*The Annals of Statistics*
1992, Vol. 20, No. 2, 929–943

# SEMIPARAMETRIC ESTIMATION OF NORMAL MIXTURE DENSITIES[1]

BY KATHRYN ROEDER

*Yale University*

A semiparametric method for estimating densities of normal mean mixtures is presented. This consistent data-driven method of estimation is based on probability spacings. The estimation technique involves iteratively fixing the standard deviation of the normal kernel that serves as a smoothing parameter, and then maximizing a function of the probability spacings over all mixing distributions. Based on the distribution of uniform spacings, a distribution free goodness-of-fit criterion is developed to guide the selection of the smoothing parameter. The result is a set of consistent estimators indexed by a range of smoothing parameters. Empirical process results are used to prove consistency.

**1. Introduction.** In this article we will be concerned with estimating densities of normal mean mixtures of the form

$$f_0(x) = f(x; Q_0, h_0) = \int \left(2\pi h_0^2\right)^{-1/2} \exp\left(-(x-\theta)^2/2h_0^2\right) Q_0(d\theta). \tag{1.1}$$

Assume that both the mixing distribution $Q_0$ and the structural parameter $h_0$ are unknown. Because the model is not identifiable, we focus attention on estimating the marginal density $f_0$. Nevertheless, to estimate $f_0$, one might consider maximizing the likelihood $\prod f(x_i; Q, h)$ over $Q$ and $h$. The maximum, however, cannot be achieved: The likelihood approaches infinity when $Q = F_n$ and $h \to 0$ ($F_n$ denotes the empirical distribution function). Consequently, in order to estimate $f_0$, the parameter space must be restricted. The class of distributions can be restricted by either forcing $Q$ to be discrete with a restricted number of support points or by bounding $h$ from below.

A subset of (1.1) is the class of finite mixture models. Let $\mathscr{Q}_r$ be the set of probability measures supported by $r$ or fewer points. If $Q_0$ happens to lie in $\mathscr{Q}_r$, for a speci<sup></sup>fied $r$, then consistent estimators can be obtained for $Q_0$ and $h_0$. For regularity conditions that lead to consistent asymptotically normal estimators for this model, see Peters and Walker (1978) and Redner and Walker (1984). No satisfactory data-based technique exists, however, by which to choose $r$ [for discussion see Titterington, Smith and Makov (1985), Chapter 5, or McLachlan and Basford (1988), Section 1.10]. Another disadvantage to this approach is that the likelihood may be multimodal, making it impossible to identify the consistent sequence of roots.

Received March 1990; revised June 1991.

[1] Research partially supported by NSF Grant DMS-90-01421.

*AMS* 1980 *subject classifications*. Primary 62G05; secondary 62G30, 62E20.

*Key words and phrases*. Confidence set of densities, normal mixtures, semiparametric density estimation, spacings.

To circumvent the problem of a possibly incorrect choice of $r$, Geman and Hwang (1982) applied Grenander's method of sieves by constraining $Q$ to lie in a family $\mathscr{Q}_{r_n}$, with $r_n$ tending to infinity with $n$. If $r_n = O(n^{\varepsilon})$ for some $\varepsilon < 1/5$, the estimator is consistent provided, the unknown density is bounded with compact support. In practice the method suffers from the defect that the estimator depends strongly on the particular choice of $r_n$.

Alternatively, one could restrict $h$. This approach is more appealing because, for a given $h$, there is a unique mixing distribution which maximizes the likelihood (Lindsay, 1983a, b). Changing $h$ changes the mixing distribution needed to provide the "best fit" to the data. A sieve can also be constructed based on the size of $h$ (Geman and Hwang, 1982). Let $\mathscr{F}_h = \{f(\cdot\,; Q, h)$: $Q$ is a probability measure$\}$. If $h_1 < h_2$, any normal mixture with variance $h_2^2$ can be constructed from a normal mixture of variance $h_1^2$ convolved with a $N(0, h_2^2 - h_1^2)$. Therefore, $\mathscr{F}_{h_2} \subset \mathscr{F}_{h_1}$ and this sieve increases in size as $h_n$ decreases. Again, the method of sieves leads to a consistent estimator of $f_0$ under certain conditions, provided $h_n \to 0$ at the appropriate rate; however, no data-based method exists by which to choose $h_n$.

Our approach is similar to Geman and Hwang's in that we will choose a normal mixture from the second type of sieve. Our sieve size, however, will be determined by a data-based selection of $h_n$. In addition, although our density estimate is consistent under certain conditions when $h_n \to 0$, because our goal is to estimate normal mixtures with $h_0 > 0$ it is not necessary that $h_n \to 0$. By taking this approach we lose flexibility, but it is hoped that our estimator will converge faster within the class of normal mixtures than a nonparametric estimator such as Geman and Hwang's.

Previous treatments of finite mixture models usually allow both the mean and variance of the components of the mixture to vary. For our method, the componentwise variance is taken to be constant for two reasons. First, this assumption is useful for mathematical tractibility. Second, because the number of components in the mixture is not prespecified, a mixture with differing variances can be approximated by a mixture with constant variance provided $h^2$ is selected to be less than or equal to the minimum variance in the true mixture.

The method, which we dub the method of spacings, is defined as follows. Let $I_k$ denote the interval between the $k$th and $(k + 1)$st order statistics. For a given continuous distribution function $F$, let $F(I_k)$ denote the probability measure of this interval. Let $F(\cdot\,; Q, h)$ be the distribution with density $f(\cdot\,; Q, h)$ and let

$$\text{(1.2)} \qquad \operatorname{LPS}(Q, h) = \sum_{k=1}^{n-1} \frac{\log F(I_k; Q, h) - \mu_n}{\sigma}$$

denote the log product spacings function evaluated at $F(\cdot\,; Q, h)$, where

$$\mu_n = -(\log(n+1) + \gamma), \qquad \sigma^2 = (\pi^2/6) - 1$$

and $\gamma$ is Euler's constant. Rather than maximize the likelihood, maximize $\mathrm{LPS}(Q, h)$ over $Q$ for a fixed $h$. [For results concerning estimation of finite dimensional parameters using spacings, see Cheng and Amin (1983) and Ranneby (1984).] Clearly the maximization is not affected by $\mu_n$ and $\sigma$; however, these constants will be important in choosing $h_n$.

