Characterization of a maximum-likelihood nonparametric density estimator of kernel type

by

Stuart Geman and Donald E. McClure

Reports in Pattern Analysis No. 114

Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

March 1982

Research supported in part by the Department of the Army under contract DAAG-80-K-0006 and by the Air Force Office of Scientific Research through grant no. 78-3514 to Brown University.

1. Introduction

As an instance of Grenander's method of sieves [2] for adapting the maximum-likelihood approach to settings where the target parameter is infinite dimensional, we have considered density functions of the form

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma} \phi((x-y)/\sigma) G(dy) = (\phi_\sigma * G)(x). \tag{1}$$

Here G is an arbitrary cdf and ϕ is the standard normal density function. In this note, we shall derive a characterization of the cdf G* that solve the maximum-likelihood equation:

$$\mathscr{L}(G^*) = \max_G \mathscr{L}(G) \tag{2}$$

where $\mathscr{L}(G)$ is the likelihood function

$$\mathscr{L}(G) = \prod_{i=1}^{n} f(x_i) \tag{3}$$

determined by a random sample $x_1, x_2, \ldots, x_n$ from an unknown population density f_0.

Geman and Hwang [1] have described the connection between this optimization problem and nonparametric maximum-likelihood estimation. In brief, if we specify a sequence $\{\sigma_m\}_{m=1}^{\infty}$ of positive values with $\sigma_m \downarrow 0$ as $m \to \infty$, then the sequence of sets

$$S_m = \{f : f = \phi_{\sigma_m} * G, \quad G \text{ an arbitrary cdf}\}$$

defines a sieve of subsets of L_1, the so-called convolution sieve. The method-of-sieves (i) fixes an index m, depending on sample size n and on the sequence $\{\sigma_m\}$, (ii) seeks the solution G_m^* of (2) determined by the sample $\{x_i\}_{i=1}^n$ and σ_m, and (iii) forms the estimator $f_m^* = \phi_{\sigma_m} * G_m^*$.

The familiar Parzen-Rosenblatt kernel estimator fits within this framework. The kernel estimator prescribes G to be the empirical cdf. One motivation for introducing the convolution sieve is to study the relationship between the kernel estimator and ones derived through the principle of maximum likelihood.

Our characterization theorem for G* exhibits a rather close relationship between f_m^* and the kernel estimator based on the Gaussian kernel. We shall show that the solution G* of (2) is a discrete cdf and that it contains no more than n points in its support. Thus, the estimator f_m^* obtained from the method-of-sieves admits a representation of the form

$$f_m^*(x) = \sum_{j=1}^{q} p_j \, \phi_{\sigma_m}(x-y_j),$$

analogous to a familiar form of the kernel estimator. In contrast to the kernel estimator, the support $\{y_j\}$ of G* does not coincide with the sample $\{x_i\}_{i=1}^n$ and, in general, the weights $\{x_j\}$ will not be identically equal to n^{-1}. Computational experiments with closely related sieves strongly indicate that the number q of points in the support of G* will typically be much smaller than sample size n.

2. Characterization Theorem

Theorem. Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with density f_0. Let $\sigma > 0$ and consider estimators f of f_0 defined by (1).

(i) There exists a solution G^* of the maximum-likelihood problem (2)-(3).

(ii) If G^* satisfies (2), then G^* is a discrete cdf with finite support. Denote $\mathrm{supp}(G) = \{s_j\}_{j=1}^q$. Then $q \leq n$.

(iii) $x_{(1)} = \min(\{x_i\}_{i=1}^n) < \max(\{x_i\}_{i=1}^n) = x_{(n)}$,

then $x_{(1)} < \min(\{s_j\}_{j=1}^q)$ and $\max(\{s_j\}_{j=1}^q) < x_{(n)}$.

Proof: We may assume, for convenience and without loss of generality, that $\sigma=1$. The sample values can be rescaled, setting $\hat{x}_i = x_i/\sigma$, if $\sigma \neq 1$.

