The Annals of Statistics
1980, Vol. 8, No. 4, 820–832

NONPARAMETRIC PROBABILITY DENSITY ESTIMATION BY DISCRETE MAXIMUM PENALIZED-LIKELIHOOD CRITERIA[1]

BY D. W. SCOTT, R. A. TAPIA AND J. R. THOMPSON
Baylor College of Medicine and Rice University

A nonparametric probability density estimator is proposed that is optimal with respect to a discretized form of a continuous penalized-likelihood criterion functional. Approximation results relating the discrete estimator to the estimate obtained by solving the corresponding infinite-dimensional problem are presented. The discrete estimator is shown to be consistent. The numerical implementation of this discrete estimator is outlined and examples displayed. A simulation study compares the integrated mean square error of the discrete estimator with that of the well-known kernel estimators. Asymptotic rates of convergence of the discrete estimator are also investigated.

1. Introduction. The problem we consider is that of estimating an unknown probability density function f given only a random sample $X_1, \cdots, X_N$ which came from this density.

The classic nonparametric density estimator is the histogram. A significant updating of the histogram approach was made by Parzen [6] and by Rosenblatt [7] in their theoretical development of the nonparametric kernel estimators. Given a kernel function K which integrates to one, the *kernel estimator at a point* x is

$$f_K(x) = \frac{1}{N}\Sigma_{i=1}^{N} K_h(x - X_i) \tag{1.1}$$

where $K_h(y) = (1/h)K(y/h)$. Observe that $f_K(x)$ actually depends on the parameter h. Convergence of $f_K(x)$ to $f(x)$ in mean square error under very mild restrictions is well known. Optimal choices for the smoothing parameter h as a function of N can be derived as a function of the true density and its second derivative f'' (which are, unfortunately, unknown). However, there has recently been some encouraging activity in the area of choosing h, most notably by Wahba [12], Silverman [9] and the authors [8].

In 1971 Good and Gaskins [2] used a modification of Fisher's maximum likelihood principle to propose an optimality criterion for obtaining a nonparametric estimate of a probability density function. Specifically, given a class of integrable functions H and a nonnegative penalty functional Φ defined on H, they proposed estimating the unknown density f from the random sample $X_1, \cdots, X_N$

Received August 1977; revised February 1979.
[1]Research supported in part by the National Heart, Lung, and Blood Institute, the U.S. Army Research Office, the U.S. Air Force Office of Scientific Research, and the Department of Energy under grants NIH 17269, DAAG29-78-G-0187, AFOSR 76-2711, and EY-76-S-05-5046, respectively.
AMS 1970 *subject classifications.* Primary G2G05; secondary G2E10.
Key words and phrases. Nonparametric density estimation, maximum likelihood estimation, kernel density estimation.

by the solution of the constrained optimization problem. Thus

(1.2) maximize $L(\varphi)$

$$= \Sigma_1^N \log(\varphi(X_i)) - \Phi(\varphi) \text{ subject to } \varphi \in H, \varphi \geqslant 0 \text{ and } \int \varphi = 1.$$

de Montricher, Tapia and Thompson [1] extended the theory and gave some important existence and uniqueness results by placing the so-called *maximum penalized-likelihood estimator* (MPLE) in the natural framework of a reproducing kernel Hilbert space. The difficult question of the statistical consistency of the MPLE has yet to be answered fully.

We begin our study with a brief presentation of two results from de Montricher, Tapia and Thompson [1] which will be needed in the sequel. Let $H(a, b)$ be a Hilbert space of functions defined on the interval $[a, b]$. We denote the inner product in $H(a, b)$ by $\langle \cdot, \cdot \rangle_H$. Recall that $H(a, b)$ is said to be a *reproducing kernel Hilbert space* (RKHS) if for each $t \in [a, b]$ the point evaluation functional E_t: $H(a, b) \to R$ defined by $E_t(f) = f(t)$ is continuous.

THEOREM 1.1. *Suppose $H(a, b)$ is a* RKHS, *integration over $[a, b]$ is a continuous operation and there exists at least one $\varphi \in H(a, b)$ which integrates to one, is nonnegative and is positive at the sample points $X_1, \cdots, X_N$. Then problem* (1.2) *with $H = H(a, b)$ and $\Phi(\varphi) = \alpha\langle \varphi, \varphi \rangle_H$ for any $\alpha > 0$ has a unique solution.*

Let $H_0^k(a, b)$ denote the Sobolev space of functions defined on the finite interval $[a, b]$ whose first $k - 1$ derivatives are absolutely continuous and vanish at a and at b and whose kth derivative is in $L^2(a, b)$. It is well known that $H_0^k(a, b)$ is a RKHS with inner product defined by

$$\langle \varphi, \xi \rangle_{H_0^k} = \int_a^b \varphi^{(k)}(t)\xi^{(k)}(t)\, dt. \tag{1.3}$$