The set of probability spacings $\{F(I_k; Q_0, h_0)\}_{k=1}^{n-1}$ has the same distribution as a set of uniform spacings. The normalizing constants $\mu_n$ and $\sigma$ are chosen so that $n^{-1/2}\mathrm{LPS}(Q_0, h_0)$ is asymptotically distributed as a standard normal (Darling, 1953). Consequently, the spacings functional can serve as both an objective function for maximization and as a goodness-of-fit test. [For results concerning goodness-of-fit tests, see Cressie (1976).] It is in this sense only that the method of spacings offers advantages over the likelihood approach.

We utilize the sum of the log of the probability spacings rather than some other function of the spacings because it yields results asymptotically equivalent to those obtained from maximizing the likelihood. In addition, the spacings functional is convex in $Q$, as is the likelihood, which facilitates computation of the maximum (Lindsay, 1983a; Roeder, 1988).

For a fixed $h$, let $\hat{Q}_h$ denote the probability measure that maximizes $\mathrm{LPS}(Q, h)$. As with the likelihood approach, joint maximization over $Q$ and $h$ will lead to an inconsistent estimator. Instead, the final mixture density estimate $f(\cdot; \hat{Q}_{\hat{h}}, \hat{h})$ can be based on the distribution of $\mathrm{LPS}(Q_0, h_0)$ (i.e., the distribution of the log spacings of a sample of uniform spacings). For example, $\hat{h}$ can always be selected so that $\mathrm{LPS}(\hat{Q}_{\hat{h}}, \hat{h}) = 0$.

As a first step in motivating the goodness-of-fit component of the estimation scheme, we construct a nonparametric confidence set of continuous distribution functions. For any continuous distribution $F$, define the log product spacings function of an arbitrary distribution function as $\mathrm{LPS}^*(F) = \Sigma(\log F(I_k) - \mu_n)/\sigma$. Assuming that $\{X_i\}_1^n$ is a random sample from the continuous distribution $F_0$, the following probability statement holds for $n$ large: $P[|n^{-1/2}\mathrm{LPS}^*(F_0)| < z_{\alpha/2}] = 1 - \alpha$ where $z_{\alpha/2}$ denotes the upper $(\alpha/2)$ 100 percentile of the normal distribution. Based on this result, we can construct a test that rejects $F$ if $|n^{-1/2}\mathrm{LPS}^*(F)| > z_{\alpha/2}$. The inverse of this test yields an asymptotically nonparametric $(1 - \alpha)100\%$ confidence set for continuous distribution functions $\{F\colon |n^{-1/2}\mathrm{LPS}^*(F)| < z_{\alpha/2}\}$. If $F_0 = F(\cdot; Q_0, h_0)$, then the set of densities $\{f(\cdot; Q, h)\colon |n^{-1/2}\mathrm{LPS}(Q, h)| < z_{\alpha/2}\}$ forms a confidence set of densities with asymptotic coverage probability of at least $1 - \alpha$. As an aid in describing this set of densities, consider a graphical presentation of a smooth subset:

$$\mathscr{C}(\alpha) = \left\{f\left(\cdot; \hat{Q}_h, h\right) \colon \left|n^{-1/2}\mathrm{LPS}\left(\hat{Q}_h, h\right)\right| < z_{\alpha/2}\right\}.$$

This subset of the confidence set, which we dub the *profile confidence set*, is easy to obtain. Call $\mathrm{LPS}(\hat{Q}_h, h)$ the *profile function*, as it represents the fit of the model as a function of $h$ after we have maximized over $Q$. This is a decreasing function because normal mixture models are nested: $\mathscr{F}_{h_2} \subset \mathscr{F}_{h_1}$

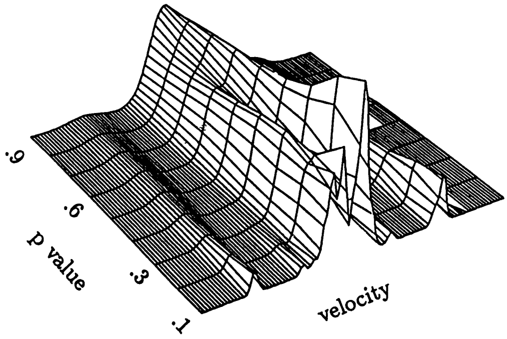

FIG. 1. *Profile confidence set* $\{f(\cdot\,;\hat{Q},h): h \in \mathscr{I}(0.20)\}$. *The data are the velocity at which each of 82 galaxies is moving away from our galaxy. For further analysis, see Roeder* (1990).

provided $h_1 \le h_2$. The monotonicity property of the profile function can be exploited to find the range of $h$ corresponding to $\mathscr{C}(\alpha)$. Let $\mathscr{I}(\alpha) = [h_1, h_2]$, where $h_1$ and $h_2$ solve $n^{-1/2}\mathrm{LPS}(\hat{Q}_h, h) = z_{\alpha/2}$ and $n^{-1/2}\mathrm{LPS}(\hat{Q}_h, h) = -z_{\alpha/2}$, respectively. It follows that $\mathscr{C}(\alpha) = \{f(\cdot\,;\hat{Q}_h, h): h \in \mathscr{I}(\alpha)\}$.

