The Annals of Statistics
1982, Vol 10, No. 2, 401-414

NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION BY THE METHOD OF SIEVES

BY STUART GEMAN AND CHII-RUEY HWANG

Brown University and Academia Sinica, Taiwan

Maximum likelihood estimation often fails when the parameter takes values in an infinite dimensional space. For example, the maximum likelihood method cannot be applied to the completely nonparametric estimation of a density function from an iid sample; the maximum of the likelihood is not attained by any density. In this example, as in many other examples, the parameter space (positive functions with area one) is too big. But the likelihood method can often be salvaged if we first maximize over a constrained subspace of the parameter space and then relax the constraint as the sample size grows. This is Grenander's "method of sieves." Application of the method sometimes leads to new estimators for familiar problems, or to a new motivation for an already well-studied technique. We will establish some general consistency results for the method, and then we will focus on three applications.

1. Introduction. Techniques for estimating finite dimensional parameters typically fail when applied to infinite dimensional problems. The difficulties encountered in moving from finite to infinite dimensions are well illustrated by the failure of maximum likelihood in nonparametric density estimation. Let $x_1, \cdots, x_n$ be an iid sample from an absolutely continuous distribution with unknown probability density function (pdf) $\alpha_0(x)$. The maximum likelihood estimator for α_0 maximizes

$$\prod_{i=1}^n \alpha(x_i) \tag{1.1}$$

over some specified set of candidates. But if this set is too large, then the method will fail to produce a meaningful estimator. For instance, in the extreme case nothing is known about α_0, and the maximum of (1.1) is not achieved. Roughly speaking, we move out of the parameter space (the space of all densities) toward a discrete distribution with jumps at the sample points.

Another example of the failure of classical methods to solve infinite dimensional problems is the breakdown of least squares in the nonparametric estimation of a regression. Let X and Y be random variables and let $(x_1, y_1), \cdots, (x_n, y_n)$ be an iid sample from the bivariate distribution of (X, Y). The least squares estimator of the regression function $E(Y \mid X = x)$ minimizes

$$\sum_{i=1}^n \{y_i - \alpha(x_i)\}^2. \tag{1.2}$$

Observe that the minimum is zero and is achieved by any function which passes through all of the points of observation, $(x_1, y_1), \cdots, (x_n, y_n)$. Excepting some very special cases, this set does not in any meaningful sense converge to the true regression.

Grenander (1981) suggests the following remedy: perform the optimization (maximization of the likelihood, minimization of the sum of square errors, etc.) within a subset of the parameter space, and then allow this subset to "grow" with the sample size. He calls this sequence of subsets from which the estimator is drawn a "sieve," and the resulting

---

Received July 1979; revised January 1980, October 1981.

[1] This work was partially supported by NSF Grant MCS76-80762, Department of the Army contract DAAG29-80-K-0006, and the National Science Council of the Republic of China.

AMS 1980 *subject classificaton.* Primary 62A10; secondary 62G05.

Key words and phrases. Method of sieves, nonparametric estimation, maximum likelihood, regression, density estimation.

estimation procedure is his "method of sieves." The method leads easily to consistent nonparametric estimators in even the most general settings, with different sieves giving rise to different estimators. Often the sieve estimator is closely related to an already well-studied estimator, and may suggest an improvement, or a new point of view and a new motivation. Numerous examples of sieve estimators are presented in Grenander (1981). A few examples here will make much more clear the technical sections which follow.

The histogram is a simple example of a sieve estimator. Consider again the problem of estimating an entirely unknown density function $\alpha_0(x)$. We have seen that unmodified maximum likelihood is not consistent for this problem. A sieve is a sequence of subsets of the parameter space, such as:

$$S_m = \left\{ \alpha : \text{is a pdf which is constant on } \left[\frac{k-1}{m}, \frac{k}{m} \right), k = 0, \pm 1, \pm 2, \cdots \right\}$$

$m = 1, 2, \cdots$. The method of sieves estimator maximizes the likelihood, $\prod_{i=1}^{n} \alpha(x_i)$, subject to $\alpha \in S_m$, allowing m to grow with the sample size. The well-known solution is the function

$$\hat{\alpha}(x) = \frac{m}{n} \#\left\{ x_i : \frac{k-1}{m} \leq x_i < \frac{k}{m} \right\} \quad \text{for} \quad x \in \left[\frac{k-1}{m}, \frac{k}{m} \right),$$

i.e. the historgram with bin width m^{-1}. Putting aside details, we know that if $m_n \uparrow \infty$ sufficiently slowly, then $\hat{\alpha}$ is consistent, e.g. in the sense that $\int | \hat{\alpha}(x) - \alpha_0(x) | \, dx \to 0$ a.s.

For the same problem, a different and more interesting sieve is the "convolution sieve":

$$S_m = \left\{ \alpha : \alpha(x) = \int \frac{m}{\sqrt{2\pi}} e^{-\frac{m^2}{2}(x-y)^2} F(dy), F \text{ an arbitrary cdf} \right\}.$$

This time, maximizing the likelihood within S_m gives rise to an estimator closely related (but not identical) to the Parzen-Rosenblatt (Gaussian) kernel estimator. In fact, the latter is in the sieve S_m: take F to be the empirical distribution function. But the maximum of the likelihood is achieved by using a different distribution. As with the Parzen-Rosenblatt estimator, if $m_n \uparrow \infty$ sufficiently slowly (i.e. the "window width" is decreased sufficiently slowly) then the estimator is consistent. A more precise discussion of this and some related sieves is in Section 6.

The inconsistency of least squares nonparametric regression can be similarly rectified by introducing sieves. Let us look again at the regression problem formulated above; recall that $(x_1, y_1), \cdots, (x_n, y_n)$ is an iid sample from the bivariate distribution of (X, Y). Given a sieve S_m, the method of sieves estimator $\hat{\alpha}$ minimizes the sum of square errors, (1.2), subject to $\hat{\alpha} \in S_m$. If, as an example,

$$S_m = \left\{ \alpha : \alpha \text{ absolutely continuous}, \int \left| \frac{d}{dx} \alpha(x) \right|^2 dx \leq m \right\},$$

then $\hat{\alpha}$ is uniquely determined; it is a first degree polynomial smoothing spline; i.e. $\hat{\alpha}$ is continuous and piecewise linear with discontinuities in $(d/dx)\ \hat{\alpha}$ at $x_1, \cdots, x_n$; see Schoenberg (1964). It is possible to show that if m_n increases sufficiently slowly, then the estimator is strongly consistent for $E(Y \mid X = x)$ in a suitable metric; details are in Geman (1981). Other sieves applied to the same problem lead to kernel estimators and still others to new estimators. Even if the squared loss function $\{y - \alpha(x)\}^2$ is replaced by a "robust" alternative, minimization over too large a set will again fail to produce a meaningful estimator. In exactly the same way, sieves offer a remedy in this case as well.

