The Annals of Statistics
1988, Vol. 16, No. 2, 629–647

ON WEAK CONVERGENCE AND OPTIMALITY OF KERNEL DENSITY ESTIMATES OF THE MODE

BY JOSEPH P. ROMANO

Stanford University

A mode of a probability density $f(t)$ is a value θ that maximizes f. The problem of estimating the location of the mode is considered here. Estimates of the mode are considered via kernel density estimates. Previous results on this problem have several serious drawbacks. Conditions on the underlying density f are imposed globally (rather than locally in a neighborhood of θ). Moreover, fixed bandwidth sequences are considered, resulting in an estimate of the location of the mode that is not scale-equivariant. In addition, an optimal choice of bandwidth depends on the underlying density, and so cannot be realized by a fixed bandwidth sequence. Here, fixed and random bandwidths are considered, while relatively weak assumptions are imposed on the underlying density. Asymptotic minimax risk lower bounds are obtained for estimators of the mode and kernel density estimates of the mode are shown to possess a certain optimal local asymptotic minimax risk property. Bootstrapping the sampling distribution of the estimates is also discussed.

1. Introduction. The problem of estimating the location of the mode of a density is considered here. A mode of a probability density $f(t)$ is a value θ that maximizes f.

Although the mode has received relatively little attention in the literature, some estimates of the mode have been studied in [2]–[6], [8] and [13]. We will consider estimates of the mode of a density via a kernel density estimate. That is, given a kernel K (a probability density on the real line), a bandwidth h and a sample $X_1, \ldots, X_n$ from a c.d.f. F on $\mathbf{R}$ having a density f, the kernel density estimate is given by

$$\hat{f}_{n,h}(t) = \hat{f}_{n,h}(t; X_1, \ldots, X_n) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right). \tag{1.1}$$

The bandwidth h may be data-dependent so that, in general, h is a measurable function of n and $X_1, \ldots, X_n$. The choice of h in the context of density estimation has been studied by many authors. A relatively new motivation for considering the problem of estimating the mode is that the mode of a density is precisely a location where the choice of bandwidth is most sensitive.

If K is bounded, continuous and $\lim_{t \to \pm\infty} K(t) = 0$, then so is $\hat{f}_{n,h}$, so there will be a point $\hat{\theta}$ such that

$$\hat{f}_{n,h}(\hat{\theta}) = \sup_t \hat{f}_{n,h}(t). \tag{1.2}$$

Received May 1986; revised May 1987.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62E20, 62G20.

Key words and phrases. Kernel density estimates, mode, weak convergence, rates of convergence, asymptotic minimax risk.

Because $\hat{\theta}$ may not be uniquely defined by this equation, consider the mode functional M defined by

$$M(f) = \inf\left\{m | f(m) = \sup_t f(t)\right\}, \tag{1.3}$$

where f is a density on R. Then the sample mode $\hat{\theta}_{n,h}$ is uniquely defined by

$$\hat{\theta}_{n,h} = M(\hat{f}_{n,h}). \tag{1.4}$$

For ease of argument, we will use this definition throughout the paper, but all the results continue to hold if $\hat{\theta}_{n,h}$ is any random variable satisfying (1.2).

Parzen [10] proved that if f is uniformly continuous with a unique mode θ, and h_n is a fixed (nonrandom) bandwidth sequence such that $nh_n^2 \to \infty$, $h_n \to 0$, then $\hat{\theta}_{n,h_n} \to \theta$ a.s. Under further conditions, Parzen also obtained the limiting distribution of the sample mode. Eddy [4] weakened Parzen's conditions to allow less stringent assumptions on the choice of the kernel K. Samanta [13] has given a multivariate version of Parzen's results. Their results on estimating the mode have several drawbacks. First, smoothness assumptions on the underlying density are imposed globally. For instance, Eddy assumes f has an absolutely continuous bounded fourth derivative. Here, a much weaker hypothesis is assumed about f. In particular, we only make an assumption concerning the local behavior of f in a neighborhood of the mode.

Next, all previous results for estimating the mode by using kernel density estimates assume fixed (nonrandom) bandwidths. This has two problems. First, neither the kernel density estimate nor the sample mode (as defined via the kernel density estimate) is equivariant under the group of scale changes. That is, if all the data are multiplied by a fixed nonzero constant c, then the kernel density estimate of the mode based on the transformed observations is not the kernel density estimate of the mode of the original observations multiplied by c. This lack of equivariance is theoretically unpleasant. By choosing the bandwidth to be scale-equivariant, and hence data-dependent, the resulting estimate of the mode is also scale-equivariant. Second, the optimal choice of bandwidth depends on the underlying density. Such a choice cannot be realized by using a nonrandom bandwidth sequence.

The results presented in this paper eliminate these problems to a certain extent. In Section 2, precise weak convergence results are obtained for both fixed and random bandwidth sequences. Bootstrap weak convergence results are also discussed, though more detail about bootstrapping functionals of a density is given in [12]. Section 3 considers the question of why one should use estimates of the mode based on kernel density estimates. In a decision-theoretic framework, it is shown that estimates of the mode based on a kernel density estimate with a bandwidth h_n proportional to $n^{-1/7}$ achieve the optimal rate of convergence in a local asymptotic minimax risk sense. However, the results are seen to be sensitive to the choice of neighborhoods used in evaluating the asymptotic minimax risk, as will be made clear. The proofs are given in Section 4. Section 5 is an appendix containing results on empirical processes indexed by classes of functions and other technical results that are invoked in Section 4. The results can be extended to estimating the mode of a density in $\mathbf{R}^k$ with minor modifications.

This section closes with a basic consistency result, which generalizes the consistency result of Parzen.

THEOREM 1.1. *Assume the kernel K is continuous and of bounded variation. Let $X_1, X_2, \ldots$ be i.i.d. F having a density f such that:*

1. *f is continuous in a neighborhood of the mode $\theta = M(f)$.*
2. *For every $\delta > 0$, $\sup_{\{t:\ |t-\theta| > \delta\}} f(t) < f(\theta)$.*

Let $S_n = S_n(X_1, \ldots, X_n)$ be any statistic such that for some real numbers s_1 and s_2,

$$0 < s_1 \leq \liminf_{n\to\infty} S_n \leq \limsup_{n\to\infty} S_n \leq s_2 < \infty \quad a.s. \tag{1.5}$$

Let v_n be a fixed sequence of numbers with $nv_n/\log(n) \to \infty$ and $v_n \to 0$, and set $h_n = v_n \cdot S_n$. Then

$$\text{(i)} \qquad \hat{\theta}_{n,h_n} \to \theta \quad a.s.,$$

$$\text{(ii)} \qquad \hat{f}_{n,h_n}\left(\hat{\theta}_{n,h_n}\right) \to f(\theta) \quad a.s.$$

REMARKS. 1. The hypothesis that f be continuous in a neighborhood of θ can be weakened. By a slight modification of the proof (deferred to Section 4), an analogous argument goes through if we assume the weaker hypothesis that f is continuous in a neighborhood (θ_1, θ) or (θ, θ_2), so that f can be discontinuous at θ.