To illustrate the method of spacings, we find the profile confidence set of densities for a data set of some practical interest [see Roeder (1990) for a more detailed analysis]. The data consist of the estimated velocity at which each of 82 galaxies in the Corona Borealis region is moving away from our galaxy. Clusters correspond to large scale structures in the pattern of expansion. In Figure 1, we present $\mathscr{C}(0.20)$. With each $h^*$, we associate the value $p^*$, such that $n^{-1/2}\mathrm{LPS}(\hat{Q}_{h^*},h^*) = z_{p^*}$. The density $f(\cdot\,;\hat{Q}_{h^*}, h^*)$ has the largest window width of any density in the one-sided profile confidence set $\{f(\cdot\,;\hat{Q}_h, h): n^{-1/2}\mathrm{LPS}(\hat{Q}_h, h) > z_{p^*}\}$. When $F(\cdot\,;\hat{Q}_{h^*}, h^*)$ is viewed as the null hypothesis, $p^*$ is the $p$ value of the data. Estimates in the foreground of the figure correspond to density estimates with smaller $h$. Typically, density estimates with $p$ values near 0 overfit the data, while estimates with $p$ values near 1 underfit the data. In this example, the confidence set suggests that there are between three and seven clusters of galaxies in the Corona Borealis region. Notice that, contrary to the behavior exhibited by nonadaptive, nonparametric density estimators, this semiparametric estimator maintains a smooth estimate of the tail of the density while still fitting the pronounced modes in the center of the data.

The subject of this article is how to construct a data-based selection procedure for choosing $h_n$ that leads to a consistent density estimator. We show that consistent results can be obtained if $h_n$ is selected based on the size of $\mathrm{LPS}(\hat{Q}_{h_n}, h_n)$. In Section 2, we state a general theorem showing that if $\hat{h}_n(n/\log\log n)^{1/4} \to \infty$ and $n^{-1}\mathrm{LPS}(\hat{Q}_{\hat{h}_n}, \hat{h}_n) \to 0$ with probability 1, then $f(\cdot\,;Q_{\hat{h}_n}, \hat{h}_n)$ is a consistent estimator of $f_0$ (a.s. in $L_1$). In Section 3, we prove a uniform law of the iterated logarithm for normal mixture likelihoods using results from empirical processes. In Section 4, consistency is proved. We also derive in Section 4 a method by which $\hat{h}_n$ can be selected so that asymptotically, with probability 1, $\hat{h}_n$ is bounded below by $ph_0$, $0 < p < 1$, and still results in a consistent estimator. Simulations, which are presented in Section 2, suggest that substantial improvements in the density estimate can be gained if $h_n$ does not go to zero.

## 2. Major results.

We first state a general theorem concerning consistency and then discuss simulations supporting the conjecture that these estimators have improved rates of convergence relative to a nonparametric estimator. Define $\mathscr{B}_a$ as the set of probability measures with support in the interval $[-a, a]$.

THEOREM 2.1. *Suppose $f_0 = f(\cdot; Q_0, h_0)$ is a normal mixture density such that $h_0 > 0$ and $Q_0 \in \mathscr{B}_a$ for some $a < \infty$. Let $\hat{h}_n$ be a bounded sequence in $\mathbb{R}^+$ (possibly random). If*

$$\text{(i)} \qquad \hat{h}_n[n/\log\log n]^{1/4} \to \infty \quad a.s.,$$

$$\text{(ii)} \qquad n^{-1}\mathrm{LPS}\left(\hat{Q}_{\hat{h}_n}, \hat{h}_n\right) \to 0 \quad a.s.,$$

*then*

$$\int \left| f\left(\cdot; \hat{Q}_{\hat{h}_n}, \hat{h}_n\right) - f_0 \right| \to 0 \quad a.s.$$

The proof of this theorem is presented in the following sections. In Section 4.2 we show that a data-based criterion can be constructed such that $\hat{h}_n$ meets these two conditions.

Before embarking upon the proof of consistency, we present some preliminary results concerning the performance of the estimator. Because our goal is to estimate the density, and since the class of normal mixtures is very rich, a natural approach would be to use nonparametric density estimation. Nevertheless, the preliminary simulations are intriguing because they suggest we can attain improved rates of convergence. We generated 50 samples of size $n = 200$ from a bimodal normal mixture $[0.67N(0,1) + 0.33N(3,1)]$. For each sample we calculated the integrated squared error $[\mathrm{ISE}(f, f_0) = \int (f - f_0)^2]$ of the density estimate $f(\cdot; \hat{Q}_h, h)$ for $h \in [0.10, 1.75]$. In Figure 2, the average ISE from the 50 samples is presented. We also calculated the ISE for normal kernel estimates for the same data sets. On the average, the best method of spacings estimate (best $h$) is twice as good as the best kernel estimate. Moreover, the spacings estimator is better than the best kernel estimate for a broad range of $h$. Hence, when the unknown density is a normal mixture, there is an interval of smoothing parameters ($[\delta, h_0 + \varepsilon]$, $\delta > 0$, $\varepsilon > 0$) that usually provides an improved fit relative to the kernel estimator. Unfortunately, the method of spacings is presently computationally intensive and large scale simulations are not feasible.

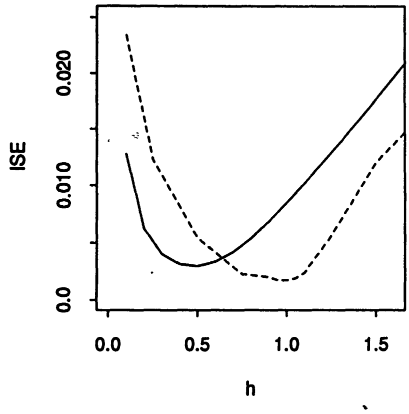

FIG. 2. *Plot of average* ISE $= \int(f(\cdot\,;\hat{Q},h) - f_0)^2$ *versus h. Fifty sample of size* 200 *were generated from the density* $0.67N(0,1) + 0.33N(3,1)$. *The smooth line is the* ISE *for the normal kernel estimator and the broken line is the* ISE *for the method of spacings estimate.*

**3. Behavior of the profile function.** In this section we prove a uniform law of the iterated logarithm (LIL) for normal mixture likelihoods. This result will be necessary in the proof of consistency.

