ON THE MINIMIZATION OF L^p ERROR IN MODE ESTIMATION

By

Birgit Grund, Dept. of Applied Statistics, University of Minnesota
Peter Hall, Mathematical Sciences Research Institute, Berkeley[1]

School of Statistics, University of Minnesota
Technical Report No. 592

August, 1993

[1] On leave from Centre for Mathematics and its Applications, Australian National University, Canberra, Australia.

ON THE MINIMIZATION OF L^p ERROR IN MODE ESTIMATION

Birgit Grund
University of Minnesota

Peter Hall
Mathematical Sciences Research Institute, Berkeley[1]

Abstract. 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 p'th moment of the sampling 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 focused 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 analyzed, 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 minimized 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.

Short Title. Mode estimation.

AMS Subject Classification. 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.

[1] On leave from Centre for Mathematics and its Applications, Australian National University, Canberra, Australia.

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 (1988) has surveyed subsequent work, including that of Eddy (1980, 1982) on kernel estimation, and of Grenander (1965) on alternative approaches. See also Tsybakov (1990). 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 (1993) 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 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 minimizing 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 a necessary and sufficient condition for L^p convergence of the mode estimator is the existence of finite p'th 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, this work requires only finite ϵ'th moment for some $\epsilon > 0$; it does not need finite p'th moment. The apparent contradiction arises because extreme values from a bootstrap resample have properties quite unlike those of extremes from the actual population. The requirement of finite p'th 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 r'th order kernel estimator $\hat{f}$ 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 summarized in Section 4.

By way of notation, $\mathcal{X} = \{X_1, \ldots, X_n\}$ 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 $\widehat{m}$ denote any quantity with the property

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

Section 2 will discuss the issue of ties for $\widehat{m}$. 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, which has order $r = 2$.

2. Convergence in probability, and in L^p, of the mode estimator. In the sense of large deviations, $\widehat{m}$ 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*

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

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

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

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

$$P(|\widehat{m} - m| > \eta) = O(n^{-\lambda})$$

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, $\widehat{m} \to m$ in probability. However, without additional regularity conditions on the tails of the sampling distribution, there can be no guarantee that $\widehat{m}$ will converge to m in any L^p metric. In part, this problem is caused by ambiguities in how $\widehat{m}$ 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 $\widehat{m}$ 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 $X_{(1)} \leq \ldots \leq X_{(n)}$, and let $x_1 < \ldots < x_n$ denote real numbers such that $x_i - x_{i-1} - 2 > 0$, $x_n > 2$ 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 $\mathcal{E}_n$ 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 $\mathcal{E}_n$ prevails, the kernel estimator $\hat{f}$ is simply a string of n non-overlapping "bumps", 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 farthest to the right. Then,

$$\begin{aligned} E|\widehat{m}|^p &\geq E\{|\widehat{m}|^p \, I(\mathcal{E}_n)\} \geq E\{(X_{(n)} - 1)^p \, I(\mathcal{E}_n)\} \\ &\geq E\{(X - 1)^p \, I(X > x_n - 1)\} \prod_{i=1}^{n-1} P\{X \in (x_i - 1, x_i + 1)\} . \end{aligned}$$

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 $E|\widehat{m}|^p < \infty$. 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 our next theorem points out.

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 $\widehat{m}$. Let $p \geq 1$. Then*

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

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

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

Our final result in this section describes an asymptotic formula for $E|\widehat{m} - m|^p$. 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 $\alpha = |E\hat{f}'(m)|$, $\beta = |E\hat{f}''(m) - f''(m)|$.