2. Compare Theorems 1.1 with Parzen [10]. He assumes f is uniformly continuous. Furthermore, h_n is nonrandom with $h_n \gg n^{-1/2}$. Specializing to the case S_n is a fixed constant, we only assume $h_n \gg \log(n)/n$.

3. If S_n is a scale-equivariant statistic, so is the sample mode $\hat{\theta}_{n,h_n}$. Assumption (1.5) on S_n is usually satisfied, because often $S_n \to s$ almost surely. A quick and simple choice might be the sample standard deviation if a second moment is assumed. Another possibility is the interquartile range or some similar difference between order statistics. More discussion on the choice of bandwidth is presented in Section 2.

Actually, the proof shows that assumption (1.5) can be weakened considerably. That is, S_n could tend to 0 or $+\infty$, but the rate at which this could happen would depend on the actual choice of v_n. Note, however, that both optimal and reasonable choices for h_n are covered by the theorem.

2. Weak convergence results. If F is a c.d.f. on $\mathbf{R}$ with density f and mode $\theta = M(f)$, consider the approximate pivot

$$R_n(X_1, \ldots, X_n; F) = \left(nh_n^3\right)^{1/2}\left(\hat{\theta}_{n,h_n} - \theta\right), \tag{2.1}$$

where $\hat{f}_{n,h_n}(t) = \hat{f}_{n,h_n}(t; X_1, \ldots, X_n)$ is given by (1.1) and $\hat{\theta}_{n,h_n} = M(\hat{f}_{n,h_n})$. Let $J_n(F)$ be the law of $R_n(X_1, \ldots, X_n)$ when $X_1, \ldots, X_n$ are i.i.d. F. In this section, the limiting behavior of $J_n(F)$ is obtained. Note that the dependence of $J_n(F)$ on h_n has been suppressed, but a choice of h_n will always be specified when reference to $J_n(F)$ has been made.

Before stating the main results, we need some weak assumptions. The assumption on the underlying density of the observations is stated as assumption (A), whereas the assumption on the kernel is given in (B).

(A) Assume f has a unique mode θ such that $\sup_{\{t:\,|t-\theta|>\delta\}} f(t) < f(\theta)$ for every $\delta > 0$. Also, f has a continuous third derivative $f^{(3)}$ in some neighborhood of θ with $f^{(3)}(\theta) < 0$.

(B) Assume the kernel K is symmetric and has a continuous second derivative of bounded variation. Also assume that, for some $p > 0$, K^{2+p} and $|K^{(1)}|^{2+p}$ are integrable. For later use, define

$$I(K) = \int [K^{(1)}(z)]^2\,dz \quad \text{and} \quad H(K) = \int z^2 K(z)\,dz.$$

For every $\delta > 0$,

$$\frac{1}{h_n^3} \int_{\{z:\,|z|>\delta/h_n\}} |K^{(j)}(z)|\,dz \to 0, \quad \text{for } j = 0,1,2.$$

Also assume $|K|^3$ and $z^2|K^{(2)}(z)|^2$ are integrable.

Assumption (A) is close to necessary, in view of the fact that the asymptotic distribution of the mode is Gaussian with variance depending on $f^{(2)}(\theta)$ and asymptotic bias that depends on $f^{(3)}(\theta)$. Other authors have made global assumptions on f in computing the asymptotic distribution of the mode.

The main theorem of this section is now given for the weak convergence of the mode based on a possibly data-dependent bandwidth. To prove limit results for random bandwidths, the idea is to consider a stochastic process representing a class of pivots indexed by a bandwidth parameter, and show this class converges weakly in an appropriate sense. Let $\mathbf{C}[u_1, u_2]$ denote the metric space of real-valued continuous functions on $[u_1, u_2]$ with the sup metric. A technical difficulty arises since the map $h \to \hat{\theta}_{n,h}$ need not be continuous.

THEOREM 2.1. *Let the distribution F have a density f and mode θ satisfying assumption* (A) *and the kernel K satisfies* (B). *Let v_n be any fixed sequence of numbers such that $nv_n^5/\log(n) \to \infty$ and $(nv_n^7)^{1/2} \to d$ for some $d < \infty$. For any $h > 0$, let $\hat{f}_{n,h}(t)$ be given by* (1.1) *and $\hat{\theta}_{n,h} = M(\hat{f}_{n,h})$. Let $0 < u_1 < u_2 < \infty$ and consider the stochastic process*

$$T_n(b) = \left(nb^3v_n^3\right)^{1/2}\left(\hat{\theta}_{n,bv_n} - \theta\right), \qquad u_1 \le b \le u_2.$$

(i) *T_n can be written as V_n/D_n, where V_n can be regarded as a random variable on $\mathbf{C}[u_1, u_2]$. The law of V_n converges weakly to the law of V, where V is a Gaussian process on $\mathbf{C}[u_1, u_2]$;*

$$\mathbf{E}V(b) = b^{7/2} \cdot c \cdot f^{(2)}(\theta),$$

where c is the constant

$$c = \frac{d}{2}\frac{f^{(3)}(\theta)}{f^{(2)}(\theta)}H(K),$$

and if $r = (b_1/b_2)^{1/2}$ *then*

$$\operatorname{Cov}(V(b_1), V(b_2)) = f(\theta)\int_{-\infty}^{\infty} K^{(1)}(ry)K^{(1)}\left(\frac{r}{y}\right) dy.$$

Also,

$$\sup_{u_1 \le b \le u_2} \left|D_n(b) - f^{(2)}(\theta)\right| \to 0 \quad a.s.$$

(ii) *Let* $S_n = S_n(X_1, \ldots, X_n)$ *be any statistic such that* $S_n \to s$ *in probability, where* s *is positive and finite. Define the bandwidth* h_n *to be* $h_n = \nu_n \cdot S_n$. *Then the law of* $(nh_n^3)^{1/2}(\hat{\theta}_{n,h_n} - \theta)$ *converges in distribution to* $Z - c \cdot s^{7/2}$, *where* Z *is Gaussian with mean* 0 *and variance given by*

$$\operatorname{Var}(Z) = \frac{f(\theta)}{\left[f^{(2)}(\theta)\right]^2} I(K).$$