Throughout the paper, we use linear functional notation whenever it can be used unambiguously. Thus we write $Pg$ for $\int g(x)P(dx)$. With a slight abuse of notation, we spell out simple functions; thus $P(x-\theta)^2 = \int(x-\theta)^2P(dx)$. We use curly brackets to denote indicator sets, so $P\{x > t\} = \int 1_{\{x>t\}}P(dx)$. When the function has two arguments, the convention is to indicate which one is to be averaged over in the following way: $Pf(\cdot, t) = \int f(x,t)P(dx)$. In the case of the simple functions above, unless otherwise specified, one can assume that $x$ is to be averaged over, which allows us to avoid unappealing notation such as $P\{\cdot < t\}$. We also use the convention that $Q$ averages over $\phi$.

Suppose $X_1, X_2, \dots, X_n$ are independent and identically distributed random variables from the distribution $P$ on the real line. Let $P_n$ denote the corresponding empirical measure, which puts mass $n^{-1}$ at each of the realizations $x_1, x_2, \dots, x_n$, and let $F_n(t) = n^{-1}\Sigma_1^n\{x_i \le t\} = P_n\{x \le t\}$ be the corresponding distribution function. In the proof of Theorem 3.2, we will use the following lemma concerning weighted empirical processes. Let $\nu_n = n^{1/2}(P_n - P)$ denote the rescaled empirical process. For a fixed $\varepsilon$, $0 < \varepsilon < 1/2$, let

$$\psi(s) = \begin{cases} [s(1-s)]^{-1/2+\varepsilon}, & \text{for } 0 < s < 1, \\ 0, & \text{elsewhere} \end{cases}$$

and let $b_n = (2\log\log n)^{1/2}$.

LEMMA 3.1. *There exists a universal constant M such that*

$$\limsup_n b_n^{-1} \sup_t |\nu_n\{x < t\}\psi(P\{x < t\})| \le M \quad a.s.$$

PROOF. This result is a consequence of a theorem due to James (1975). □

We now present the main result of this section: Uniform convergence of the log likelihood function over all mixing distributions with $Q \in \mathscr{B}_a$ and $h$ in a bounded interval.

THEOREM 3.2. *Suppose* $X_1, X_2, \dots, X_n$ *are independent, identically distributed random variables from a normal mixture* $f(\cdot\,; Q_0, h_0)$ *where* $h_0 > 0$ *and* $Q_0 \in \mathscr{B}_a$. *Then the corresponding empirical process satisfies*

$$\limsup_n \sup_{\substack{Q \in \mathscr{B}_a \\ 0<h<K}} h^2 |\nu_n \log f(\cdot\,; Q, h)| = O(b_n) \quad a.s.$$

PROOF. In the proof it will be understood that $Q$ is restricted to $\mathscr{B}_a$ and $h$ to $(0, K)$. Define $\theta_Q$ as the mean of $Q$. Let $f(\cdot\,; \phi, h)$ be the $N(\phi, h^2)$ density, which corresponds to $f(\cdot\,; Q, h)$ with $Q$ degenerate at $\phi$. Now

$$\begin{aligned}
&\sup_{Q,h} h^2 |\nu_n \log f(\cdot\,; Q, h)| \\
&\quad = \sup_{Q,h} h^2 \left| \nu_n \log\left( \frac{f(\cdot\,; Q, h)}{f(\cdot\,; \theta_Q, h)} \right) + \nu_n \log f(\cdot\,; \theta_Q, h) \right| \\
&\quad \le \sup_{Q,h} h^2 \left| \nu_n \log\left( \frac{f(\cdot\,; Q, h)}{f(\cdot\,; \theta_Q, h)} \right) \right| + \sup_{\phi \in [-a,a],\, h} h^2 |\nu_n \log f(\cdot\,; \phi, h)|.
\end{aligned}$$

We will deal separately with the two terms after the inequality. The second term equals

$$\tfrac{1}{2} \sup_{\phi \in [-a,a]} \left| \nu_n (x - \phi)^2 \right| = \tfrac{1}{2} \sup_{\phi \in [-a,a]} |\nu_n x^2 - 2\phi \nu_n x|.$$

Both $X^2$ and $X$ have finite means since $Q_0$ has bounded support. From the LIL for independent identically distributed random variables, we conclude that

$$\limsup_n \sup_{\phi \in [-a,a],\, h} h^2 |\nu_n \log f(\cdot\,; \phi, h)| = O(b_n) \quad \text{a.s.} \tag{3.1}$$

To handle the first term we will first show that $h^2 \log(f(x; Q, h)/f(x; \theta_Q, h))$ is convex in $x$, bounded below by $-Q(\phi - \theta_Q)^2$ and bounded above by $(x - \theta_Q)^2/2$. For convexity in $x$, notice that $h^{-2}$ times the second derivative with respect to $x$ equals

$$\frac{d^2}{dx^2} \log\left( \frac{f(x; Q, h)}{f(x; \theta_Q, h)} \right) = \frac{Q\phi^2 f(x; \phi, h)}{Qf(x; \cdot\,, h)} - \left[ \frac{Q\phi f(x; \phi, h)}{Qf(x; \cdot, h)} \right]^2,$$

which is the variance of the conditional distribution of $\phi$ given $x$. Without loss of generality we can assume strict convexity because otherwise $f(\cdot\,; Q, h) = f(\cdot\,; \theta_Q, h)$ (in this case the desired result is obviously true). Next, let $\alpha = \sup_Q Q(\phi - \theta_Q)^2$. Because $Q$ has compact support, $\alpha < \infty$. To obtain a lower bound of $-\alpha$, apply Jensen's inequality

$$h^2[\log f(x;Q,h) - \log f(x;\theta_Q,h)] \geq h^2[Q\log f(x;\cdot,h) - \log f(x;\theta_Q,h)]$$

$$= -\tfrac{1}{2}Q(x-\phi)^2 + \tfrac{1}{2}(x-\theta_Q)^2 \geq -\alpha$$

for all $Q$. The upper bound is easily established.