THEOREM 2.3. *Assume that f has a "unique largest peak" at m; that f''' exists in a neighborhood 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 for some $\eta > 0$,*

$$\sup_{n \geq 0} (nh^5)^{-1} (\log n)^{1+\eta} < \infty .$$

Then for each $p \geq 1, r \geq 3$,

$$\begin{aligned}(E|\widehat{m} - m|^p)^{1/p} = \{E|N_1 - N_1 N_2 \, f''(m)^{-1} + N_1 N_2^2 \, f''(m)^{-2}|^p\}^{1/p} \, |f''(m)|^{-1} \\ + O\big[\{(nh^3)^{-1/2} + \alpha\}\{(nh^3)^{-1/2} + (nh^5)^{-3/2} + \alpha + \beta^3\}\big],\end{aligned} \tag{2.3}$$

and for each $p \geq 1, r \geq 2$,

$$(E|\widehat{m} - m|^p)^{1/p} = \{E|N_1 - N_1 N_2\, f''(m)^{-1} + N_1 N_2^2\, f''(m)^{-2} + \tfrac{1}{2}\, N_1^2\, f'''(m) f''(m)^{-2} + \tfrac{1}{2}\, N_1^2 N_3\, f''(m)^{-2}|^p\}^{1/p}\, |f''(m)|^{-1} + O\big[\{(nh^3)^{-1/2} + \alpha\}\{(nh^3)^{-1} + (nh^5)^{-3/2} + \alpha^2 + \beta^3\}\big(\log n\big)^2\big] \quad (2.4)$$

as $n \to \infty$.

Formula (2.4) implies that

$$(E|\widehat{m} - m|^p)^{1/p} \sim \big[E|\{\mathrm{var}\, \hat{f}'(m)\}^{1/2}\, N + E\hat{f}'(m)|^p\big]^{1/p}\, |f''(m)|^{-1}\,,$$

where N denotes a Standard Normal random variable. This weaker form of (2.4) may be used to derive first-order asymptotic properties of the bandwidth that is optimal for minimizing $E|\widehat{m} - m|^p$. To appreciate this point, suppose that K is an r'th order kernel; i.e., for some $r \geq 2$,

$$\int y^j\, K(y)\, dy = \begin{cases} 1 & \text{if} \quad j = 0 \\ 0 & \text{if} \quad 1 \leq j \leq r-1 \\ (-1)^r\, r!\, \kappa \neq 0 & \text{if} \quad j = r\,. \end{cases} \quad (2.5)$$

Assume, in addition to the conditions of Theorem 2.3, that $f^{(r+1)}$ exists in a neighborhood of m and is continuous at m. Then,

$$\mathrm{var}\, \hat{f}'(m) = (nh^3)^{-1}\, f(m) \int (K')^2 + o\{(nh^3)^{-1}\}\,, \quad (2.6)$$
$$E\hat{f}'(m) = \kappa h^r\, f^{(r+1)}(m) + o(h^r)\,.$$

If we define $c_1^2 = f(m) \int (K')^2$ and $c_2 = |\kappa f^{(r+1)}(m)|$, we have

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

It follows that the bandwidth h_0 that minimizes $A_1(h) \equiv \{E|\widehat{m} - m|^p\}^{1/p}$, 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 minimizes

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

and that for this choice of h,

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

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 $r \geq 2$,*

$$h_0 = u_0\, n^{-1/(2r+3)}\, \{1 + o(n^{-2/(2r+7)})\}\,. \tag{2.7}$$

The proof for even $r \geq 4$ is outlined below. It follows directly from (2.3). The case of odd $r \geq 3$ may be treated similarly, but requires more care. For $r = 2$ the remainder term "$O(\ldots)$" in (2.3) is not sufficiently small to yield the desired result, and that case should be treated separately. However, using (2.4) instead of (2.3) we may show that when $r = 2$, (2.7) follows as before.