REMARK 2.1. If $f^{(3)}(\theta) = 0$, then it is of interest to determine the rate for which the asymptotic bias of $\hat{\theta}_n$ is not negligible. Specifically, assume $f^{(2j+1)}(\theta)$ is the first odd order derivative of f at θ which is nonzero, and $f^{(2j+1)}$ is continuous at θ. Then, under further tail assumptions on K, if $(nh_n^{4j+3})^{1/2} \to d$, then $(nh_n^3)^{1/2}(\hat{\theta}_n - \theta)$ converges weakly to the law of $Z - c$, where

$$c = \frac{d}{(2j)!}\frac{f^{(2j+1)}(\theta)}{f^{(2)}(\theta)}\int_{-\infty}^{\infty} z^{2j}K(z)\, dz.$$

REMARK 2.2. Under similar conditions on the bandwidth, one can study the asymptotic behavior of $[\hat{f}_{n,h_n}(\hat{\theta}_{n,h_n}) - f(\theta)]$. Upon division by h_n^2, its distribution converges weakly to the degenerate distribution concentrated at $f^{(2)}(\theta)H(K)/2$. Details are given in [12].

REMARK 2.3. By forcing the asymptotic bias to be negligible by assuming $d = 0$, one can weaken the assumption that f has a third derivative in a neighborhood of θ.

Discussion.

2.1. *Assumptions on the kernel.* Since the statistician is free to choose the kernel K, no serious attempt has been made to weaken the assumptions imposed on K. Furthermore, it is generally believed that the choice of kernel in density estimation is not as crucial (or as sensitive) as the choice of bandwidth. Further, in the more interesting case $(nh_n^7)^{1/2} \to d > 0$, the optimal choice of K depends on the unknown f, so no fixed optimal choice of K exists.

2.2. *Assumptions on the bandwidth.* All the results presented in this section assume the bandwidth h_n satisfies $nh_n^5/\log(n) \to \infty$ as $n \to \infty$ (where this convergence is almost sure in the case h_n is random). This assumption is needed

to be able to estimate $f^{(2)}(\theta)$ consistently. In fact, Silverman [14] has shown that if f has a uniformly continuous second derivative, then, for fixed bandwidth sequences h_n, it is necessary and sufficient that $h_n \to 0$ and $nh_n^5/\log(n) \to \infty$ as $n \to \infty$ in order for

$$\sup_t \left| \hat{f}^{(2)}_{n,h_n}(t) - f^{(2)}(t) \right| \to 0,$$

in probability and almost surely. Proposition 2.1 of [12] shows that the assumption $nh_n^5/\log(n) \to \infty$ can, at best, be weakened to $nh_n^5 \to \infty$ in order to estimate $f^{(2)}(\theta)$ consistently. The assumption $(nh_n^7)^{1/2} \to d$ is exact in allowing one to calculate the asymptotic bias of the location of the sample mode.

2.3. *Choice of bandwidth.* As discussed in Section 1, the choice of a bandwidth that is scale-equivariant results in a scale-equivariant estimator of the location of the mode. The next issue is to determine the rate at which the bandwidth tends to 0.

One method for choosing the bandwidth to estimate the location of the mode is to minimize the (asymptotic) mean-squared error. For a fixed bandwidth sequence h_n, the asymptotic mean-squared error of $\hat{\theta}_{n,h_n}$ is

$$\frac{1}{nh_n^3} \frac{f(\theta)}{\left[f^{(2)}(\theta)\right]^2} I(K) + \left[\frac{h_n^2}{2} \frac{f^{(3)}(\theta)}{f^{(2)}(\theta)} H(K)\right]^2. \tag{2.2}$$

The rate at which this tends to 0 is fastest when the asymptotic variance and the square of the asymptotic bias are of the same order. If $f^{(3)}(\theta) \neq 0$, this happens when $(nh_n^7)^{1/2} \to d$ for some positive number d. In Section 3, it will be seen that this choice of bandwidth corresponds to the best achievable rate of convergence in a local asymptotic minimax risk sense. Theorem 2.1 covers this assumption on h_n. Furthermore, for fixed K, (2.2) is minimized (asymptotically) if h_n is chosen so

$$nh_n^7 \to \frac{3f(\theta)I(K)}{\left[f^{(3)}(\theta)H(K)\right]^2}. \tag{2.3}$$

Such a choice is possible if h_n is data-dependent. Specifically, if

$$h_n = S_n \cdot n^{-1/7},$$

where S_n^7 is a consistent estimate of the right-hand side of (2.3), then the minimum asymptotic mean-squared error is attained without knowledge of f (by Theorem 2.1). To construct a consistent estimate of the right-hand side of (2.3), apply Proposition 2.1. Note, however, $\hat{f}^{(3)}_{n,h_n}(\hat{\theta}_{n,h_n})$ is not a consistent estimate of $f^{(3)}(\theta)$ under the assumption $(nh_n^7)^{1/2} \to d$ (see Proposition 2.1 of [12]), so different bandwidths must be used to estimate $f^{(3)}(\theta)$ and θ. In any case, an optimal kernel estimator of θ exists based on a random bandwidth, assuming $f^{(3)}(\theta) \neq 0$. If $f^{(3)}(\theta) = 0$, then the preceding choice results in a faster rate of convergence of $\hat{\theta}_n$ to θ since $nh_n^7 \to \infty$ with probability 1, depending on the first odd order derivative of f at θ, which is nonzero (see Remark 2.1). Whether or

not this order can actually be estimated consistently is not pursued here. By Theorem 3.1, the best achievable rate (in the sense described there) would not be increased anyway.

PROPOSITION 2.1. *Let f be a bounded density with a jth continuous derivative in some neighborhood of x. Let $X_1, X_2, \ldots$ be i.i.d. with density f. Suppose $v_n \to 0$ and $nv_n^{2j+1}/\log(n) \to \infty$. Assume the kernel K has a continuous integrable jth derivative of bounded variation and such that, for every $\delta > 0$,*

$$\frac{1}{v_n^j} \int_{\{z:\, |z| > \delta/v_n\}} \left| K^{(j)}(z) \right| dz \to 0.$$

Let $S_n = S_n(X_1, \ldots, X_n)$ satisfy

$$0 < s_1 \le \liminf_{n\to\infty} S_n \le \limsup_{n\to\infty} S_n \le s_2 < \infty \quad a.s. \tag{2.4}$$

Let $h_n = v_n \cdot S_n$ and set

$$\hat{f}_{n,h} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right).$$

Then there exists $\delta_0 > 0$ such that

$$\sup_{\{t:\, |t-x| < \delta_0\}} \left| \hat{f}_{n,h_n}^{(j)}(t) - f^{(j)}(t) \right| \to 0 \quad a.s. \tag{2.5}$$

Hence, if $f^{(j)}$ is continuous in a neighborhood of ξ and $\xi_n^* \to \xi$ a.s., then $\hat{f}_{n,h_n}^{(j)}(\xi_n^*) \to f^{(j)}(\xi)$ a.s.