Because this same method produces a variety of estimators, certain properties (existence and consistency, at the least) can be given a unified rather than case-by-case treatment. This paper is a first step toward such an approach. So that the paper will have sufficient focus, our theorems are about maximum likelihood estimation only. It will be obvious that much of the discussion also applies to least squares regression, or to other estimators

similarly derived from optimization problems. Following a section devoted to notation and definitions, Section 3 contains the main results. These are two theorems declaring the existence and consistency of maximum likelihood sieve estimators under the condition that the sieve grow sufficiently slowly with the sample size. Then, in Sections 4, 5, and 6 we apply these general results to some specific examples. The examples were chosen for illustration; they represent simple applications of the results in Section 3. We believe that some of these estimators, particularly in Section 6, have good practical potential, but this was not a consideration in their selection.

There are numerous well-studied techniques, in both numerical analysis and statistics, that are closely related to the method of sieves. So as to put the method in better perspective, let us list some (but far from all) of these related approaches. The finite element and the Rayleigh-Ritz-Galerkin approximations, most commonly applied to the solutions of partial differential equations, are close analogues in the deterministic setting; see, for example, Strang and Fix (1973). For density estimation with the maximum likelihood criterion, the method of penalized maximum likelihood (Good and Gaskins, 1971; Tapia and Thompson, 1978) is a sort of "dual" of the method of sieves. This is because the problem of choosing α from a suitable class of densities (ψ) to maximize

$$\sum_{i=1}^{n} \log \alpha(x_i) + \lambda\phi(\alpha) \tag{1.3}$$

for some penalty function ϕ is the Lagrange multiplier version of the following constrained optimization problem: maximize $\prod_{i=1}^{n} \alpha(x_i)$ subject to $\alpha \in \psi$ and $\phi(\alpha) \leq m$. And, the solution to this is the method of sieves estimator when employing the sieve

$$S_m = \{\alpha \in \psi : \phi(\alpha) \leq m\}.$$

For the regression problem, a similar relation exists between the least squares polynomial smoothing splines and certain sieve estimators. Fix $p = 1, 2, \cdots$, and let ψ be the collection of functions having $p - 1$ absolutely continuous derivatives. The sieve

$$S_m = \{\alpha \in \psi : \int \left| \frac{d^p}{dx^p} \alpha(x) \right|^2 dx \leq m\}$$

applied to the criterion (1.2) gives rise to a $2p - 1$ degree polynomial smoothing spline. The latter, widely studied as an estimator for nonparametric regressions (see Craven and Wahba, 1979, and references therein), is usually arrived at by solving a least squares analogue to problem (1.3): minimize

$$\sum_{i=1}^{n} \{y_i - \alpha(x_i)\}^2 + \lambda \int \left| \frac{d^p}{dx^p} \alpha(x) \right|^2 dx$$

over ψ. Finally, we should also mention the truncated orthogonal series estimators, especially as treated by Kronmal and Tarter (1968) and Tarter and Kronmal (1970), and the "maximum likelihood admissible" estimator introduced by Wegman (1975). If the coefficients in an orthogonal series estimator are chosen by optimizing some criterion, then the estimator has an obvious interpretation as an example of the method of sieves. And, if we are willing to relax the definition of a sieve so that it may depend on the random sample, then Wegman's estimator also permits this interpretation.

2. Definitions and notation. We will assume that the parameter space, A, is a metric space, with metric d. α_0 will refer to the "true" (and unknown) parameter. The value space is a measure space, $(X, \mathscr{B}, dx)$, with σ-finite measure dx. On $(X, \mathscr{B})$, we have a family of probability measures, $\{P_\alpha : \alpha \in A\}$, with the properties that $P_\alpha \neq P_\beta$ if $\alpha \neq \beta$, and that P_α is absolutely continuous w.r.t. dx and $(dP_\alpha/dx)(x) = f(x, \alpha)$.

A sieve for the parameter space A is a sequence $\{S_m\}$ of subsets of A. (Usually S_m is compact, $S_m \subseteq S_{m+1}$ and $\cup S_m$ is dense in A.) We will use the following notations and definitions

(a) For $\alpha \in S_m$, $B_m(\alpha, \epsilon) = \{\beta : \beta \in S_m \text{ and } d(\alpha, \beta) < \epsilon\}$.
(b) $E_\alpha g(x) = \int g(x)\, dP_\alpha = \int g(x) f(x, \alpha)\, dx$. The "formal entropy" is $H(\alpha, \beta) = E_\alpha \ln f(x, \beta)$. $H(\alpha, \alpha) - H(\alpha, \beta)$ is the familiar Kullback-Leibler information.
(c) For any extended real-valued function g on A, and any $B \subset A$, $g(B) = \sup_{\beta \in \mathscr{B}} g(\beta)$.
(d) P will denote the infinite produce measure, $P_{\alpha_0} \times P_{\alpha_0} \times \cdots$, on $(\Omega, \mathscr{F})$, where $\Omega = X^\infty$ and $\mathscr{F}$ = completion of $\mathscr{B}^\infty$ w.r.t. P. Unless otherwise specified, "almost sure" will be understood to mean w.r.t. P. ω will denote the typical point in Ω: $\omega = (x_1, x_2, \cdots)$.
(e) For each n, the likelihood function based on $(x_1, \cdots, x_n)$ is $L_n(\omega, \alpha) = \prod_{i=1}^n f(x_i, \alpha)$.
(f) The set of all maximum likelihood estimators in S_m, given a sample of size n, is defined by

$$M_m^n(\omega) = \{\alpha \in S_m : L_n(\omega, \alpha) = L_n(\omega, S_m)\}.$$

The maximum entropy set in S_m is

$$A_m = \{\alpha : \alpha \in S_m \quad \text{and} \quad H(\alpha_0, \alpha) = H(\alpha_0, S_m)\}.$$

(g) For $C_m \subseteq A$, $C_m \to \alpha$ means $\sup_{\beta \in C_m} d(\alpha, \beta) \to 0$.

3. General results. Let us first settle the question of the *existence* of a sequence m_n for which the maximum likelihood set, $M_{m_n}^n$, is consistent. Shortly thereafter, we will discuss the more important question of identifying such a sequence.