Let $l_{Q,h}(x) = \log(f(x;Q,h)/f(x;\theta_Q,h)) + \alpha$. Note that $l_{Q,h}(x)$ is a positive, strictly convex function. Consequently

$$\nu_n l_{Q,h} = n^{1/2}\int_0^\infty \big[P_n\{l_{Q,h}(x) > t\} - P\{l_{Q,h}(x) > t\}\big]\,dt. \tag{3.2}$$

By strict convexity and continuity, there exists an $R_{Q,h}(t) \geq L_{Q,h}(t)$ such that

$$\{l_{Q,h}(x) > t\} = \{x < L_{Q,h}(t)\} \cup \{x > R_{Q,h}(t)\}.$$

The right-hand side of equality (3.2) becomes

$$\int_0^\infty \big[\nu_n\{x < L_{Q,h}(t)\} + \nu_n\{x > R_{Q,h}(t)\}\big]\,dt.$$

We consider only the first integral; treatment of the second integral is similar. It is bounded in absolute value by

$$\int_0^\infty \big|\nu_n\{x < L_{Q,h}(t)\}\psi\big(P\{x < L_{Q,h}(t)\}\big)\big|\big[\psi\big(P\{x < L_{Q,h}(t)\}\big)\big]^{-1}\,dt.$$

From Lemma 3.1 it follows that

$$\sup_{Q,h,s}\big|\nu_n\{x < L_{Q,h}(s)\}\psi\big(P\{x < L_{Q,h}(s)\}\big)\big| = O(b_n) \quad \text{a.s.} \tag{3.3}$$

It remains only to show that

$$\sup_{Q,h} h^2\int_0^\infty \big[\psi\big(P\{x < L_{Q,h}(t)\}\big)\big]^{-1}\,dt < \infty. \tag{3.4}$$

By definition,

$$\{x < L_{Q,h}(s)\} \subseteq \{l_{Q,h}(x) > s\}.$$

Furthermore,

$$\{l_{Q,h}(x) > s\} \subseteq \left\{\alpha + \frac{1}{2h^2}(x-\theta_Q)^2 > s\right\}.$$

When $s > 2\alpha$, the last set is contained in $\{|x| > -|\theta_Q| + hs^{1/2}\}$. The left-hand side of (3.4) is therefore less than

$$\sup_{Q,h} h^2\int_0^\infty P\{x < L_{Q,h}(s)\}^{1/2-\varepsilon}\,ds$$

$$\leq \sup_h 2\alpha h^2 + \sup_h h^2\int_{2\alpha}^\infty P\{|x| > -a + hs^{1/2}\}^{1/2-\varepsilon}\,ds.$$

The change of variable $z = h^2 s$ gives the bound

$$\int_0^\infty P\{|x| > -a + z^{1/2}\}^{1/2-\varepsilon}\,dz.$$

The tail probabilities of the $P$ distribution are bounded by the sum of tail probabilities for the $N(a, h_0^2)$ and $N(-a, h_0^2)$ distributions. Exponential decrease of normal tails therefore ensures finiteness of the last integral. Then the assertion of the theorem follows directly. □

COROLLARY 3.3. *If $\hat{h}_n$ is a sequence of (possible random) bandwidths for which eventually $\hat{h}_n < K$ a.s., and $\hat{h}_n(n/\log\log n)^{1/4} \to \infty$ a.s., then*

$$\sup_{Q \in \mathscr{B}_a} (P_n - P)\log f(\cdot\,; Q, \hat{h}_n) \to 0 \quad a.s.$$

PROOF. Use the fact that $n^{1/2}\hat{h}_n^2/b_n \to \infty$ a.s. and, from the theorem,

$$n^{1/2}\hat{h}_n^2 \sup_{Q \in \mathscr{B}_a} \left|(P_n - P)\log f(\cdot\,; Q, \hat{h}_n)\right| = O(b_n) \quad \text{a.s.} \qquad \square$$

## 4. Consistency proofs.

4.1. *General consistency results.* To prove consistency, we wish to show $f(\cdot\,; \hat{Q}_{\hat{h}_n}, \hat{h}_n) \to f_0$ a.s., in a suitable metric. By expanding the log product spacings function (1.2), we obtain a relationship between it and the log likelihood function. This reveals a natural metric for establishing consistency. Notice

$$\log F(I_k) = \log\int_{I_k} f(x)/f_0(x)\,dF_0(x), \tag{4.1}$$

which by the mean value theorem, for some $X_k^* \in I_k$, equals

$$\log F(I_k) = \log\frac{f(X_k^*)}{f_0(X_k^*)} + \log F_0(I_k) \approx \log\frac{f(X_{(k)})}{f_0(X_{(k)})} + \log F_0(I_k).$$

It follows that

$$\frac{\sigma}{n}\mathrm{LPS}(Q, h) \approx P_n \log\left(\frac{f(\cdot\,; Q, h)}{f_0}\right) + \frac{1}{n}\sum \log F_0(I_k) - \mu_n. \tag{4.2}$$

Let $\mathrm{KL}(f_1, f_0) = \int f_0 \log(f_1/f_0)$ denote the Kullback–Leibler information and $\widehat{\mathrm{KL}}(f(\cdot\,; Q, h), f_0)$ denote the first term on the right-hand side of (4.2), which is an empirical estimate of $\mathrm{KL}(f(\cdot\,; Q, h), f_0)$. It is apparent from (4.2) that Kullback–Leibler information generates a natural measure of fit. From Jensen's inequality, it follows that $\mathrm{KL}(f(\cdot\,; Q, h), f_0) \le 0$ for all $f(\cdot\,; Q, h)$ with equality if and only if $f(\cdot\,; Q, h) = f_0$ with probability 1. By choosing $h_n$ such that $n^{-1}\mathrm{LPS}(\hat{Q}_{h_n}, h_n) \to 0$, we hope to obtain a consistent estimator. We will formalize this heuristic argument in this section. Details pertaining to remainder terms will be relegated to the Appendix.

As an aside, note that if we choose $h_n^*$ to maximize $\mathrm{LPS}(\hat{Q}_h, h)$, then $\mathrm{KL}(f(\cdot\,; \hat{Q}_{h_n^*}, h_n^*),\ f_0)$ will converge to a positive constant; hence, the corresponding density $f(\cdot\,; \hat{Q}_{h_n^*}, h_n^*)$ cannot converge to a density without contradicting Jensen's inequality. The distribution maximizing LPS places equal probability in each interval.