REMARK 2.4. Proposition 2.1 would typically be applied when S_n is scale-equivariant and $S_n \to s$ a.s. so that (2.4) is easily satisfied. The assumption that f be bounded can be removed if K is assumed to have compact support (by an easy modification of Proposition 5.1). As in Theorem 1.1, assumption (2.4) can be weakened, depending on the choice of v_n. Also, the weak convergence properties of $\hat{f}_{n,h_n}^{(j)}(x)$ can be derived by the method used to prove Theorem 2.2. Finally, the sup in (2.5) can be replaced by a sup over all t if $f^{(j)}(\cdot)$ is uniformly continuous.

2.4. *Bootstrapping the distribution of the mode.* In order to apply Theorem 2.1 to construct a confidence interval for the mode θ, one must explicitly estimate $f^{(2)}(\theta)$ and $f^{(3)}(\theta)$, which is quite an intricate problem. In contrast, the bootstrap approach is automatic, so it is of interest to investigate its properties. We now indicate the nature of the behavior of the bootstrap with discussion confined to the fixed bandwidth case; see [12] for details.

Let $J_{n,h_n}(F)$ be the distribution of the pivot (2.1) under F, where the dependence of h_n is made explicit. The bootstrap approach is to estimate $J_{n,h_n}(F)$ by $J_{n,h_n}(\hat{G}_n)$, where $\hat{G}_n$ is some estimate of F, and confidence intervals can then be constructed by using the appropriate quantiles from this estimated sampling distribution. The following can be shown: If $\hat{G}_n$ is a distribution that has a density $\hat{g}_n$ such that with probability 1 the first three derivatives of $\hat{g}_n$

converge uniformly to f in some neighborhood of $M(f)$, then the bootstrap is consistent in the sense

$$\rho\big(J_{n,h_n}(F), J_{n,h_n}(\hat{G}_n)\big) \to 0 \quad \text{a.s.},$$

where ρ is any metric metrizing weak convergence. A proof of this can be based on Proposition 4.1 by first showing it holds when G_n is a nonrandom sequence. Thus, if $\hat{G}_n$ is the distribution with density $\hat{f}_{n,b_n}$ with $nb_n^7/\log(n) \to \infty$, the conditions are satisfied by Proposition 2.1 and the bootstrap is consistent. However, this specifically rules out $h_n = b_n$ when h_n satisfies the optimal rate $nh_n^7 \to d < \infty$. In fact, the bootstrap consistency result fails in this case. The essential reason is that $\hat{f}^{(3)}_{n,h_n}(\theta)$ is not a consistent estimator of $f^{(3)}(\theta)$, and the limiting distribution of the sample mode depends on the parameter $f^{(3)}(\theta)$. Therefore, the bootstrap is a consistent method, but only if one resamples from a density that is necessarily different from the original kernel density estimate used to estimate the mode. In summary, straightforward application of a naive bootstrap procedure may result in invalid inferences.

3. Minimax rates for estimating the mode. This section addresses the question: Why use estimates of the mode based on kernel density estimates? To begin, they are easy to understand conceptually and are easy to compute, their statistical properties are understood and they possess certain optimality properties in the context of density estimation. This section specifically focuses on optimality properties of kernel density estimates of the mode, and considers the more general question of how well can the mode be estimated by any procedure. In a decision-theoretic framework, the result given in the following discussion will establish a certain optimal local asymptotic minimax property of estimates of the mode via kernel density estimates.

We have seen that, by choosing a bandwidth sequence h_n proportional to $n^{-1/7}$, the kernel density estimate of the mode $\hat{\theta}_n$ converges to θ at a rate $(nh_n^3)^{1/2}$, which is proportional to $n^{2/7}$; that is, $\hat{\theta}_n$ is a $n^{2/7}$-consistent estimator of θ. Do their exist estimators that converge to θ at a faster rate? We formulate this question as an asymptotic minimax property, as in Hasminskii [7]. To start, let $\mathbf{L}$ be the class of subconvex loss functions on the line [that is, $l \in \mathbf{L}$ if $l \geq 0$, l is symmetric and for each c the set $\{x \in \mathbf{R}: l(x) \leq c\}$ is closed and convex]. Fix a density f_0 and, based on n observations, consider the problem of estimating the mode for a family of densities in some neighborhood N_n of f_0. Neighborhood is used informally here for purposes of motivation, but will be made more specific in the discussion that follows. (It may be helpful to recall the classical parametric estimation problem where the neighborhoods shrink with n to the true parameter value at rate $n^{-1/2}$. See Section XIII.2 of Millar [9].) Let $l \in \mathbf{L}$ and let δ_n be a normalizing sequence so that the loss of using an estimate d of θ based on n observations is $l[\delta_n(d - \theta)]$. Then the minimax risk for this problem, which depends on δ_n, N_n and l, is just

$$R_n(\delta_n, N_n, l) = \inf_{T_n} \sup_{f \in N_n} \mathbf{E}_f\{l[\delta_n(T_n - M(f))]\},$$

where the infimum is over all estimators T_n of the mode based on a sample of size n (not just estimates based on kernel density estimates). If $\liminf_{n\to\infty} R_n(\delta_n, N_n, l) > 0$, then no sequence of estimates T_n converges to $M(f)$ faster than rate δ_n uniformly over the neighborhood sequence N_n; for if $\gamma_n/\delta_n \to \infty$ and $\liminf_{n\to\infty} R_n(\delta_n, N_n, l) > 0$, then $\liminf_{n\to\infty} R_n(\gamma_n, N_n, l) = \sup_x l(x)$.

Hasminskii [7] considered this problem and for a certain choice of neighborhoods N_n found an upper bound for the fastest achievable minimax rate to be $n^{1/5}$. At first, this might seem inconsistent with the fact that we can achieve a faster rate of $n^{2/7}$ based on kernel density estimates. However, estimates of the mode based on kernel density estimates misbehave for Hasminskii's choice of N_n (due to the bias of these estimators, which depends on the third derivative of the underlying density at the mode). On the other hand, we will see that for a choice of neighborhoods that is, in some sense, slightly smaller than those used by Hasminskii, the best achievable rate is increased to $n^{2/7}$ and estimates of the mode based on kernel density estimates achieve this rate for such a choice of neighborhoods. The result actually considers various choices for N_n, depending on a parameter p, and the best achievable rate is seen to depend on p. The case $p = 2$ approximately corresponds to Hasminskii's choice of neighborhoods, whereas the case $p = 3$ corresponds to neighborhoods for which kernel density estimates of the mode enjoy the optimal achievable minimax rate.