Proof of Lemma 2.1. We show (2.7) for even $r \geq 4$. For symmetric K, the distribution of (N_1, N_2) may be elucidated relatively easily. It may be shown that the quantity h_0' that minimizes $A_2(h) \equiv \{E|N_1 - N_1 N_2\, f''(m)^{-1} + N_1 N_2^2\, f''(m)^{-2}|^p\}^{1/p}$ satisfies $h_0'/(u_0\, n^{-1/(2r+3)}) = 1 + O(n^{-2/(2r+5)}) = 1 + o(n^{-2/(2r+7)})$. Similarly, noting the usual quadratic Taylor expansion in the neighborhood of a minimum,

$$A_2(h_0') = G(u_0)^{1/p}\, n^{-r/(2r+3)} \{1 + o(n^{-4/(2r+7)})\}\,. \tag{2.8}$$

Write $h_0 = h_0'(1 + \delta)$, where $\delta = \delta(n) \to 0$. Again noting the quadratic Taylor expansion in the neighborhood of a minimum, we see that

$$A_1(h_0')/A_1(h_0) = 1 + C\delta^2 + o(\delta^2)\,, \tag{2.9}$$

where $C > 0$. 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),

$$\begin{aligned} A_1(h_0') &= A_2(h_0')\{1 + O(n^{-r/(2r+3)})\} \\ &= A_2(h_0')\{1 + o(n^{-4/(2r+7)})\}, \end{aligned}$$

the last identity requiring only $r \geq 3$. Hence, by (2.8),

$$A_1(h_0') = G(u_0)^{1/p}\, n^{-r/(2r+3)}\{1 + o(n^{-4/(2r+7)})\},$$

which in view of (2.9) implies

$$A_1(h_0) = G(u_0)^{1/p}\, n^{-r/(2r+3)}\{1 - C\delta^2 + o(n^{-4/(2r+7)})\}.$$

Since the left-hand side is minimized with $\delta = 0$, the right-hand side must be too, which entails $\delta = o(n^{-2/(2r+7)})$. This proves (2.7). □

We conclude this section by outlining proofs of Theorems 2.1–2.3.

Proof of Theorem 2.1. Observe that

$$\begin{aligned} P(|\widehat{m} - m| > \eta) &= P\Big\{ \sup_{x:|x-m|\leq\eta} \hat{f}(x) \leq \sup_{x:|x-m|>\eta} \hat{f}(x) \Big\} \\ &\leq P\Big\{ \sup_{x:|x-m|\leq\eta} E\hat{f}(x) - \sup_{x:|x-m|\leq\eta} |\hat{f}(x) - E\hat{f}(x)| \\ &\qquad \leq \sup_{x:|x-m|>\eta} E\hat{f}(x) + \sup_{x:|x-m|>\eta} |\hat{f}(x) - E\hat{f}(x)| \Big\} \\ &\leq P\Big\{ 2 \sup_{-\infty<x<\infty} |\hat{f}(x) - E\hat{f}(x)| \\ &\qquad > \sup_{x:|x-m|\leq\eta} E\hat{f}(x) - \sup_{x:|x-m|>\eta} E\hat{f}(x) \Big\}. \end{aligned} \tag{2.10}$$

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

$$\sup_{x:|x-m|\leq\eta} E\hat{f}(x) - \sup_{x:|x-m|>\eta} E\hat{f}(x) \geq \eta' \tag{2.11}$$

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). □

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,

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

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

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

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

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

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 $Z_n = \max_{i \le n} |X_i|$,

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

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

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

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

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

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. □

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

LEMMA 2.2. *Assume the conditions of Theorem 2.3. Then for $\eta > 0$ sufficiently small, and all $\xi, \lambda > 0$,*

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

and for $j = 1, 2$,

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

Proof of Theorem 2.3. By Taylor expansion,

$$\begin{aligned} 0 = \hat{f}'(\widehat{m}) &= \hat{f}'(m) + (\widehat{m} - m)\, \hat{f}''(m) \\ &\quad + \tfrac{1}{2}\, (\widehat{m} - m)^2\, \hat{f}'''\{m + \theta(\widehat{m} - m)\}, \end{aligned}$$

