The Annals of Statistics
1995, Vol. 23, No. 6, 2264–2284

ON THE MINIMISATION OF L^p ERROR IN MODE ESTIMATION

BY BIRGIT GRUND AND PETER HALL

University of Minnesota and Mathematical Sciences Research Institute, Berkeley

We show that, for L^p convergence of the mode of a nonparametric density estimator to the mode of an unknown probability density, finiteness of the pth moment of the underlying distribution is both necessary and sufficient. The basic requirement of existence of finite variance has been overlooked by statisticians, who have earlier considered mean square convergence of nonparametric mode estimators; they have focussed on mean squared error of the asymptotic distribution, rather than on asymptotic mean squared error. The effect of bandwidth choice on the rate of L^p convergence is analysed, and smoothed bootstrap methods are used to develop an empirical approximation to the L^p measure of error. The resulting bootstrap estimator of L^p error may be minimised with respect to the bandwidth of the nonparametric density estimator, and in this way an empirical rule may be developed for selecting the bandwidth for mode estimation. Particular attention is devoted to the problem of selecting the appropriate amount of smoothing in the bootstrap algorithm.

1. Introduction. The study of nonparametric mode estimation is now three decades old, having its roots in Parzen's (1962) article on kernel density estimation. Romano (1988a) has surveyed subsequent work, including that of Eddy (1980, 1982) on kernel estimation and of Grenander (1965) on alternative approaches. See also Tsybakov (1990). Romano (1988b) has discussed bootstrap methods in the context of mode estimation. Recent work on estimating peaks nonparametrically includes that of Müller (1989), in the context of nonparametric regression. Mammen, Marron and Fisher (1992) and Fisher, Mammen and Marron (1994) have discussed nonparametric estimation of the number of modes in a multimodal distribution. It is well-known that the asymptotically optimal bandwidth for mode estimation is an order of magnitude larger than that which is appropriate for point estimation of a probability density.

In the special case where asymptotic mean squared error is used to describe performance of the mode estimator, the optimal bandwidth could, in principle, be estimated empirically using plug-in methods. These would require pilot estimators to be developed for a number of quantities in the

Received August 1993; revised February 1995.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20.

Key words and phrases. Bandwidth, bootstrap, convergence in L^p, kernel density estimator, mean squared error, mode, smoothed bootstrap, smoothing parameter.

formula for the optimal bandwidth, including the mode itself, the value of the density at the mode and the value of a high-order derivative at the mode. However, this is a very complex procedure, and that unattractiveness is undoubtedly an important reason for the lack of information which exists about its theoretical and numerical properties.

In the present paper we propose a much simpler approach to bandwidth selection. We suggest a bootstrap method for estimating the mean squared error of the mode estimator, and propose selecting the bandwidth by minimising this estimator.

The simplicity of our procedure enables us to treat L^p measures of error in mode estimation, not just mean squared error. Therefore, we introduce our techniques in this general context. We show in Section 2 that if the underlying distribution is smooth, a necessary and sufficient condition for L^p convergence of the mode estimator is the existence of finite pth absolute moment of the underlying distribution. A reader who is familiar with classical L^2 theory for mode estimation may doubt the correctness of this claim, since the assumption of finite variance is never imposed in that work. However, one should remember that classical L^2 theory is concerned only with *asymptotic* mean squared error—that is, with mean squared error of the asymptotic distribution of the mode estimator. By way of contrast, we study the actual mean L^p error, for finite n and for general $p \geq 1$. Hitherto, not even the problem of mean square convergence has been treated with the degree of explicitness and detail offered in the present paper.

Section 3 describes a smoothed bootstrap estimator of mean L^p error. Curiously, that approach requires only finite εth moment for some $\varepsilon > 0$; it does not need finite pth moment. The apparent contradiction arises because extreme values from a bootstrap resample have properties quite unlike that of extremes from the actual population. The requirement of finite pth moment in Section 2 arises because of properties of extreme values.

Bootstrap methods have been used before to estimate mean squared error in the context of curve estimation. See, for example, Taylor (1989), Faraway and Jhun (1990), Hall (1990) and Hall, Marron and Park (1992). Unlike Hall (1990), but like Faraway and Jhun (1990), we use a resample size that is identical to sample size. One of our aims is to solve, at least theoretically, the difficult problem of selecting the correct bandwidth for the resampling part of the bootstrap algorithm. This problem is not addressed by Faraway and Jhun (1990), and requires significantly more detailed results about convergence rates than are available from classical literature on mode estimation. The new results are derived in Section 2, in the general context of mean L^p error, and in Section 3 for our smoothed bootstrap method. By combining the resulting formulae we show in Section 3 that if an rth order kernel [as defined at (2.5)] is employed when estimating the mode and a second-order kernel estimator $\tilde{f}$ is used in the resampling operation, then the bandwidth for $\tilde{f}$ should be taken to be of size $n^{-1/(2r+7)}$ if our aim is to develop an empirical approximation to the optimal bandwidth for $\hat{f}$. This size is very much larger than that required for optimal point estimation using $\tilde{f}$. Hence,

the bootstrap algorithm should involve substantial oversmoothing when resampling. The results of a simulation study, illustrating these conclusions, are summarised in Section 4.

By way of notation, $\mathscr{X} = \{X_1, \ldots, X_n\}$ represents a random sample from a population with density f, which we assume has a unique largest mode m. Write X for a generic X_i. Given a continuous kernel function K and a bandwidth h satisfying $0 < h \leq 1$, define the kernel estimator

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}.$$

Let $\hat{m}$ denote any quantity with the property

$$\hat{f}(\hat{m}) = \sup_{-\infty < x < \infty} \hat{f}(x).$$

Section 2 will discuss the issue of ties for $\hat{m}$. We assume throughout that kernel functions are supported on the interval $(-1, 1)$. This condition is imposed to simplify technical arguments, and may be removed at the expense of longer proofs. In particular, all our results are valid if we take all kernels to equal the standard normal density, but the present proofs would need modification. Versions of our results may be derived for estimators other than those based on kernels, although it seems particularly difficult to obtain the detail of our second-order theory for estimators based, for example, on log splines and penalized log-likelihood methods.

2. Convergence in probability, and in L^p, of the mode estimator. In the sense of large deviations, $\hat{m}$ converges to m at a geometrically fast rate, as our first result shows.

THEOREM 2.1. *Assume that f is bounded, continuous at a point m and satisfies*

$$\sup_{x : |x - m| > \eta_1} f(x) < f(m) \tag{2.1}$$

for all $\eta_1 > 0$. Assume that K is of bounded variation, is supported on $(-1, 1)$, is continuous and satisfies $\int K = 1$, and that for some $\eta_2 > 0$, $1 \geq h = h(n) \to 0$,

$$\sup_{n \geq 0} (nh)^{-1} (\log n)^{2 + \eta_2} < \infty.$$

Then for each $\eta_3, \lambda > 0$,

$$P\big(|\hat{m} - m| > \eta_3\big) = O(n^{-\lambda})$$

as $n \to \infty$.

Condition (2.1) defines $f(m)$ as the "unique largest peak" of f.

Theorem 2.1 implies that, under the assumptions there, $\hat{m} \to m$ in probability. However, without additional regularity conditions on the tails of the

sampling distribution, there can be no guarantee that $\hat{m}$ will converge to m in any L^p metric. In part, this problem is caused by ambiguities in how $\hat{m}$ should be defined when $\hat{f}$ has two or more modes at which $\hat{f}$ achieves the same height. While this is, in a sense, a pathological issue, the matter of whether $\hat{m}$ converges to m in L^p is fraught with difficulties caused by pathological arrangements of the data.