For $p \geq 2$, let $\mathbf{D}_p$ denote the collection of densities f satisfying: f has a unique mode θ; for every $\delta > 0$, $\sup_{\{t:\,|t-\theta|>\delta\}} f(t) < f(\theta)$; and f has a bounded derivative of order p is some neighborhood of θ with $f^{(2)}(\theta) < 0$. Fix $p \geq 2$ and let $\beta_n = n^{-\beta}$, where $\beta = (2p+1)^{-1}$. If $f_0 \in \mathbf{D}_p$, define a neighborhood of $N_n(\varepsilon, p, f_0)$ of f_0 as the set of densities f in $\mathbf{D}_p$ satisfying

(i) $|M(f) - M(f_0)| < \beta_n/\varepsilon$,
(ii) $\sup|f^{(p)}(x)| < 1/\varepsilon$,
(iii) $\sup[\sum_{j=0}^{p-1}|f^{(j)}(x) - f_0^{(j)}(x)|] < \beta_n/\varepsilon$, where the supremum in (ii) and (iii) is over the set $A_\varepsilon = \{x: |x - M(f_0)| < \varepsilon\}$,
(iv) $\sup_{A_\varepsilon^c} f(x) \leq \sup_{A_\varepsilon^c} f_0(x)$.

The following theorem asserts that $M(f)$ can be estimated no faster than rate $\delta_n = n^{-r}$ over this choice of neighborhoods, where $r = (p-1)/(2p+1)$.

THEOREM 3.1. *Fix $p \geq 2$ and let $f_0 \in \mathbf{D}_p$. For any loss function l in $\mathbf{L}$ and for every sufficiently small $\varepsilon > 0$,*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in N_n(\varepsilon,p,f_0)} \mathbf{E}_f\{l[\delta_n(T_n - M(f))]\} > 0, \tag{3.1}$$

where $\delta_n = n^r$, $r = (p-1)/(2p+1)$, and the infimum is over all estimators T_n of the mode, i.e., all measurable functions of a sample $X_1, \ldots, X_n$ from a density f. In particular, for all $\lambda > 0$ and all sufficiently small $\varepsilon > 0$,

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in N_n(\varepsilon,p,f_0)} \mathbf{P}\{[\delta_n(T_n - M(f))] > \lambda\} > 0. \tag{3.2}$$

REMARK 3.1. Condition (iv) in the definition of $N_n(\varepsilon, p, f_0)$ is a sort of identifiability constraint to allow for perturbations of f_0 outside an ε neighborhood of $M(f_0)$. Note, however, that this condition could simply be removed from the definition of $N_n(\varepsilon, p, f_0)$ and the theorem would still be true. Alternatively, one could look at only those f that are equal to f_0 outside some ε neighborhood of $M(f_0)$. In fact, many variations on the choice of neighborhoods are possible. The ones given here were designed to readily yield an optimal property for kernel density estimates of the mode (see Remark 3.3).

REMARK 3.2. The neighborhoods $N_n(\varepsilon, 2, f_0)$ approximately correspond to Hasminskii's choice of neighborhoods in the sense that both neighborhoods contain densities f such that f and its first derivative are uniformly close to f_0 and its first derivative in some neighborhood of $M(f_0)$. Hasminskii, however, does not allow for perturbations of f_0 outside a neighborhood of $M(f_0)$. Furthermore, the neighborhoods $N_n(\varepsilon, 2, f_0)$ shrink with n, whereas Hasminskii's do not, yielding an improved result since the same bound for the minimax rate, namely $\delta_n = n^{2/5}$, is still obtained.

REMARK 3.3. It follows from Proposition 4.1 (also see the proof of Theorem 2.1) that a kernel density estimate $\hat{\theta}_n$ based on a bandwidth sequence h_n proportional to $n^{-1/7}$ achieves the optimal minimax rate of $n^{2/7}$ over the neighborhood sequence $N_n(\varepsilon, 3, f_0)$. This means that, if $f_0 \in \mathbf{D}_3$ and $\delta_n = n^{2/7}$, then

$$0 < \limsup_{n\to\infty} \sup_{f \in N_n(\varepsilon, 3, f_0)} \mathbf{P}\{[\delta_n(\hat{\theta}_n - M(f))] > \lambda\} < 1. \tag{3.3}$$

In fact, this result can be modified to yield an asymptotic minimax result over a fixed (nonshrinking) neighborhood of densities. Specifically, let $N = N(\varepsilon, \delta)$ be a family of densities f in D_3 satisfying $f^{(2)}(M(f)) > \varepsilon$, $f^{(3)}(M(f)) < \varepsilon^{-1}$ and

$$\sup_{A_\varepsilon^c} |f(x) - f(M(f))| > \delta,$$

where A_ε is the set $\{x: |x - M(f)| < \varepsilon\}$. Then (3.3) holds with N_n replaced by N.

4. Proofs.

PROOF OF THEOREM 1.1. (i) The first step is to show the following. Let a_n and b_n be fixed sequences of numbers such that $(nb_n^2)/[\log(n)a_n] \to \infty$ and $a_n \to 0$. Then

$$\sup_{\{h:\, b_n \le h \le a_n\}} |\hat{\theta}_{n,h} - \theta| \to 0 \quad \text{a.s.} \tag{4.1}$$

Define

$$f_{n,h}(t) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{t-y}{h}\right) f(y)\, dy.$$

Choose δ such that $0 < \delta < \delta_0$, where δ_0 is obtained from Proposition 5.1 so that

$$\lim_{n\to\infty} \sup_{\{h:\, 0<h\le a_n\}} \sup_{\{t:\, |t-\theta|<\delta_0\}} \left| f_{n,h}(t) - f(t) \right| = 0. \tag{4.2}$$

By Corollary 5.1,

$$\sup_{\{h:\, b_n\le h\le a_n\}} \sup_t \left| \hat{f}_{n,h}(t) - f_{n,h}(t) \right| \to 0 \quad \text{a.s.} \tag{4.3}$$

Combining (4.2) and (4.3) yields

$$\sup_{\{h:\, b_n\le h\le a_n\}} \sup_{\{t:\, |t-\theta|<\delta\}} \left| \hat{f}_{n,h}(t) - f(t) \right| \to 0 \quad \text{a.s.} \tag{4.4}$$

The assumptions imply for any $\delta > 0$,

$$\limsup_{n\to\infty} \sup_{\{h:\, 0<h\le a_n\}} \sup_{\{t:\, |t-\theta|>\delta\}} f_{n,h}(t) < f(\theta). \tag{4.5}$$