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

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

then for a random variable T satisfying $|T| \le C_1 \log n$,

$$\begin{aligned} \widehat{m} - m &= \hat{f}''(m)\,[\{1 - 2\hat{f}'(m)\,\hat{f}''(m)^{-2}\,T\}^{1/2} - 1]\,T^{-1} \\ &= -\hat{f}'(m)\,\hat{f}''(m)^{-1} - \tfrac{1}{2}\,\hat{f}'(m)^2\,\hat{f}'''(m)\,\hat{f}''(m)^{-3} + R_1\,, \end{aligned} \tag{2.12}$$

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

$$\begin{aligned} \widehat{m} - m = &- \hat{f}'(m)\,f''(m)^{-1} + \hat{f}'(m)\,\{\hat{f}''(m) - f''(m)\}\,f''(m)^{-2} \\ &- \hat{f}'(m)\,\{\hat{f}''(m) - f''(m)\}^2\,f''(m)^{-3} \\ &- \tfrac{1}{2}\,\hat{f}'(m)^2\,f'''(m)\,f''(m)^{-3} \\ &- \tfrac{1}{2}\,\hat{f}'(m)^2\,\{\hat{f}'''(m) - f'''(m)\}\,f''(m)^{-3} + R_2\,, \end{aligned}$$

where

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

and C_2 depends only on C_1 and $|f^{(j)}(m)|$, $j = 2, 3$. In view of Lemma 2.2, if C_1 is sufficiently large and η sufficiently small then for all $\lambda > 0$,

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

Therefore,

$$|\{E|\widehat{m} - m|^p\, I(|\widehat{m} - m| \le \eta)\}^{1/p}$$

$$
\begin{aligned}
&- [E|\hat{f}'(m)\, f''(m)^{-1} - \hat{f}'(m)\,\{\hat{f}''(m) - f''(m)\}\, f''(m)^{-2} \\
&\qquad + \hat{f}'(m)\,\{\hat{f}''(m) - f''(m)\}^2\, f''(m)^{-3} - \tfrac{1}{2}\,\hat{f}'(m)^2\, f'''(m)\, f''(m)^{-3} \\
&\qquad - \tfrac{1}{2}\,\hat{f}'(m)^2\,\{\hat{f}'''(m) - f'''(m)\}\, f''(m)^{-3}|^p\, I(|\widehat{m} - m| \le \eta)]^{1/p}| \\
&= O[\{E|\hat{f}'(m)|^{3p}\}^{1/p} + \{E|\hat{f}'(m)|^{2p}\, E|\hat{f}''(m) - f''(m)|^{6p} (\log n)^{4p}\}^{1/(2p)}] \\
&= O[(nh^3)^{-3/2} + \alpha^3 + \{(nh^3)^{-1} + \alpha^2\}^{1/2}\,\{(nh^5)^{-1} + \beta^2\}^{3/2}\,(\log n)^2] \\
&= O[\{(nh^3)^{-1/2} + \alpha\}\,\{(nh^3)^{-1} + (nh^5)^{-3/2} + \alpha^2 + \beta^3\}\,(\log n)^2]\,.
\end{aligned}
$$

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,

$$
\begin{aligned}
&(E|\widehat{m} - m|^p)^{1/p} - [E|\hat{f}'(m)\, f''(m)^{-1} - \hat{f}'(m)\,\{\hat{f}''(m) - f''(m)\}\, f''(m)^{-2} \\
&\qquad + \hat{f}'(m)\,\{\hat{f}''(m) - f''(m)\}^2\, f''(m)^{-3} - \tfrac{1}{2}\,\hat{f}'(m)^2\, f'''(m)\, f''(m)^{-3} \\
&\qquad - \tfrac{1}{2}\,\hat{f}'(m)^2\,\{\hat{f}'''(m) - f'''(m)\}\, f''(m)^{-3}|^p]^{1/p} \\
&\quad = O[\{(nh^3)^{-1/2} + \alpha\}\,\{(nh^3)^{-1} + (nh^5)^{-3/2} + \alpha^2 + \beta^3\}\,(\log n)^2]\,.
\end{aligned}
$$

The proof of (2.4) may be completed by applying a result on the rate of convergence of moments in the bivariate central limit theorem; see Theorem 15.1, p.145 of Bhattacharya and Rao (1976).