THEOREM 1.2. *If in problem* (1.2) *we let $H = H_0^k(a, b)$ and $\Phi(\varphi) = \alpha\langle \varphi, \varphi \rangle_{H_0^k}$ for any $\alpha > 0$, then the solution (maximum penalized-likelihood estimate) exists, is unique and is a polynomial spline of degree $2k$. Moreover, if the estimate is positive in the interior of an interval then on this interval it is a polynomial spline of degree $2k$ with knots exactly at the sample points and with continuous derivatives up to order $2k - 2$.*

2. Discrete maximum penalized-likelihood estimators. At the present time it is not computationally feasible to calculate the spline density estimators described in Theorem 1.2 since complete identification of knot locations is an open problem. For this reason and others pertaining to numerical efficiency we are led to propose and investigate the following finite dimensional problem as an approximation to the infinite dimensional penalized-likelihood problem corresponding to the Sobolev space $H_0^1(a, b)$.

Let $\Omega = (a, b)$ be a finite interval covered by an equally spaced mesh $a = t_0, t_1, \cdots, t_m = b$ with $t_{k+1} - t_k = h$. Further let $s(\cdot)$ and $p(\cdot)$ be a simple function and a continuous piecewise linear function, respectively, defined over the mesh $\{t_k\}$ and vanishing outside Ω. The discrete maximum penalized-likelihood

estimates (DMPLE) will be of this form. Defining the kth mesh interval to be $I_k = [t_k, t_{k+1})$ the two estimators are completely determined by their values at the mesh nodes,

$$s_k = s(t_k), \qquad k = 0, \cdots, m \tag{2.1}$$

and

$$p_k = p(t_k), \qquad k = 0, \cdots, m. \tag{2.2}$$

We require $p_0 = p_m = 0$, so that the linear spline p will belong to $H_0^1(a, b)$. Note that it is also continuous on the entire real line. There is no way that we can make the simple function s a member of $H_0^1(a, b)$ unless we require it to be a constant function. However for the sake of convenience we let $s_m = 0$. It is a straightforward matter to show that

$$\int_a^b s(t)\,dt = h\Sigma_0^{m-1} s_k, \tag{2.3}$$

$$\int_a^b p(t)\,dt = h\Sigma_1^{m-1} p_k \tag{2.4}$$

and

$$s(t) \geqslant 0, p(t) \geqslant 0 \Leftrightarrow s_k \geqslant 0, p_k \geqslant 0, \qquad k = 0, \cdots, m. \tag{2.5}$$

For $X_1, \cdots, X_N \in (a, b)$ consider the following constrained optimization problems modeled after problem (1.2): (defining $s_{-1} = 0$)

$$\text{maximize } L(s_0, \cdots, s_{m-1}) = \Sigma_1^N \log(s(X_i)) - \alpha h \Sigma_0^m \left[\frac{s_k - s_{k-1}}{h}\right]^2 \tag{2.6}$$

$$\text{subject to } h\Sigma_0^{m-1} s_k = 1 \text{ and } s_k \geqslant 0, \quad k = 0, \cdots, m-1;$$

and

$$\text{maximize } L(p_1, \cdots, p_{m-1}) = \Sigma_1^N \log(p(X_i)) - \alpha h \Sigma_1^m \left[\frac{p_k - p_{k-1}}{h}\right]^2 \tag{2.7}$$

$$\text{subject to } h\Sigma_1^{m-1} p_k = 1 \text{ and } p_k \geqslant 0, \quad k = 1, \cdots, m-1.$$

Solutions to problems (2.6) and (2.7) are called *discrete maximum penalized-likelihood estimates*. Here we have replaced the Sobolev inner product with a discrete Sobolev inner product in the penalty terms. The extension of (2.6) and (2.7) to penalty terms with higher order derivatives is straightforward. A different approach has been taken by Lii and Rosenblatt [4] who considered a cubic spline estimator.

THEOREM 2.1. *The* DMPLE *defined by* (2.6) *or* (2.7) *exists and is unique.*

PROOF. The proof follows in a straightforward manner from Theorem 1.1 once we note that any finite dimensional inner product space is a RKHS and that all linear functionals defined on finite dimensional spaces are continuous.

3. Approximation results. We have proposed the DMPLE as an approximation to the spline MPLE given by Theorem 1.2 for the Sobolev space $H_0^1(a, b)$. This was accomplished by replacing the infinite dimensional problem by the finite

dimensional problem which arises when we restrict our attention to piecewise constant (simple functions) or piecewise linear functions defined using a regular mesh or partition of the interval (a, b). In this section we establish the important fact that the DMPLE approaches the spline MPLE as the mesh size approaches zero. Specifically, we prove the following theorem.