To appreciate this point, let us order the data values as $X_{(1)} \leq \cdots \leq X_{(n)}$ and let $x_1 < \cdots < x_n$ denote real numbers such that $x_i - x_{i-1} - 4 > 0$, $x_n > 2$ and $P\{X \in (x_i - 1, x_i + 1)\} > 0$ for each i. Numbers x_i with these properties exist if the distribution of X is unbounded to the right. Consider the event $\mathscr{E}_n$ that $X_{(i)} \in (x_i - 1, x_i + 1)$ for $1 \leq i \leq n - 1$ and $X_{(n)} > x_n - 1$. Since for large n the bandwidth h employed to construct $\hat{f}$ is taken to be no greater than 1, and since K vanishes outside $(-1, 1)$, then when $\mathscr{E}_n$ prevails, the kernel estimator $\hat{f}$ is simply a string of n nonoverlapping "humps," each with the shape of $(nh)^{-1}K(\cdot/h)$ and centered at respective values $X_1, \ldots, X_n$. Suppose that in such cases, when there is a tie for the mode of $\hat{f}$, we agree to take as our mode estimator that candidate which is furthest to the right. Then,

$$E|\hat{m}|^p \geq E\{|\hat{m}|^p I(\mathscr{E}_n)\} \geq E\left\{\left(X_{(n)} - 1\right)^p I(\mathscr{E}_n)\right\}$$

$$\geq E\{(X - 1)^p I(X > x_n - 1)\} \prod_{i=1}^{n-1} P\{X \in (x_i - 1, x_i + 1)\}.$$

The right-hand side is infinite if $E(X^+)^p = \infty$. Arguing thus, the condition $E(X^+)^p < \infty$ is seen to be necessary for $E|\hat{m}|^p < \infty$. Similarly, if we choose randomly among tied modes, then $E|X|^p < \infty$ is a necessary condition. The latter constraint is also sufficient for convergence in L^p, as Theorem 2.2 will point out.

Of course, our proof of the necessity of finite pth moments relies on a highly unusual, indeed pathological case. We are not claiming that such pathologies arise with significant frequency, only that they have positive probability of occurring. Our argument shows, in effect, that the limit of the pth moment of the mode equals the pth moment of the limit of the mode if and only if the sampling distribution has finite pth moment. In that context our first-order limit theory has been anticipated by earlier workers. However, note that unlike those earlier contributors, we address the limiting properties of moments, rather than moments of the limiting distribution. In order to be technically correct in the former setting one must ask that the underlying distribution have finite pth moments. The case of kernels with unbounded support—for example, when K is the standard normal density—may be treated similarly. The proof is longer there, but ties never arise and moment conditions are still necessary for L^p convergence of the mode.

The spirit of our proof is reminiscent of work of Mammen, Marron and Fisher (1992), which describes how the number of modes of a kernel density estimator depends on bandwidth, and of Devroye [(1985), page 248], which addresses the number of isolated bumps in a kernel density estimator.

THEOREM 2.2. *Assume the conditions of Theorem* 2.1. *If there are two or more modes of* $\hat{f}$ *with the same height, select among them randomly when defining* $\hat{m}$. *Let* $p \geq 1$. *Then*

$$E|\hat{m} - m|^p \to 0$$

if and only if $E|X|^p < \infty$. *Furthermore, if* $E|X|^p < \infty$, *then for each* $\eta, \lambda > 0$,

$$E\{|\hat{m} - m|^p I(|\hat{m} - m| > \eta)\} = O(n^{-\lambda}). \tag{2.2}$$

Our final result in this section describes an asymptotic formula for $E|\hat{m} - m|^p$. We assume that any tie for the mode estimator is broken at random. Let (N_1, N_2, N_3) denote a trivariate normal random vector with the same mean vector and covariance matrix as $(\hat{f}'(m), \hat{f}''(m) - f''(m), \hat{f}'''(m) - f'''(m))$, and put $\alpha = |E\hat{f}'(m)|$, $\beta = |E\hat{f}''(m) - f''(m)|$, $g_1(x_1, x_2; f, m) = x_1 - x_1 x_2 f''(m)^{-1} + x_1 x_2^2 f''(m)^{-2}$ and $g_2(x_1, x_2, x_3; f, m) = g_1(x_1, x_2; f, m) + \frac{1}{2}x_1^2 f'''(m) f''(m)^{-2} + \frac{1}{2}x_1^2 x_3 f''(m)^{-2}$.

THEOREM 2.3. *Assume that* f *has a "unique largest peak" at* m, *that* f''' *exists in a neighbourhood of* m *and is continuous at* m, *that* $f''(m) \neq 0$, *that* $E|X|^p < \infty$ *and that* K *is supported on* $(-1, 1)$, *has four bounded derivatives and satisfies* $\int K = 1$ *and* $\int yK(y)\,dy = 0$. *Suppose too that* $h = h(n) \to 0$ *and that* $(nh^7)^{-1} = O(1)$. *Then for each* $p \geq 1$,

$$\begin{aligned}
&(E|\hat{m} - m|^p)^{1/p} \\
&\quad = \{E|g_1(N_1, N_2; f, m)|^p\}^{1/p} |f''(m)|^{-1} \\
&\qquad + O\Big[\{(nh^3)^{-1/2} + \alpha\} \\
&\qquad\quad \times \{(nh^3)^{-1/2} + (nh^5)^{-3/2} + \alpha + \beta^3\}(\log n)^2\Big].
\end{aligned} \tag{2.3}$$

If, in addition, $f^{(4)}$ *exists and is bounded in a neighbourhood of* m, *and* K *has five derivatives, then for each* $p \geq 1$,

$$\begin{aligned}
&(E|\hat{m} - m|^p)^{1/p} \\
&\quad = \{E|g_2(N_1, N_2, N_3; f, m)|^p\}^{1/p} |f''(m)|^{-1} \\
&\qquad + O\Big(\{(nh^3)^{-1/2} + \alpha\} \\
&\qquad\quad \times \Big[(nh^3)^{-1} + (nh^5)^{-3/2} + \alpha^2\{1 + (nh^9)^{-1/2}\} + \beta^3\Big](\log n)^2\Big)
\end{aligned} \tag{2.4}$$

as $n \to \infty$.

If K is an rth-order kernel, meaning that

$$\int y^j K(y)\,dy = \begin{cases} 1, & \text{if } j = 0, \\ 0, & \text{if } 1 \le j \le r-1, \\ (-1)^r r!\kappa \ne 0, & \text{if } j = r, \end{cases} \tag{2.5}$$

and if f has sufficiently many continuous derivatives, then the mean and variance of N_j are asymptotic to $\kappa f^{(r+j)}(m)h^r$ and $\kappa_j' f(m)(nh^{2j+1})^{-1}$, respectively, where $\kappa_j' = \int K^{(j)^2}$. In particular this implies that α and β are both of size h^r and, by (2.3), writing N for a standard normal random variable, that

$$(E|\hat{m} - m|^p)^{1/p} \sim \left\{E|c_1(nh^3)^{-1/2}N + c_2 h^r|^p\right\}^{1/p} |f''(m)|^{-1}, \tag{2.6}$$

where $c_1^2 = \kappa_1' f(m)$ and $c_2 = |\kappa f^{(r+1)}(m)|$. The longer expansion at (2.4) is needed to derive second-order terms in such approximations when $r = 2$. Surprisingly, the simpler result (2.3) is adequate when $r \ge 3$. It follows from (2.6) that the bandwidth h_0 that minimises $A_1(h) \equiv \{E|\hat{m} - m|^p\}^{1/p}$, over values of h in the range prescribed by Theorem 2.3, satisfies $h_0 \sim u_0 n^{-1/(2r+3)}$, where u_0 minimises

$$G(u) = E|c_1 u^{-3/2}N + c_2 u^r|^p, \tag{2.7}$$

and that for this choice of h,

$$E|\hat{m} - m|^p \sim n^{-pr/(2r+3)} G(u_0)|f''(m)|^{-p}.$$

The form of the remainder terms in (2.3) and (2.4) is carefully chosen so as to capture as much as possible of the effect of bootstrap estimation of mean L^p error. This point will be elucidated in Section 3. For that purpose we provide now a high-order approximation to h_0.

LEMMA 2.1. *Assume the conditions of Theorem* 2.3 *and let K be symmetric. Then, for* $r \ge 2$,

$$h_0 = u_0 n^{-1/(2r+3)}\{1 + o(n^{-2/(2r+7)})\}. \tag{2.8}$$

PROOF. We show (2.8) for even $r \ge 4$. For symmetric K, the distribution of (N_1, N_2) may be elucidated relatively easily. Put $h_0'' = u_0 n^{-1/(2r+3)}$. We claim that the quantity h_0' that minimises $A_2(h) \equiv \{E|N_1 - N_1 N_2 f''(m)^{-1} + N_1 N_2^2 f''(m)^{-2}|^p\}^{1/p}$ satisfies $h_0'/h_0'' = 1 + O(n^{-2/(2r+3)}) = 1 + o(n^{-2/(2r+7)})$. To appreciate why, observe first that when $p = 2$, a simple Taylor expansion gives $A_2(h)^2 = \{C_1 h^{2r} + C_2(nh^3)^{-1}\}\{1 + O(h^2)\}$, where $C_1, C_2 > 0$. This produces $h_0' = h_0''\{1 + O(h_0''^2)\}$, as required. The case of general p may be treated similarly, after expansion of the covariance of (N_1, N_2).