Formula (2.3) may be proven by the same technique, just using fewer terms in the Taylor expansion for $\widehat{m} - m$ in (2.12). We omit the details. □

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

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

We showed that if K is an r'th order kernel then this quantity is asymptotically minimized by taking $h = u_0\, n^{-1/(2r+3)}$ in the definition of $\widehat{m}$. Here, u_0 minimizes the function $G(u)$, $u > 0$, which depends on the unknowns $f(m)$ 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 r'th order kernel; see (2.5) for a definition of the "r'th 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 $\mathcal{X} = \{X_1, \ldots, X_n\}$, 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 $\mathcal{X}$; $Y_1, \ldots, Y_n$ are independent and identically distributed with density L; and conditional on $\mathcal{X}$, 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 $\widehat{m}^*$ and $\widetilde{m}$ denote the modes of $\hat{f}^*$ and $\tilde{f}$, respectively, defined by breaking ties randomly when necessary; and set

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

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

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

Conditional on $\mathcal{X}$, let (N_1', N_2', N_3') denote a trivariate Normal random vector with the same conditional mean and conditional variance matrix as $(\hat{f}^{*\prime}(\widetilde{m}),\ \hat{f}^{*\prime\prime}(\widetilde{m}) - \tilde{f}''(\widetilde{m}),\ \hat{f}^{*\prime\prime\prime}(\widetilde{m}) - \tilde{f}'''(\widetilde{m}))$. Put $\alpha' = |E'\,\hat{f}^{*\prime}(\widetilde{m})|$, $\beta' = |E'\,\hat{f}^{*\prime\prime}(\widetilde{m}) - \tilde{f}''(\widetilde{m})|$.

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 neighborhood of m; that $f''(m) \neq 0$; that $E|X|^\epsilon < \infty$ for some $\epsilon > 0$; 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 $0 < h$, $h_1 \leq 1$, $h + h_1 \to 0$ and that for some $\eta > 0$,*

$$\sup_{n \geq 1} \left\{(nh^5)^{-1} + (nh_1^7)^{-1}\right\} (\log n)^{1+\eta} < \infty\,.$$

Let $p \geq 1$. Then for each $\eta, \lambda > 0$,

$$E'\{|\widehat{m}^* - \widetilde{m}|^p\, I(|\widehat{m}^* - \widetilde{m}| > \eta)\} = O(n^{-\lambda}) \tag{3.1}$$

with probability one, and for $r \geq 3$,

$$\begin{aligned}(E'|\widehat{m}^* - \widetilde{m}|^p)^{1/p} = {} & \{E'|N_1' - N_1' N_2'\, \tilde{f}''(\widetilde{m})^{-1} + N_1' N_2'^2\, \tilde{f}''(\widetilde{m})^{-2}|^p\}^{1/p} \\ & \times |\tilde{f}''(\widetilde{m})|^{-1} + O[\{(nh^3)^{-1/2} + \alpha'\} \\ & \times \{(nh^3)^{-1/2} + (nh^5)^{-3/2} + \alpha' + \beta'^3\}]\end{aligned} \tag{3.2}$$

with probability one. For $r \geq 2$,

$$\begin{aligned}(E'|\widehat{m}^* - \widetilde{m}|^p)^{1/p} = {} & \{E'|N_1' - N_1' N_2'\, \tilde{f}''(\widetilde{m})^{-1} + N_1' N_2'^2\, \tilde{f}''(\widetilde{m})^{-2} \\ & + \tfrac{1}{2}\, N_1'^2\, \tilde{f}'''(\widetilde{m})\, f''(\widetilde{m})^{-2} + \tfrac{1}{2}\, N_1'^2\, N_3'\, \tilde{f}''(\widetilde{m})^{-2}|^p\}^{1/p} \\ & \times |\tilde{f}''(\widetilde{m})|^{-1} + O[\{(nh^3)^{-1/2} + \alpha'\} \\ & \times \{(nh^3)^{-1} + (nh^5)^{-3/2} + \alpha'^2 + \beta'^2\}\,(\log n)^2]\end{aligned} \tag{3.3}$$

with probability one.

