The Annals of Statistics
1974, Vol. 2, No. 3, 454-463

ESTIMATION OF DISTRIBUTIONS USING ORTHOGONAL EXPANSIONS

BY BRADFORD R. CRAIN

University of Oklahoma

Let $f(x)$ be a continuous, strictly positive probability density function over an interval $[a, b]$ and $F(x)$ its associated cdf. Suppose $\{\phi_i(x)\}_{i=0}^{\infty}$ is a complete orthonormal basis for $L_2[a, b]$ and that $f(x)$ and $\log f(x)$ have orthogonal series expansions, in the ϕ_i's, over $[a, b]$. Estimators for $f(x)$ and $F(x)$ are chosen from the canonical exponential family of distributions generated by $\{\phi_i(x)\}_{i=0}^{\infty}$, and convergence theorems are presented for these estimators in the special case of Legendre polynomials over $[-1, 1]$.

1. Introduction. Let $f(x)$ be a continuous, strictly positive density over the interval $[a, b]$ and $F(x)$ its cdf. We assume $f(x)$ simultaneously has the following orthogonal series expansions:

$$f(x) = \sum_{i=0}^{\infty} \theta_i \phi_i(x) \qquad a \leqq x \leqq b \tag{1.1}$$

$$= \exp[\sum_{i=1}^{\infty} \tau_i \phi_i(x) - \Psi(\tau)] \qquad a \leqq x \leqq b\,. \tag{1.2}$$

The general problem is to estimate $f(x)$ and $F(x)$ over the support set $[a, b]$ within the assumed model (1.1) and (1.2). The approach is a two-phase process: 1. We approximate $f(x)$ and $F(x)$ in a very natural way using the canonical exponential family of distributions. 2. We estimate the approximations by taking advantage of their "intrinsic estimability." The approximations are of the form

$$p_m(x \mid \tau^*) = \exp[\sum_{i=1}^{m} \tau_i^* \phi_i(x) - \Psi_m(\tau^*)] \qquad \text{and} \qquad F_m^*(x) = \int_a^x p_m(y \mid \tau^*)\, dy\,,$$

where the vector $\tau^* \in R^m$ arises from restricted maximum likelihood considerations and is shown to be the almost sure limit of a sequence $\{\tau_n^*\}$ of vector estimates based on a random sample of size n.

This procedure efficiently reduces an infinite-dimensional problem to one of finite-dimension, and allows the simultaneous estimation of the distribution over the entire interval of support, rather than at just a point. This latter property is due to the one-to-one correspondence between the class of all finite-dimensional canonical exponential distributions and the set of all finite-dimensional vectors (equivalently, the set of all finitely nonzero elements of R^∞). Consequently, certain parametric procedures can be adapted to attack a basically non-parametric problem.

Several examples of the approximation method are illustrated in Figures 1–3. Computations were done for the following densities over $[-1, 1]$: $f_1(x)$ was the uniform density, $f_2(x)$ was a truncated normal density, $f_3(x) = \frac{3}{4} - (|x|/2)$, $f_4(x)$

Received June 1972; revised June 1973.

AMS 1970 *subject classifications.* Primary 62G05; Secondary 62G99, 41A10, 42A08.

Key words and phrases. Densities, estimation of densities, cumulative distribution functions, estimation of distributions, restricted maximum likelihood estimation, exponential families.

was proportional to $(.5\cos(10x) + 1)$, and $f_5(x)$ was split-uniform, taking the value $\frac{2}{3}$ over $[-1, 0]$ and $\frac{1}{3}$ over $(0, 1]$. In each example the interval of support was $[-1, 1]$ because the computations were done in terms of Legendre polynomials. Since $f_1(x)$ and $f_2(x)$ are exactly of the exponential class, they furnished a convenient computer check and were approximated exactly, hence the results do not appear in the figures. All five examples are piecewise smooth and positive so that the model (1.1) and (1.2) will apply at points of continuity. In every case the approximation procedure worked well, surprisingly even for $f_5(x)$ which is discontinuous at the origin.

For convenience, the series expansions in (1.1) and (1.2) are assumed to converge uniformly on $[a, b]$ (the principal reason is to ensure the conclusions of Lemma 2.1). The system of functions $\{\phi_i(x)\}_{i=0}^{\infty}$ is any complete orthonormal basis for the space $L_2[a, b]$ such that each $\phi_i(x)$ is continuous and $\phi_0(x)$ is a constant. For $1 \leqq i, j < \infty$, $\int_a^b \phi_i(x)\phi_j(x)\,dx = \delta_{ij}$, where δ_{ij} is the Kronecker delta. The coefficients θ_i and τ_i in the model are $\int_a^b \phi_i(x)f(x)\,dx$ and $\int_a^b \phi_i(x)\log f(x)\,dx$ respectively. The normalizing function $\Psi(\tau) = \Psi(\tau_1, \tau_2, \tau_3, \cdots)$ is determined by the condition that $\int_a^b f(x)\,dx = 1$, thus $\exp[\Psi(\tau)] = \int_a^b \exp[\sum_{i=1}^{\infty} \tau_i\phi_i(x)]\,dx$. Furthermore, we impose a regularity condition on $\{\phi_i(x)\}_{i=0}^{\infty}$:

REGULARITY CONDITION. Assume that every non-degenerate generalized polynomial $\sum_{i=1}^{m} c_i\phi_i(x)$ achieves its supremum over $[a, b]$ at a finite number of points at most; let $\Phi(m)$ denote that number of points.