The maximum of $\mathscr{L}(G)$, if it exists, will be attained by a cdf with support in $[x_{(1)}, x_{(n)}]$. To see this, consider an arbitrary right-continuous cdf G and defined G_0 in terms of G by

$$G_0((-\infty,x]) = \begin{cases} 0 & , \text{ for } x < x_{(1)} \\ G((-\infty,x]), & \text{ for } x_{(1)} \leq x < x_{(n)} \\ 1 & , \text{ for } x_{(n)} \leq x. \end{cases}$$

G_0 is designed so that $G_0(\{x_{(1)}\}) = G((-\infty, x_{(1)}])$ and $G_0(\{x_{(n)}\}) = G([x_{(n)}, \infty))$. Since ϕ is monotone on the separate intervals $(-\infty,0]$ and $[0,\infty)$, we have

$$\phi(x_i - x_{(n)}) G_0(\{x_{(n)}\}) \geq \int_{x_{(n)}-0}^{\infty} \phi(x-y) G(dy) \quad \text{and}$$

$$\phi(x_i - x_{(1)}) G_0(\{x_{(1)}\}) \geq \int_{-\infty}^{x_{(1)}+0} \phi(x-y) G(dy).$$

Consequently $(\phi * G_0)(x) \geq (\phi * G)(x)$ for all x in $[x_{(1)}, x_{(n)}]$ and hence $\mathscr{L}(G_0) \geq \mathscr{L}(G)$.

The existence of a solution G* of (2) follows from (i) the compactness of the (tight) family of cdfs having support in $[x_{(1)}, x_{(n)}]$, and (ii) the observation that $\mathscr{L}(G)$ is a bounded and continuous functional on this set of cdfs, i.e. continuous with respect to the topology of weak convergence.

Let G* be a solution of (2) and set $f^* = (\phi * G^*)$. A variational argument characterizes the points in the support of G* as roots of a transcendental equation. Let s be an arbitrary point in the support of G*. For any $\varepsilon > 0$ and for any z, define a measure $H_{s,\varepsilon,z}$ by

$$H_{s,\varepsilon,z}(B) = G^*((s-\varepsilon, s+\varepsilon] \cap (B-z))$$

$H_{s,\varepsilon,z}$ is a rigid shift through distance z of G* restricted to $(s-\varepsilon, s+\varepsilon]$. Define $G^*_{s,\varepsilon} = G^* - H_{s,\varepsilon,0}$. Then $G^*_{s,\varepsilon} + H_{s,\varepsilon,z}$ is a cdf for any z, and it may be regarded as a local perturbation near s of G*.

Set $f_{s,\varepsilon,z} = \phi * [G^*_{s,\varepsilon} + H_{s,\varepsilon,z}]$ and observe that $f^* = f_{s,\varepsilon,0}$. Since $\Pi f^*(x_i)$ is maximal, we have

$$0 = \frac{d}{dz} \sum_{i=1}^{n} \log f_{s,\varepsilon,z}(x_i) \Big|_{z=0} .$$

Evaluation of the derivative gives

$$0 = \sum_{i=1}^{n} \frac{1}{f^*(x_i)} \frac{d}{dz} (\phi * H_{s,\varepsilon,z})(x_i)\Big|_{z=0}$$

$$= \sum_{i=1}^{n} \frac{1}{f^*(x_i)} \frac{d}{dz} \int_{s-\varepsilon}^{s+\varepsilon} \phi(x_i - y - z) G^*(dy)\Big|_{z=0}$$

$$= \sum_{i=1}^{n} \frac{1}{f^*(x_i)} \int_{s-\varepsilon}^{s+\varepsilon} (x_i - y)\phi(x_i - y) G^*(dy).$$

Dividing this expression by $G^*((s-\varepsilon, s+\varepsilon])$ and letting $\varepsilon \to 0$ yields

$$\sum_{i=1}^{n} \frac{(x_i - s)}{f^*(x_i)} \phi(x_i - s) = 0,$$

for any s in the support of G^*.