THEOREM 1. *Assume that a sieve, $\{S_m\}$, is chosen such that:*
B1. *For every m, every $\alpha \in S_m$, and every $\epsilon > 0$, $f(x, B_m(\alpha, \epsilon))$ is measurable in x; for every m and almost every $x(dx)$, $\lim_{\epsilon \to 0} f(x, B_m(\alpha, \epsilon)) = f(x, \alpha)$ for all $\alpha \in S_m$, i.e. $f(x, \alpha)$ is upper-semicontinuous in α on S_m.*
B2. *For every m and every $\alpha \in S_m$, there exists $\epsilon > 0$ such that*

$$E_{\alpha_0} \ln f(x, B_m(\alpha, \epsilon)) < \infty.$$

B3. *S_m is compact for each m.*
B4. *$A_m \to \alpha_0$ as $m \to \infty$.*
Then, for every n, m, and almost every ω, $M_m^n(\omega)$ is nonempty, and for every sequence m_n increasing slowly enough, $M_{m_n}^n \to \alpha_0$ a.s.

REMARK 1. B4 is usually easy to verify: Quite often $H(\alpha_0, \alpha_n) \to H(\alpha_0, \alpha_0)$ implies $\alpha_n \to \alpha_0$. (Implicit is the assumption that if $H(\alpha_0, \alpha_0) = +\infty$, then $H(\alpha_0, \alpha) < +\infty$ for all $\alpha \neq \alpha_0$, and that $H(\alpha_0, \alpha_0) > -\infty$.) If, in addition, $\{S_m\}$ is chosen so that there exists $\alpha_m \in S_m$ with $H(\alpha_0, \alpha_m) \to H(\alpha_0, \alpha_0)$, then $A_m \to \alpha_0$. (Roughly speaking, the condition guarantees that the sieve is sufficiently dense in the parameter space.)

REMARK 2. The set M_m^n may be replaced by $\{\alpha : \alpha \in S_m, L_n(\omega, \alpha) \geq q L_n(\omega, S_m)\}$ where q is any fixed constant, with $0 < q \leq 1$. Theorem 1 still holds (i.e. if α is chosen so that $L_n(\omega, \alpha)$ is always greater than a fixed proportion of the maximum, then α will converge to α_0). Theorem 2, below, can be similarly reformulated.

A proof of Theorem 1 can be obtained by a straightforward adaptation of the techniques of Wald (1949), Bahadur (1967), and others. Since the theorem is not used in what follows, we omit the details.

In Theorem 1, "slowly enough" may depend on α_0. The method has no application if we cannot identify a "universal" sequence m_n which guarantees consistency whatever the target parameter. Theorem 2 enables us to identify such a sequence. Its application will be illustrated by three examples, to be presented in Sections 4, 5, and 6.

The conditions for Theorem 2 will include the following.

CONDITION C1. For every m and every n, M_m^n is almost surely (dP) nonempty.

CONDITION C2. (a) If, for some sequence $\alpha_m \in S_m$, $H(\alpha_0, \alpha_m) \to H(\alpha_0, \alpha_0)$, then $\alpha_m \to \alpha_0$. (b) There exists a sequence $\alpha_m \in S_m$ such that $H(\alpha_0, \alpha_m) \to H(\alpha_0, \alpha_0)$.

For each $\delta > 0$ and each m, define

$$D_m = \{\alpha \in S_m : H(\alpha_0, \alpha) \leq H(\alpha_0, \alpha_m) - \delta\}$$

where α_m is the sequence in C2(b). Given ℓ sets $\mathcal{O}_1, \cdots \mathcal{O}_\ell$ in S_m such that $f(\cdot, \mathcal{O}_k)$ is measurable for each k, define

$$\rho_m = \sup_k \inf_{t \geq 0} E_{\alpha_0} \exp\left[t \ln \left\{ \frac{f(x, \mathcal{O}_k)}{f(x, \alpha_m)} \right\} \right].$$

THEOREM 2. *Assume $\{S_m\}$ is chosen so that conditions C1 and C2 are in force, and let $\{m_n\}$ be a sequence diverging to ∞. Suppose that for each $\delta > 0$ we can find $\mathcal{O}_1^m, \cdots \mathcal{O}_{\ell_m}^m$ in S_m, $m = 1, 2, \cdots$, such that*

(i) $$D_m \subseteq \cup_{k=1}^{\ell_m} \mathcal{O}_k^m,$$

(ii) $$f(\cdot, \mathcal{O}_k^m) \text{ is measurable}$$

(iii) $$\sum_{n=1}^{\infty} \ell_{m_n} (\rho_{m_n})^n < \infty.$$

Then $M_{m_n}^n \to \alpha_0$ a.s.

REMARK 1. The condition on m_n depends on α_0. The theorem is applied by demonstrating that once given m_n, the condition holds for arbitrary $\alpha_0 \in A$. An example is worked through in detail in Section 4.

REMARK 2. We expect $\rho_m < 1$ since, at least if $\mathcal{O}_k$ is small and if there is some $\alpha \in \mathcal{O}_k \cap D_m$, then the function

$$\phi(t) \equiv E_{\alpha_0} \exp\left[t \ln \left\{ \frac{f(x, \mathcal{O}_k)}{f(x, \alpha_m)} \right\} \right]$$

satisfies $\phi(0) = 1$, and

$$\phi'(0) = E_{\alpha_0}\{\ln f(x, \mathcal{O}_k)\} - H(\alpha_0, \alpha_m) \approx H(\alpha_0, \alpha) - H(\alpha_0, \alpha_m) \leq -\delta.$$

Notice that smaller sets $\mathcal{O}_1, \cdots, \mathcal{O}_{\ell_m}$ will in general lead to smaller ρ_m, but larger ℓ_m.

REMARK 3. The theorem can be reformulated so as to replace the exponential bound, ρ_m, by a more general moment bound. But in all of our examples, the functions $f(x, \alpha)$, $\alpha \in S_m$, are "sufficiently regular" to permit application of Theorem 2 in its present form. When an exponential bound is possible, it should be used; the weaker moment bounds lead to severely restricted rates of growth for m_n.

REMARK 4. A metric must first be chosen for the parameter space A. Consistency is then in the sense of this metric. Often (as in the examples of the next two sections) C2(a) suggests the "natural" metric for a problem.

REMARK 5. When A is separable, we can take S_m to be *finite* (say, the first m points of a countable dense subset). But this will typically necessitate an awkward procedure for calculating the maximum likelihood solution (especially when m is large), whereas a more carefully chosen sieve will often define an easily computed estimator.

