Numerical techniques for estimating probabilities. *J. Statist. Comp. Sim.* To appear.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1-38.

EVERITT, B. S. and HAND, D. J. (1981). *Finite Mixture Distributions.* Chapman and Hall, London.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments.* Academic, New York.

HILL, D. L., et al. (1980). Maximum likelihood estimation for mixtures. *Canad. J. Statistics* **8** 87-93.

JEWELL, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10** 479-484.

KARLIN, S. (1968). *Total Positivity.* Stanford University Press, Stanford, California.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887-906.

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805-811.

LINDSAY, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. *Statistical Distributions in Scientific Work* (Editor, C. Taillie, *et al.*) **5** 95-110.

PHELPS, R. R. (1966). *Lectures on Choquet's Theorem.* Van Nostrand, Princeton.

POLYA, G. and SZEGO, G. (1925). *Aufgaben und Lehrsätze aus der Analysis,* 2. Springer, Berlin.

PUKELSHEIM, F. (1980). On linear regression designs which maximize information. *J. Statist. Plann. Inference* **4** 339-364.

ROBERTS, A. W. and VARBERG, D. E. (1973). *Convex Functions.* Academic, New York.

SILVEY, S. D. (1980). *Optimal Design.* Chapman and Hall, London.

SILVEY, S. D. and TITTERINGTON, D. M. (1973). A geometric approach to optimal design theory. *Biometrika* **60** 21-32.

SILVEY, S. D. and TITTERINGTON, D. M. (1974). A Largrangian approach to optimal design. *Biometrika* **61** 299-302.

SILVEY, S. D., TITTERINGTON, D. M., and TORSNEY, B. (1978). An algorithm for optimal designs on a finite design space. *Commun. Statist. Theor. Meth.* **A7** 1379-1389.

SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson Process. *Ann. Statist.* **4** 1200-1209.

TEICHER, H. (1963). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244-248.

WHITTLE, P. (1973). Some general points in the theory of optimal experimental design. *J. Roy. Statist. Soc. Ser. B* **35** 123-130.

WU, C.-F. (1978). Some algorithmic aspects of the theory of optimal designs. *Ann. Statist.* **6** 1286-1301.

WYNN, H. P. (1970). The sequential generation of *D*-optimum experimental designs. *Ann. Math. Statist.* **41** 1655-1664.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
219 POND LABORATORY
UNIVERSITY PARK, PENNSYLVANIA 16802