The Annals of Statistics
1984, Vol. 12, No. 1, 372–379

A SIMPLE SOLUTION TO A NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION PROBLEM

BY GILBERT G. WALTER AND JULIUS R. BLUM[1]

University of Wisconsin-Milwaukee and University of California-Davis

Three approaches to a nonparametric maximum likelihood problem are considered. One, based on the method of "sieves", is shown to include the other two. The sieve considered is a double exponential convolution sieve. A closed form solution is given for certain values of the sieve parameter.

1. Introduction. Maximum likelihood estimation, while widely and successfully used for parameter estimation, has had mixed success as a tool in the estimation of density. Indeed the MLE of a density $f(x)$, based on an iid sample $x_1, x_2, \cdots x_n$, is

$$\hat{f}(x) = (1/n) \sum_{i=1}^{n} \delta(x - x_i) \tag{1.1}$$

which is itself not a density. This conclusion is usually based on a heuristic argument since the likelihood is infinity for this $\hat{f}$.

In order to get around these problems, a modification of MLE is often used. One such modification is the introduction of a penalty functional in the definition of likelihood. Another is the restriction of the allowable estimators to a subspace of an appropriate space of densities. Both of these methods are described by Tapia and Thompson (1978). Another approach, based on "sieves," is described by Grenander (1981) and Geman (1981).

In this work we shall consider one such sieve based on a two-sided exponential distribution. We obtain a closed form expression for the MLE in this case which we show to be consistent. This basic result allows us to consider two other modifications of the MLE. One will restrict the space of allowable densities while the other will enlarge the space to include expressions such as (1.1).

2. Three approaches to MLE. One of the difficulties with the use of MLE of a nonparametric density is that the likelihood functional

$$L(f) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \langle \delta_{x_i}, f \rangle \tag{2.1}$$

has no finite maximum as f is allowed to range over an appropriate space (usually a Sobolev space H^p). This may be rectified by changing the linear functional $\langle \delta_{x_i}, f \rangle$ in some way. One such way is to approximate δ by a delta sequence $\{\delta_\lambda\}$, then maximize the approximate functional

$$L_\lambda(f) = \prod_{i=1}^{n} \langle \delta_{\lambda, x_i}, f \rangle. \tag{2.2}$$

This will lead to the method of sieves.

Received October 1982; revised September 1983.

[1] Deceased, research on this paper in progress at the time of death.

AMS 1980 *subject classification.* 62605.

Key words and phrases. Sieves, Sobolev space, maximum likelihood, nonparametric estimation.

A sieve is defined by Geman (1981) as a sequence $\{S_n\}$ of sets of functions, indexed by sample size, from which the estimator is taken. In this work we use the "convolution sieve"

$$S_n = \left\{ \alpha(x) : \alpha(x) = \int_{-\infty}^{\infty} \frac{\lambda_n}{2} e^{-\lambda_n |x-y|} \, dF(y), \; F \text{ a pdf} \right\}$$

where $\lambda_n \to \infty$ as $n \to \infty$. The associated maximum likelihood problem:

$$\text{"maximize } \prod_{i=1}^{n} \alpha(x_i) \text{ subject to } \alpha \in S_n\text{"}$$

is the one we are interested in. It should be noted that $\{(\lambda_n/2) \exp(-\lambda_n | x - y |)\}$ is a positive delta-sequence (see Walter and Blum, 1979). Hence for fixed F, the corresponding element of S_n, α_n satisfies $\alpha_n(x) \to \int_{-\infty}^{\infty} \delta(x - y) \, dF(y) = F'(x)$ as $n \to \infty$ at points of continuity of $F'(x)$.