PROOF OF THEOREM 2. Fix $\delta > 0$. We want to show that

(3.1) $$P(D_{m_n} \cap M_{m_n}^n \neq \phi \text{ i.o.}) = 0.$$

For, if (3.1) holds, then with probability one

$$\inf_{\alpha \in M_{m_n}^n} H(\alpha_0, \alpha) \geq H(\alpha_0, \alpha_{m_n}) - \delta$$

for all n sufficiently large. Since δ is arbitrary, and since $H(\alpha_0, \alpha_m) \to H(\alpha_0, \alpha_0)$ by condition C2(b),

$$\liminf_{n\to\infty} \inf_{\alpha \in M^n_{m_n}} H(\alpha_0, \alpha) \geq H(\alpha_0, \alpha_0) \text{ a.s.}$$

Then, combining with the well-known inequality $H(\alpha_0, \alpha) \leq H(\alpha_0, \alpha_0)$,

$$\lim_{n\to\infty} \sup_{\alpha \in M^n_{m_n}} |H(\alpha_0, \alpha) - H(\alpha_0, \alpha_0)| = 0 \text{ a.s.} \tag{3.2}$$

Fix $\epsilon > 0$, and for each n choose $\beta_n \in M^n_{m_n}$ such that

$$\frac{d(\alpha_0, \beta_n)}{1 + d(\alpha_0, \beta_n)} > \sup_{\alpha \in M^n_{m_n}} \frac{d(\alpha_0, \alpha)}{1 + d(\alpha_0, \alpha)} - \epsilon.$$

Condition C2(a), combined with (3.2), implies that $d(\alpha_0, \beta_n) \to 0$ a.s. Hence

$$\limsup \sup_{\alpha \in M^n_{m_n}} \frac{d(\alpha_0, \alpha)}{1 + d(\alpha_0, \alpha)} \leq \epsilon \text{ a.s.}$$

Since ϵ is arbitrary, $M^n_{m_n} \to \alpha_0$ a.s., and so it is enough to prove (3.1).

For now, fix m and n. Then

$$\begin{aligned}
\{D_m \cap M^n_m \neq \phi\} &\subseteq \{\sup_{\alpha \in D_m} L_n(\omega, \alpha) \geq L_n(\omega, \alpha_m)\} \\
&\subseteq \cup_{k=1}^{\ell_n} \{\sup_{\alpha \in \mathcal{O}^n_k} \textstyle\prod_1^n f(x_i, \alpha) \geq \prod_1^n f(x_i, \alpha_m)\} \\
&\subseteq \cup_{k=1}^{\ell_n} \{\textstyle\prod_1^n f(x_i, \mathcal{O}^m_k) \geq \prod_1^n f(x_i, \alpha_m)\}.
\end{aligned}$$

We will now bound the probability, call it π, of this latter set.

$$\pi \leq \sum_{k=1}^{\ell_n} P\{\textstyle\prod_1^n f(x_i, \mathcal{O}^m_k) \geq \prod_1^n f(x_i, \alpha_m)\} = \sum_{k=1}^{\ell_n} P\left(\exp\left[\sum_1^n t_k \ln\left\{\frac{f(x_i, \mathcal{O}^m_k)}{f(x_i, \alpha_m)}\right\}\right] \geq 1\right)$$

$$\leq \sum_{k=1}^{\ell_n} \left(E_{\alpha_0} \exp\left[t_k \ln\left\{\frac{f(x, \mathcal{O}^m_k)}{f(x, \alpha_m)}\right\}\right]\right)^n,$$

for any nonnegative $t_1, \cdots t_k$. Since $t_1, \cdots, t_k$ are arbitrary, $\pi \leq \ell_m(\rho_m)^n$, and then (3.1) follows from the Borel-Cantelli lemma. □

Quite obviously, the theorem does not provide a simple recipe for computing m_n. Most of the work is left to the application of rather complicated conditions to specific examples. But the approach is versatile, and can be applied without essential change to most infinite dimensional estimation problems.

A comment should be made concerning the connection with maximum likelihood estimation in conventional parameter spaces. Suppose that A is a finite dimensional Euclidean space, and that the target parameter, α_0, is contained in S_m for all m sufficiently large. Suppose also that Theorem 1 applies. Following some well known arguments (e.g., Cramér, 1966, Chapter 33), we can conclude, under typical local regularity assumptions, that the maximum likelihood set, $M^n_{m_n}$, will eventually contain only one element, α^*_n. Furthermore, $\sqrt{n}\ (\alpha^*_n - \alpha_0)$ is asymptotically normal with optimal covariance matrix.

4. Estimation of a regression function. Our first example is about nonparametric estimation of a regression function. Stone (1977) has presented a non-likelihood-based solution to this problem in a more general setting. The sieve method, too, extends to the completely general formulation, when likelihood is replaced by squared error (see Geman, 1981). But to illustrate Theorem 2, we will stick with maximum likelihood estimation and accept some otherwise unnecessary assumptions.

Our model here is

$$y = \alpha_0(x) + \mathcal{N} \tag{4.1}$$

with the assumptions:

A1. x and $\mathcal{N}$ are independent random variables.
A2. F, the distribution of x, concentrates on [0, 1].
A3. $\mathcal{N} \sim N(0, \sigma^2)$, σ^2 possibly unknown.
A4. $\int_0^1 \exp\{t \mid \alpha_0(x)|\}F(dx) < \infty$, for some $t > 0$.

A4, which is for the exponential bound required by Theorem 2, can be relaxed to a moment condition (see Remark 3 following Theorem 2), but this necessitates a far more restrictive bound on the growth of m_n.

The parameter space is

$$A = \{\alpha(x)x \in [0, 1]: Ee^{t|\alpha(x)|} < \infty \quad \text{for some} \quad t > 0\}.$$

It is of no consequence that A depends on F, which may not be known. The observations are of iid random variables (x_1, y_1), (x_2, y_2), $\cdots$ in the value space $[0, 1] \times R$, these variables being generated by (4.1) from iid random variables $(x_1, \mathcal{N}_1)$, $(x_2, \mathcal{N}_2)$, $\cdots$. If we assign the measure $F \times \lambda$ (λ = Lebesgue measure) to $[0, 1] \times R$, then for any $\alpha \in A$

$$f(x, y\ \alpha) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\{y - \alpha(x)\}^2/(2\sigma^2)].$$

Maximizing likelihood amounts to choosing $\alpha \in A$ to minimize $\sum_{i=1}^n \{y_i - \alpha(x_i)\}^2$. Since any function which passes through each of the observations (x_1, y_1), $\cdots$, (x_n, y_n) achieves the minimum (zero), the maximum likelihood set is not consistent. Let us introduce a sieve.