THEOREM 3.1. *Suppose $\Omega = (a, b)$ is a finite interval, $x_1, \cdots, x_N$ is a fixed sample and that data outside Ω is ignored. Let h be the size of the mesh used to obtain the* DMPLE *guaranteed by Theorem* 2.1. *Then the simple function* DMPLE *converges to the $H_0^1(a, b)$ spline* MPLE *(guaranteed by Theorem* 1.2*) in the sup norm as $h \to 0$.*

PROOF. The proof is quite lengthy and is given in three steps. We will denote a simple function determined by the regular partition of (a, b) of size h by $s_h(\cdot)$ in order to emphasize the dependence on the mesh spacing h. The criterion functionals for the finite dimensional problem and the infinite dimensional problems are respectively (letting $s_h(t_{-1}) = 0$)

$$L_m(s_h) = \Sigma_{i=1}^{N} \log(s_h(x_i)) - \frac{\alpha}{h}\Sigma_0^m [s_h(t_k) - s_h(t_{k-1})]^2 \tag{3.1}$$

and

$$L_\infty(f) = \Sigma_{i=1}^{N} \log(f(x_i)) - \alpha \int_a^b f'(t)^2 \, dt. \tag{3.2}$$

Again, the subscripts m and ∞ on L indicate the dimension of the problem we are considering. Let f^* denote the quadratic spline MPLE given by Theorem 1.2 for the criterion functional (3.2) and let s_h^* denote the simple function DMPLE given by Theorem 2.1 for the criterion functional (3.1). The proof of the theorem will follow from the following three lemmas.

LEMMA 3.1. *The quadratic spline f^* is infinitely differentiable on (a, b) except for at most $3N$ points.*

PROOF. We show that f^* has at most 2 knots between any two data points. Suppose that the graph of f^* looks like that given in Figure 1. If f^* consists of the solid line then it is easy to show that $\hat{f}^*$ can be constructed using the dotted line so that it has the same definite integral on $[x_{i-1}, x_i]$ and the same log likelihood term as f^* but with a smaller penalty term in (3.2), contradicting the optimality of f^*. It follows from the second part of Theorem 1.2 that if f^* was constant and nonzero on a subinterval of (x_{i-1}, x_i), then it would be necessarily constant on all of (x_{i-1}, x_i). In essence this proves the lemma.

LEMMA 3.2. *For $h > 0$ it is possible to construct a family of simple functions $s_{f^*, h}$ which are piecewise constant on the intervals I_k, are nonnegative, integrate to one and have the property that*

$$L_m(s_{f^*, h}) \to L_\infty(f^*) \quad \text{as} \quad h \to 0. \tag{3.3}$$

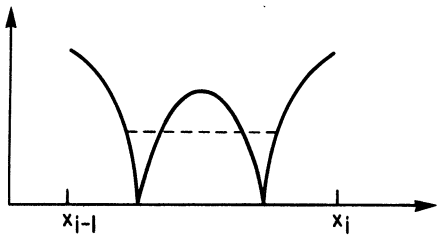

FIG. 1

PROOF. We actually construct the desired simple function $s_{f^*,h}$. Let

$$s_{f^*,h}(t_k) = \frac{1}{h}\int_{I_k} f^*(t)\,dt \quad , \qquad k = 0, \cdots, m-1. \tag{3.4}$$

Then $s_{f^*,h}$ is nonnegative and integrates to one. From (3.4) we see that for some $c_k \in I_k$ we have

$$s_{f^*,h}(x) = f^*(c_k), \forall x \in I_k. \tag{3.5}$$

Now by Lemma 3.1 we can apply the fundamental theorem of calculus to obtain

$$f^*(x) - f^*(c_k) = \int_{c_k}^{x} f^{*\prime}(y)\,dy. \tag{3.6}$$

Using Cauchy-Schwarz on (3.6) and using (3.5) leads to

$$|f^*(x) - s_{f^*,h}(x)| \leqslant (h)^{\frac{1}{2}} \|f^*\|_{H_0^1(a,b)}. \tag{3.7}$$

It follows from (3.7) that $s_{f^*,h}(x) \to f^*(x)$ as $h \to 0$ for all $x \in (a, b)$. This shows that the log likelihood term in (3.1) converges to the log likelihood term in (3.2) as $h \to 0$. We now consider the convergence of the penalty term. From (3.5) and the mean value theorem, there exists $\hat{x}_k \in I_k$ such that

$$\begin{aligned} \frac{1}{h}\Sigma_k\left[s_{f^*,h}(t_k) - s_{f^*,h}(t_{k-1})\right]^2 &= \frac{1}{h}\Sigma_k\left[f^*(\hat{x}_k) - f^*(\hat{x}_{k-1})\right]^2 \\ &= \int_a^b f^{*\prime}(t)^2\,dt + O(h). \end{aligned} \tag{3.8}$$

The second quality follows from a straightforward application of the fundamental theorem of calculus. We have ignored the finite number of intervals where $f^{*\prime}$ has a discontinuity, since for the penalty term this contribution goes to zero with h. This proves the lemma.