Combining (4.3) and (4.5) yields

$$\limsup_{n\to\infty} \sup_{\{h:\, b_n\le h\le a_n\}} \sup_{\{t:\, |t-\theta|\ge\delta\}} \hat{f}_{n,h}(t) < f(\theta) \quad \text{a.s.} \tag{4.6}$$

It follows from (4.4) and (4.6) that $P\{\sup_{\{h:\ b_n\le h\le a_n\}} |\hat{\theta}_{n,h} - \theta| < \delta \text{ e.v.}\} = 1$. This proves (4.1). To prove (i), let $a_n = \nu_n \cdot s_2$ and $b_n = \nu_n \cdot s_1$. Then $b_n \le h_n \le a_n$ eventually with probability 1. Hence, by (4.1),

$$\limsup_{n\to\infty} |\hat{\theta}_{n,h_n} - \theta| \le \limsup_{n\to\infty} \sup_{\{h:\, b_n\le h\le a_n\}} |\hat{\theta}_{n,h} - \theta| \to 0 \quad \text{a.s.}$$

(ii) Combine (4.4) and (4.6) to get

$$\sup_{\{h:\, b_n\le h\le a_n\}} \sup_t \hat{f}_{n,h}(t) \to f(\theta) \quad \text{a.s.}$$

The result follows. □

Before proving Theorem 2.1, we first give a *triangular* version of the asymptotic distribution of the sample mode for fixed bandwidth sequences. The result is given with F_n varying with n for use in the context of bootstrapping.

PROPOSITION 4.1. *Fix F with density f and mode θ satisfying* (A) *and assume K satisfies* (B). *Also, assume $nh_n^5/\log(n) \to \infty$ and $h_n \to 0$. Let F_n be a sequence of distributions on* **R**. *Suppose F_n has a density f_n with mode $\theta_n = M(f_n)$. Assume the following*:

(a) *f_n and $f_n^{(2)}$ converge to f and $f^{(2)}$, respectively, uniformly in some neighborhood of θ.*

(b) *For every $\delta > 0$, $\limsup_{n\to\infty} \sup_{\{t:\, |t-\theta|>\delta\}} f_n(t) < f(\theta)$.*

Let $X_{n,1}, \ldots, X_{n,n}$ be i.i.d. F_n. Set

$$\hat{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - X_{n,i}}{h_n} \right)$$

and $\hat{\theta}_n = M(\hat{f}_n)$. *Let*

$$\mu_n = \left(\frac{n}{h_n}\right)^{1/2} \mathbf{E}_{F_n}\left(K^{(1)}\left(\frac{\theta_n - X_{n,1}}{h_n}\right)\right).$$

Then $(nh_n^3)^{1/2}(\hat{\theta}_n - \theta_n)$ *can be represented as the law of* $W_n - \mu_n/Y_n$, *where the law of* W_n *converges weakly to the law of* Z (*as defined in the statement of Theorem* 2.1) *and the law of* Y_n *converges in* F_n^n *probability to* $f^{(2)}(\theta)$. *Hence, if we also assume*

(c) $\mu_n \to \mu < \infty$,

then $J_n(F_n)$ *converges weakly to the law of* $Z - \mu/f^{(2)}(\theta)$.

PROOF. First note that assumptions (a) and (b) imply $\theta_n \to \theta$. By an argument analogous to that of Theorem 1.1, $\hat{\theta}_n - \theta_n \to 0$ in F_n^n probability. By Taylor's theorem, for some random variable θ_n^* between θ_n and $\hat{\theta}_n$,

$$(4.7) \qquad (nh_n^3)^{1/2}(\hat{\theta}_n - \theta_n) = -\frac{(nh_n^3)^{1/2}\hat{f}_n^{(1)}(\theta_n)}{\hat{f}_n^{(2)}(\theta_n^*)}, \quad \text{if } \hat{f}_n^{(2)}(\theta_n^*) \neq 0.$$

The result will follow by showing:

(1) The law of $S_n = -(nh_n^3)^{1/2}\hat{f}_n^{(1)}(\theta_n)/f^{(2)}(\theta)$ converges weakly to the law of $Z - \mu/f^{(2)}(\theta)$.

(2) $\hat{f}_n^{(2)}(\theta_n^*) \to f^{(2)}(\theta)$ in F_n^n probability for any sequence $\theta_n^* \to \theta$.

Proof of (1).

$$-\hat{f}_n^{(1)}(\theta_n) = \frac{1}{n}\sum_{j=1}^{n} V_{n,j},$$

where the $V_{n,j}$ are independent and identically distributed as

$$V_{n,1} = \frac{-1}{h_n^2}K^{(1)}\left(\frac{\theta_n - X_{n,1}}{h_n}\right).$$

By Proposition 5.1, we have, for $m = 1, 2, 2 + p$,

$$h_n^{2m-1}\mathbf{E}_{F_n}|V_{n,1}|^m \to f(\theta)\int_{-\infty}^{\infty}|K^{(1)}(y)|^m\,dy.$$

By assumption (c),

$$(nh_n^3)^{1/2}\mathbf{E}_{F_n}V_{n,1} \to -\mu.$$

Hence

$$(nh_n^3)\mathrm{Var}_{F_n}(\hat{f}_n^{(1)}(\theta_n)) \to f(\theta)\int_{-\infty}^{\infty}|K^{(1)}(y)|^2\,dy$$

and

$$\frac{\mathbf{E}_{F_n}|V_{n,1} - \mathbf{E}_{F_n}(V_{n,1})|^{2+p}}{n^{p/2}\sigma^{2+p}(V_{n,1})} = O(nh_n)^{-p/2} = o(1).$$

By Liapounov's CLT, (1) is proved.