A simple sieve for this problem, good for the purpose of illustrating the method, is a set of trigonometric polynomials with bounded coefficients:

$$S_m = \{\textstyle\sum_{k=0}^m a_k \cos k\pi x: \sum_{k=0}^m \mid a_k \mid \le K \ln m\}$$

for some constant K. $\alpha \in M_m^n$ if and only if $\alpha \in S_m$ and

$$\textstyle\sum_{i=1}^n \{y_i - \alpha(x_i)\}^2 = \inf_{\beta \in S_m} \sum_{i=1}^n \{y_i - \beta(x_i)\}^2.$$

(An alternative sieve,

$$S'_m = \{\alpha: \alpha \quad \text{absolutely continuous}, \quad \int_0^1 \mid \dot{\alpha}(x)|^2\, dx \le m\},$$

leads us to the polynomial smoothing spline of degree one (cf. Schoenberg, 1964), and can be treated by the same techniques. In fact the conclusion of the theorem below, a strong consistency result, holds with S_m replaced by S'_m and $m_n = O(n^{1/4-\epsilon})$; see Geman (1981). The estimator derived from S_m is perhaps less attractive, certainly from a computational viewpoint, but it offers a more elementary illustration of Theorem 2.)

Let us look at the conditions for Theorem 2.

C1: For fixed x and y, $f(x, y, \alpha)$ can be viewed as a continuous function on a compact subset of R^{m+1}:

$$\{(a_0, \cdots, a_m): \textstyle\sum_{k=0}^m \mid a_k \mid \le K \ln m\}.$$

It follows that for fixed (x_1, y_1), $\cdots$, (x_n, y_n) the same can be said for $L_n(\omega, \alpha)$, and therefore M_m^n is nonempty.

C2(a): Examination of the condition leads to the "natural" metric for A:

$$\begin{aligned} H(\alpha_0, \alpha_0) - H(\alpha_0, \alpha) &= \frac{1}{2\sigma^2}[E\{y - \alpha(x)\}^2 - E\{y - \alpha_0(x)\}^2] \\ &= \frac{1}{2\sigma^2} E[E\{\alpha^2(x) - 2\alpha(x)y + 2\alpha_0(x)y - \alpha_0^2(x) \mid x\}] \\ &= \frac{1}{2\sigma^2} E\{\alpha_0(x) - \alpha(x)\}^2 = \frac{1}{2\sigma^2} \| \alpha_0 - \alpha \|_{L_2}^2, \end{aligned}$$

where $L_2 = L_2([0, 1], B, F)$. Taking $d(\alpha, \beta) = \|\alpha - \beta\|_{L_2}$, C2(a) is trivially satisfied. C2(b): Clearly in force, because $\cup_{m=1}^{\infty} S_m$ is dense in L_2, and d is the L_2 norm.

In the sense of this metric, an application of Theorem 2 establishes consistency of the maximum likelihood set as follows.

THEOREM 3. *If $m_n \to \infty$ in such a way that $m_n = O(n^{1-\epsilon})$ for some $\epsilon > 0$, then*

$$\sup_{\alpha \in M_{m_n}^n} \int |\alpha(x) - \alpha_0(x)|^2 F(dx) \to 0 \quad \text{a.s.}$$

PROOF. We have already checked C1 and C2. In the calculations below (and in the following sections as well), "c" will refer ambiguously to many different constants.

Fix $\delta > 0$. We will first define suitable sets $\mathcal{O}_1^m, \cdots, \mathcal{O}_{\ell_m}^m$ covering D_m. Consider the set of functions $\alpha \in S_m$ of the form

$$\alpha(x) = \textstyle\sum_{k=0}^m a_k \cos k\pi x,$$

where for each k

$$a_k = -K \ln m + \frac{p}{m^2} \tag{4.2}$$

for some $p = 0, 1, \cdots$. Since $|a_k| \le K \ln m$, there are no more than

$$\left(\frac{2K \ln m}{1/m^2}\right)^{m+1} \le (cm)^{cm} \tag{4.3}$$

such α. Associate with each such α the set of all functions $\beta \in S_m$ satisfying

$$\sup_x |\alpha(x) - \beta(x)| \le \frac{2}{m}.$$

Call the resulting collection of sets $\hat{\mathcal{O}}_1^m, \cdots, \hat{\mathcal{O}}_{\ell_m}^m$, where by (4.3), $\ell_m \le (cm)^{cm}$. Notice that for any $\beta \in S_m$, $\beta = \sum_{k=0}^m b_k \cos k\pi x$, we can find a_k, $0 \le k \le m$, of the form in (4.2), such that $|a_k - b_k| \le m^{-2}$ for each k. If $\alpha = \sum_{k=0}^m a_k \cos k\pi x$, then

$$\sup_x |\alpha(x) - \beta(x)| \le \textstyle\sum_{k=0}^m |a_k - b_k| \le \frac{2}{m}.$$

Hence $\hat{\mathcal{O}}_1^m, \cdots, \hat{\mathcal{O}}_{\ell_m}^m$ covers S_m. Now define $\mathcal{O}_1^m, \cdots, \mathcal{O}_{\ell_m}^m$ by

$$\mathcal{O}_k^m = \hat{\mathcal{O}}_k^m \cap D_m,$$

and observe that these cover D_m.

Fix k and fix $\alpha \in \mathcal{O}_k^m$. Then

$$E \ln f(x, y, \mathcal{O}_k^m) - E \ln f(x, y, \alpha) = \frac{1}{2\sigma^2} E \sup_{\beta \in \mathcal{O}_k^m} [\{\alpha(x) - \beta(x)\}\{\alpha(x) + \beta(x) - 2y\}].$$

For any $\gamma \in S_m$, $\sup_x |\gamma(x)| \le K \ln m$, and for α and β both in $\mathcal{O}_k^m$, $\sup_x |\alpha(x) - \beta(x)| \le 4/m$. Hence

$$|E \ln f(x, y, \mathcal{O}_k^m) - E \ln f(x, y, \alpha)| < \frac{c \ln m}{m}.$$

Using this and the definition of D_m, we have

$$E \ln \left\{\frac{f(x, y, \mathcal{O}_k^m)}{f(x, y, \alpha_m)}\right\} \le H(\alpha_0, \alpha) - H(\alpha_0, \alpha_m) + \frac{c \ln m}{m} \le \frac{c \ln m}{m} - \delta \tag{4.4}$$

for all $k = 1, 2, \cdots, \ell_m$ and all m.