Then for every integer m and $c \in R^m$ $(c \neq 0)$,

$$\int_a^b [\textstyle\sum_{i=1}^{m} c_i\phi_i(x)]\,dP(x) < \max_{a \leq x \leq b} [\textstyle\sum_{i=1}^{m} c_i\phi_i(x)]\,, \tag{1.3}$$

where dP is any probability measure over the Lebesgue subsets of $[a, b]$ which has more than $\Phi(m)$ points of support in $[a, b]$.

REMARK 1.1. The regularity condition on the system $\{\phi_i(x)\}_{i=0}^{\infty}$, more specifically relation (1.3), is needed in the proofs of Theorems 3.1 and 3.3. Since in this paper we eventually choose $\{\phi_i(x)\}_{i=0}^{\infty}$ to be the Legendre polynomials over $[-1, 1]$, we define $\Phi(m)$ to be $[(m/2) + 1]$, the greatest integer less than or equal to $(m/2) + 1$.

Now suppose $X_1, X_2, \cdots, X_n$ is a random sample from $f(x)$.

DEFINITION 1.1. For $i = 1, 2, 3, \cdots$, the sample mean $\bar{\phi}_i$ of $\phi_i(x)$ is defined to be $\bar{\phi}_i = (1/n)\sum_{j=1}^{n} \phi_i(X_j)$.

The likelihood function $L(X_1, X_2, \cdots, X_n; \tau)$ is given by

$$L(X_1, X_2, \cdots, X_n; \tau) = \textstyle\prod_{j=1}^{n} f(X_j) = \exp\{n[\sum_{i=1}^{\infty} \tau_i\bar{\phi}_i - \Psi(\tau)]\}\,. \tag{1.4}$$

The method of maximum likelihood is not adequate to estimate the τ_i's since there are too many of them. Alternately, the likelihood function can be made arbitrarily large by choosing $\tau \in R^{\infty}$ which concentrates the probability mass of $f(x)$ about the sample points.

We consider a restricted maximum likelihood approach. Set all but the first

m of the τ_i's equal to zero and consider maximizing over all $\tau \in R^m$ the function $L_m(X_1, X_2, \cdots, X_n; \tau)$ given by

$$(1.5) \qquad L_m(X_1, X_2, \cdots, X_n; \tau) = \prod_{j=1}^{n} \exp[\sum_{i=1}^{m} \tau_i \phi_i(X_j) - \Psi_m(\tau)] = \exp\{n[\sum_{i=1}^{m} \tau_i \bar{\phi}_i - \Psi_m(\tau)]\},$$

where $\Psi_m(\tau) = \Psi_m(\tau_1, \tau_2, \cdots, \tau_m) = \Psi(\tau_1, \tau_2, \cdots, \tau_m, 0, 0, \cdots)$. It will be shown that the restricted maximum likelihood problem has a solution (unique) with probability one whenever $n > [(m/2) + 1]$.

DEFINITION 1.2. The solution vector to the restricted maximum likelihood problem, whenever it exists, will be denoted by $\tau_n^* = (\tau_{n1}^*, \cdots, \tau_{nm}^*)'$.

REMARK 1.2. τ_n^* has a number of useful properties to be presented later. Note that τ_n^* lies in R^m by (1.5). Its components depend on both m and n; this dependence has been suppressed to ease notation.

2. The canonical exponential family.

DEFINITION 2.1. The family of canonical exponential densities (of size m) generated by $\{\phi_i(x)\}_{i=1}^m$ is the collection of densities $\{p_m(x \mid \tau) : \tau \in R^m\}$ over $[a, b]$, where for each $\tau \in R^m$,

$$(2.1) \qquad p_m(x \mid \tau) = \exp[\sum_{i=1}^{m} \tau_i \phi_i(x) - \Psi_m(\tau)], \qquad a \leqq x \leqq b.$$

The integral $\exp[\Psi_m(\tau)] = \int_a^b \exp[\sum_{i=1}^{m} \tau_i \phi_i(x)]\, dx$ possesses derivatives of all orders with respect to the τ_i's which may be passed through the integral sign. Consequently, by differentiating $\exp[\Psi_m(\tau)]$, and a little manipulation, one obtains

$$(2.2) \qquad \partial\Psi_m(\tau)/\partial\tau_i = E[\phi_i(X)], \qquad \partial^2\Psi_m(\tau)/\partial\tau_i\partial\tau_j = \text{Cov}[\phi_i(X), \phi_j(X)],$$

where expectation is taken with respect to $p_m(x \mid \tau)$ in (2.1). The function $\Psi_m(\tau)$ is well-defined, analytic and strictly convex throughout R^m (see Barndorff-Nielsen [1]).