LEMMA 3.3. *For $h > 0$ it is possible to construct a family of $H_0^1(a, b)$ functions f_h^* which are nonnegative, integrate to one and have the properties that*

$$\|f_h^* - s_h^*\|_{\sup} \to 0 \quad \text{as} \quad h \to 0 \tag{3.9}$$

and

$$L_m(s_h^*) - L_\infty(f_h^*) \to 0 \quad \text{as} \quad h \to 0. \tag{3.10}$$

PROOF. Recall that s_h^* is the unique maximizer of (3.1). Thus $L_m(s_h^*) \geqslant L_m(s_{f^*,h})$ for all h. From Lemma 3.2 we know that $L_m(s_{f^*,h}) \to L_\infty(f^*)$ as $h \to 0$. We know from Theorem 1.2 that $L_\infty(f^*) > -\infty$, implying that

$$|s_h^*(t_k) - s_h^*(t_{k-1})| \to 0 \quad \text{as} \quad h \to 0. \tag{3.11}$$

This follows from the fact that if (3.11) did not hold, then from (3.1) $L_m(s_h^*)$ would become unbounded as $h \to 0$. This would be a contradiction. We construct $f_h^* \in H_0^1(a, b)$ by considering the piecewise linear function from the points $(t_0, 0)$ and $(t_m, 0)$, interpolating the simple function s_h^* at the midpoints of the intervals I_k and then dividing this function by $1 - \varepsilon$ where $\varepsilon = h(s_0 + s_{m-1})/4$ and $s_k = s_h^*(t_k)$. Now $\varepsilon = O(h)$ since $s_0 \to 0$ and $s_{m-1} \to 0$ as $h \to 0$ by (3.11). With this choice f_h^* is nonnegative and integrates to one. A straightforward calculation shows that

$$\int_a^b f_h^{*\prime}(t)^2\,dt = \frac{1}{h[1 - \varepsilon(s_0 + s_{m-1})]} \cdot \left\{\sum_{k=0}^{m}(s_k - s_{k-1})^2 - \tfrac{1}{2}(s_0^2 + s_{m-1}^2)\right\}; \tag{3.12}$$

hence the penalty terms coincide as $h \to 0$. By adding and subtracting the (unscaled interpolating) function $(1 - \varepsilon)f_h^*$ and using the triangle inequality we have

$$\|f_h^* - s_h^*\|_{\sup} \leqslant \sup_k |s_h^*(t_k) - s_h^*(t_{k-1})| + \varepsilon \cdot \sup f_h^*. \tag{3.13}$$

Since $s_h^* \leqslant 1/h$ by constraint (2.3) and using the definitions of f_h^* and ε, $\varepsilon \cdot \sup f_h^* \leqslant (s_0 + s_{m-1})/4(1 - \varepsilon)$ which vanishes as $h \to 0$. Thus (3.11) and (3.13) imply (3.9) since $\varepsilon = 0(h)$. Combining (3.9) with (3.12) gives (3.10) and the lemma is established.

PROOF OF THEOREM 3.1. By their respective optimality properties

$$L_\infty(f_h^*) \leqslant L_\infty(f^*) \tag{3.14}$$

and

$$L_m(s_{f^*,h}) \leqslant L_m(s_h^*). \tag{3.15}$$

Combining (3.3) from Lemma 3.2, (3.10) from Lemma 3.3, (3.14) and (3.15) it follows that

$$L_\infty(f_h^*) \to L_\infty(f^*) \quad \text{as} \quad h \to 0. \tag{3.16}$$

From [1] the functional $\hat{L} = -L_\infty$ is uniformly convex on the subset of $H_0^1(a, b)$ which consists of the nonnegative functions which integrate to one. It follows from the definition of uniform convexity (see page 83 of [5]) that there exists $C > 0$ such that for all $0 \leqslant \beta \leqslant 1$. Thus

$$\beta\hat{L}(f_h^*) + (1 - \beta)\hat{L}(f^*) - \hat{L}(\beta f_h^* + (1 - \beta)f^*) \geqslant C\beta(1 - \beta)\|f_h^* - f^*\|_{H_0^1(a, b)}. \tag{3.17}$$

By the optimality of f^* (3.17) implies

$$\hat{L}(f_h^*) - \hat{L}(f^*) \geqslant C\beta(1 - \beta)\|f_h^* - f^*\|_{H_0^1(a, b)}. \tag{3.18}$$

Coupling (3.16) with (3.18) we see that

$$\|f_h^* - f^*\|_{H_0^1(a,b)} \to 0 \qquad \text{as} \quad h \to 0. \tag{3.19}$$

Now convergence in $H_0^1(a, b)$ implies sup norm convergence. Therefore, from (3.9), (3.19) and the triangle inequality we arrive at

$$\|s_h^* - f^*\|_{\sup} \to 0 \qquad \text{as} \quad h \to 0. \tag{3.20}$$

This proves the theorem.