Define δ_1 by $h_0' = h_0''(1 + \delta_1)$. By the usual quadratic Taylor expansion in the neighbourhood of a minimum,

$$A_2(h_0'')/A_2(h_0') = 1 + O(\delta_1^2) = 1 + o(n^{-4/(2r+7)}). \tag{2.9}$$

Define δ_2 by $h_0 = h_0'(1 + \delta_2)$. By the same Taylor expansion argument,

$$A_1(h_0')/A_1(h_0) = 1 + C\delta_2^2 + o(\delta_2^2), \tag{2.10}$$

where $C > 0$. In the next step of the proof, suppose first that $r \geq 3$. When h is of size $n^{-1/(2r+3)}$, the quantity $(nh^3)^{-1/2} + (nh^5)^{-3/2} + \alpha + \beta^3$ appearing in (2.3) is of size $n^{-r/(2r+3)} + n^{-3(r-1)/(2r+3)} = O(n^{-r/(2r+3)})$. By this fact and (2.3), $A_1(h_0')/A_2(h_0') = 1 + O\{n^{-r/(2r+3)}(\log n)^2\} = 1 + o(n^{-4/(2r+7)})$, the last identity requiring $r \geq 3$. Hence, by (2.9), $A_1(h_0')/A_2(h_0'') = 1 + o(n^{-4/(2r+7)})$, which in view of (2.10) implies $A_1(h_0)/A_2(h_0'') = 1 - C\delta_2^2 + o(n^{-4/(2r+7)})$. Since the left-hand side is minimised with $\delta_2 = 0$, the right-hand side must be too, which entails $\delta_2 = o(n^{-2/(2r+7)})$. This proves (2.8). The argument is the same when $r = 2$, except that (2.4) is used instead of (2.3) to bound the difference between A_1/A_2 and 1. □

We conclude this section by outlining proofs of Theorems 2.1 and 2.2. That of Theorem 2.3 is based on a relatively intricate Taylor expansion argument and is deferred to the Appendix.

PROOF OF THEOREM 2.1. Observe that

$$\begin{aligned}
&P(|\hat{m} - m| > \eta) \\
&\quad = P\Big\{ \sup_{x:\,|x-m|\leq\eta} \hat{f}(x) \leq \sup_{x:\,|x-m|>\eta} \hat{f}(x) \Big\} \\
&\quad \leq P\Big\{ \sup_{x:\,|x-m|\leq\eta} E\hat{f}(x) - \sup_{x:\,|x-m|\leq\eta} |\hat{f}(x) - E\hat{f}(x)| \\
&\qquad\qquad \leq \sup_{x:\,|x-m|>\eta} E\hat{f}(x) + \sup_{x:\,|x-m|>\eta} |\hat{f}(x) - E\hat{f}(x)| \Big\} \\
&\quad \leq P\Big\{ 2 \sup_{-\infty<x<\infty} |\hat{f}(x) - E\hat{f}(x)| \\
&\qquad\qquad > \sup_{x:\,|x-m|\leq\eta} E\hat{f}(x) - \sup_{x:\,|x-m|>\eta} E\hat{f}(x) \Big\}.
\end{aligned} \tag{2.11}$$

For each $\eta > 0$ there exists $\eta' > 0$ such that

$$\sup_{x:\,|x-m|\leq\eta} E\hat{f}(x) - \sup_{x:\,|x-m|>\eta} E\hat{f}(x) \geq \eta' \tag{2.12}$$

for all sufficiently large n. Therefore, it suffices to prove that for all $\eta, \lambda > 0$,

$$P\Big\{ \sup_{-\infty<x<\infty} |\hat{f}(x) - E\hat{f}(x)| > \eta \Big\} = O(n^{-\lambda}).$$

This may be achieved by applying the so-called Hungarian embedding [Komlós, Major and Tusnády (1975)], and modifying arguments of Silverman (1978), as follows. There exist constants $C_1, C_2, C_3 > 0$ and a Brownian

bridge W^0 such that, with F denoting the distribution function corresponding to f and with

$$Z_1(x) = -n^{-1/2}h^{-1}\int W^0\{F(t)\}\,d_t\,K\{h^{-1}(x-t)\}$$

and $Z_2(x) = \hat{f}(x) - E\hat{f}(x) - Z_1(x)$, we have for each n and each $y > 1/C_2$,

$$P\left\{n(\log n)^{-1}\sup_{-\infty<x<\infty}|Z_2(x)| > C_1 + y\right\} \le C_3\exp(-C_2 y).$$

Hence, $P\{\sup_x|Z_2(x)| > \frac{1}{2}\eta\} = O(n^{-\lambda})$ for all $\lambda > 0$. Let V denote the modulus of continuity of W^0. Then there exists another constant $C_4 > 0$ such that

$$\begin{aligned} n^{1/2}h\sup_{-\infty<x<\infty}|Z_1(x)| &= \sup_{-\infty<x<\infty}\left|\int\left[W^0\{F(x-ht)\} - W^0\{F(x)\}\right]dK(t)\right| \\ &\le \left(\int|dK|\right)\sup_{x:\,|t|\le 1}|W^0\{F(x-ht)\} - W^0\{F(x)\}| \\ &\le V(C_4 h). \end{aligned}$$

Now, $V(u) \le C_5(u\log u^{-1})^{1/2}V_0$, where $E(\exp V_0^2) < \infty$ [Garsia (1970)], and so with $\eta' = \eta/(3C_4^{1/2}C_5)$ and n large,

$$\begin{aligned} P\left\{\sup_{-\infty<x<\infty}|Z_1(x)| > \tfrac{1}{2}\eta\right\} &\le P\left[V_0 > \eta'\{nh/(\log h^{-1})\}^{1/2}\right] \\ &= O\left[\exp\{-\eta'^2(nh)/(\log h^{-1})\}\right] \\ &= O(n^{-\lambda}) \end{aligned}$$

for all $\lambda > 0$. □

PROOF OF THEOREM 2.2. We may assume without loss of generality that $m = 0$. The proof of "necessity" was outlined earlier. It is enough to show that $E|X|^p < \infty$ is sufficient for (2.2). To this end, observe that for any $\alpha, \eta, \lambda > 0$ and sufficiently large n,

$$\begin{aligned} E\{|\hat{m}|^p I(|\hat{m}| > \eta)\} &\le 2^{p-1}n^{\alpha p}P(|\hat{m}| > \eta) + 2^{p-1}E\{|\hat{m}|^p I(|\hat{m}| > n^\alpha)\} \\ &= O(n^{-\lambda}) + 2^{p-1}E\{|\hat{m}|^p I(|\hat{m}| > n^\alpha)\}, \end{aligned}$$

the last identity following from Theorem 2.1. Let Y_n denote the second-largest value of $|X_i|$ and note that for the large n and $\beta = \alpha - 1$,

$$\begin{aligned} &2^{-p}E\{|\hat{m}|^p I(n^\alpha < |\hat{m}| \le Y_n + 1)\} \\ &\quad\le 2^{-p}E\{(Y_n+1)^p I(Y_n + 1 > n^\alpha)\} \\ &\quad\le n^2\int_{x>n^\beta} x^p P(|X| \le x)^{n-2}P(|X| > x)\,dP(|X| \le x) \\ &\quad\le n^2\int_{x>n^\beta} x^p(x^{-p}E|X|^p)\,dP(|X| \le x) \\ &\quad= n^2E|X|^pP(|X| > n^\beta) \le n^{2-p\beta}(E|X|^p)^2 = O(n^{-\lambda}), \end{aligned}$$

provided $\alpha > \{(\lambda + 2)/p\} + 1$. It remains only to show that for all $\lambda > 0$,

$$t_n \equiv E\{|\hat{m}|^p I(|\hat{m}| > Y_n + 1)\} = O(n^{-\lambda}).$$

Let $X_{y1}, \ldots, X_{y,n-1}$ be independent, and independent of $X_1, \ldots, X_n$, with the (conditional) distribution of X given that $|X| \le y$, and put $Z_n = \max_{i \le n} |X_i|$,

$$\mathscr{H}_{y,n} = \left\{ \sup_{-\infty < x < \infty} \sum_{i=1}^{n-1} K\{(x - X_{y,i})/h\} \le \sup K \right\}.$$