LEMMA 2.1. *Let* $f(x) = \exp[\sum_{i=1}^{\infty} \tau_i \phi_i(x) - \Psi(\tau)]$, *where the series converges uniformly on* $[a, b]$. *Then*

(i) $\exp[\sum_{i=1}^{m} \tau_i \phi_i(x)] \to \exp[\sum_{i=1}^{\infty} \tau_i \phi_i(x)]$, *uniformly in* x, *as* $m \to \infty$.

(ii) $\Psi_m(\tau) = \Psi(\tau_1, \tau_2, \cdots, \tau_m, 0, 0, \cdots) \to \Psi(\tau) = \Psi(\tau_1, \tau_2, \cdots)$ *as* $m \to \infty$.

(iii) $\exp[\sum_{i=1}^{m} \tau_i \phi_i(x) - \Psi_m(\tau)] \to \exp[\sum_{i=1}^{\infty} \tau_i \phi_i(x) - \Psi(\tau)]$, *uniformly in* x, *as* $m \to \infty$.

(iv) $\int_a^y \exp[\sum_{i=1}^{m} \tau_i \phi_i(x) - \Psi_m(\tau)]\, dx \to \int_a^y \exp[\sum_{i=1}^{\infty} \tau_i \phi_i(x) - \Psi(\tau)]\, dx$, *uniformly in* y, *as* $m \to \infty$ $(a \leqq y \leqq b)$.

(v) $\int_a^b \phi_j(x) \exp[\sum_{i=1}^{m} \tau_i \phi_i(x) - \Psi_m(\tau)]\, dx \to \int_a^b \phi_j(x) \exp[\sum_{i=1}^{\infty} \tau_i \phi_i(x) - \Psi(\tau)]\, dx$, *uniformly in* j, *as* $m \to \infty$.

PROOF. The uniform limit of a sequence of continuous functions on $[a, b]$ is continuous, and hence $\sum_{i=1}^{\infty} \tau_i \phi_i(x)$ is uniformly continuous over $[a, b]$. The remainder of the proof follows standard lines and is omitted.

By Lemma 2.1, the canonical exponential family provides distributions which uniformly approximate $f(x)$ and $F(x)$ over $[a, b]$.

3. Existence of the restricted maximum likelihood estimator. By (1.5) the restricted MLE τ_n^* exists iff $Q_n^m(\tau) = \sum_{i=1}^m \tau_i \bar{\phi}_i - \Psi_m(\tau)$ achieves its supremum over R^m. The function $\Psi_m(\tau)$ is strictly convex so that $Q_n^m(\tau)$ is strictly concave on R^m (see Barndorff-Nielsen [1]) and τ_n^* will be unique when it exists (Zangwill [11]).

Let $h(\tau)$ be a class C^2 function from R^m to R^m. The Hessian matrix for $h(\tau)$ evaluated at τ is defined to be the $m \times m$ matrix $H_h(\tau)$ whose ijth element is $\partial^2 h(\tau)/\partial\tau_i \partial\tau_j$.

Now with probability one, $\bar{\phi}_i \to \theta_i$ ($\theta_i = \int_a^b \phi_i(x) f(x)\,dx$) as $n \to \infty$ (n = sample size), so that corresponding to an infinite sample, the function $Q_\theta^m(\tau) = \sum_{i=1}^m \tau_i \theta_i - \Psi_m(\tau)$ is introduced. $Q_\theta^m(\tau)$ is a strictly concave function of τ, since its Hessian matrix is the negative of the Hessian matrix for $\Psi_m(\tau)$, and is thus negative definite (a condition sufficient for strict concavity, Zangwill [11]).

The next theorem is a result which can be found in Barndorff-Nielsen. An independent proof is given which is of interest by itself.

THEOREM 3.1. *The function $Q_\theta^m(\tau)$ achieves its supremum over R^m for every positive integer m.*

PROOF. $Q_\theta^m(\tau)$ is continuous, strictly concave and $Q_\theta^m(0)$ is finite. Choose an integer n such that $Q_\theta^m(0) \geqq -n$. The set $S_n^m = \{\tau \in R^m \mid Q_\theta^m(\tau) \geqq -n\}$ is a closed, convex set containing the origin. We show S_n^m is bounded, hence compact, and thus $Q_\theta^m(\tau)$ achieves its supremum on S_n^m.