COROLLARY 3.1. *Under the conditions of Theorem* 3.1 *the piecewise linear* DMPLE (*guaranteed by Theorem* 2.1) *converges to the* $H_0^1(a, b)$ *spline* MPLE (*guaranteed by Theorem* 1.2) *in* $H_0^1(a, b)$ *as* $h \to 0$.

PROOF. The proof consists of making the appropriate modifications to the proof of Theorem 3.1.

4. Consistency of the DMPLE. In practice, the consistency question is overshadowed by the variances of the small sample properties of an estimator (provided, of course, that asymptotic consistency has been theoretically demonstrated). Furthermore, with limited prior information, poor estimates of the density f in the tails are expected due to the presence of few sample points there. If f has infinite support or unknown finite support, then it is easy to show by an appropriate modification of Theorem 4.1 below that the DMPLE defined on the interval $\Omega = (a, b)$ converges to the truncated density $\bar{f}(x)$ defined by

$$\begin{aligned} \bar{f}(x) &= \frac{f(x)}{1-\varepsilon} && x \in \Omega \\ &= 0 && x \notin \Omega \end{aligned} \tag{4.1}$$

where

$$\varepsilon = 1 - \int_\Omega f(x)\, dx.$$

Thus we shall assume that f has known finite support (a, b) in view of this remark.

Given a sample $X_1, \cdots, X_N \in (a, b)$, let

$$\nu_k = \text{number of samples in } I_k = [t_k, t_{k+1}), \qquad k = 0, \cdots, m-1. \tag{4.2}$$

Then $\sum_{k=0}^{m-1} \nu_k = N$. Using (4.2), problem (2.6) (see also (3.1)) becomes

$$\text{maximize } L(s_0, \cdots, s_{m-1}) = \sum_0^{m-1} \nu_k \log(s_k) - \frac{\alpha}{h} \sum_0^m (s_{k-1} - s_k)^2 \tag{4.3}$$

$$\text{subject to } h\sum_0^{m-1} s_k = 1 \text{ and } s_k \geqslant 0, \quad k = 0, \cdots, m-1.$$

In the sequel we shall slightly abuse notation and denote the solution of (4.3) also by $(s_0, \cdots, s_{m-1})$ and the simple function representing this solution by s_h. We now prove the following important consistency result for the simple function DMPLE.

THEOREM 4.1. *Let $X_1, \cdots, X_N$ be a random sample from a continuous density f of bounded support (a, b). Consider the simple function* DMPLE *with the number of partitions given by $m = [cN^q]$ where $c > 0$, $0 < q < \frac{1}{4}$ and $[d]$ denotes the integer closest to d. Then for $x \in (a, b)$, $\lim_{N\to\infty} s_h(x) = f(x)$ almost surely* (a.s.).

PROOF. In order to emphasize the dependence on N we write $m(N)$ for the number of mesh nodes. In this case the mesh spacing will be $h = h_N = (b - a)/(m(N) - 1) = 0(N^{-q})$. We have from the theory of Lagrange multipliers that there exist multipliers $\lambda, \mu_0, \cdots, \mu_{m-1}$ such that

$$\frac{\partial L(s_h)}{\partial s_i} + \lambda \frac{\partial}{\partial s_i}\left[\Sigma_k s_k - \frac{1}{h}\right] + \mu_i = 0, \quad i = 0, \cdots, m - 1 \tag{4.4}$$

and $\mu_i s_i = 0$.

For our problem, (4.4) becomes

$$\mu_i + \frac{\nu_i}{s_i} + \frac{2\alpha}{h}\delta^2 s_i + \lambda = 0, \qquad i = 0, \cdots, m - 1 \tag{4.5}$$

where

$$\delta^2 s_i = s_{i+1} - 2s_i + s_{i-1}. \tag{4.6}$$

Multiplying by s_i, summing over i, and using the first constraint in (4.3) we obtain the following expression for the Lagrange multiplier λ:

$$\lambda = -Nh - 2\alpha\Sigma_{i=0}^{m-1} s_i \delta^2 s_i. \tag{4.7}$$

Substituting (4.7) into (4.5), our necessary conditions upon dividing by N, become

$$\frac{\mu_i}{N} + \frac{\nu_i}{Ns_i} + \frac{2\alpha}{Nh}\delta^2 s_i - h - \frac{2\alpha}{N}\Sigma_{j=0}^{m-1} s_j \delta^2 s_j = 0. \tag{4.8}$$

Before solving for s_i, let us approximate the third and fifth terms in (4.8). From the integral constraint we know that $s_i \leqslant 1/h$ so that

$$\delta^2 s_j = O\left(\frac{1}{h}\right) \tag{4.9}$$

and

$$\Sigma_{j=0}^{m-1} s_j \delta^2 s_j = O\left(\frac{1}{h^3}\right). \tag{4.10}$$