Our proof of Theorem 3.1 remains valid if we take h to be a function of the data, $\mathcal{X}$. 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 one, and $\sup(nh^5)^{-1}\,(\log n)^{1+\eta} < \infty$ with probability one.

The principal application of the bootstrap estimator $\hat{\mu}_p(h)$ is to calculating an empirical version of the bandwidth h_0 that minimizes $\mu_p(h)$. We discussed h_0 briefly in Section 2, where we showed that if K is an r'th 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 minimizes

$$G(u) = E|c_1\, u^{-3/2}\, N + c_2\, u^r|^p\,, \qquad u > 0\,,$$

where $c_1 = \{f(m)\int (K')^2\}^{1/2}$ and $c_2 = |\kappa\, f^{(r+1)}(m)|$. 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 minimizes $\hat{\mu}_p(h)$ is given by

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

with probability one, where $\hat{u}_0$ minimizes

$$\widehat{G}(u) = E'|\hat{c}_1\, u^{-3/2}\, N' + \hat{c}_2\, u^r|^p\,, \qquad u > 0\,,$$

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 $\mathcal{X}$. A formal derivation of this result requires $\tilde{f}^{(r+1)}$ to be strongly consistent for $f^{(r+1)}$ in a neighborhood 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 minimizes 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 $j \leq 2r+3$, 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 $\widetilde{m} - m$ are respectively of size $(nh_1)^{-1/2} + h_1^2$, $(nh_1^{2r+3})^{-1/2} + h_1^2$, $(nh_1^3)^{-1/2} + h_1^2$. 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 $\widehat{G}$ 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 minimized 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 minimizes the mean squared error of $\tilde{f}^{(r+1)}(m)$.

We conclude this section with a derivation of Theorem 3.1.

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, 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 $\mathcal{X}$, and the latter identity is interpreted as holding with probability one. From (3.6) it follows, via the Borel-Cantelli lemma, that

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

with probability one. 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 one. Replacing $(K, h, \hat{f})$ by $(L, h_1, \tilde{f})$ in Theorem 2.1, we deduce that $\widetilde{m} \to m$ with probability one. Noting the remark that contains (2.11) we conclude that for each $\eta > 0$ there exists $\eta' > 0$ such that with probability one, for all sufficiently large n,

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

Arguing as at (2.10) we may now deduce that with probability one, and for all sufficiently large n,

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

the last line following from (3.7).

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

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

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

$$\begin{aligned}
&\sum_{i=1}^{\infty} n^{\alpha} E'\{|\widehat{m}^* - \widetilde{m}|^p I(|\widehat{m}^* - \widetilde{m}| > \eta)\} \\
&\le \sum_{i=1}^{\infty} n^{\alpha} \Big(2 \max_{1 \le i \le n} |X_i| + 4\Big)^p P'(|\widehat{m}^* - \widetilde{m}|\eta) \\
&= O\left(\sum_{i=1}^{\infty} n^{\alpha-\lambda} \sum_{i=1}^{n} |X_i|^p\right) = O\left(\sum_{i=1}^{\infty} n^{\alpha-\lambda+1} |X_n|^p\right),
\end{aligned}$$

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

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

with probability one. It follows that

$$E'\{|\widehat{m}^* - \widetilde{m}|^p I(|\widehat{m}^* - \widetilde{m}| > \eta)\} = O(n^{-\alpha}).$$