By a ray in R^m we mean the point set $\{y + \rho c \mid \rho \geqq 0\}$ where $\|c\| = 1$, $y, c \in R^m$, ρ real. A closed, convex subset of R^m is bounded iff it contains no rays (Rockafellar [7]). Thus we show that an arbitrary ray $\{y + \rho c \mid \rho \geqq 0\}$ cannot lie entirely in S_n^m. Define $\phi\colon R^1 \to R^1$ by $\phi(\rho) = Q_\theta^m(y + \rho c)$. Then $\phi(\rho)$ is strictly concave, hence $\phi'(\rho)$ is strictly decreasing. If ρ_0 is a number such that $\phi'(\rho_0) < 0$, then by the Taylor's series expansion of $\phi(\rho)$ for $\rho > \rho_0$,

$$\phi(\rho) = \phi(\rho_0) + \phi'(\xi)(\rho - \rho_0) \tag{3.1}$$

for some $\xi \in (\rho_0, \rho)$, and

$$\phi(\rho) < \phi(\rho_0) + \phi'(\rho_0)(\rho - \rho_0)\,, \qquad \rho > \rho_0\,. \tag{3.2}$$

This would imply that $\phi(\rho) \to -\infty$ as $\rho \to \infty$ and S_n^m could not contain the ray $\{y + \rho c \mid \rho \geqq 0\}$.

To finish the theorem we show $\phi'(\rho) < 0$ for all large ρ. With $y = (y_1, y_2, \cdots, y_m)'$ $c = (c_1, c_2, \cdots, c_m)'$, $\phi(\rho) = \sum_{i=1}^m (y_i + \rho c_i)\theta_i - \Psi_m(y + \rho c)$, by the chain rule we have

$$\phi'(\rho) = \sum_{i=1}^m c_i \theta_i - \sum_{i=1}^m c_i\, \partial\Psi_m(y + \rho c)/\partial\tau_i\,. \tag{3.3}$$

Using (2.1) and (2.2) and interchanging summation and integration, we get

$$\sum_{i=1}^m c_i\, \partial\Psi_m(y + \rho c)/\partial\tau_i = E[\textstyle\sum_{i=1}^m c_i \phi_i(X)] \tag{3.4}$$

where the expectation is taken with respect to $p_m(x \mid y + \rho c)$. As $\rho \to \infty$, the probability distribution $p_m(x \mid y + \rho c)$ becomes more and more peaked or concentrated about the points $x \in [a, b]$ for which $\sum_{i=1}^m c_i \phi_i(x)$ is maximum, and so

$$\sum_{i=1}^m c_i \partial \Psi_m(y + \rho c)/\partial \tau_i \to \sup_{a \leq x \leq b} \sum_{i=1}^m c_i \phi_i(x) \qquad \text{as } \rho \to \infty . \tag{3.5}$$

In view of (1.3), (3.4) and the observation that $\sum_{i=1}^m c_i \theta_i = \int_a^b [\sum_{i=1}^m c_i \phi_i(x)] f(x)\, dx$, it follows that $\phi'(\rho) < 0$ for ρ large.

DEFINITION 3.1. For each m, the unique vector in R^m at which $Q_\theta{}^m(\tau)$ achieves its supremum will be denoted by τ^*.

The dimension and components of τ^* all depend on m.

The vector τ^* maximizes $Q_\theta{}^m(\tau)$ over R^m iff

$$\nabla Q_\theta{}^m(\tau) = 0 \qquad \text{(Zangwill [11])}, \qquad \text{or} \tag{3.6}$$

$$\nabla \Psi_m(\tau^*) = \theta , \tag{3.7}$$

where $\nabla \Psi_m(\tau) = (\partial \Psi_m(\tau)/\partial \tau_1, \cdots, \partial \Psi_m(\tau)/\partial \tau_m)'$ is the gradient mapping and $\theta = (\theta_1, \theta_2, \cdots, \theta_m)'$. When the restricted MLE $\tau_n{}^*$ exists, it must satisfy

$$\nabla Q_n{}^m(\tau_n{}^*) = 0 , \qquad \text{or} \tag{3.8}$$

$$\bar{\phi} = \nabla \Psi_m(\tau_n{}^*) , \qquad \text{where} \quad \bar{\phi} = (\bar{\phi}_1, \bar{\phi}_2, \cdots, \bar{\phi}_m)' . \tag{3.9}$$

DEFINITION 3.2. The symbol $\mathscr{R}(\nabla \Psi_m)$ represents the range of the gradient mapping $\nabla \Psi_m(\tau)$, that is, $\mathscr{R}(\nabla \Psi_m) = (\nabla \Psi_m(\tau) \mid \tau \in R^m\}$.

REMARK 3.1. Let $a = (a_1, a_2, \cdots, a_m)' \in R^m$. The function $Q_a(\tau) = \sum_{i=1}^m \tau_i a_i - \Psi_m(\tau)$ achieves its supremum over R^m iff $\nabla Q_a(\tau) = 0$ has a solution iff $a \in \mathscr{R}(\nabla \Psi_m)$. By Theorem 3.1 we always have $\theta = (\theta_1, \theta_2, \cdots, \theta_m)' \in \mathscr{R}(\nabla \Psi_m)$. Also, $\tau_n{}^*$ will exist iff $\bar{\phi} \in \mathscr{R}(\nabla \Psi_m)$.