Since $|\hat{m}| \le Z_n + 1$, then

$$\begin{aligned} t_n &\le \int_0^\infty P(\mathscr{H}_{y,n}) E(|\hat{m}|^p | Z_n = y)\, dP(Z_n \le y) \\ &\le \int_0^\infty P(\mathscr{H}_{y,n})(y + 1)^p\, dP(Z_n \le y). \end{aligned}$$

The methods used to prove Theorem 2.1 may be employed to show that for all $\lambda > 0$,

$$\sup_{y>0} P(\mathscr{H}_{y,n}) = O(n^{-\lambda}).$$

Hence,

$$\begin{aligned} t_n &= O(n^{-\lambda}) E\{(Z_n + 1)^p\} \\ &\le O(n^{-\lambda}) nE\{(|X| + 1)^p\} = O(n^{-\lambda+1}), \end{aligned}$$

as had to be shown. □

3. Bootstrap estimation of mean L_p error. In Section 2 we discussed the convergence to zero of

$$\mu_p = \mu_p(h) = E|\hat{m} - m|^p.$$

We showed that if K is an rth-order kernel, then this quantity is asymptotically minimised by taking $h = u_0 n^{-1/(2r+3)}$ in the definition of $\hat{m}$. Here, u_0 minimises the function $G(u)$, $u > 0$, which depends on the unknown $f(m)$ and $f^{(r+1)}(m)$. Both these quantities are unknown, and so this prescription for selecting h is not really practical. In the present section we show that bootstrap methods may be employed to estimate $\mu_p(h)$, and thus to empirically select a bandwidth for estimating m.

Let K, used in the construction of $\hat{f}$, denote a compactly supported rth-order kernel; see (2.5) for a definition of the "rth-order" property. Let L be a compactly supported, symmetric density with $r + 1$ derivatives. Define

$$\tilde{f}(x) = (nh_1)^{-1} \sum_{i=1}^{n} L\{(x - X_i)/h_1\}.$$

Our bootstrap sampling will be from the distribution that has this density. The assumption that L is a density, in particular, that it is nonnegative, is necessary if the sampling part of the operation is to be feasible, since we cannot easily sample from a "distribution" whose density takes negative values [although, see Hall and Murison (1991)]. The quantity $\tilde{f}$ is, in a sense, a pilot estimator of f, with its own bandwidth h_1. We shall discuss choice of h_1 later in this section.

Conditional on the sample $\mathscr{X} = \{X_1, \ldots, X_n\}$, draw a sample $\{X_1^*, \ldots, X_n^*\}$ from the distribution with density $\tilde{f}$. We may take $X_i^* = X_i' + h_1 Y_i$, where $X_1', \ldots, X_n'$ are drawn randomly, with replacement, from $\mathscr{X}$, $Y_1, \ldots, Y_n$ are independent and identically distributed with density L and, conditional on $\mathscr{X}$, the variables $X_1', \ldots, X_n', Y_1, \ldots, Y_n$ are stochastically independent. Put

$$\hat{f}^*(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i^*)/h\}.$$

Let $\hat{m}^*$ and $\tilde{m}$ denote the modes of $\hat{f}^*$ and $\tilde{f}$, respectively, defined by breaking ties randomly when necessary, and set

$$\hat{\mu}_p = \hat{\mu}_p(h) = E'|\hat{m}^* - \tilde{m}|^p,$$

where here and below, E' denotes expectation conditional on $\mathscr{X}$. Note particularly that $\tilde{m}$, not $\hat{m}$, is employed in the definition of $\hat{\mu}_p$.

Our main result in this section follows. It provides bootstrap versions of portions of Theorem 2.2 and 2.3. Note particularly that, in terms of moment conditions, we assume only that $E|X|^\varepsilon < \infty$ for some $\varepsilon > 0$, not that $E|X|^p < \infty$.

Conditional on $\mathscr{X}$, let (N_1', N_2', N_3') denote a trivariate normal random vector with the same conditional mean and conditional variance matrix as $(\hat{f}^{*\prime}(\tilde{m}), \hat{f}^{*\prime\prime}(\tilde{m}) - \tilde{f}''(\tilde{m}), \hat{f}^{*\prime\prime\prime}(\tilde{m}) - \tilde{f}'''(\tilde{m}))$. Put $\alpha' = |E'\hat{f}^{*\prime}(\tilde{m})|$, $\beta' = |E'\hat{f}^{*\prime\prime}(\tilde{m}) - \tilde{f}''(\tilde{m})|$. The conditional means and variances of (N_1', N_2', N_3') are asymptotic to the unconditional means and variances of (N_1, N_2, N_3), discussed in the paragraph following Theorem 2.3.

THEOREM 3.1. *Assume that f has a "unique largest peak" at m, that f is uniformly continuous on $(-\infty, \infty)$ and f''' exists and is continuous in a neighbourhood of m, that $f''(m) \neq 0$, that $E|X|^\varepsilon < \infty$ for some $\varepsilon > 0$, that $\int K = 1$, that L is a symmetric probability density and that K, L are of bounded variation, are supported on $(-1, 1)$ and have three derivatives. Suppose too that $0 < h$, $h_1 \leq 1$, $h + h_1 \to 0$, $(nh^7)^{-1} = O(1)$ and for some $\eta_1 > 0$, $\sup_{n \geq 1}(nh_1^7)^{-1}(\log n)^{1+\eta_1} < \infty$. Let $p \geq 1$. Then for each $\eta_2, \lambda > 0$,*

$$E'\{|\hat{m}^* - \tilde{m}|^p I(|\hat{m}^* - \tilde{m}| > \eta_2)\} = O(n^{-\lambda}) \tag{3.1}$$

with probability 1, *and for* $r \geq 3$,

$$
\begin{aligned}
&\bigl(E'|\hat{m}^* - \tilde{m}|^p\bigr)^{1/p} \\
&\quad = \Bigl\{E'\Bigl|g_1\bigl(N_1', N_2'; \tilde{f}', \tilde{m}\bigr)\Bigr|^p\Bigr\}^{1/p} |\tilde{f}''(\tilde{m})|^{-1} \\
&\qquad + O\Bigl[\Bigl\{(nh^3)^{-1/2} + \alpha'\Bigr\} \\
&\qquad\quad \times \Bigl\{(nh^3)^{-1/2} + (nh^5)^{-3/2} + \alpha' + \beta'^3\Bigr\}(\log n)^2\Bigr]
\end{aligned}
\tag{3.2}
$$

with probability 1. *For* $r \geq 2$,

$$
\begin{aligned}
&\bigl(E'|\hat{m}^* - \tilde{m}|^p\bigr)^{1/p} \\
&\quad = \Bigl\{E'\Bigl|g_2\bigl(N_1', N_2', N_3'; \tilde{f}, \tilde{m}\bigr)\Bigr|^p\Bigr\}^{1/p} |\tilde{f}''(\tilde{m})|^{-1} \\
&\qquad + O\Bigl(\Bigl\{(nh^3)^{-1/2} + \alpha'\Bigr\} \\
&\qquad\quad \times \Bigl[(nh^3)^{-1} + (nh^5)^{-3/2} \\
&\qquad\qquad + \alpha'^2\Bigl\{1 + (nh^9)^{-1}\Bigr\} + \beta'^3\Bigr](\log n)^2\Bigr)
\end{aligned}
\tag{3.3}
$$

with probability 1.

Our proof of Theorem 3.1 remains valid if we take h to be a function of the data, $\mathscr{X}$. In this case, the conditions imposed on h in the statement of Theorem 3.1 should be interpreted as asking that $O < h \leq 1$, $h \to 0$ with probability 1 and $\sup(nh^5)^{-1}(\log n)^{1+\eta} < \infty$ with probability 1.