A second method is to replace the functional $\langle \delta_{x_i}, f \rangle$ by an inner product in a Hilbert space which contains both δ and f and whose value is close to that of the functional. Clearly the L^2 inner product or the inner product of the Sobolev spaces H^p, $p > 0$ won't do since δ does not belong to these spaces. However both δ and f are elements of the Sobolev space H^{-s} for $s > \frac{1}{2}$. A description of such spaces may be found in Rudin (1973). The inner product for $f, g \in H^{-s}$ is given by

$$\langle f, g \rangle_s = \int_{-\infty}^{\infty} \tilde{f}(x) \overline{\tilde{g}(x)} (1 + x^2)^{-s} \, dx \tag{2.3}$$

where $\tilde{f}$ is the Fourier transform of f, $\tilde{f}(x) = \sqrt{2\pi}^{-1} \int_{-\infty}^{\infty} f(t) e^{-ixt} \, dt$. This inner product exists whenever $\tilde{f}, \tilde{g} \in L^2(1 + x^2)^{-s}$ or equivalently whenever $f, g \in H^{-s}$. In particular since $\tilde{\delta} = (2\pi)^{-1/2}$, $\delta \in H^{-s}$ as is every density f and indeed every generalized derivative of a distribution function, for $s > \frac{1}{2}$.

The inner product given by (2.3) may be used to approximate $\langle f, g \rangle$ by introducing a scale parameter λ. By modifying (2.3) to

$$\langle f, g \rangle_{s,\lambda} = \int_{-\infty}^{\infty} \tilde{f}(x) \tilde{g}(x) (1 + x^2/\lambda^2)^{-s} \, dx \tag{2.3a}$$

we obtain an expression which is approximately scale invariant for large values of λ. Again if $g \in H^p$ for some $p > \frac{1}{2}$, then

$$\langle \delta, g \rangle_{s,\lambda} \to \langle \delta, g \rangle, \quad \lambda \to \infty.$$

In particular for $s = 1$, the inner product assumes a tractable and simple form. We shall concentrate on it in the sequel. The problem then is to maximize

$$L_\lambda^{(1)}(f) = \prod_{i=1}^{n} \langle \delta_{x_i}, f \rangle_{1,\lambda} \tag{2.4}$$

for those $f \in H^{-1}$ which are the derivative of a distribution function. Fortunately this has a simple characterization.

LEMMA 2.1. $L_1^{(1)}(f) = 2^{-n} \prod_{i=1}^{n} \int \exp(-|x_i - \phi|) f(\phi) \, d\phi$.

PROOF. The inner product $\langle \delta_x, f \rangle_1$ may be given by $\int \tilde{\delta}_x(t) \overline{\tilde{f}(t)} (t^2 + 1)^{-1} \, dt$

or

$$\langle \delta_x, f \rangle_1 = \int (2\pi)^{-1/2} e^{-ixt} \overline{\tilde{f}(t)} (t^2 + 1)^{-1} \, dt$$

$$= \mathcal{F}(\overline{\tilde{f}}(t^2 + 1)^{-1} = f^* \mathcal{F}(t^2 + 1)^{-1} \tag{2.5}$$

$$= f^* e^{-|x|}/2 = \frac{1}{2} \int e^{-|x-\phi|} f(\phi) \, d\phi$$

since the Fourier transform of $(t^2 + 1)^{-1}$ is $e^{-|x|}/2$.

Thus we see that this second method is a special case of the first corresponding to the sieve for $\lambda_n = 1$, and with the scale parameter $\lambda \neq 1$, is exactly the same.

The third method is a standard maximum likelihood method which uses (2.1) with $f \in H^1$ but with the added condition that $D^2 f$ belongs to a bounded set in H^{-1}.

Since $e^{-|x|}$ satisfies the differential equation

$$(1 - D^2) e^{-|x|} = 2\delta(x), \tag{2.6}$$

we may express the functional $\langle \delta, f \rangle$ as

$$\langle \delta, f \rangle = \langle (1 - D^2) e^{-|x|}/2, f \rangle = \langle e^{-|x|}/2, (1 - D^2) f \rangle \tag{2.7}$$

provided $(1 - D^2) f \in H^{-1}$. This last expression is the value of an element of H^1 (the dual space of H^{-1}) on an element of H^{-1}.