Although the Kullback–Leibler information provides a natural measure of fit, we would like to establish that $f(\cdot\,; \hat{Q}_{\hat{h}_n}, \hat{h}_n)$ converges in a standard metric such as the $L_1$ norm. From Kemperman's (1969) inequality,

$$\int f_0 \log(f_0/f) \geq \frac{1}{2}\left(\int |f - f_0|\right)^2,$$

it is sufficient to establish convergence of the Kullback–Leibler information.

We can now prove consistency quite easily.

Proof of Theorem 2.1. Clearly

$$\begin{aligned}\left|P\log\left(f_0/f\left(\cdot\,;\hat{Q}_{\hat{h}_n},\hat{h}_n\right)\right)\right| &\leq \left|(P_n - P)\log f\left(\cdot\,;\hat{Q}_{\hat{h}_n},\hat{h}_n\right)\right| \\ &\quad + \left|P\log f_0 - P_n\log f\left(\cdot\,;\hat{Q}_{\hat{h}_n},\hat{h}_n\right)\right|.\end{aligned} \tag{4.3}$$

The first term on the right-hand side in (4.3) goes to zero a.s. by uniform convergence provided $\hat{h}_n[n/\log\log n]^{1/4} \to \infty$ a.s. (Corollary 3.3). Assumption (ii) means that $n^{-1}\sum \log F(I_k; \hat{Q}_{\hat{h}_n}, \hat{h}_n) - \mu_n \to 0$. The second term in (4.3) goes to zero provided

$$\begin{aligned}\lim_n n^{-1} \sum \log F\left(I_k; \hat{Q}_{\hat{h}_n}, \hat{h}_n\right) &- \mu_n \\ - P\log f_0 + P_n \log f\left(\cdot\,; \hat{Q}_{\hat{h}_n}, \hat{h}_n\right) &= 0 \quad \text{a.s.},\end{aligned}$$

but this holds by Theorem A.4 provided $\hat{h}_n n^2/\log n \to \infty$ a.s. (see Appendix). □

4.2. *Using the profile function to adjust $h$.* In this subsection we show how the profile function can be used to select the bandwidth. In particular, the bandwidth is selected so that, with probability 1, $\hat{h}_n$ is eventually greater than $ph_0$ for any preselected $p$, $0 < p \leq 1$. In addition, this sequence of random bandwidths meets the conditions of Theorem 2.1. Consequently, this result is substantially stronger than the asymptotic requirement for consistency of Theorem 2.1. (In Section 2 it was suggested that, as $h_n \to 0$, the ISE of the estimator tends to increase substantially. Hence it is useful to ensure $\hat{h}_n$ does not go to zero.)

As a preliminary step, we explore the behavior of the profile function. It is established that this function is nonincreasing for all $h$, and strictly decreasing for $h \geq h_0$ when $n$ is sufficiently large. This result provides a method for ensuring the randomly selected bandwidth $\hat{h}_n$ is asymptotically bounded below. We can do this by ensuring that $\mathrm{LPS}(\hat{Q}_{\hat{h}_n}, \hat{h}_n)$ is not too large.

THEOREM 4.1. *$\mathrm{LPS}(\hat{Q}_h, h)$ is nonincreasing in $h$; moreover, if $\mathrm{LPS}(\hat{Q}_{h_1}, h_1)$ is less than the global maximum, then $\mathrm{LPS}(\hat{Q}_{h_1}, h_1) > \mathrm{LPS}(\hat{Q}_{h_2}, h_2)$ for $h_2 > h_1$. In addition, provided $Q_0 \in \mathscr{B}_a$,*

$$\liminf_n n^{-1}\left[\mathrm{LPS}\left(\hat{Q}_{h_0}, h_0\right) - \mathrm{LPS}\left(\hat{Q}_{h_1}, h_1\right)\right] > 0 \quad a.s.$$

*for any $h_1 > h_0$.*

PROOF. Let $Q^*$ be the convolution of $\hat{Q}_{h_1}$ and a $N(0, h_1^2 - h_0^2)$. Clearly

$$\mathrm{LPS}\left(\hat{Q}_{h_1}, h_1\right) = \mathrm{LPS}(Q^*, h_0) \le \mathrm{LPS}\left(\hat{Q}_{h_0}, h_0\right).$$

Relying on the geometric results of Lindsay (1983a, b), we know that $\hat{Q}_{h_0}$ is unique and has a discrete number of support points provided $\mathrm{LPS}(\hat{Q}_{h_0}, h_0)$ has not achieved the global maximum $(n\gamma/\sigma)$. From this we conclude that the second inequality is strict, provided $n^{-1}\mathrm{LPS}(\hat{Q}_{h_0}, h_0) < \gamma/\sigma$. Clearly,

$$n^{-1}\mathrm{LPS}\left(\hat{Q}_{h_0}, h_0\right) \le n^{-1}\left|\mathrm{LPS}\left(\hat{Q}_{h_0}, h_0\right) - \mathrm{LPS}(Q_0, h_0)\right| + n^{-1}\left|\mathrm{LPS}(Q_0, h_0)\right|.$$

The second term converges to 0 a.s. (Lemma A.3, see Appendix). The first term converges to zero a.s. because (i) $\hat{Q}_{h_0}$ converges weakly to $Q_0$ a.s. (Roeder, 1988), (ii) $\sup_{Q \in \mathscr{B}_a}(P_n - P)\log f(\cdot; Q, h_0) \to 0$ a.s. (Corollary 3.3) and (iii) the remainder relating log likelihood to log spacings goes to zero a.s. (Theorem A.4). □

Consider a selection procedure where $h_n$ is chosen such that $\mathrm{LPS}(\hat{Q}_h, h) = d_n$. The maximum LPS can always be achieved by taking $h_n$ small enough that the probability spacings are equal. By continuity, for any $d_n$ less than the maximum, there exists a solution $F(\cdot; \hat{Q}_{h_n}, h_n)$ satisfying this requirement. In the following theorem, we use a selection procedure to choose $\hat{h}_n$ so that it is eventually bounded below by $h_0$ a.s. The remark following the theorem explains the practical implications of the result.