The principal application of the bootstrap estimator $\hat{\mu}_p(h)$ is to calculate an empirical version of the bandwidth h_0 that minimises $\mu_p(h)$. We discussed h_0 briefly in Section 2, where we showed that if K is an rth-order kernel [see (2.5) for a definition of kernel order] and if f has $r + 1$ derivatives, then

$$
h_0 = u_0 n^{-1/(2r+3)}\{1 + o(n^{-2/(2r+7)})\}. \tag{3.4}
$$

In this formula, u_0 is defined to be that quantity which minimises $G(u)$, defined at (2.7). An almost identical argument, based on (3.2) and (3.3) rather than (2.3) and (2.4), shows that the bandwidth $\hat{h}_0$ which minimises $\hat{\mu}_p(h)$ is given by

$$
\hat{h}_0 = \hat{u}_0 n^{-1/(2r+3)}\{1 + o(n^{-2/(2r+7)})\}, \tag{3.5}
$$

with probability 1, where $\hat{u}_0$ minimises $\hat{G}(u) = E'|\hat{c}_1 u^{-3/2} N' + \hat{c}_2 u^r|^p$, $u > 0$, and $\hat{c}_1 = \{\tilde{f}(\tilde{m})\int (K')^2\}^{1/2}$, $\hat{c}_2 = |\kappa \tilde{f}^{(r+1)}(\tilde{m})|$. In this formula we take N' to be a standard normal random variable independent of $\mathscr{X}$. A formal derivation of this result requires $\tilde{f}^{(r+1)}$ to be strongly consistent for $f^{(r+1)}$ in a neighbourhood of m, and for that we ask that $\sup(nh_1^{2r+3})^{-1}(\log n)^{1+\eta} < \infty$ for some $\eta > 0$.

We claim the following consequences of (3.4) and (3.5): if we choose the bandwidth h_1, employed to construct $\tilde{f}$, such that it minimises the relative error $(\hat{h}_0 - h_0)/h_0$, then h_1 is asymptotic to a constant multiple of $n^{-1/(2r+7)}$ and the relative error is of size $n^{-2/(2r+7)}$. Now, the value of the best constant in the formula $h_1 \simeq \text{const} \cdot n^{-1/(2r+7)}$ depends on the unknowns $f^{(j)}(m)$ for $j \le 2r+3$ and also on the metric in which the error $(\hat{h}_0 - h_0)/h_0$ is measured (e.g., whether it is asymptotic mean squared error or some other asymptotic L^q metric). Hence, there seems to be little point in being more specific about the constant, and so we shall not pursue that matter further here. However, knowing that the optimal size is $n^{-1/(2r+7)}$ does indicate that the bandwidth for constructing $\tilde{f}$ for our present purpose should be substantially larger than that for point estimation of f. As is well known [see, e.g., Silverman (1986), Chapter 3], the latter is of size $n^{-1/5}$.

To verify our claim, observe that the quantities $\tilde{f} - f$, $\tilde{f}^{(r+1)} - f^{(r+1)}$ and $\tilde{m} - m$ are, respectively, of size $(nh_1)^{-1/2} + h_1^2$, $(nh_1^{2r+3})^{-1/2} + h_1^2$ and $(nh_1^3)^{-1/2} + h_1^2$. Therefore, $\hat{c}_1 - c_1$ and $\hat{c}_2 - c_2$ are of sizes $(nh_1^3)^{-1/2} + h_1^2$ and $(nh_1^{2r+3})^{-1/2} + h_1^2$, respectively. Comparing the formulae for G and $\hat{G}$ we see that $\hat{u}_0 - u_0$ is of the same size as $\hat{c}_2 - c_2$ and that the size of this error is minimised at $n^{-2/(2r+7)}$ by selecting h_1 to be of size $n^{-1/(2r+7)}$. For example, when $p = 2$ we have

$$\hat{u}_0 - u_0 \sim -\{\tilde{f}^{(r+1)}(m) - f^{(r+1)}(m)\}(4/3)r(2r+3)^{-1}u_0 f^{(r+1)}(m)^{-1},$$

so that the asymptotically optimal bandwidth h_1 is that which minimises the mean squared error of $\tilde{f}^{(r+1)}(m)$.

We conclude this section with a derivation of (3.1) in Theorem 3.1. Proofs of (3.2) and (3.3) are similar to that of Theorem 2.3 and so are not given here.

PROOF OF (3.1). The conditions imposed on K, L, f, h and h_1 are sufficient to enable us to prove, via the "Hungarian embedding" [see Komlós, Major and Tusnády (1975) and Silverman (1978)] that for each $\eta, \lambda > 0$,

$$P\Big\{\sup_{-\infty<x<\infty} |\tilde{f}(x) - f(x)| > \eta\Big\} = O(n^{-\lambda}), \tag{3.6}$$

$$P'\Big\{\sup_{-\infty<x<\infty} |\hat{f}^*(x) - E'\hat{f}^*(x)| > \eta\Big\} = O(n^{-\lambda}), \tag{3.7}$$

where P' denotes probability conditional on $\mathscr{X}$, and the latter identity is interpreted as holding with probability 1. From (3.6) it follows, via the Borel–Cantelli lemma, that

$$\sup_{-\infty<x<\infty} |\tilde{f}(x) - f(x)| \to 0$$

with probability 1. Therefore,

$$\sup_{-\infty<x<\infty} |E'\hat{f}^*(x) - E\hat{f}(x)| \le \Big(\int |K|\Big) \sup_{-\infty<x<\infty} |\tilde{f}(x) - f(x)| \to 0$$

with probability 1. Replacing $(K, h, \hat{f})$ by $(L, h_1, \tilde{f})$ in Theorem 2.1, we deduce that $\tilde{m} \to m$ with probability 1. Noting the remark that contains (2.12), we conclude that for each $\eta > 0$ there exists $\eta' > 0$ such that with probability 1, for all sufficiently large n,

$$\sup_{x:\,|x-\tilde{m}|\le\eta} E'\hat{f}^*(x) - \sup_{x:\,|x-\tilde{m}|>\eta} E'\hat{f}^*(x) \ge \eta'.$$

Arguing as at (2.11), we may now deduce that with probability 1 and for all sufficiently large n,

$$\begin{aligned} P'\bigl(|\hat{m}^* - \tilde{m}| > \eta\bigr) &\le P'\Bigl\{2 \sup_{-\infty<x<\infty} |\hat{f}^*(x) - E'\hat{f}^*(x)| \\ &\qquad > \sup_{x:\,|x-\tilde{m}|\le\eta} E'\hat{f}^*(x) - \sup_{x:\,|x-\tilde{m}|>\eta} E'\hat{f}^*(x)\Bigr\} \\ &\le P'\Bigl\{\sup_{-\infty<x<\infty} |\hat{f}^*(x) - E'\hat{f}^*(x)| > \tfrac{1}{2}\eta'\Bigr\} \\ &= O(n^{-\lambda}), \end{aligned}$$

the last line following from (3.7).

Observe that since $h, h_1 \le 1$ and K, L vanish outside $(-1, 1)$,

$$|\hat{m}^*|, |\tilde{m}| \le \max_{1\le i\le n} |X_i| + 2.$$

Therefore, if $\alpha > 0$, $\lambda > \alpha + 1$ and $P'(|\hat{m}^* - \tilde{m}| > \eta) = O(n^{-\lambda})$,

$$\begin{aligned} &\sum_{i=1}^{\infty} n^{\alpha} E'\bigl\{|\hat{m}^* - \tilde{m}|^p I\bigl(|\hat{m}^* - \tilde{m}| > \eta\bigr)\bigr\} \\ &\quad \le \sum_{i=1}^{\infty} n^{\alpha}\Bigl(2 \max_{1\le i\le n} |X_i| + 4\Bigr)^p P'\bigl(|\hat{m}^* - \tilde{m}| > \eta\bigr) \\ &\quad = O\Bigl(\sum_{i=1}^{\infty} n^{\alpha-\lambda} \sum_{i=1}^{n} |X_i|^p\Bigr) = O\Bigl(\sum_{i=1}^{\infty} n^{\alpha-\lambda+1} |X_n|^p\Bigr), \end{aligned}$$

with probability 1. Since $E|X|^{\varepsilon} < \infty$, then with probability 1, $|X_n| \le n^{2/\varepsilon}$ for all sufficiently large n. Therefore, if $\lambda > \alpha + (2p/\varepsilon) + 2$,

$$\sum_{i=1}^{\infty} n^{\alpha} E'\bigl\{|\hat{m}^* - \tilde{m}|^p I\bigl(|\hat{m}^* - \tilde{m}| > \eta\bigr)\bigr\} < \infty$$

with probability 1. It follows that

$$E'\bigl\{|\hat{m}^* - \tilde{m}|^p I\bigl(|\hat{m}^* - \tilde{m}| > \eta\bigr)\bigr\} = O(n^{-\alpha}).$$