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

Proof of (3.2) and (3.3). The proof is similar to that of Theorem 2.3, and so we give it only in outline. Note that in the former proof, $(f, \hat{f})$ should be replaced by $(\tilde{f}, \hat{f}^*)$, and probability measures and expectations should be interpreted conditionally on $\mathcal{X}$. Our assumption that f''' exists and is continuous in a neighborhood of m, and that $(nh_1^7)^{-1} (\log n)^{1+\eta}$ is bounded for some $\eta > 0$, ensures that for some $\eta > 0$ and all $\xi, \lambda > 0$, and for $j = 0, \ldots, 3$,

$$P\Big\{\sup_{x:|x-m|\le\eta} |\tilde{f}^{(j)}(x) - f^{(j)}(x)| > \xi\Big\} = O(n^{-\lambda}).$$

This result may be proved much as in Silverman (1978). The version of Lemma 2.2 for this setting asks that for $\eta > 0$ sufficiently small, and for all $\xi, \lambda > 0$,

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

and for $j = 1, 2$,

$$P'\{|\hat{f}^{*(j)}(\widetilde{m}) - E'\,\hat{f}^{*(j)}(\widetilde{m})| > \xi\} = O(n^{-\lambda}),$$

both identities holding with probability one. Proofs of these results are little more than conditional versions of the arguments employed to establish Lemma 2.2. From this point the proof of (3.2) and (3.3) may be conducted straightforwardly, using conditional versions of arguments in the proof of Theorem 2.3.

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 $D_3 = D(0.4, 0.75, \sigma_3^2)$. 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 0.50.

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.

[Insert Table 4.1 about here]

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 favorable effect of strong oversmoothing can be explained by D_1 being symmetric and unimodal; mode and median coincide. Since kernel estimators with large bandwidth tend to have their mode near the sample median, large values of c produce (smooth) resampling distributions with modes usually close to the true mode.

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 towards the median, 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 resample 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_3. 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 4.1, which summarizes our numerical results. The fact that the mean squared error does depend on the choice of h_1 indicates that a good adaptive procedure is required. For now, we leave this as an open problem.

References

BHATTACHARYA, R.N. AND RAO, R.R. (1976). *Normal Approximation and Asymptotic Expansions*. 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. Wahrscheinlichkeitstheorie verw. Geb.* **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. (1993). Testing for multimodality. *Computat. Statist. Data Anal.*

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. Multivar. Anal.* **32**, 177-203.

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

HALL, P. AND MURISON, R.D. (1991). Correcting the negativity of high-order kernel density estimators. *Centre for Mathematics and its Applications Research Report No.* CMA–SR21–91, Australian National University, Canberra, Australia.

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

MAMMEN, E., MARRON, J.S. AND FISHER, N.I. (1992). Some asymptotics for multimodality tests based on kernel estimates. *Probab. Thy Rel. 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. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16**, 629–647.

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). Bootstrap 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. *Prob. Informat. Transmission* **26**, 31–37.

TABLE 4.1. *Mean squared error of mode estimator.* The table gives Monte Carlo approximations to mean squared errors, and also their standard deviations (in parentheses). These quantities were computed as the sample mean and standard deviation, respectively, of M independently simulated values of the squared error of the mode estimator, where $M = 1,000$ for $n = 50$ and $M = 500$ for $n = 100$ and 200. The number of bootstrap simulations was $B = 50$ in each case. The distributions D_2, D_2 and D_3 are defined in the text.

n	D_1		D_2		D_3	
	$c = 2.5$	$c = 3$	$c = 1$	$c = 1.5$	$c = 1$	$c = 1.5$
50	0.0240	0.0235	0.1473	0.1411	0.1899	0.1183
	(0.0012)	(0.0012)	(0.0173)	(0.0081)	(0.0084)	(0.0049)
100	0.0162	0.0141	0.0865	0.0595	0.1465	0.1073
	(0.0012)	(0.0009)	(0.0171)	(0.0041)	(0.0090)	(0.0060)
200	0.0079	0.0078	0.0613	0.0255	0.0963	0.0728
	(0.0005)	(0.0006)	(0.0139)	(0.0018)	(0.0057)	(0.0095)