Again fix k, and define

$$\phi(t) = E \exp\left[t \ln \left\{ \frac{f(x, y, \mathcal{O}_k^m)}{f(x, y, \alpha_m)} \right\} \right].$$

Then $\phi(0) = 1$, and by (4.4) $\phi'(0) \leq m^{-1} c \ln m - \delta$. By similar computations (and here is where A4 comes in) it is possible to show that $\phi''(t) \leq c(\ln m)^c$ provided $t \leq 1/(\ln m)^c$. But then

$$\phi'(t) \leq \frac{c \ln m}{m} - \delta + c(\ln m)^c t,$$

which implies

$$\phi(t) \leq 1 + \frac{c \ln m}{m} t - \delta t + \frac{1}{2} c(\ln m)^c t^2.$$

Hence if $t = 1/(\ln m)^p$, and if p, q and m are sufficiently large, then

$$\phi(t) \leq 1 - \frac{\delta}{(\ln m)^q}.$$

At least for large m then,

$$\rho_m \leq 1 - \frac{\delta}{(\ln m)^q}.$$

And, finally,

$$\sum_{n=1}^{\infty} \ell_{m_n} (\rho_{m_n})^n \leq \sum_{n=1}^{\infty} (cm)^{cm} \left\{ 1 - \frac{\delta}{(\ln m)^q} \right\}^n$$

which is finite if $m_n = O(n^{1-\epsilon})$. □

5. Estimation of the mean function of a Gaussian process. The following example is discussed in Grenander (1981, Chapter 8), where consistency in the sense of our Theorem 1 is proven. Here we will identify explicitly a sequence m_n which guarantees this consistency.

Suppose that we make repeated and independent observations of the process

$$x(t) = \int_{-1/2}^{t} \alpha_0(s)\, ds + w(t), \quad t \in \left[-\frac{1}{2}, \frac{1}{2} \right],$$

where $\alpha_0 \in L_2[-1/2, 1/2]$ is unknown and w is the Wiener process with unit variance per unit time. For simplicity, let us assume that α_0 is real and even. The parameter space is

$$A = \left\{ \alpha : \alpha \quad \text{is even and} \quad \alpha \in L_2 \left[-\frac{1}{2}, \frac{1}{2} \right] \right\}.$$

Define, for $k \neq 0$,

$$a_k = \sqrt{2} \int_{-1/2}^{1/2} \cos(2\pi k i) \alpha(t)\, dt,$$

$$w_k = \sqrt{2} \int_{-1/2}^{1/2} \cos(2\pi k t) w(dt),$$

$$x_k = \sqrt{2} \int_{-1/2}^{1/2} \cos(2\pi k t) x(dt) = a_k^o + w_k,$$

where a_k^o = kth Fourier coefficient for α_0; for $k = 0$, use "1" in place of "$\sqrt{2}$" to define a_0, w_0, and x_0. Observe that the w_k's are iid $N(0, 1)$ random variables. The Radon-Nikodym derivative, w.r.t. the measure with $\alpha = 0$, is

$$\frac{dP_\alpha}{dP_0}(x) = \exp \sum_{-\infty}^{\infty} \left(a_k x_k - \frac{1}{2} a_k^2 \right).$$

Since $| \sum_{-\infty}^{\infty} a_k x_k | < \infty$ a.s., P_α is equivalent to P_0. Using P_0 as "dx", we define

$$f(x, \alpha) = \exp \sum_{-\infty}^{\infty} \left(a_k x_k - \frac{1}{2} a_k^2 \right).$$

It is easily seen that the maximum likelihood solution does not exist in A. Hence, we introduce a sieve.

We shall use the notation $\sum k^p c_k$ with the understanding that summation is over $-\infty$ to $+\infty$, and at $k = 0$, $k^p c_k$ means c_0. Let

$$S_m = \{\alpha : \alpha \in A \quad \text{and} \quad \sum k^2 a_k^2 \leq m\}.$$

It is not hard to see that S_m is compact, $S_m \subseteq S_{m+1}$ and $U_m S_m$ is dense in A. In S_m, the maximum likelihood solution given n iid samples $x^1(\cdot), \cdots, x^n(\cdot)$ is

$$\hat{\alpha}_m^n(t) = \sum_{k=-\infty}^{\infty} \hat{\alpha}_k \cos(2\pi k t),$$

where

$$\hat{\alpha}_k = \frac{\sum_{i=1}^n x_k^i}{n + \lambda k^2}, \quad \sum \frac{k^2 (\sum_{i=1}^n x_k^i)^2}{(n + \lambda k^2)^2} = m.$$

An application of Theorem 2 leads to the following.

THEOREM 4. *If $m_n \to \infty$ in such a way that $m_n = O(n^{1/3-\varepsilon})$ for some $\varepsilon > 0$, then*

$$\int_{-1/2}^{1/2} | \hat{\alpha}_{m_n}^n(t) - \alpha_0(t) |^2 \, dt \to 0 \quad \text{a.s.}$$

The proof is by an argument entirely analogous to the one presented for Theorem 3 in the previous section. To avoid unnecessary repetition, we will mention just a few of the details. We have already demonstrated the unique maximum likelihood solution $(\hat{\alpha}_m^n)$, and hence condition C1 is satisfied. When we write out condition C2(a) we arrive at the metric $d(\alpha, \beta) = \| \alpha - \beta \|_{L_2}$, much as we did in Section 4 (this time using Lebesgue measure to define L_2). For C2(b) we observe that $\cup_{m=1}^{\infty} S_m$ is again dense in A.

Now, for each $\delta > 0$ we must define sets $\mathcal{O}_1^m, \cdots, \mathcal{O}_{\ell_m}^m$, suitably small, which cover D_m. Observe that if $\alpha \in S_m$ then $| a_k | \leq \sqrt{m} / | k |$ for all $k \neq 0$, and $| a_0 | \leq \sqrt{m}$. For each $k \neq 0$, divide $[-m/|k|, m/|k|]$ into $[m^2/|k| + 1]$ intervals of equal length, where $[x]$ is the greatest integer less than or equal to x, and let I_k denote the set of all endpoints of these intervals. Similarly, let I_0 denote the endpoints of the intervals obtained by dividing $[-\sqrt{m}, \sqrt{m}]$ into $[m^2 + 1]$ equal lengths. Notice that these intervals are all of length less than $2m^{-1.5}$. Associate with each collection $\{b_k : b_k \in I_k, k = 0 \pm 1, \pm 2, \cdots, [m^{1+\varepsilon}]\}$ a set

$$\hat{\mathcal{O}}^m(\{b_k\}) = \{\alpha \in S_m : | a_k - b_k | \leq 2m^{-1.5}, \quad k = 0, \pm 1, \cdots, \pm [m^{1+\varepsilon}]\},$$

where ε is the same as in the theorem statement. If $\hat{\mathcal{O}}_1^m, \cdots, \hat{\mathcal{O}}_{\ell_m}^m$ is the collection of all such sets, then $\cup_{k=1}^{\ell_m} \hat{\mathcal{O}}_k^m$ covers S_m and $\ell_m \leq (cm)^{cm^{1+\epsilon}}$. Finally, we define $\mathcal{O}_k^m = \hat{\mathcal{O}}_k^m \cap D_m$, $k = 1, 2, \cdots, \ell_m$, and these clearly cover D_m.