Using the bounds (4.9) and (4.10) in (4.8) and dividing by h we obtain

$$\frac{\nu_i}{Nhs_i} - 1 + O\left(\frac{1}{Nh^4}\right) = 0. \tag{4.11}$$

Now as $N \to \infty$ we are interested in the behavior of the quantity ν_i/Nhs_i corresponding to the interval containing the particular x in question. In this context the subscript i is very misleading, since the dependence is actually on x, h and N. However, recall that h is itself a function of N. We therefore introduce more meaningful notation. Let $[x\colon N]$ denote the mesh interval of length h containing x

for a particular value of N and let $\nu[x: N]$ denote the number of samples in $[x: N]$. In order to complete our proof we need the following lemma which describes the behavior of the random variable $\nu[x: N]/Nh$.

The proof of the following lemma is due to Silverman and is significantly shorter than the authors' original proof.

LEMMA 4.1. *Under the conditions of Theorem* 4.1

$$\lim_{N\to\infty} \frac{\nu[x: N]}{Nh_N} = f(x) \quad \text{a.s.}$$

PROOF. Define the triangular array of random variables

$$Y_{Nj} = \frac{I[x: N](X_j) - p[x: N]}{h_N}$$

where $I[x: N](\cdot)$ denotes the indicator function of the interval $[x: N]$ and

$$p[x: N] = \int_{[x: N]} f(t)\, dt. \tag{4.12}$$

Now $\{Y_{Nj}: j = 1, \cdots, N\}$ forms an independent sequence for each N. Also $I[x: N](X_j)$ is a binomial random variable with expectation given by $p[x: N]$; hence, $EY_{Nj} = 0$, and $EY_{Nj}^2 = p[x: N](1 - p[x: N])/h_N^2$. Let

$$S_N = \Sigma_{j=1}^N Y_{Nj}.$$

To prove the lemma, we wish to show that $S_N/N \to 0$ almost surely.

If T is a $B(n, p)$ random variable, it is easily shown that $E(T - ET)^4 < 3pn^2$. Hence, by the fourth moment generalization of Chebyshev's inequality, for any $\varepsilon > 0$,

$$P\{|S_N| > N\varepsilon\} < \frac{3p[x: N]}{N^2 h_N^4 \varepsilon^4}. \tag{4.13}$$

From the mean value theorem, $p[x: N] = h_N f(\xi_N)$ where $\xi_N \in [x: N]$. Combining this fact with (4.13) leads to

$$\Sigma_{N=1}^\infty P\{|S_N| > N\varepsilon\} < \Sigma_{N=1}^\infty \frac{3f(\xi_N)}{N^2 h_N^3 \varepsilon^4} = \Sigma_{N=1}^\infty \frac{1}{0(N^{2-3q})} < \infty$$

since $q < \frac{1}{4}$; hence the Borel-Cantelli lemma implies

$$\frac{S_N}{N} \equiv \frac{\nu[x: N]}{Nh_N} - \frac{p[x: N]}{h_N} \to 0 \quad \text{a.s.}$$

Since f is continuous, (4.12) implies $p[x: N]/h_N \to f(x)$, completing the proof of the lemma.

We now return to the proof of the theorem. Replacing v_i with $\nu[x: N]$ and s_i with $s_h(x)$ in (4.11), observing that $Nh^4 \to \infty$, recalling Lemma 4.1, and letting $N \to \infty$ in (4.11) we obtain

$$s_h(x) \to f(x) \quad \text{a.s.},$$

which proves the theorem.

REMARK. If in Theorem 4.1 we consider a discrete kth order derivative in the penalty term of our criterion functional, then the analogous consistency result would require $0 < q < (2k + 2)^{-1}$; however, numerical work indicates that this requirement is an artifact of the method of proof and not necessary for consistency. This parallels the conjecture that the MPLE is consistent.

5. Numerical implementation and Monte Carlo simulation results. In this section we investigate the small-sample properties as well as the asymptotic properties of the discrete maximum penalized-likelihood estimator. In presenting the numerical solution for the DMPLE we chose the continuous piecewise linear solution rather than the simple function solution for its smoothness and derivative approximation properties. We also chose the penalty functional using second differences to approximate second derivatives.