Hence (2.1) may be expressed as

$$L(f) = \prod_{i=1}^n \left\langle \frac{e^{-|x-x_i|}}{2}, g \right\rangle, \quad g = (1 - D^2) f, f \in H^1 \tag{2.8}$$

which, if g is restricted as f was in the second method, reduces to (2.4) with $\lambda = 1$. Without this restriction on g, (2.8) would merely be the standard MLE problem with no maximum in H^1. Thus all three approaches lead to the problem of maximizing

$$L_\lambda(f) = \prod_{i=1}^n \int \frac{\lambda}{2} e^{-\lambda |x_i - \theta|} f(\theta) \, d\theta \tag{2.9}$$

for $f \in H^{-1}, f \geq 0$ and $\int f = 1$.

3. The solution to the maximization problem. The problem of maximizing (2.9) is a familiar one in a number of other settings. It arises in mixture problems in which a random sample with density

$$h(x) = \int h(x|\theta) f(\theta) \, d\theta$$

must be used to estimate f. The MLE may be estimated by the well-known *E–M* algorithm due, in our context, to Hasselblad (1969).

THEOREM 3.1. *The problem of maximizing $L_\lambda(f)$ given by* (2.9) *for $f \in H^{-1}$,*

$f \geq 0$, and $\tilde{f}(0) = (2\pi)^{-1/2}$, has a unique solution given by

$$\tilde{f}(\theta) = \sum_{i=1}^{n} p_i \delta(\theta - x_i) \tag{3.1}$$

where $\{p_i\}$ satisfies, for $\lambda^{-1} \leq m$, the distance between nearest neighbors of $\{x_i\}$,

$$(1/n) \sum_{i=1}^{n} e^{-\lambda|x_i - x_j|} / \sum_{k=1}^{n} e^{-\lambda|x_i - x_k|} p_k = 1; \tag{3.2}$$

and $p_i > 0$, $i = 1, \cdots, n$.

REMARK 3.1. The element of the sieve S_n that corresponds to $\tilde{f}(\theta)$ is of course

$$\hat{\alpha}(x) = (\lambda/2) \sum_{i=1}^{n} p_i e^{-\lambda|x - x_i|}.$$

This was already shown to be consistent by Kiefer and Wolfowitz (1956) for $h(x)$ the true density of the X_i as $n \to \infty$, λ fixed. However, we are interested in letting $\lambda \to \infty$ as well and in showing that $\hat{\alpha}(x) \to f(x)$ is some sense.

PROOF. We first observe that f is a probability measure, since it is a non-negative (Schwartz) distribution normalized to total mass 1. Laird (1978) has shown that $\tilde{f}(\theta)$, the MLE, is self consistent, i.e. is of the form

$$\frac{1}{n} \sum_{i=1}^{n} \frac{e^{-\lambda|x_i - \theta|} \tilde{f}(\theta)}{\int e^{-\lambda|x_i - \phi|} \tilde{f}(\phi)\, d\phi} = \tilde{f}(\theta). \tag{3.3}$$

We observe that $e^{-\lambda|\theta|}$ satisfies the differential equation

$$(D^2 - \lambda^2) e^{-\lambda|\theta|} = -2\lambda\delta(\theta). \tag{3.4}$$

Hence if $\hat{f}(\theta) > 0$ in an interval, say (a, b), then we may divide both sides of (3.3) by $\hat{f}(\theta)$ and then operate on both sides with D^2 to obtain, for $\theta \in (a, b)$

$$D^2 \frac{1}{n} \sum_{i=1}^{n} \frac{e^{-\lambda|x_i - \theta|}}{h_\lambda(x_i)} = \frac{-2\lambda}{n} \sum_{i=1}^{n} \frac{\delta(x_i - \theta)}{h_\lambda(x_i)} + \frac{\lambda^2}{n} \sum_{i=1}^{n} \frac{e^{-\lambda|x_i - \theta|}}{h_\lambda(x_i)} = 0 \tag{3.5}$$

where $h_\lambda(x_i) = \int \exp(-\lambda \mid x_i - \theta \mid) \hat{f}(\theta)\, d\theta$. This is clearly a contradiction, since for $\theta \neq x_i$, the left side of (3.5) is positive. Hence $f(\theta)$ cannot be nonzero everywhere in an interval. The same argument (with a modification of the definition of derivative) shows that $\hat{f}$ cannot be nonzero in a set which includes a limit point. Hence $\hat{f}$ must be discrete.