Since $\alpha > 0$ may be chosen arbitrarily large, then (3.1) is proved. □

4. Numerical results. We applied the methods in Section 3 to three different distributions, denoted by D_1, D_2 and D_3. With $D(\alpha, \mu, \sigma^2)$ representing the normal mixture $\alpha N(-\mu, \sigma^2) + (1 - \alpha)N(\mu, \sigma^2)$, the three distributions were the standard normal $D_1 = D(1, 0, 1)$, $D_2 = D(0.4, 1, \sigma_2^2)$ and $D_3 = D(0.4, 0.75, \sigma_3^2)$. In each case, we computed the bootstrap estimators $\hat{\mu}_p(h)$ on a grid of h-values and minimised over h. We chose $\sigma_2 = 0.6$ and $\sigma_3 = 0.8$, which are the unique values such that D_2 and D_3 have unit variance. With this selection D_2 is markedly bimodal, with the unique largest peak being on the right at 0.98, and D_3 is unimodal and skewed to the right, with a relatively flat top and the mode at 0.49. For the sake of brevity we focus on describing the distance of $\hat{m}$ from m. Alternative topics would include the distance of $\hat{\mu}_p(h)$ from $\mu_p(h)$, of $\hat{h}_0$ from h_0 and of $E|\hat{m} - m|^p$ from the minimum of $\mu_p(h)$ (with $\hat{h}_0$ employed to construct this particular $\hat{m}$). However, we regard $|\hat{\mu} - \mu|$ as the most informative measure of the quality of the mode estimator.

We took $n = 50$, 100 or 200 for all three distributions. The bootstrap procedure from Section 3 was implemented with the oversmoothing bandwidth $h_1 = cn^{-1/(2r+7)}$, for a variety of values of c. We used the standard normal kernel, so that $r = 2$.

From the point of view of mode estimation, the distributions D_1, D_2 and D_3 are increasingly sensitive to choice of the value of c in the oversmoothing bandwidth, and the modes are increasingly difficult to estimate. These features are reflected in our simulation study.

We found that the amount of oversmoothing which gives good performance for D_1 is significantly more than is appropriate for either D_2 or D_3. In particular, in the case of D_1, mean squared error decreases monotonically with increasing oversmoothing, over quite a wide range. The value $c = 2.5$ or 3 provides performance close to the best obtainable. The favourable effect of strong oversmoothing can be explained by D_1 being symmetric and unimodal; mode and median coincide.

For skewed and/or multimodal distributions, such as D_2 and D_3, the choice of c is much more delicate. Strong oversmoothing may shift the mode of the resampling distribution away from the true mode. Another effect, even more severe, showed in the simulations for distribution D_2. If we smooth too much, then the heights of the two peaks in the empirical study are quite close to each other, and the positions of the highest and the lowest peaks in the bootstrap resamples may occasionally interchange, leading to a serious degradation of performance of the mode estimator. The distribution D_3 is even more sensitive to smoothing, owing to its "flat top" characteristic. Even a small degree of smoothing can result in changing the estimated mode location by a significant amount. In the case of D_2 and D_3, the optimum value of c is near 1 or 1.5, much lower than for D_1. Values larger than 2.5 lead to a dramatic increase in the mean squared error.

Even with careful choice of c, the mean squared error of the mode estimator increases steadily as we pass from D_1 to D_2 and then to D_3. These properties are apparent from Table 1 and Figure 1, which summarise our

TABLE 1

Standard deviations, biases and mean squared errors of mode estimators. Within each box the first row gives Monte Carlo approximations to standard deviation, bias and mean squared error; and the second row gives approximations (in parentheses) to the standard errors of quantities in the first row. Distributions D_1, D_2 and D_3 are defined in the text. Each Monte Carlo approximation was computed by averaging over M independently simulated samples, where $M = 1000$ when $n = 50$ and $M = 500$ for $n = 100$ and 200. Throughout, $B = 50$ bootstrap simulations were used

(i) Distribution D_1

	$c = 2.5$			$c = 3$		
n	sd	bias	mse	sd	bias	mse
50	0.155	1.72×10^{-3}	2.40×10^{-2}	0.153	-6.16×10^{-3}	2.35×10^{-2}
	(0.004)	(4.90×10^{-3})	(1.2×10^{-3})	(0.004)	(4.84×10^{-3})	(1.2×10^{-3})
100	0.127	3.20×10^{-4}	1.62×10^{-2}	0.119	-4.16×10^{-3}	1.41×10^{-2}
	(0.005)	(5.70×10^{-3})	(1.2×10^{-3})	(0.004)	(5.30×10^{-3})	(9.0×10^{-4})
200	8.87×10^{-2}	-4.16×10^{-3}	7.88×10^{-3}	8.80×10^{-2}	-3.76×10^{-3}	7.77×10^{-3}
	(2.9×10^{-3})	(3.97×10^{-3})	(5.0×10^{-4})	(3.23×10^{-2})	(3.94×10^{-3})	(6.0×10^{-4})

(ii) Distribution D_2

	$c = 1$			$c = 1.5$		
n	sd	bias	mse	sd	bias	mse
50	0.377	-7.38×10^{-2}	0.147	0.299	-0.227	0.141
	(0.021)	(1.19×10^{-2})	(0.017)	(0.009)	(0.009)	(0.008)
100	0.294	-1.86×10^{-2}	8.65×10^{-2}	0.225	-9.44×10^{-2}	5.95×10^{-2}
	(0.029)	(1.31×10^{-2})	(1.71×10^{-2})	(0.007)	(1.01×10^{-2})	(4.1×10^{-3})
200	0.247	1.71×10^{-2}	6.13×10^{-2}	0.160	-5.76×10^{-3}	2.56×10^{-2}
	(0.029)	(1.10×10^{-2})	(1.39×10^{-2})	(0.006)	(7.14×10^{-3})	(1.8×10^{-3})

(iii) Distribution D_3

	$c = 1$			$c = 1.5$		
n	sd	bias	mse	sd	bias	mse
50	0.393	-0.187	0.190	0.264	-0.220	0.118
	(0.009)	(0.012)	(0.008)	(0.007)	(0.008)	(0.005)
100	0.352	-0.149	0.146	0.243	-0.220	0.107
	(0.012)	(0.016)	(0.009)	(0.009)	(0.011)	(0.006)
200	0.286	-0.119	9.63×10^{-2}	0.196	-0.183	7.49×10^{-2}
	(0.009)	(0.013)	(5.7×10^{-3})	(0.010)	(0.009)	(3.3×10^{-3})

main numerical results. To clearly define the quantities appearing in the table, let $\hat{m}_l, 1 \le l \le M$, denote the value of the mode estimator computed from the lth sample $\mathscr{X}_l$ drawn from a given distribution with mode m and set $\hat{m}_\cdot = M^{-1}\sum_l \hat{m}_l$. We took $B = 50$ bootstrap resamples throughout and either $M = 1000$ (when $n = 50$) or $M = 500$ (when $n = 100$ or 200). Each box in the

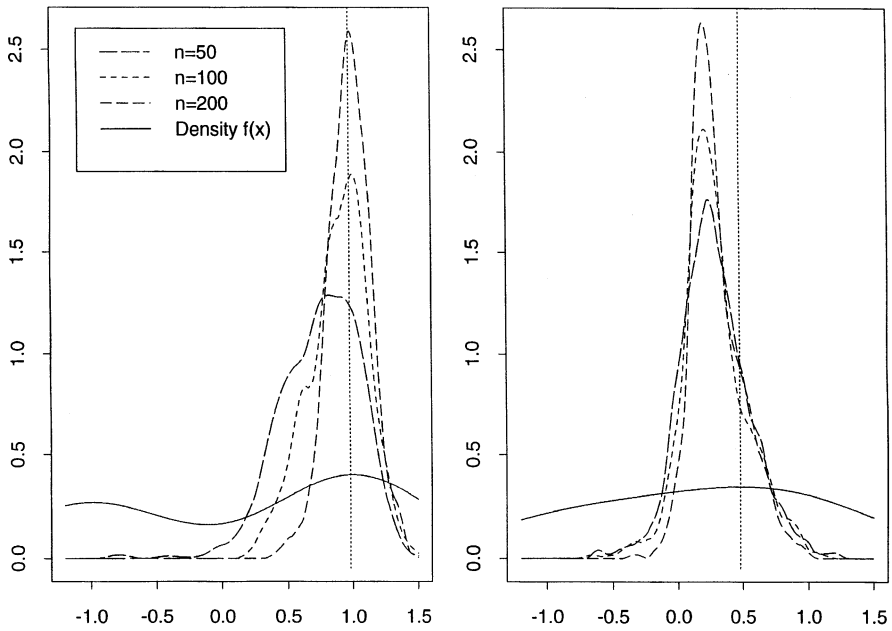

FIG. 1. *Density of $\hat{m}$. Panels* (a) *and* (b) *depict estimated densities of $\hat{m}$ under the distributions D_2 and D_3, respectively. The densities were estimated by the kernel method, based on the same simulated values for $\hat{m}$ as Table* 1. *The underlying densities D_2 and D_3 are shown as solid lines. The dotted vertical lines indicate the positions of the true global modes.*