Specifically, given $x_1, \cdots, x_N$, an interval (a, b), a positive scalar α, and a positive integer m we let

$$h = (b - a)/m, \tag{5.1}$$

$$t_i = a + ih, \qquad i = 0, \cdots, m, \tag{5.2}$$

$$p_{-1} = p_0 = p_m \doteq p_{m+1} = 0 \tag{5.3}$$

and solve the $m - 1$ dimensional constrained optimization problem

(5.4)

$$\text{maximize } L(p_1, \cdots, p_{m-1}) = \Sigma_{i=1}^{N} \log p(x_i) - \frac{\alpha}{h^3} \Sigma_{k=0}^{m} [\, p_{k+1} - 2p_k + p_{k-1}]^2$$

$$\text{subject to } h\Sigma_{k=1}^{m-1} p_k = 1 \quad \text{and} \quad p_k \geqslant 0, \quad k = 1, \cdots, m - 1,$$

where

$$p(t) = p_k + \frac{p_{k+1} - p_k}{h}(t - t_k) \qquad t \in [\, t_k, t_{k+1}) \tag{5.5}$$

$$= 0 \qquad t \notin [\, t_0, t_m).$$

A computer program has been written to solve problem (5.4) and is contained in the IMSL Library [3]. This program uses a modification due to Tapia [10], [11] of Newton's method. This modification takes advantage of the special banded structure of the Hessian matrix of the Lagrangian functional for problem (5.4). Thus the amount of work turns out to be $O(m^2)$ instead of the expected $O(m^3)$ per iteration. Notice that the sample size N enters only in evaluating the gradient and Hessian and not in the matrix inversions. No initial guesses for Newton's method are required. Rather a bootstrap algorithm is employed. First the problem is solved for $m = 7$, an easy problem to solve. This estimate provides initial guesses for $m = 11$, then for $m = 19$, and so on. The bootstrap algorithm takes advantage of the stability of the DMPLE with respect to h for a fixed choice of α.

The choice of α is very important and more difficult than the selection of the mesh spacing h. Asymptotically, of course, any positive value gives consistent

results. For finite sample sizes, however, the choice is critical. The α parameter is analogous to the kernel scaling parameter h in (1.1). Unfortunately, (as is customary) we have also denoted the mesh spacing by h. Hopefully, this will not lead to confusion in the sequel where we will be discussing the roles of the penalty parameter α and the kernel scaling parameter h. For a fixed finite sample there are values of α and h which give the "best" approximations for the DMPLE and kernel estimates, respectively. For values smaller than "best", the corresponding estimates peak sharply at the sample points. On the other hand, values larger than "best" correspond to depressed and oversmoothed estimates. For the kernel estimator, asymptotically, "optimal" choices for h are known; however, knowledge of the true (unknown) density is required to evaluate this h. Wahba [12] has given some insight to this problem. Also, Scott, et al., [8] have developed an iterative data-based approach for estimating h which only requires the prior knowledge that the unknown density has a square integrable second derivative. An interactive approach is often used where the smallest value of α (or h for the kernel estimator) is chosen that reveals fine structure without "too much" oscillatory behavior (consistent with prior knowledge).

To compare the approximation properties of the DMPLE with those of the kernel estimators we performed a Monte Carlo simulation. A "reasonable" value for α was chosen for each density (e.g., $\alpha = 10$ for the standard Gaussian and $\alpha = 30$ for the Gaussian mixture). For comparison purposes we weighted the simulation study in favor of the kernel estimator by using the optimal choice of h (since in this case the optimal h is known). The highly popular Gaussian kernel $K(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)$ was used, although kernels with finite support enjoy computational savings. The optimal choice for the scaling parameter h as a function of N in this case is

$$h(N) = \left\{ \frac{\int K(x)^2 \, dx}{\left[\int x^2 K(x) \, dx\right]^2 \int f''(x)^2 \, dx} \right\}^{\frac{1}{5}} N^{-\frac{1}{5}}.$$

Random samples were generated on the computer [3] and the integrated mean square error (IMSE) evaluated numerically over the truncated interval $(-5, 5)$. The Monte Carlo technique is to report the mean and standard deviation of the IMSE of 25 generated samples from a fixed distribution for a fixed sample size N. We also calculated the kernel estimate (using the "optimal" choice of h) for the same random samples and evaluated the IMSE numerically in the same manner. These results are given in Table 5.1.

We used Monte Carlo methods to estimate the rate of convergence of the DMPLE as a function of N, using Gaussian random samples. When plotted on log-log paper, the values of the estimated IMSE given in Table 5.2 fall on a straight line with slope $-.773$. The actual regression analysis gave

$$\log_{10}(\text{IMSE}) = -.773 \log_{10} N - .873$$

TABLE 5.1.
Monte Carlo estimation of integrated mean square error of DMPLE *and Gaussian kernel estimator.**

Sampling Density	Sample Size	DMPLE	Gaussian Kernel
$N(0, 1)$	$N = 25$	.010	.016
		(.008)	(.012)
$N(0, 1)$	$N = 100$	.0037	.0050
		(.0021)	(.0027)
$N(0, 1)$	$N = 400$	.0015	.0021
		(.0008)	(.0009)
Bimodal	$N = 25$	.010	.009
		(.003)	(.007)
Bimodal	$N = 100$	.0036	.0036
		(.0007)	(.0020)

*Each row represents the mean of the IMSE for 25 trials of the DMPLE and the Gaussian kernel estimator based on 25 random samples from the density in question for fixed N; the standard deviation is given in parentheses; $\alpha = 10$ was used for the $N(0, 1)$, $\alpha = 30$ was used for the bimodal and the bimodal density is the mixture $.5N(-1.5, 1) + .5N(1.5, 1)$.