Moreover, $\hat{f}$ has support in the convex hull C of $x_1, x_2, \cdots, x_n$. If $\hat{f}$ placed positive probability at a point θ_k not in C, then

$$L_\lambda(f) = \prod_i (\lambda/2)(\sum_{j \neq k} p_j e^{-\lambda|x_i - \theta_j|} + p_k e^{-\lambda|x_i - \theta_k|})$$

could be increased by moving θ_k toward the set C while leaving the other θ_j and p_j, the same. This would make $\exp(-\lambda \mid x_i - \theta_k \mid)$ larger for each i, but would not affect the other terms nor p_k. Hence $\hat{f}$ has the form

$$\hat{f}(\theta) = \sum_{j=1}^{m} p_j \delta(\theta - \theta_j). \tag{3.6}$$

The problem may now be treated as a parametric problem with parameters

$\theta_1, \cdots, \theta_m, p_1, \cdots, p_{m-1}$ and (2.9) becomes

$$L_\lambda(f) = L_\lambda(\theta, \mathbf{p}) = \prod_{i=1}^{n} (\lambda/2) \sum_{j=1}^{m} p_j e^{-\lambda|x_i-\theta_j|}. \tag{3.7}$$

Possible critical points of L_λ with respect to θ_k are at $\theta_k = x_1, x_2, \cdots, x_n$ since the derivative fails to exist at those points. The derivative may be zero at intermediate points, but the second derivative (logarithmic) in any interval excluding $x_1, x_2, \cdots, x_n$ is easily shown to be positive. Hence the only possible maxima are at the points $x_1, x_2, \cdots, x_n$, and $\tilde{f}(\theta)$ must have the form

$$\hat{f}(\theta) = \sum_{j=1}^{n} p_j \delta(\theta - x_j). \tag{3.8}$$

The p_j in turn may be found by substituting (3.8) in (3.3) to obtain

$$(1/n) \sum_{i=1}^{n} e^{-\lambda|x_i-x_j|} p_j / h_\lambda(x_i) = p_j. \tag{3.9}$$

This equation has a number of solutions, for example $p_1 = 1, p_2 = \cdots = p_n = 0$. However the only solution associated with the MLE has all positive components for $\lambda \geq m^{-1}$. To see this, we return to the log likelihood function, $\log L_\lambda$, whose derivative with respect to p_k has a local extremum at the points satisfying (3.2) and hence (3.9). The second derivative with respect to p_k is always negative. Hence a local solution is global and we need merely show that (3.2) has for such λ, a solution with all positive components. This and the last conclusion of the theorem follows from

LEMMA 3.2. *The solution to* (3.2) *is given by*

$$p_i = (1/n) + (1/n)\, \Delta e_i, \quad i = 1, 2, \cdots, n \tag{3.10}$$

where

$$e_i = \frac{1 + C_i}{1 - C_i}\left[\frac{1}{1 - C_i C_{i-1}} - \frac{1}{1 - C_{i+1} C_i}\right] \tag{3.11}$$

and where

$$C_i = e^{\lambda(X_{(i)} - X_{(i+1)})}, \quad i = 1, \cdots, n-1;$$

$X_{(1)}, X_{(2)} \cdots X_{(n)}$, *the order statistics,*

$$C_{-1} = C_0 = C_n = C_{n+1} = 0.$$

PROOF. Let $X_{(1)}, \cdots, X_{(n)}$ be the order statistics; then we have

$$\begin{aligned} e^{-\lambda|X_{(i)} - X_{(j)}|} &= e^{\lambda(X_{(i)} - X_{(i+1)})} \cdot e^{\lambda(X_{(i+1)} - X_{(i+2)})} \cdots e^{\lambda(X_{(j-1)} - X_{(j)})} \\ &= C_i \cdot C_{i+1} \cdots C_{j-1}; \quad i < j. \end{aligned} \tag{3.12}$$