Proof of (2). By Corollary 5.1 and Proposition 5.1, there exists a $\delta_0 > 0$ so that

$$\sup_{\{t:\,|t-\theta|<\delta_0\}} \left|\hat{f}_n^{(2)}(t) - f^{(2)}(t)\right| \to 0, \quad \text{in } F_n^n \text{ probability.}$$

Now use the fact that $\theta_n^* \to \theta$ in F_n^n probability. □

PROOF OF THEOREM 2.1. First, we prove part (ii) of the theorem for the fixed bandwidth case where S_n is identically 1. Apply Proposition 4.1 taking $F_n = F$. Conditions (a) and (b) are trivial. To verify (c),

$$\left(\frac{n}{h_n}\right)^{1/2} \mathbf{E}_{F_n}\left(K^{(1)}\left(\frac{\theta_n - X}{h_n}\right)\right) = \left(\frac{n}{h_n}\right)^{1/2} \int_{-\infty}^{\infty} K^{(1)}\left(\frac{\theta - y}{h_n}\right) f(y)\, dy.$$

By assumption (B), for and $\delta > 0$, this integral over the set $\{y\colon |y - \theta| > \delta\}$ tends to 0. On the set $\{y\colon |y - \theta| \le \delta\}$, we expand f about θ and, since $f^{(1)}(\theta) = 0$ and $K^{(1)}$ is odd, the preceding integral becomes

$$\left(\frac{n}{h_n}\right)^{1/2} \int_{\{y:\,|y-\theta|\le\delta\}} K^{(1)}\left(\frac{\theta - y}{h_n}\right) w(y)\, dy, \tag{4.8}$$

where $w(y) = f(y) - f(\theta) - \frac{1}{2}f^{(2)}(\theta)(y - \theta)^2$ and δ in (4.8) is chosen so that $f(y)$ has a continuous third derivative in $\{y\colon |y - \theta| \le \delta\}$. Now $w(y) = \frac{1}{6}f^{(3)}(v(y))(y - \theta)^3$, where $v(y)$ is between y and θ. Let $z = (\theta - y)/h_n$. Then (4.8) becomes

$$\frac{\left(nh_n^7\right)^{1/2}}{6} \int_{\{z:\,|z|\le\delta/h_n\}} K^{(1)}(z) z^3 f^{(3)}(v(\theta - h_n z))\, dz. \tag{4.9}$$

Note that, for any z, $f^{(3)}(v(\theta - h_n z)) \to f^{(3)}(\theta)$. Apply dominated convergence and then integration by parts to conclude that expression (4.9) converges to μ, where

$$\mu = \frac{d}{2} f^{(3)}(\theta) \int_{-\infty}^{\infty} z^2 K(z)\, dz.$$

We now turn to the general case. As in the proof of Proposition 4.1 [see (4.7)],

$$T_n(b) = \frac{V_n(b)}{D_n(b)},$$

where $V_n(b) = -(nb^3 v_n^3)^{1/2} \hat{f}_{n,\,bv_n}(\theta)$ and $D_n(b) = \hat{f}^{(2)}_{n,\,bv_n}(\theta^*_{n,\,bv_n})$ for some random variable $\theta^*_{n,\,bv_n}$ between θ and $\hat{\theta}_{n,\,bv_n}$. V_n can be regarded as a random variable on $\mathbf{C}[u_1, u_2]$.

Step 1. Show D_n converges (in the sup norm) to the constant function $f^{(2)}(\theta)$ almost surely. To see why, apply Corollary 5.3 (making the identifications $b_n = u_1 \cdot v_n$ and $a_n = u_n \cdot v_n$), and Proposition 5.1 to get there exists $\delta_0 > 0$ such that

$$\sup_{\{t:\,|t-\theta|<\delta_0\}} \sup_{\{b:\,u_1\le b\le u_2\}} \left|\hat{f}^{(2)}_{n,\,bv_n}(t) - f^{(2)}(t)\right| \to 0 \quad \text{a.s.} \tag{4.10}$$

By the proof of Theorem 1.1 [namely (4.1)],

$$\sup_{\{b:\, u_1 \le b \le u_2\}} |\hat{\theta}_{n,\, bv_n} - \theta| \to 0 \quad \text{a.s.} \tag{4.11}$$

and thus the same is true with $\hat{\theta}_{n,\, bv_n}$ replaced by $\theta^*_{n,\, bv_n}$. Thus

$$\sup_{\{b:\, u_1 \le b \le u_2\}} \left| \hat{f}^{(2)}_{n,\, bv_n}\left(\theta^*_{n,\, bv_n}\right) - f^{(2)}\left(\theta^*_{n,\, bv_n}\right) \right| \to 0 \quad \text{a.s.}$$

Apply (4.11) again the use the fact that $f^{(2)}(t)$ is continuous at $t = \theta$ to get the desired conclusion.

Step 2. Show the law of V_n converges weakly to the law of V. $V_n(b)$ can be represented as

$$\left(\frac{b^3 v_n^3}{n}\right)^{1/2} \sum_{i=1}^{n} V_{n,i}(b),$$

where, for each b, $V_{n,i}(b)$ are independent and identically distributed as

$$\frac{-1}{b^2 v_n^2} K^{(1)}\left(\frac{\theta - X_1}{bv_n}\right).$$

By the proof of Proposition 4.1 (i.e., the fixed bandwidth case),

$$\mathbf{E}\left[V_n(b)\right] \to b^{7/2} \frac{d}{2} f^{(3)}(\theta) \int_{-\infty}^{\infty} z^2 K(z)\, dz = \mathbf{E}\left[V(b)\right].$$

Furthermore,

$$\begin{aligned} \operatorname{Cov}\left[V_n(b_i), V_n(b_j)\right] = {} & \frac{1}{(b_i b_j)^{1/2} v_n} \int_{-\infty}^{\infty} K^{(1)}\left(\frac{\theta - x}{b_i v_n}\right) K^{(1)}\left(\frac{\theta - x}{b_j v_n}\right) f(x)\, dx \\ & - \frac{1}{(b_i b_j)^{1/2} v_n} \int_{-\infty}^{\infty} K^{(1)}\left(\frac{\theta - x}{b_i v_n}\right) f(x)\, dx \\ & \times \int_{-\infty}^{\infty} K^{(1)}\left(\frac{\theta - x}{b_j v_n}\right) f(x)\, dx. \end{aligned}$$

Make a change of variables by letting $z = v_n^{-1}(b_i b_j)^{-1/2}(\theta - x)$ and then apply dominated convergence to conclude the last expression converges to

$$\operatorname{Cov}\left[V(b_i), V(b_j)\right].$$

As in the proof of Proposition 4.1, it follows (by a multivariate version of Liapounov's CLT) that the law of $(V_n(b_i), \ldots, V_n(b_k))$ converges weakly to the law of $(V(b_i), \ldots, V(b_k))$.

To complete step 2, it must be shown that the sequence of laws of V_n is tight. By Theorem 12.3 of Billingsley [1], it suffices to show

$$\mathbf{E}\left[V_n(b_1) - V_n(b_2)\right]^2 \le A^2 (b_1 - b_2)^2,$$

for some finite A which is independent of n, b_1 and b_2. This can be verified by a straightforward, but tedious, calculation; see [12].