THEOREM 4.2. *Let $\hat{h}_n$ solve $\mathrm{LPS}(\hat{Q}_h, h) = -c(n \log n)^{1/2}$, $c > 0$. Then eventually $\hat{h}_n > h_0$ a.s.*

PROOF. By definition,

$$\mathrm{LPS}\left(\hat{Q}_{h_0}, h_0\right) \ge \mathrm{LPS}(Q_0, h_0).$$

Hence, by Lemma A.3,

$$n^{-1}\mathrm{LPS}\left(\hat{Q}_{h_0}, h_0\right) > -c\left(n^{-1}\log n\right)^{1/2}, \tag{4.4}$$

with probability 1, for $n$ large. If $\hat{h}_n$ is selected such that

$$n^{-1}\mathrm{LPS}\left(\hat{Q}_{\hat{h}_n}, \hat{h}_n\right) = -c\left(n^{-1}\log n\right)^{1/2},$$

it follows that

$$\mathrm{LPS}\left(\hat{Q}_{h_0}, h_0\right) > \mathrm{LPS}\left(\hat{Q}_{\hat{h}_n}, \hat{h}_n\right)$$

a.s. for $n$ large. By Theorem 4.1, $\mathrm{LPS}(\hat{Q}_h, h)$ is a decreasing function in $h$ when $h \geq h_0$; it follows that for $n$ large, $h_0 < \hat{h}_n$ a.s. □

Remark. Note that although $\hat{h}_n$ is asymptotically bounded below by $h_0$, it will generally provide an overly smooth estimate according to the spacings functional ($h$ too large; see Section 4.4). To remedy this problem, consider the following correction factor: For any $p$, $0 < p \leq 1$, the sequence of smoothing parameters $p\hat{h}_n$ is eventually bounded below by $ph_0$. Any sequence $\tilde{h}_n \geq p\hat{h}_n$, selected so that $n^{-1}\mathrm{LPS}(\hat{Q}_{\tilde{h}_n}, \tilde{h}_n) \to 0$, satisfies thè conditions of Theorem 2.1. In practice, it is desirable to obtain a selection of plausible estimates (e.g., the profile confidence set) as well as a point estimate of the density. Every $h \in \mathscr{I}(\alpha)$ meets condition (ii) of Theorem 2.1. A modest number of simulations [based on a preliminary estimate of $f_0$, say $f(\cdot; \hat{Q}_{h^{**}}, h^{**})$, where $h^{**}$ was chosen so that $\mathrm{LPS}(\hat{Q}_{h^{**}}, h^{**}) = 0$] could be used to determine a sensible choice of the arbitrary constant $p$. For example, because simulations suggest that good estimates can be obtained whenever $h \in [0.5h_0, h_0]$, one could use simulations to choose $p$ so that a lower bound of approximately $0.5h_0$ was obtained. Finally, any estimates in $\mathscr{E}(\alpha)$ with $h$ greater than the lower bound could be considered plausible. A point estimate could be based on a cross-validation procedure with $h$ restricted to the interval just obtained [see Roeder (1990)].

## APPENDIX

In this Appendix we address a recurrent technical nuisance. The spacings are not independently distributed; therefore, we cannot apply any of the standard theory for sums of independent, identically distributed random variables to the problem. Through a series of lemmas we establish that the difference between average log likelihood and log spacings functions (properly centered),

$$P_n \log f\left(\cdot; Q, \hat{h}_n\right) - P \log f_0 - n^{-1} \sum \log F\left(I_k; Q, \hat{h}_n\right) + \mu_n,$$

converges to zero uniformly in $Q$ in the limit, provided $\hat{h}_n^2 n/\log n \to \infty$.

Consider a random sample from a normal mixture for which the mixing distribution has a bounded range. In Lemma A.1, we prove the number of observations outside an interval increasing in $n$ is zero for $n$ sufficiently large.

Lemma A.1. *Suppose $\{X_i\}_{i=1}^n$ is a random sample from a normal mixture $f(\cdot; Q, h)$, where $Q$ has support on a compact set $A \subset [-a, a]$, $a > 0$. Let $\gamma_n^2 = (2 + \varepsilon)\log n$. With probability* 1, *all observations eventually lie in $[-a - h\gamma_n, a + h\gamma_n]$.*

PROOF. Let $P_Q(dx) = f(x; Q, h)\,dx$, where $Q$ has support on $[-a, a]$. As in the proof of Theorem 3.2, we use the fact that the upper tail probability $P_Q\{x > \beta\}$ is uniformly bounded (in $Q$) by the upper tail probability of a $N(a, h^2)$. Choose $\beta = a + \gamma_n$. The result now follows from the first Borel–Cantelli lemma. □

LEMMA A.2. *Suppose that* $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ *are order statistics from* $f_0$. *Let* $y$ *be a nonnegative integer and let* $X_i^*$ *be any point between* $X_{(i)}$ *and* $X_{(i+1)}$. *Then*

$$\text{(A.1)}\qquad \sup_{Q \in \mathscr{B}_a} \left| \sum_{y+1}^{n-1-y} \log\left(\frac{f(X_i^*; Q, h)}{f_0(X_i^*)}\right) - \log\left(\frac{f(X_i; Q, h)}{f_0(X_i)}\right) \right|$$
$$\leq \left[a + \max\left\{|X_{(y+1)}|, |X_{(n-y)}|\right\}\right](X_{(n-y)} - X_{(y+1)})h^{-2}.$$

PROOF. This result is based on a Taylor series expansion of each summand about $x_i$. □

Before completing our discussion of the relationship between log spacings and log likelihood, we need to state a LIL for spacings.