table gives the values of

$$\begin{array}{ccc} \mathrm{sd}_0 & \mathrm{bias}_0 & \mathrm{mse}_0 \\ (e_1) & (e_2) & (e_3), \end{array}$$

where $\mathrm{sd}_0 = \{M^{-1}\Sigma_l(\hat{m}_l - \hat{m}_\cdot)^2\}^{1/2}$, $\mathrm{bias}_0 = \hat{m}_\cdot - m$ and $\mathrm{mse}_0 = (\mathrm{sd}_0)^2 + (\mathrm{bias}_0)^2$ are Monte Carlo approximations to the true standard deviation, bias and mean squared error and e_1, e_2 and e_3 are Monte Carlo approximations to the standard errors of sd_0, bias_0 and mse_0, respectively.

To provide further information about the distribution of bootstrap mode estimates, Figure 1 illustrates the density of $\hat{m}$ for the distribution of D_2 [panel (a)] and D_3 [panel (b)] and all three sample sizes. The densities were estimated by the kernel method and were based on the same simulated data values $\hat{m}_l$, $1 \le l \le M$, as Table 1. The figure illustrates properties noted two paragraphs earlier. In panel (a), the distribution of the mode estimator is skewed toward the secondary mode, with noticeable shoulders for lower sample sizes. In panel (b), even the peak of the mode estimator distribution is shifted toward the shoulder of the underlying distribution. The reason is not a possibly poor choice of h_1, since variance and squared bias appear to be balanced. However, it indicates particular difficulties in estimating modes of densities with flat tops.

For the sake of comparison with the theory we also calculated first-order asymptotic values of sd_0, bias_0 and mse_0, using the formulae

$$\begin{aligned}\mathrm{sd}_{\mathrm{as}} &= (2\pi^{1/4})^{-1}\bigl(nh_1^3\bigr)^{-1/2} f(m)^{1/2}|f''(m)|^{-1},\\ \mathrm{bias}_{\mathrm{as}} &= \tfrac{1}{2}h_1^2 f'''(m)|f''(m)|^{-1} \text{ and}\\ \mathrm{mse}_{\mathrm{as}} &= (\mathrm{sd}_{\mathrm{as}})^2 + (\mathrm{bias}_{\mathrm{as}})^2.\end{aligned}$$

Agreement with the "true" values derived by Monte Carlo simulation was not high, being in error by several fold in some instances. The closest agreement of mean squared error was in the case of the distribution D_3, where true and theoretical values differed by only 10% when $c = 1$. Overall, the difficulty in obtaining good agreement between theory and "practice" for the parameter settings treated here confirms our belief that particularly large sample sizes are required for accurate and reliable estimation of a mode.

APPENDIX

Before passing to a proof of Theorem 2.3, we note the following lemma. The first and last parts of the lemma may be derived using methods employed to establish Theorem 2.1. The second part follows via Bernstein's inequality.

LEMMA A.1. *Assume the conditions of Theorem* 2.3, *preceding* (2.3). *Then for* $\eta > 0$ *sufficiently small and each* $\lambda > 0$, *we may choose* $\xi = \xi(\lambda) > 0$ *so large that*

$$P\biggl\{\sup_{x:\,|x-m|\le\eta} |\hat f'''(x) - E\hat f'''(x)| > \xi\log n\biggr\} = O(n^{-\lambda}),$$

and for $j = 1, 2$ *and all* $\xi > 0$,

$$P\bigl\{|\hat f^{(j)}(m) - E\hat f^{(j)}(m)| > \xi\bigr\} = O(n^{-\lambda}).$$

Additionally, *under the conditions preceding* (2.4), *we may choose* $\xi = \xi(\lambda)$ *so large that*

$$P\biggl\{\sup_{x:\,|x-m|\le\eta} |\hat f^{(4)}(x) - E\hat f^{(4)}(x)| > \xi\bigl(nh^9\bigr)^{-1/2}\log n\biggr\} = O(n^{-\lambda}).$$

PROOF OF THEOREM 2.3. By Taylor expansion

$$0 = \hat f'(\hat m) = \hat f'(m) + (\hat m - m)\hat f''(m) + \tfrac{1}{2}(\hat m - m)^2\hat f'''\{m + \theta_1(\hat m - m)\},$$

where $0 \le \theta_1 \le 1$. From this formula we may conclude that if $|\hat m - m| \le \eta$ and, for a constant $C_1 > 0$,

$$\begin{aligned}&\sup_{x:\,|x-m|\le\eta} |\hat f'''(x)| \le C_1\log n \quad\text{and}\\ &|\hat f'(m)|\hat f''(m)^{-2} \le (20C_1\log n)^{-1},\end{aligned} \tag{A.1}$$

then with $T = \hat{f}'''\{m + \theta_1(\hat{m} - m)\}$, which satisfies $|T| \le C_1 \log n$,

$$\begin{aligned}\hat{m} - m &= \hat{f}''(m)\Big[\big\{1 - 2\hat{f}'(m)\hat{f}''(m)^{-2}T\big\}^{1/2} - 1\Big]T^{-1} \\ &= -\hat{f}'(m)\hat{f}''(m)^{-1} - \tfrac{1}{2}\hat{f}'(m)^2 T\hat{f}''(m)^{-3} + R_1,\end{aligned}$$

where

$$|R_1| \le |\hat{f}''(m)|\,|\hat{f}'(m)\hat{f}''(m)^{-2}T|^3|T|^{-1}.$$

Hence, if $f''(m) \ne 0$ and $|\hat{f}''(m) - f''(m)| \le \frac{1}{2}|f''(m)|$, then

$$\begin{aligned}\hat{m} - m = {} & -\hat{f}'(m)f''(m)^{-1} + \hat{f}'(m)\big\{\hat{f}''(m) - f''(m)\big\}f''(m)^{-2} \\ & - \hat{f}'(m)\big\{\hat{f}''(m) - f''(m)\big\}^2 f''(m)^{-3} \\ & - \tfrac{1}{2}\hat{f}'(m)^2 f'''(m)f''(m)^{-3} \\ & - \tfrac{1}{2}\hat{f}'(m)^2\{T - f'''(m)\}f''(m)^{-3} + R_2,\end{aligned} \tag{A.2}$$

where

$$|R_2| \le C_2\big\{|\hat{f}'(m)|^3 + |\hat{f}'(m)|\,|\hat{f}''(m) - f''(m)|^3\big\}(\log n)^2$$

and C_2 depends only on C_1 and $|f^{(j)}(m)|$, $j = 2, 3$.