TABLE 5.2.
Asymptotic rate of convergence for DMPLE *based on the* $N(0, 1)$ *sampling density.*

Sample size	*Number of samples*	*Estimated IMSE*
25	50	.0110
100	100	.00347
400	50	.00151
800	50	.000843
1000	50	.000545
2000	84	.000360

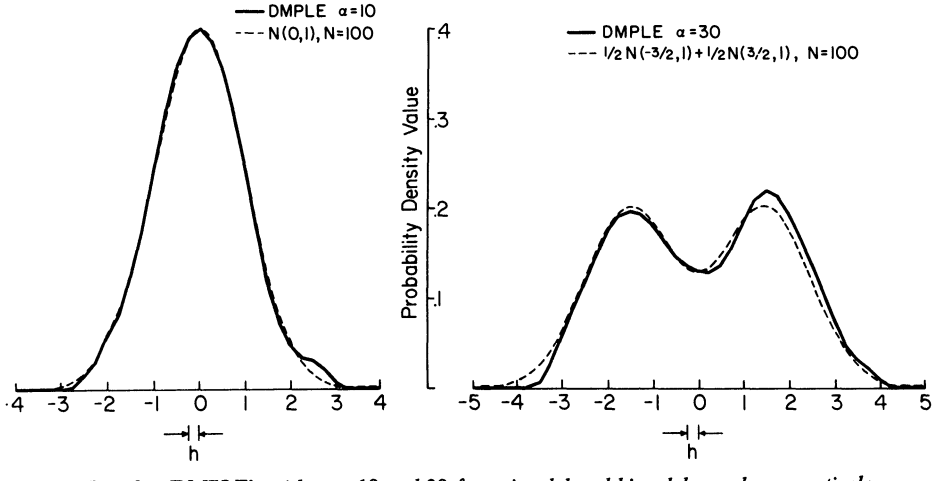

FIG. 2. DMPLE's *with* $\alpha = 10$ *and* 30 *for unimodal and bimodal samples respectively.*

with a sample correlation of $r = -.996$. Thus in this case the IMSE $\approx O(N^{-.773})$ which is about the same as that for kernel estimators, namely $O(N^{-\frac{4}{5}})$ as is discussed in [6]. See Figure 2 for DMPLE examples of selected random samples.

Acknowledgments. The authors would like to thank R. H. Byrd, P. E. Pfeiffer, and W. E. Veech for helpful discussions. We would also like to thank the referees whose constructive criticisms have been particularly appreciated.

REFERENCES

[1] DE MONTRICHER, G. F., TAPIA, R. A. and THOMPSON, J. R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalized function methods. *Ann. Statist.* **3** 1329–1348.

[2] GOOD, I. J. and GASKINS, R. A. (1972). Global nonparametric estimation of probability densities. *Virginia J. Sci.* **23** 171–193.

[3] International Mathematical and Statistical Libraries Subroutines NDMPLE, NDXEST, and GGNOR. Houston, Texas.

[4] LII, K. and ROSENBLATT, M. (1975). Asymptotic behavior of a spline estimate of a density function. *Comput. Math. Appl.* **1** 223–235.

[5] ORTEGA, J. M. and RHEINBOLDT, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York.

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

[7] ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–835.

[8] SCOTT, D. W., TAPIA, R. A. and THOMPSON, J. R. (1977). Kernel density estimation revisited. *J. Nonlinear Analysis, Theory, Methods and Applications* **1** 339–372.

[9] SILVERMAN, B. W. (1978). Choosing the window width when estimating a density. *Biometrika* **65** 1–11.

[10] TAPIA, R. A. (1974). A stable approach to Newton's method for general mathematical programming problems in R^n. *J. Optimization Theory Appl.* **14** 453–476.

[11] TAPIA, R. A. (1977). Diagonalized multiplier methods and quasi-Newton methods for constrained optimization. *J. Optimization Theory Appl.* **22** 135–194.

[12] WAHBA, G. (1978). Data based optimal smoothing of orthogonal series density estimates. TR. #509, Statist. Dept., Univ. Wisconsin.

D. W. SCOTT
DEPARTMENT OF COMMUNITY MEDICINE
BAYLOR COLLEGE OF MEDICINE
TEXAS MEDICAL CENTER
HOUSTON, TEXAS 77030

R. A. TAPIA AND J. R. THOMPSON
DEPARTMENT OF MATHEMATICAL SCIENCES
RICE UNIVERSITY
HOUSTON, TEXAS 77001