The following results can be found in Barndorff-Nielsen [1]:

LEMMA 3.1. *$\mathscr{R}(\nabla \Psi_m)$ is an open, bounded, convex set.*

LEMMA 3.2. *The gradient mapping $\nabla \Psi_m \colon R^m \to \mathscr{R}(\nabla \Psi_m)$ is 1–1 and has a continuously differentiable inverse on $\mathscr{R}(\nabla \Psi_m)$.*

DEFINITION 3.3. The inverse mapping $(\nabla \Psi_m)^{-1}$ is denoted by $\phi_m \colon \mathscr{R}(\nabla \Psi_m) \to R^m$.

THEOREM 3.2. *For each m and for all sufficiently large n, $\tau_n{}^*$ exists and $\tau_n{}^* \to \tau^*$ almost surely as $n \to \infty$.*

PROOF. $\mathscr{R}(\nabla \Psi_m)$ is open and contains θ. Since $\bar{\phi} \to \theta$ a.s. as $n \to \infty$, with probability one we have $\bar{\phi} \in \mathscr{R}(\nabla \Psi_m)$ for all large n. But $\tau_n{}^*$ exists iff $\bar{\phi} \in \mathscr{R}(\nabla \Psi_m)$, so that for n sufficiently large, $\nabla \Psi_m(\tau_n{}^*) = \bar{\phi}$, and by Lemma 3.2,

$$\tau_n{}^* = \phi_m(\nabla \Psi_m(\tau_n{}^*)) = \phi_m(\bar{\phi}) , \tag{3.10}$$

which implies

$$\begin{aligned} \lim_{n\to\infty} \tau_n{}^* &= \lim_{n\to\infty} \phi_m(\bar{\phi}) = \phi_m(\theta) \quad \text{(a.s.)} \\ &= \phi_m(\nabla \Psi_m(\tau^*)) = \tau^* . \end{aligned} \tag{3.11}$$

DEFINITION 3.4. Θ_m is the set of vectors $v \in R^m$ with $v_i = \int_a^b \phi_i(x)\, dP(x)$, $1 \leqq i \leqq m$, where $dP(x)$ is any probability measure on the Lebesgue subsets of $[a, b]$ having $[m/2] + 2$ or more points of support.

THEOREM 3.3. *For each m the sets* $\mathscr{R}(\nabla\Psi_m)$ *and* Θ_m *are identical.*

PROOF. Suppose $v \in \mathscr{R}(\nabla\Psi_m)$. Then $v = \nabla\Psi_m(\tau)$ for some $\tau \in R^m$. By (2.2), $v_i = \int_a^b \phi_i(x)\, dP(x)$ where $dP(x) = p_m(x \mid \tau)\, dx$ and hence $v \in \Theta_m$.

Next suppose $v \in \Theta_m$. The function $Q_v(\tau) = \sum_{i=1}^m \tau_i v_i - \Psi_m(\tau)$ can be shown to achieve its supremum over R^m by using (1.3) and the identical proof of Theorem 3.1. Hence $v \in \mathscr{R}(\nabla\Psi_m)$, and so $\Theta_m = \mathscr{R}(\nabla\Psi_m)$.

Let $X_1, X_2, \cdots, X_n$ be a random sample with common marginal density $f(x)$ as given in (1.1) and (1.2). If $F_n(x)$ is the empirical cdf and $n > [m/2] + 1$ then $dF_n(x)$ a.s. has at least $[m/2] + 2$ points of support in $[a, b]$. Since $\bar{\phi}_i = \int_a^b \phi_i(x)\, dF_n(x)$, $1 \leqq i \leqq m$, by Theorem 3.3 we have $\bar{\phi} = (\bar{\phi}_1, \bar{\phi}_2, \cdots, \bar{\phi}_m)' \in \mathscr{R}(\nabla\Psi_m)$ a.s. whenever $n > [m/2] + 1$. But $\bar{\phi} \in \mathscr{R}(\nabla\Psi_m)$ iff τ_n^* exists, so the next result is proven.

THEOREM 3.4. *The restricted maximum likelihood estimate* τ_n^* *exists almost surely whenever* $n > [m/2] + 1$.

The next theorem gives an appealing property of τ^*.

THEOREM 3.5. *Suppose that* $f(x)$ *is of the canonical exponential family of densities, i.e.,* $f(x) = \exp[\sum_{i=1}^s \tau_i \phi_i(x) - \Psi_s(\tau)]$ *for some positive integer s. Then* $p_m(x \mid \tau^*) \equiv f(x)$ *on* $[a, b]$ *whenever* $m \geqq s$.

PROOF. For $m \geqq s$ we have $\theta = \nabla\Psi_m(\tau^*)$ by (3.7), where $\theta \in R^m$. Set $\tau = (\tau_1, \tau_2, \cdots, \tau_m)'$ where $\tau_i = 0$ for $s + 1 \leqq i \leqq m$. Then $\nabla\Psi_m(\tau) = \theta$ by (2.2), so $\nabla\Psi_m(\tau) = \nabla\Psi_m(\tau^*)$. Lemma 3.2 then implies $\phi_m(\nabla\Psi_m(\tau)) = \phi_m(\nabla\Psi_m(\tau^*))$ or $\tau = \tau^*$.