First we derive (2.3). In view of Lemma A.1, if C_1 is sufficiently large and η sufficiently small, then for all $\lambda > 0$,

$$\begin{aligned}&P\Big\{\sup_{x:\,|x-m|\le\eta}|\hat{f}'''(x)| > C_1\log n\Big\} + P\big\{|\hat{f}'(m)|\hat{f}''(m)^{-2} > (20C_1\log n)^{-1}\big\} \\ &\quad + P\big\{|\hat{f}''(m) - f''(m)| > \tfrac{1}{2}|f''(m)|\big\} = O(n^{-\lambda}).\end{aligned}$$

Therefore,

$$\begin{aligned}&\Big|\big\{E|\hat{m} - m|^p I(|\hat{m} - m| \le \eta)\big\}^{1/p} \\ &\quad - \Big[E\Big|\hat{f}'(m)f''(m)^{-1} - \hat{f}'(m)\big\{\hat{f}''(m) - f''(m)\big\}f''(m)^{-2} \\ &\qquad + \hat{f}'(m)\big\{\hat{f}''(m) - f''(m)\big\}^2 f''(m)^{-3} \\ &\qquad\quad - \tfrac{1}{2}\hat{f}'(m)^2 f'''(m)f''(m)^{-3}\Big|^p I(|\hat{m} - m| \le \eta)\Big]^{1/p}\Big| \\ &\quad = O\Big[\big\{E|\hat{f}'(m)|^{2p}\big\}^{1/p}\log n \\ &\qquad + \big\{E|\hat{f}'(m)|^{2p}E|\hat{f}''(m) - f''(m)|^{6p}\big\}^{1/(2p)}(\log n)^2\Big] \\ &\quad = O\Big[\big\{(nh^3)^{-1} + \alpha^2\big\}\log n \\ &\qquad + \big\{(nh^3)^{-1} + \alpha^2\big\}^{1/2}\big\{(nh^5)^{-1} + \beta^2\big\}^{3/2}(\log n)^2\Big]\end{aligned} \tag{A.3}$$

$$= O\Big[\Big\{(nh^3)^{-1/2} + \alpha\Big\} \times \Big\{(nh^3)^{-1/2} + (nh^5)^{-3/2} + \alpha + \beta^3\Big\}(\log n)^2\Big].$$

In view of Theorem 2.2, and the fact that for $j \le 3$ and $q \ge 1$, $E|\hat{f}^{(j)}(m)|^q$ is bounded in n, the indicator function may be dropped throughout the left-hand side above. Therefore,

$$(E|\hat{m} - m|^p)^{1/p} - \Big[E\Big|\hat{f}'(m)f''(m)^{-1} - \hat{f}'(m)\{\hat{f}''(m) - f''(m)\}f''(m)^{-2} + \hat{f}'(m)\{\hat{f}''(m) - f''(m)\}^2 f''(m)^{-3} - \tfrac{1}{2}\hat{f}'(m)^2 f'''(m) f''(m)^{-3}\Big|^p\Big]^{1/p}$$
$$= O\Big[\Big\{(nh^3)^{-1/2} + \alpha\Big\}\Big\{(nh^3)^{-1/2} + (nh^5)^{-3/2} + \alpha + \beta^3\Big\}(\log n)^2\Big].$$

The proof of (2.3) may be completed by applying a result on the rate of convergence of moments in the bivariate central limit theorem; see Bhattacharya and Rao [(1976), Theorem 15.1, page 145].

Next we derive (2.4). The argument is similar to that above, except that in treating the term $\frac{1}{2}\hat{f}'(m)^2\{T - f'''(m)\}f''(m)^{-2}$ in (A.2) we divide it into two parts: $S_1 = \frac{1}{2}\hat{f}'(m)^2\{\hat{f}'''(m) - f'''(m)\}f''(m)^{-2}$ and $S_2 = \frac{1}{2}\hat{f}'(m)^2\{T - \hat{f}'''(m)\}f''(m)^2$. The second of these goes into the remainder, and for the purpose of treating it, we add to the sequence of events at (A.1) the requirement that

$$\sup_{x:\,|x-m|\le\eta} |\hat{f}^{(4)}(x) - E\hat{f}^{(4)}(x)| \le (nh^9)^{-1/2}\log n.$$

In view of the last part of Lemma A.1, this inequality holds with probability $1 - O(n^{-\lambda})$ for all $\lambda > 0$. When it is valid,

$$|S_2| I(|\hat{m} - m| \le \eta) \le C_3|\hat{f}'(m)|^2\Big\{1 + (nh^9)^{-1/2}\log n\Big\}|\hat{m} - m| I(|\hat{m} - m| \le \eta).$$

Arguing thus we may deduce instead of (A.3) that

$$\Big|\big\{E|\hat{m} - m|^p I(|\hat{m} - m| \le \eta)\big\}^{1/p} - \Big[E\Big|\hat{f}'(m)f''(m)^{-1} - \hat{f}'(m)\{\hat{f}''(m) - f''(m)\}f''(m)^{-2} + \hat{f}'(m)\{\hat{f}''(m) - f''(m)\}^2 f''(m)^{-3} - \tfrac{1}{2}\hat{f}'(m)^2 f'''(m) f''(m)^{-3} - \tfrac{1}{2}\hat{f}'(m)^2\{\hat{f}'''(m) - f'''(m)\}f''(m)^{-3}\Big|^p I(|\hat{m} - m| \le \eta)\Big]^{1/p}\Big|$$

$$= O\Big[\big\{E|\hat{f}'(m)|^{3p}\big\}^{1/p} + \big\{E|\hat{f}'(m)|^{2p}E|\hat{f}''(m) - f''(m)|^{6p}\big\}^{1/(2p)}(\log n)^2$$
$$+\big\{E|\hat{f}'(m)|^{4p}E|\hat{m} - m|^{2p}I(|\hat{m} - m| \le \eta)\big\}^{1/(2p)}$$
$$\times\Big\{1 + (nh^9)^{-1/2}\log n\Big\}\Big]$$
$$= O\Big[(nh^3)^{-3/2} + \alpha^3 + \Big\{(nh^3)^{-1} + \alpha^2\Big\}^{1/2}\Big\{(nh^5)^{-1} + \beta^2\Big\}^{3/2}(\log n)^2$$
$$+\Big\{(nh^3)^{-3/2} + \alpha^3\Big\}\Big\{1 + (nh^9)^{-1/2}\log n\Big\}\Big]$$
$$= O\Big(\Big\{(nh^3)^{-1/2} + \alpha\Big\}$$
$$\times\Big[(nh^3)^{-1} + (nh^5)^{-3/2} + \alpha^2\Big\{1 + (nh^9)^{-1/2}\Big\} + \beta^3\Big](\log n)^2\Big).$$

Result (2.4) follows from this formula in the same way that (2.3) did from (A.3). □

Acknowledgments. We are grateful to three referees and an Associate Editor for their helpful and constructive criticism.

REFERENCES

BHATTACHARYA, R. N. and RAO, R. R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York.

DEVROYE, L. (1985). *Nonparametric Density Estimation: The L^1 View*. Wiley, New York.

EDDY, W. (1980). Optimal kernel estimators of the mode. *Ann. Statist.* **8** 870–882.

EDDY, W. (1982). The asymptotic distributions of kernel estimators of the mode. *Z. Wahrsch. Verw. Gebiete* **59** 279–290.

FARAWAY, J. J. and JHUN, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85** 1119–1122.

FISHER, N. I., MAMMEN, E. and MARRON, J. S. (1994). Testing for multimodality. *Comput. Statist. Data Anal.* **18** 499–512.

GARSIA, A. M. (1970). Continuity properties of Gaussian processes with multidimensional time parameter. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **2** 269–274. Univ. California Press, Berkeley.

GRENANDER, U. (1965). Some direct estimates of the mode. *Ann. Math. Statist.* **36** 131–138.

HALL, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.* **32** 177–203.

HALL, P., MARRON, J. S. and PARK, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* **92** 1–20.

HALL, P. and MURISON, R. D. (1991). Correcting the negativity of high-order kernel density estimators, Report CMA-SR21-91, Centre for Mathematics and its Applications, Australian National Univ.

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent rv's and the sample df. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131.

MAMMEN, E., MARRON, J. S. and FISHER, N. I. (1992). Some asymptotics for multimodality tests based on kernel estimates. *Probab. Theory Related Fields* **91** 115–132.

MÜLLER, H.-G. (1989). Adaptive nonparametric peak regression. *Ann. Statist.* **17** 1053–1069.

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

ROMANO, J. P. (1988a). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647.

ROMANO, J. P. (1988b). Bootstrapping the mode. *Ann. Inst. Statist. Math.* **40** 565–586.

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6** 177–184.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

TAYLOR, C. C. (1989). Boostrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **76** 705–712.

TSYBAKOV, A. B. (1990). Recursive estimation of the mode of a multivariate distribution. *Problems Inform. Transmission* **26** 31–37.

DEPARTMENT OF APPLIED STATISTICS
UNIVERSITY OF MINNESOTA
352 CLASSROOM-OFFICE BUILDING
ST. PAUL, MINNESOTA 55108-6042

CENTRE FOR MATHEMATICS AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
G.P.O. BOX 4
CANBERRA ACT 2601
AUSTRALIA