Fix k, and define

$$\phi(t) = E \exp\left[t \ln \left\{ \frac{f(x, \mathcal{O}_k^m)}{f(x, \alpha_m)} \right\} \right].$$

Calculations analogous to those performed in the previous example lead to (i) $\phi(0) = 1$, (ii)

$\phi'(0) \le cm^{-\epsilon/2} - \delta$, and, with the help of a result in Hwang (1980), (iii) $\phi''(t) \le cm^2$ for $t \in [0, 1/(cm)]$. From this we conclude that $\phi(m^{-2}) \le 1 - \delta/(cm^2)$ for all m sufficiently large. For large m then,

$$\ell_m(\rho_m)^n \le (cm)^{cm^{1+\epsilon}}(1 - \delta/\mathrm{cm}^2)^n,$$

and the latter is summable when $m = m_n = O(n^{1/3-\epsilon})$.

6. A variation on the kernel estimator. To motivate our last example, let us return to the "convolution sieve" defined in Section 1:

$$S_m = \left\{\alpha : \alpha(x) = \int_{-\infty}^{\infty} \frac{m}{\sqrt{2\pi}} \exp\left\{-\frac{m^2}{2}(x-y)^2\right\} F(dy),\ F \text{ an arbitrary cdf}\right\}.$$

Recall our description of the associated maximum likelihood estimator as being closely related to the Parzen-Rosenblatt kernel estimator. More specifically:

PROPOSITION 1. *For every n and m, M_m^n is nonempty, and $\alpha \in M_m^n$ implies*

$$\alpha(x) = \textstyle\sum_{i=1}^n p_i \dfrac{m}{\sqrt{2\pi}} \exp\left\{-\dfrac{m^2}{2}(x-y_i)^2\right\}$$

for some $y_1, \cdots, y_n$ and $p_1, \cdots, p_n$ satisfying $p_i \ge 0$, $1 \le i \le n$, $\sum_{i=1}^n p_i = 1$. Furthermore, if $\min(x_1, \cdots, x_n) < \max(x_1, \cdots, x_n)$ then $\min(x_1, \cdots, x_n) < \min(y_1, \cdots, y_n)$ and $\max(y_1, \cdots, y_n) < \max(x_1, \cdots, x_n)$. (The proof, by S. Geman and D. E. McClure, is in Geman (1981). Since the proposition is not directly related to Theorem 2 or its application, we will not reproduce it here.)

It is interesting to note that the kernel estimator with Gaussian kernel, i.e.

$$\beta(x) = \frac{1}{n}\textstyle\sum_{i=1}^n \dfrac{m}{\sqrt{2\pi}} \exp\left\{-\dfrac{m^2}{2}(x-x_i)^2\right\}$$

is in S_m, but the last statement in the proposition indicates that β is not among the maximum likelihood solutions, i.e. $\beta \notin M_m^n$.

Although we have characterized the maximum likelihood set up to the $2n$ parameters $y_1, \cdots, y_n, p_1, \cdots, p_n$, its actual computation is difficult. The proposition suggests a smaller and computationally more attractive sieve,

$$\hat{S}_m = \left\{\alpha : \alpha(x) = \frac{1}{n}\textstyle\sum_{i=1}^n \dfrac{m}{\sqrt{2\pi}} \exp\left[-\dfrac{m^2}{2}(x-y_i)^2\right]\right\}$$

i.e., we give equal mass to each kernel, but allow the locations to move in such a way as to maximize the likelihood. (Here again, it is easy to show that for $\alpha \in M_m^n$, which is the maximum likelihood set, $\min(x_1, \cdots, x_n) < \min(y_1, \cdots, y_n)$ and $\max(y_1, \cdots, y_n) < \max(x_1, \cdots, x_n)$ provided $\min(x_1, \cdots, x_n) < \max(x_1, \cdots, x_n)$; and so, again, the kernel estimator is not among the maximum likelihood solutions.) We have experimented with $\hat{S}_m$ and have found, as a rule, that the number of *distinct* y's in a maximum likelihood solution is considerably smaller than n. In other words, the kernels will often coalesce to achieve an increased likelihood. Sometimes this results in strikingly accurate density estimators, while at other times this "maximum likelihood" solution is a poor second to the corresponding (same window width) kernel estimator. In either case, this estimator suffers the very same stability problem as the kernel estimator: the results are critically dependent on the choice of the kernel width (which is here governed by the sieve parameter m).

One approach to this critical dependence on window width (σ) is to include σ as a free parameter within the sieve, and thus allow it to be chosen by maximum likelihood. But we must be somewhat careful; we cannot merely replace $\hat{S}_m$ by

$$\left\{\alpha : \alpha(x) = \frac{1}{n}\textstyle\sum_{i=1}^n \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\dfrac{1}{2\sigma^2}(x-y_i)^2\right]\right\},$$

since then the maximum of the likelihood is achieved with $\sigma = 0$ and the kernels centered at the sample points. Let us instead define the sieve parameter m to be the number of kernels, restricting this to be smaller than n, and consider

$$\tilde{S}_m = \left\{ \alpha : \alpha(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x - y_i)^2 \right] \right\}.$$

The associated maximum likelihood estimator has performed well in our simulations, but it is still true that the extreme possible values of the sieve parameter produce radically different estimators: with moderate sample sizes ($n \approx 50$), $m = 1$ generally oversmooths and $m = n - 1$ will almost always drastically undersmooth.

As a final example of the application of Theorem 2, we will obtain an asymptotic bound on the growth of the sieve $\tilde{S}_m$ which will guarantee strong consistency in the L_1 metric (S_m or $\hat{S}_m$ can be similarly treated). But first let us briefly discuss, in general terms, the important issue raised in the previous paragraphs: the dependence of a sieve estimator on the precise choice of sieve size. We have developed a general approach to obtaining asymptotic bounds on the growth rate of seives so as to ensure consistent estimation. But among the important practical questions that remain unanswered (including relative efficiency, asymptotic distributions, good sieves for robust estimation, etc.), perhaps most pressing is the problem of choosing an appropriate sieve size when given a fixed finite collection of observations. In one form or another, this "smoothing" problem faces all nonparametric estimators of densities and regressions. For kernel estimators it is the problem of choosing the right kernel width. For the maximum penalized likelihood estimators, it is the problem of choosing an appropriate weight to be given the penalty function. In each case the problem is one of choosing the right degree of smoothing when given finite data for a potentially infinite dimensional problem.