4. Estimation. The system of normalized Legendre polynomials is a complete orthonormal basis for $L_2[-1, 1]$ and satisfies all previous assumptions on the system $\{\phi_i(x)\}_{i=0}^\infty$, including (1.3). Throughout this section we take $\{\phi_i(x)\}_{i=0}^\infty$ to be the normalized Legendre polynomials and $[a, b] = [-1, 1]$. The set $\{\phi_i(x)\}_{i=0}^m$ is an orthonormal basis for the vector space $\mathscr{P}_m$ of polynomials with real coefficients and of degree $\leqq m$.

Now for $f(x)$ given by (1.1) and $F(x)$ its cdf, we have, by (3.7)

$$\theta_i = \int_{-1}^1 \phi_i(x) f(x)\, dx = \int_{-1}^1 \phi_i(x) p_m(x \mid \tau^*)\, dx\,, \qquad 0 \leqq i \leqq m\,.$$

LEMMA 4.1. *If* $0 \leqq k \leqq m$ *then*

$$\int_{-1}^1 x^k f(x)\, dx = \int_{-1}^1 x^k p_m(x \mid \tau^*)\, dx\,. \tag{4.1}$$

PROOF. For $0 \leqq k \leqq m$, write the monomial $x^k = \sum_{i=0}^k \alpha_i^k \phi_i(x)$. Interchanging summation and integration,

$$\begin{aligned}\int_{-1}^1 x^k f(x)\, dx &= \sum_{i=0}^k \alpha_i^k \int_{-1}^1 \phi_i(x) f(x)\, dx \\ &= \sum_{i=0}^k \alpha_i^k \int_{-1}^1 \phi_i(x) p_m(x \mid \tau^*)\, dx = \int_{-1}^1 x^k p_m(x \mid \tau^*)\, dx\,.\end{aligned}$$

DEFINITION 4.1. $F_m^*(x)$ is the cdf corresponding to $p_m(x \mid \tau^*)$.

THEOREM 4.1. *$F_m^*(x) \to F(x)$ uniformly as $m \to \infty$.*

PROOF. By Lemma 4.1, $\lim_{m\to\infty} \int_{-1}^{1} x^s \, dF_m^*(x) = \int_{-1}^{1} x^s \, dF(x)$ for all positive integers s. By the Weierstrass Approximation Theorem it follows that $\lim_{m\to\infty} \int_{-1}^{1} g(x) \, dF_m^*(x) = \int_{-1}^{1} g(x) \, dF(x)$ for all continuous functions $g(x)$. The sequence of probability measures $\{dF_m^*\}$ then converges both weakly and vaguely to dF, and so $F_m(x)$ converges everywhere on R^1 to $F(x)$. The proof of uniform convergence is straightforward.

DEFINITION 4.2. $F_{mn}(x)$ is the cdf corresponding to $p_m(x \mid \tau_n^*)$.

COROLLARY 4.1. *With probability one,*

$$\lim_{m\to\infty} \lim_{n\to\infty} F_{mn}(x) = F(x) , \qquad \textit{uniformly in} \quad x . \tag{4.2}$$

PROOF. $\tau_n^* \to \tau^*$ as $n \to \infty$ implies $F_{mn}(x) \to F_m^*(x)$ as $n \to \infty$ by the Dominated Convergence Theorem, and $F_m^*(x) \to F(x)$ as $m \to \infty$ by Theorem 4.1. The uniform part is straightforward.

The θ_i's in (1.1) can be estimated unbiasedly by the $\bar{\phi}_i$'s where $\bar{\phi}_i = \int_{-1}^{1} \phi_i(x) \, dF_n(x)$ and $F_n(x)$ is the empirical cdf. If $F(x)$ is expanded as

$$F(x) = \sum_{i=0}^{\infty} \beta_i \phi_i(x) , \qquad -1 \leqq x \leqq 1 , \tag{4.3}$$

where $\beta_i = \int_{-1}^{1} \phi_i(x) F(x) \, dx$, then $\hat{\beta}_i = \int_{-1}^{1} \phi_i(x) F_n(x) \, dx$ is an unbiased estimate of β_i (Kronmal and Tarter [4]).

If the cdf $F_m^*(x)$ is expanded in an orthogonal series also by

$$F_m^*(x) = \sum_{i=0}^{\infty} \beta_i^m \phi_i(x) , \qquad -1 \leqq x \leqq 1 , \tag{4.4}$$

then using Lemma 4.1 and integration by parts it can be shown that

$$\beta_i = \beta_i^m , \qquad i = 0, 1, \cdots, m-1 . \tag{4.5}$$

Similarly if we expand $p_m(x \mid \tau^*)$ by

$$p_m(x \mid \tau^*) = \sum_{i=0}^{\infty} \theta_i^m \phi_i(x) , \tag{4.6}$$

we have by (2.2) and (3.7) that

$$\theta_i = \theta_i^m , \qquad i = 0, 1, 2, \cdots, m . \tag{4.7}$$

THEOREM 4.2. *The function $[F_m^*(x) - F(x)]$ has at least $m-1$ zeroes in the interval $(-1, 1)$. The function $[p_m(x \mid \tau^*) - f(x)]$ has at least m zeroes in the interval $(-1, 1)$.*