Equation (3.2) may be expressed as

$$A\begin{bmatrix}1\\ \overline{A\mathbf{p}}\end{bmatrix} = n\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix} \tag{3.13}$$

in terms of matrices and vectors where A is the matrix

$$A = [e^{-\lambda|X_{(i)}-X_{(j)}|}]$$

and $p_1, p_2, \cdots, p_n$ have the same order as the order statistics. Then

$$A = \begin{bmatrix} 1 & C_1 & C_1C_2 & \cdots & C_1C_2 \cdots C_{n-1} \\ C_1 & 1 & C_2 & \cdots & C_2C_3 \cdots C_{n-1} \\ \cdots & & & & \cdots \\ C_1C_2 \cdots C_{n-1} & \cdots & \cdots & \cdots & 1 \end{bmatrix}. \tag{3.14}$$

The reciprocal of the vector in (3.13) denotes the vector of reciprocals. Since each C_i satisfies $0 < C_i < 1$ almost surely, the matrix A is invertible and (3.13) may be solved for $\mathbf{p}$

$$\mathbf{p} = A^{-1}\begin{bmatrix}1\\ \overline{nA^{-1}\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}}\end{bmatrix}. \tag{3.15}$$

Indeed, by means of elementary row operations, A^{-1} may be shown to be

$$A^{-1} = \begin{bmatrix} \frac{1}{1-C_1^2} & \frac{-C_1}{1-C_1^2} & & \cdots & & & 0 \\ \frac{-C_1}{1-C_1^2} & \frac{1}{1-C_2^2}+\frac{C_1^2}{1-C_1^2} & & \cdots & & & 0 \\ \cdots & & & & & & \\ 0 & 0 \cdots 0 & & \frac{-C_{i-1}}{1-C_{i-1}^2} & \frac{1-C_{i-1}^2C_i^2}{(1-C_{i-1}^2)(1-C_i^2)} & \frac{-C_i}{1-C_i^2} & 0 \cdots 0 \\ \cdots & & & & & & \\ 0 & 0 & & \cdots & & & \frac{1}{1-C_{n-1}^2} \end{bmatrix} \tag{3.16}$$

Hence we find

$$nA^{-1}\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix} = n\begin{bmatrix}1/(1+C_1)\\ \vdots\\ (1-C_{i-1}C_i)/(1+C_{i-1})(1+C_i)\\ \vdots\\ 1/(1+C_{n-1})\end{bmatrix}$$

which may be inverted and multiplied by A^{-1} again to obtain

$$\mathbf{p} = A^{-1}\begin{bmatrix} \dfrac{1}{nA^{-1}\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}} \end{bmatrix}$$

$$= \frac{1}{n}\begin{bmatrix} 1 + \dfrac{1+C_1}{1-C_1}\left\{1 - \dfrac{1}{1-C_1C_2}\right\} \\ \cdots \\ 1 - \dfrac{1+C_{i-1}}{1-C_{i-1}}\left\{\dfrac{1}{1-C_{i-1}C_{i-2}} - \dfrac{1}{1-C_iC_{i-1}}\right\} + \dfrac{1+C_i}{1-C_i}\left\{\dfrac{1}{1-C_iC_{i-1}} - \dfrac{1}{1-C_iC_{i+1}}\right\} \\ \cdots \\ 1 - \dfrac{1+C_{n-1}}{1-C_{n-1}}\left\{\dfrac{1}{1-C_{n-2}C_{n-1}} - 1\right\} \end{bmatrix} \tag{3.17}$$

$$= \frac{1}{n}\begin{bmatrix} 1+\Delta e_1 \\ \cdots \\ 1+\Delta e_i \\ \cdots \\ 1+\Delta e_n \end{bmatrix}$$

COROLLARY 3.3. *The solution to* (3.2) *satisfies*

(i) $p_i > 0$, $i = 1, 2, \cdots n$, $\lambda \geq m^{-1}$
(ii) $p_i = p_i(\lambda) \to (1/n)$ *as* $\lambda \to \infty$ a.s.