LEMMA A.3. *Let* $\{Z_i\}_{i=1}^n$ *be a random sample of* $n$ *uniform random variables, and let* $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ *be the set of order statistics. Let* $\{D_i = Z_{(i)} - Z_{(i-1)}\}_{i=1}^n$ *be the set of uniform spacings. Then, for any* $\varepsilon > 0$,

$$\limsup_n \left| n^{-1/2} \sum (\log D_i - \mu_n) \right| = O(b_n) \quad a.s.$$

PROOF. The proof is arduous, but familiar, as it borrows the arguments usually applied to sums of independent random variables. First apply the usual spacings trick by noting that the distribution of a set of $n$ uniform spacings is equivalent to the distribution of $\{Y_1/\Sigma Y_i, \dots, Y_n/\Sigma Y_i\}$, where $Y_1, \dots, Y_n$ are a set of $n$ exponential random variables. We can use the distribution of these independent, identically distributed random variables to get an exponential bound on a subsequence of probabilities. Finally we are able to apply the usual proof to get the LIL. □

In Theorem A.4 we establish that the remainder term encountered in the proof of Theorem 2.1 goes to zero a.s.

THEOREM A.4. *Let* $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ *be order statistics from* $F(\cdot; Q_0, h_0) = F_0$. *Let*

$$R_n(Q, h) = \frac{1}{n} \sum_{1}^{n-1} \log F(I_k; Q, h) - \mu_n - P_n \log f(\cdot; Q, h) + P \log f_0.$$

*Provided* $0 < \hat{h}_n < K$ *eventually, a.s. and* $\hat{h}_n^2 n/\log n \to \infty$ *a.s.*,

$$\lim_n \sup_{Q \in \mathscr{B}_a} R_n(Q, \hat{h}_n) = 0 \quad a.s.$$

Proof. By the usual LIL, eventually, on a set of probability 1,

$$|P_n \log f_0 - P \log f_0| = o(n^{-1} \log n)^{1/2}. \tag{A.2}$$

Similarly, by the spacings LIL (Lemma A.3), eventually,

$$\left| \frac{1}{n} \sum_1^{n-1} \log F_0(I_i) - \mu_n \right| = o(n^{-1} \log n)^{1/2} \quad \text{a.s.} \tag{A.3}$$

Define

$$\begin{aligned} R_n^*(Q, h) = \frac{1}{n} \sum_1^{n-1} [&\log F_0(I_k) - \log F(I_k; Q, h) \\ &-\log f_0(x_k) + \log f(x_k; Q, h)]. \end{aligned} \tag{A.4}$$

From (A.2) and (A.3), it is sufficient to show $\lim_n \sup_{Q \in \mathscr{B}_a} R_n^*(Q, \hat{h}_n) = 0$ a.s. From (4.1) we have

$$\log F(I_k; Q, h) = \log(f(X_k^*; Q, h)/f_0(X_k^*)) + \log F_0(I_k) \tag{A.5}$$

for $X_k^* \in I_k$. Let $Y_n$ be the number of observations outside the interval $[-a - h\gamma_n, a + h\gamma_n]$. Substituting (A.5) into (A.4) and dividing the sum into two parts we get

$$\begin{aligned} R_n^*(Q, h) = \frac{1}{n} \sum_{Y_n+1}^{n-1-Y_n} &\left[ \log\left( \frac{f(X_{(i)}; Q, h)}{f_0(X_{(i)})} \right) - \log\left( \frac{f(X_i^*; Q, h)}{f_0(X_i^*)} \right) \right] \\ + \frac{1}{n} &\left( \sum_1^{Y_n} + \sum_{n-Y_n}^{n-1} \right) \Bigg[ \log F_0(I_i) - \log F(I_i; Q, h) \\ &+ \log\left( \frac{f(X_i; Q, h)}{f_0(X_i)} \right) \Bigg]. \end{aligned}$$

By Lemma A.1, eventually $Y_n = 0$ a.s., so we can ignore the second term. Apply Lemma A.2 to the first sum to obtain an upper bound of

$$h^{-2} n^{-1} (X_{(n-Y_n)} - X_{(Y_n+1)}) \left[ a + \max\{|X_{(Y_n+1)}|, |X_{(n-Y_n)}|\} \right].$$

Thus $h^2$ times the first term is eventually bounded by $2n^{-1}(a + K\gamma_n)[2a + K\gamma_n]$ uniformly in $h$ and $Q$. Conclude that provided $\hat{h}_n^2 n/\log n \to \infty$ a.s., $\sup_{Q \in \mathscr{B}_a} R_n^*(Q, \hat{h}_n) \to 0$ a.s. □

**Acknowledgments.** A portion of this work was presented in the author's doctoral dissertation under Bruce Lindsay at the Pennsylvania State University. The author is grateful to him for his guidance. Thanks are also due to

David Pollard, the referees and an Associate Editor for recommendations that greatly improved the presentation of the article.

## REFERENCES

Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J. Roy. Statist. Soc. Ser. B* **45** 394–403.

Cressie, N. (1976). On the logarithms of high-order spacings. *Biometrika* **63** 343–355.

Darling, D. A. (1953). On a class of problems relating to the random division of an interval. *Ann. Math. Statist.* **24** 239–353.

Geman, S. and Hwang, C. R. (1982). Nonparametric and maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

James, B. R. (1975). A functional law of the iterated logarithm for weighted empirical distributions. *Ann. Probab.* **3** 762–772.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting. *Probability and Information Theory. Lecture Notes in Math.* **89**. Springer, New York.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods, Part I: A general theory. *Ann. Statist.* **11** 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, Part II: The exponential family. *Ann. Statist.* **11** 783–792.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Application to Clustering*. Dekker, New York.

Peters, B. C. and Walker, H. F. (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.* **35** 362–378.

Ranneby, B. (1984). The maximum spacings method: An estimation method. *Scand. J. Statist.* **11** 93–112.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239.

Roeder, K. (1988). Method of spacings for semiparametric inference. Ph.D. dissertation, Pennsylvania State Univ.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85** 617–624.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Department of Statistics
Box 2179 Yale Station
New Haven, Connecticut 06520