PROOF. By (4.5), $\int_{-1}^{1} \phi_i(x)[F_m^*(x) - F(x)] \, dx = 0$ for $0 \leqq i \leqq m-1$. Since $\{\phi_i(x)\}_{i=0}^{m-1}$ forms an orthonormal basis for the space $\mathscr{P}_{m-1}$ of polynomials of degree $\leqq m-1$, $\int_{-1}^{1} p(x)[F_m^*(x) - F(x)] \, dx = 0$ for any $p(x) \in \mathscr{P}_{m-1}$, a condition sufficient for the first part of the theorem. Since $F_m^*(x) = F(x)$ at $x = -1, 1$, the second part follows from Rolle's Theorem.

REMARK 4.1. We see a good approximation to $F(x)$ can be chosen from the canonical exponential family, if τ^* is known. The vectors τ^* $(m = 1, 2, \cdots)$ are in general unknown, but are intrinsically estimable by Theorem 3.2.

DEFINITION 4.3. The Kullback-Leibler information number $I[f; g]$ is defined by

$$I[f; g] = \int f(x) \log [f(x)/g(x)]\, dx \,. \tag{4.8}$$

$I[f; g]$ is nonnegative and equals zero iff $f(x) = g(x)$ a.e.-Lebesgue measure (Kullback [5]).

THEOREM 4.3. *$I[f(x); p_m(x \mid \tau^*)]$ converges to zero as $m \to \infty$.*

PROOF. With $f(x)$ given by (1.2) and $p_m(x \mid \tau') = \exp[\sum_{i=0}^m \tau_i'\phi_i(x) - \Psi_m(\tau')]$, where τ' is an arbitrary vector in R^m, we have

$$\begin{aligned} &0 \leqq \textstyle\int_{-1}^1 f(x) \log [f(x)/p_m(x \mid \tau')]\, dx \,, \qquad \tau' \in R^m \\ &0 \leqq \textstyle\int_{-1}^1 f(x)\{[\sum_{i=1}^\infty \tau_i \phi_i(x) - \Psi(\tau)] - [\sum_{i=1}^m \tau_i'\phi_i(x) - \Psi_m(\tau')]\}\, dx \\ &0 \leqq [\textstyle\sum_{i=1}^\infty \tau_i \theta_i - \Psi(\tau)] - [\sum_{i=1}^m \tau_i'\theta_i - \Psi_m(\tau')] \,. \end{aligned} \tag{4.9}$$

Now τ^* maximizes $Q_\theta{}^m(\tau) = \sum_{i=1}^m \tau_i \theta_i - \Psi_m(\tau)$ over R^m as was proved in Theorem 3.1, so for any $\tau' \in R^m$,

$$\begin{aligned} 0 &\leqq [\textstyle\sum_{i=1}^\infty \tau_i \theta_i - \Psi(\tau)] - [\sum_{i=1}^m \tau_i^*\theta_i - \Psi_m(\tau^*)] \\ &\leqq [\textstyle\sum_{i=1}^\infty \tau_i \theta_i - \Psi(\tau)] - [\sum_{i=1}^m \tau_i'\theta_i - \Psi_m(\tau')] \end{aligned} \tag{4.10}$$

or equivalently

$$0 \leqq I[f(x); p_m(x \mid \tau^*)] \leqq I[f(x); p_m(x \mid \tau')] \,, \qquad \tau' \in R^m \,. \tag{4.11}$$

In particular, with $\tau' = \tau$ where $\tau = (\tau_1, \tau_2, \cdots, \tau_m)'$ and $\tau_i = \int_{-1}^1 \phi_i(x) \log f(x)\, dx$ as in (1.2), we have

$$0 \leqq I[f(x); p_m(x \mid \tau^*)] \leqq I[f(x); p_m(x \mid \tau)] \,. \tag{4.12}$$

By assumption, in (1.2) the series $\sum_{i=1}^\infty \tau_i \phi_i(x)$ converges uniformly, and using Lemma 2.1 we have

$$\log [f(x)/p_m(x \mid \tau)] \to 0 \,, \qquad \text{uniformly in } x \,, \qquad \text{as } m \to \infty \,. \tag{4.13}$$

By (4.12), this proves the theorem.

REMARK 4.2. From (4.11) it is evident that τ^* minimizes $I[f(x); p_m(x \mid \tau')]$ over all $\tau' \in R^m$, thus $p_m(x \mid \tau^*)$ is that density of the canonical exponential family of size m which most resembles $f(x)$ in the sense of Kullback-Leibler information.

THEOREM 4.4. *The sequence $p_m(x \mid \tau^*)$ converges to $f(x)$ in the L_1-norm,* i.e., $\int_{-1}^1 |p_m(x \mid \tau^*) - f(x)|\, dx \to 0$ *as* $m \to \infty$.