Now consider the function

$$T(y) = \sum_{i=1}^{n} \frac{(x_i - y)}{f^*(x_i)} \phi(x_i - y).$$

The support of G^* is a subset of the set of roots of T. Properties of this set follow from the connection of T with an extended Tchebycheff system. We can re-express T as

$$T(y) = \frac{e^{-y^2/2}}{\sqrt{2\pi}} \sum_{i=1}^{n} [x_i e^{-x_i^2/2} e^{x_i y} - e^{-x_i^2/2} y e^{x_i y}]$$

$$\doteq e^{-y^2/2} \left\{ \sum_{i=1}^{n} (a_i e^{x_i y} + b_i \, y e^{x_i y}) \right\}.$$

The expression in braces is a simple linear combination of the 2n functions $\left\{ e^{x_i y}, \; y e^{x_i y} \right\}_{i=1}^{n}$. When the x_i's are distinct, this set is an extended Tchebycheff system of order 2n. (And of course if $\{x_i\}_{i=1}^{n}$ is a random sample from population density f_0, then the x_i's are distinct w.p.1. If the x_i's were not distinct, we could reduce the order of the system accordingly to express T(y) in terms of an extended Tchebycheff system with fewer than 2n elements.) The Tchebycheff property implies:

(i) $Z^0 = \{y : T(y)=0\}$ has at most 2n-1 elements, and
(ii) $Z^{+-} = \{y : T(y)=0,\ T'(y) \leq 0\}$ has at most n elements
(Karlin and Studden [3]).

Since the support of G* is contained in Z^0, G* is discrete with at most 2n-1 jumps.

In order to show that G* has at most n jumps, it suffices to show that the support of G* is actually contained in Z^{+-}, i.e. that $T'(s) \leq 0$ for any s in the support of G*. For f*, we can now write

$$f^*(x) = \sum_{j=1}^{q} p_j\ \phi(x-s_j)$$

where $\{s_j\}_{j=1}^{q}$ is the support of G*, $q \leq 2n-1$, $p_j > 0$, and $\sum_1^q p_j = 1$. Set $s=s_\ell$, for fixed ℓ between 1 and q. Let $\varepsilon > 0$ and define a perturbation f_ε of f* by

$$f_\varepsilon(x) = \sum_{j\neq\ell} p_j\phi(x-s_j) + \frac{p_\ell}{2}\ \phi(x-s+\varepsilon) + \frac{p_\ell}{2}\ \phi(x-s-\varepsilon).$$

The density f_ε admits a representation of the form (1) and $f^* = f_0$. Since $\Pi f^*(x_i)$ is maximal,

$$\frac{d^2}{d\varepsilon^2} \sum_{i=1}^{n} \log f_\varepsilon(x_i)\Big|_{\varepsilon=0} \leq 0.$$

Straightforward calculation yields

$$\frac{d^2}{d\varepsilon^2} \sum_{i=1}^{n} \log f_\varepsilon(x_i)\Big|_{\varepsilon=0} = p_\ell \; T'(s),$$

and hence, as claimed, $T'(s) \leq 0$.

Finally, to confirm the last statement in the theorem, observe that if $s \leq x_{(1)}$ for some s in the support of G*, then $\phi(x_i - s)$ is strictly increasing for sufficiently small increases in s and for all x_i, except perhaps $x_{(1)}$. Further, $\frac{d}{ds}\,\phi(x_{(1)} - s) \geq 0$ as long as $s \leq x_{(1)}$; hence $\Pi f^*(x_i)$ is a strictly increasing function of s, contradicting the maximum-likelihood property of G* and f*. The same reasoning precludes $s > x_{(n)}$. □

3. Concluding Remarks

The characterization theorem was announced in the paper by Geman and Hwang [1], where consistency questions for f^* are analyzed. The consistency results guarantee that $f^* \to f_0$ in L_1-norm, with probability one, provided that $\sigma \to 0$ sufficiently slowly as sample size $n \to \infty$.

H. Robbins recently restimulated interest in the maximum-likelihood problem per se during his lecture at the NASA Workshop on Density Estimation and Function Smoothing at Texas A&M University, March 11-13, 1982. Professor Robbins recalled his 1950 formulation of the maximum-likelihood problem (1)-(3) in [4] wherein connections are made with statistical decision problems.

References

[1] S. Geman and C-R. Hwang, Nonparametric maximum-likelihood estimation by the method of sieves, to appear in Ann. Statist.

[2] U. Grenander, Abstract Inference, John Wiley & Sons, New York (1981).

[3] S. Karlin and W.J. Studden, Tchebycheff systems: with applications in analysis and statistics, Interscience, John Wiley & Sons, New York (1966).

[4] H. Robbins, A generalization of the method of maximum likelihood: estimating a mixing distribution (abstract), Ann. Math. Statist. 21 (1950), 314-15.