We first observe that for $\lambda m \geq 1$, we have

$$C_i = e^{-\lambda(X_{(i+1)}-X_{(i)})} \leq e^{-\lambda m} \leq e^{-1}. \tag{3.18}$$

Hence e_i satisfies

$$|e_i| = \left| \frac{1+C_i}{1-C_i}\left\{\frac{C_i(C_{i-1}C_{i+1})}{(1-C_iC_{i-1})(1-C_iC_{i+1})}\right\} \right|$$

$$\leq e^{-2\lambda m}\frac{1+e^{-\lambda m}}{1-e^{-\lambda m}}\frac{1}{(1-e^{-2\lambda m})^2} \leq e^{-2\lambda m}\frac{1+e^{-1}}{1-e^{-1}}\frac{1}{(1-e^{-2})^2} \tag{3.19}$$

$$\leq 3e^{-2\lambda m} < .5$$

and therefore $|\Delta e_i| = |e_i - e_{i-1}| < 1$ and $\Delta e_i \to 0$ as $\lambda \to \infty$. Thus both conclusions hold and the proof of the theorem is complete.

REMARK 3.2. As a consequence of this corollary we see that our MLE

$$\hat{f}(\theta) = \sum_{i=1}^n p_i\delta(\theta - x_i) \to (1/n)\sum_{i=1}^n \delta(\theta - x_i) = f^*(\theta) \quad \text{as} \quad \lambda \to \infty$$

where f^* is the (derivative of the) empiric distribution function. Similarly for $\hat{\alpha}(x)$, the element of the sieve S_n, we see that it is approximately given by the

kernel estimator

$$\hat{\beta}(x) = (\lambda/2) \sum (1/n) e^{-\lambda|x-x_i|}$$

for λ sufficiently large. In fact the difference can be shown to converge to zero weakly as $\lambda\ m \to \infty$. The kernel estimator in turn converges to the underlying density.

THEOREM 3.4. *Let $X_1, X_2, \cdots, X_n$ be an* iid. *sample with density $f(x) \in H^p$, $p \geq 0$, then for each $\varepsilon > 0$, $s \leq p - 1$, $s < 3/2$*

$$E \| \hat{\beta} - f \|_s^2 \leq c_1 \lambda^{1+\varepsilon+2s)}/n + c_2 \lambda^{-1}$$

where c_1 and c_2 are constants.

The proof is similar to others involving kernel estimators and will be omitted.

REMARK 3.3. The case of the L^2 norm corresponds to $s = 0$. In this case the hypothesis is satisfied for $f \in H^1$, i.e. density in L^2 whose derivative is also in L^2.

REFERENCES

BLUM, J., SUSARLA, V., and WALTER, G. (1980). Estimation of the prior distribution using differential operators. *Colloquia Math. Soc. J. Bolyai* **32** (Non Parametric Inference), 57–75.

GEMAN, S. and HWANG, C-R. (1982). Non-parametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

GRENANDER, V. (1981). *Abstract Inference.* Wiley, New York.

HASSELBLAD, V. (1969). Estimation of finite mixtures of distributions for the exponential family. *J. Amer. Statist. Assoc.* **64** 1459–1471.

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

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

RUDIN, W. (1973). *Functional Analysis.* McGraw, New York.

TAPIA, R. N. and THOMPSON, J. R. (1978). *Non-parametric Probability Density Estimation.* Johns Hopkins, Baltimore.

WALTER, G. and BLUM, J. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328–340.

GILBERT G. WALTER
THE UNIVERSITY OF WISCONSIN-MILWAUKEE
DEPARTMENT OF MATHEMATICAL SCIENCES
P.O. BOX 413
MILWAUKEE, WISCONSIN 53201