PROOF. See Kullback [5], page 390.

Acknowledgment. This research was performed while the author was a graduate student in the Department of Statistics, Oregon State University, Corvallis, and was done under the direction of Professor H. D. Brunk, whose guidance and encouragement are gratefully acknowledged.

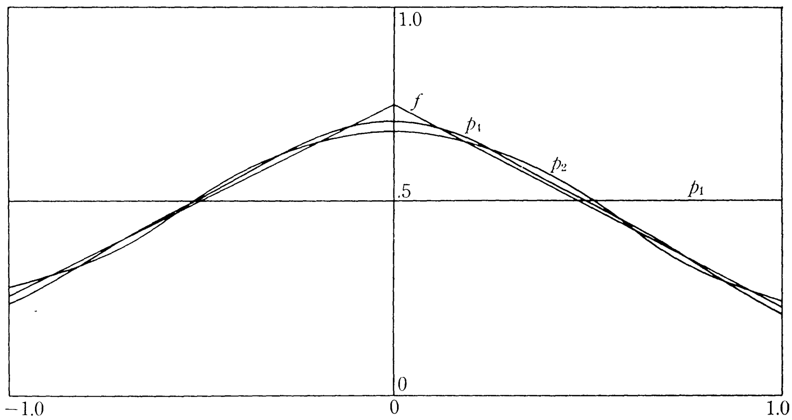

FIG. 1. $p_m(x \mid \tau^*)$ for $f(x) = ((\frac{3}{4}) - |x|/2)$.

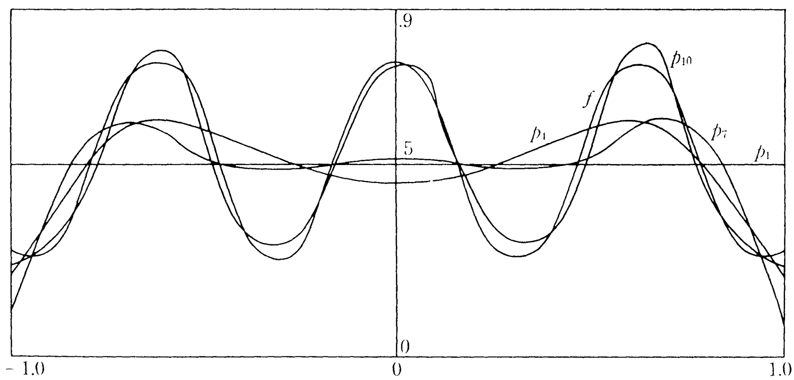

FIG. 2. $p_m(x \mid \tau^*)$ for $f(x) \propto ((\frac{1}{2}) \cos (10x) + 1)$.

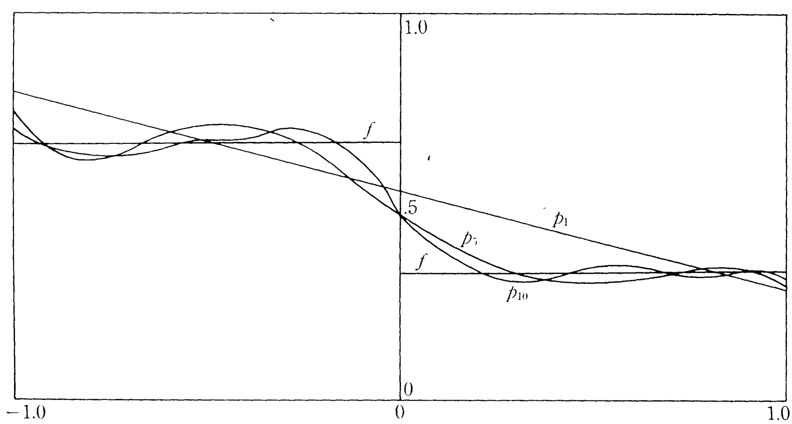

FIG. 3. $p_m(x \mid \tau^*)$ for $f(x) = (\frac{2}{3})I_{[-1,0]}(x) + (\frac{1}{3})I_{(0,1]}(x)$.

REFERENCES

[1] Barndorff-Nielsen, O. (1970). Exponential families, exact theory. Various Publication Series No. 19, Aarhus Universitet.

[2] Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.* **43** 193-204.

[3] Čencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3** 1559-1562.

[4] Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925-952.

[5] Kullback, S. (1968). *Information Theory and Statistics*. Dover, New York.

[6] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065-1076.

[7] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.

[8] Schwartz, S. C. (1967). Estimation of a probability density by an orthogonal series. *Ann. Math. Statist.* **38** 1262-1265.

[9] Watson, G. S. (1969). Density estimation by orthogonal series. *Ann. Math. Statist.* **40** 1496-1498.

[10] Wilks, S. S. (1962). *Mathematical Statistics*. Wiley, New York.

[11] Zangwill, W. I. (1969). *Nonlinear Programming: a Unified Approach*. Prentice-Hall, Englewood Cliffs, N.J.

Department of Mathematics
University of Oklahoma
Norman, Oklahoma 73069