Among the general solutions proposed for the smoothing problem, there are at least two which have proven widely successful and which can be applied directly to the choice of sieve size. These are the methods of cross-validation—see Stone (1978) and Wahba (1981), and the many references therein—and Akaike's (1977) information criterion. We have experimented extensively with the former, and have found what many others have found (see, e.g., Scott and Factor, 1981; Utreras, 1979; Wahba and Wold, 1975; Wahba, 1981): that cross-validation is often a strikingly effective means of choosing an appropriate degree of smoothing. But aside from these promising simulation results, we have no real mathematical evidence to support the application of these techniques to the method of sieves. Indeed, the properties of estimators employing data-driven smoothing are almost entirely unknown, whether the application be to sieves or to any other nonparametric estimation technique. In our opinion, the identification of these properties stands as an unusually challenging and relevant problem for mathematical statistics.

Let us return to the easier task of guaranteeing consistent estimation.

THEOREM 5. *Assume that α_0 is a bounded density with compact (but possibly unknown) support. If $m_n \to \infty$ in such a way that $m_n = O(n^{1/5-\epsilon})$ for some $\epsilon > 0$, and $m_n \leq n - 1$ for all n, then*

$$\sup_{\alpha \in \tilde{M}^n_{m_n}} \int_{-\infty}^{\infty} |\alpha(x) - \alpha_0(x)| \, dx \to 0 \text{ a.s.},$$

where $\tilde{M}^n_m$ is the maximum likelihood set associated with $\tilde{S}_m$.

Most of the proof is a repeat of the calculations performed for the examples in Sections 4 and 5. But there are two new aspects which are perhaps worth mentioning. The first is the relation between the condition C2(a) and the metric for convergence (L_1 in the theorem); C2(a) does not directly translate into a "natural" metric for this problem, as it can be said to have done in the previous examples. Instead, one must first establish a

relation between convergence of the Kullback-Leibler information and L_1 convergence. In this regard, we have the following.

PROPOSITION 2. *Let α_0 be a density function satisfying $\int_{-\infty}^{\infty} \alpha_0(x) \ln \alpha_0(x)\, dx < \infty$. If, for each n, T_n is a collection of density functions, and if*

$$\lim_{n\to\infty} \sup_{\alpha\in T_n} \int_{-\infty}^{\infty} \alpha_0(x) \ln\left\{\frac{\alpha_0(x)}{\alpha(x)}\right\} dx = 0,$$

then also

$$\lim_{n\to\infty} \sup_{\alpha\in T_n} \int_{-\infty}^{\infty} |\alpha(x) - \alpha_0(x)|\, dx = 0,$$

and hence C2(a) *holds.*

The proof is in Geman (1981), and will not be repeated here.

Direct application of Theorem 2 to the sieve $\tilde{S}_m$ is not possible; for each $\delta > 0$, $\rho_m = 1$ no matter what the choice of covering sets $\mathcal{O}_1^m, \cdots, \mathcal{O}_{l_m}^m$. The underlying intuitive reason is that σ can be made arbitrarily small, and therefore each sieve contains estimators arbitrarily ill-behaved. But the maximum likelihood set $\tilde{M}_{m_n}^n$ consists of relatively smooth functions. This allows us to define a smaller and more regular "dummy" sieve which is guaranteed to contain $\tilde{M}_{m_n}^n$ for all n sufficiently large. Theorem 2 is then applied to this substitute sieve. (An analogous procedure can be used for the sieves S_m and $\hat{S}_m$ as well.) Specifically, let k be such that $[-k, k]$ contains the support of α_0. Then we first show (and this is not difficult) that with probability one, $\tilde{M}_{m_n}^n \subseteq \tilde{S}'_{m_n}$ for all n sufficiently large, where

$$\tilde{S}'_m = \left\{\alpha \in \tilde{S}_m : |y_i| \le k\ \forall i, \quad \text{and} \quad \frac{1}{m \ln m} \le \sigma \le 2k\right\}.$$

Theorem 2 applies directly to $\tilde{S}'_m$ (it does not matter that k is not known), and when n is large, $\tilde{S}'_{m_n}$ has the same maximum likelihood set $(\tilde{M}_{m_n}^n)$ as $\tilde{S}_{m_n}$.

Acknowledgements. Suggestions by Professor Bahadur and two referees led to a complete reorganization and reemphasis of the paper, all for the better. We gratefully acknowledge these contributions. Professor Grenander introduced us to the method of sieves, taught us the techniques of Bahadur, Wald, and others, and suggested that results like those in Section 3 might be available.

REFERENCES

AKAIKE, H. (1977). On entropy maximization principle. In: *Applications of Statistics*, Ed. P. R. Krishnaiah. North-Holland, New York.

BAHADUR, P. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

CRAMÉR, H. (1966). Mathematical Methods of Statistics. Princeton University Press, Princeton, N.J.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

GEMAN, S. (1981). Sieves for nonparametric estimation of densities and regressions. Repts. in Pattern Analysis, No. 99, D.A.M., Brown University.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

GRENANDER, U. (1981). Abstract Inference. Wiley, New York.

HWANG, C.R. (1980). Gaussian measure of large balls in a Hilbert space. *Proc. Amer. Math. Soc.* **78** no. 1, 107–110.

KRONMAL, R. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925–952.

SCHOENBERG, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci.* **52** 947–950.

SCOTT, D. W. and FACTOR, L. E. (1981). Monte Carlo study of three data-based nonparametric probability density estimators. *J. Amer. Statist. Assoc.* **76** 9–15.
STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.
STONE, M. (1978). Cross-validation: A review. *Math. Oper. Sch. Statist.*, Ser. Statistics **9** 127–139.
STRANG, G. and FIX, G. J. (1973). An Analysis of the Finite Element Method. Prentice-Hall, Englewood Cliffs, N.J.
TAPIA, R. A. and THOMPSON, J. R. (1978). Nonparametric Probability Density Estimation. Johns Hopkins University Press, Baltimore.
TARTER, M. and KRONMAL, R. (1970). On multivariate density estimates based on orthogonal expansions. *Ann. Math. Statist.* **41** 718–722.
UTRERAS, F. (1979). Cross-validation techniques for smoothing spline functions in one or two dimensions. In: Smoothing Techniques for Curve Estimation, Ed. T. Gasser and M. Rosenblatt. *Lecture Notes in Mathematics* **757**, Springer-Verlag, Berlin.
WAHBA, G. and WOLD, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Comm. Statist.* **4** 1–17.
WAHBA, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Ann. Statist.* **9** 146–156.
WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.
WEGMAN, E. J. (1975). Maximum likelihood estimation of a probability density function. *Sankhya* Ser. A **37** 211–224.

DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912

INSTITUTE OF MATHEMATICS
ACADEMIA SINICA
TAIPEI, TAIWAN, R.O.C.