Finally, $D_n(S_n) \to f^{(2)}(\theta)$ with probability 1. From part (i) of the theorem, the law of $V_n(S_n)$ converges weakly to the law of $V(s)$. The result now follows by Slutsky. □

PROOF OF PROPOSITION 2.1. Let $a_n = \nu_n \cdot s_2$ and $b_n = \nu_n \cdot s_1$. Note that $b_n \le h_n \le a_n$ eventually with probability 1. By Corollary 5.1 and Proposition 5.1, there exists δ_0 such that

$$\sup_{\{b_n \le h \le a_n\}} \sup_{\{t:\, |t-x|<\delta_0\}} \left| \hat{f}^{(j)}_{n,h}(t) - f^{(j)}(t) \right| \to 0 \quad \text{a.s.} \qquad \square$$

PROOF OF THEOREM 3.1. It will suffice to establish the result for bounded l; if l is unbounded, replace it by $\min(l, c)$ (another subconvex loss function) and then let $c \to +\infty$. Without loss of generality, we may assume $M(f_0) = 0$ and set $a = -f_0^{(2)}(0)$. As in Hasminskii [7], consider an auxiliary function g satisfying g has a bounded derivative of order p, g is symmetric with compact support and $g(x) = x$ if $|x| < a^{-1}$. Let $f_n(x; \theta)$, $\theta \in [0,1]$, be the parametric family of functions defined by

$$f_n(x; \theta) = f_0(x) + \theta n^{-\beta p} g(x \cdot n^{\beta}),$$

where $\beta = (2p+1)^{-1}$. An easy, though slightly tedious, analysis (similar to Lemma 3.1 of Hasminskii [7]) shows that for any fixed $\theta \in [0,1]$, all sufficiently large n and all sufficiently small ε, $f_n(x; \theta)$ is a density belonging to $N_n(\varepsilon, p, f_0)$, and

$$M(f_n(\cdot, \theta)) = \theta a^{-1} \delta_n^{-1} + o(\delta_n), \quad \text{as } n \to \infty, \tag{4.12}$$

where $\delta_n = n^r$ and $r = (p-1)/(2p+1)$. Define P_θ^n to be the n-fold product measure of $f_n(\cdot, \theta)$ if $f_n(\cdot, \theta)$ is a density and, say, the n-fold product of f_0 otherwise. Then, the left-hand side of (3.1) is bounded below by

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{\theta} \int l\left[\delta_n(T_n - a^{-1}\theta)\right] dP_\theta^n. \tag{4.13}$$

Now, it is easy to see that the experiments $\{P_\theta^n, \theta \in [0,1]\}$ converge in the sense of Le Cam to a Gaussian shift experiment $\{P_\theta, \theta \in [0,1]\}$, either by application of Theorem 3.1′, page 178, of Ibragimov and Hasminskii [8], or by Proposition II.2.3 of Millar [9]. In fact, the information $\Psi^2(n, \theta)$ in experiment $\{P_\theta^n, \theta \in [0,1]\}$ (as defined in I.7 of Ibragimov and Hasminskii [8]), is given by

$$\Psi^2(n, \theta) = \int_{-\infty}^{\infty} \frac{g^2(y)}{f_0(yn^{-\beta}) + \theta n^{-\beta p} g(y)}\, dy.$$

Thus

$$\Psi^2(n, \theta) \to \frac{1}{f_0(0)} \int_{-\infty}^{\infty} g^2(y)\, dy = I,$$

and so P_θ is the Gaussian measure with mean θ and variance I^{-1}. Therefore, by the Hájek–Le Cam asymptotic minimax theorem and (4.12), (4.13) is bounded

below by

$$\inf_T \sup_{\theta \in [0,1]} \int l(T - a^{-1}\theta)\, dP_\theta = \inf_T \sup_{\theta \in [0,1]} \int l_a(T - \theta)\, dP_\theta,$$

where $l_a(x) = l(ax)$ is also a subconvex loss function. Finally, the minimax risk for a Gaussian shift experiment is strictly positive, and the result follows. □

5. Appendix. The proofs of the results involve the use of empirical processes indexed by classes of functions. The reader may refer to Pollard [11] for background.

THEOREM 5.1 (Pollard [11]). *Let $(S, \mathbf{S})$ be a measure space. Let $\mathbf{F}_n$ be a countable class of real-valued measurable functions f on S, with $|f| \leq 1$, and whose covering numbers satisfy*

$$\sup_Q N_1(\varepsilon, Q, \mathbf{F}_n) \leq A\varepsilon^{-W}, \quad \text{for } 0 < \varepsilon < 1, \tag{5.1}$$

with constants A and W independent of n. Let $X_{n,1}, \ldots, X_{n,n}$ be i.i.d. S-valued random variables with distribution P_n; let $\hat{P}_n$ denote the empirical measure of the $X_{n,i}$, $1 \leq i \leq n$. Choose δ_n so that

$$\sup_{f \in F_n} P_n f^2 \leq \delta_n^2.$$

Let α_n be a bounded sequence of positive numbers for which $n\delta_n^2\alpha_n^2/\log(n) \to \infty$. Then

$$\text{(i)} \qquad \sup_{f \in \mathbf{F}_n} \frac{|\hat{P}_n f - P_n f|}{\delta_n^2 \alpha_n} \to 0, \quad \text{in } P_n^n \text{ probability},$$

$$\text{(ii)} \qquad \sup_{f \in \mathbf{F}_n} \frac{|\hat{P}_n f - P_n f}{\delta_n^2 \alpha_n} \to 0, \quad a.s. \text{ if } P_n = P.$$

PROOF. See Theorem 37, and its proof in Pollard [11], page 34. He has $P_n = P$, but the proof carries over yielding convergence in P_n^n probability without this assumption. We will need to use the more general version. Pollard also assumes α_n is nonincreasing, but his proof goes through if we assume α_n is bounded. □

REMARK 5.1. The only classes of functions considered in this paper are classes (or subclasses) of functions of the form

$$\mathbf{F} = \left\{ g\left(\frac{\cdot - \theta}{\sigma} \right), -\infty < \theta < \infty, \sigma > 0 \right\},$$

where g is a Borel-measurable real-valued function on $\mathbf{R}$ of bounded variation. In this case, condition (5.1) in Theorem 5.1 is satisfied for $\mathbf{F}$ and hence for any sequence $\mathbf{F}_n$ of subclasses of functions of $\mathbf{F}$, by an argument similar to that of

Pollard [11], page 29, as given in Theorem 5.2 of Romano [12]. Furthermore, the function g will also be assumed continuous. Thus we can replace the supremum over $\mathbf{F}$ by a supremum over a countable subcollection of functions in $\mathbf{F}$ (by a dominated convergence argument), but more to the point, Theorem 5.1 will apply.

COROLLARY 5.1. *Let f be a density with mode θ satisfying* (A1); *let the kernel K have a continuous integrable jth derivative of bounded variation; Let $X_{n,1},\dots,X_{n,n}$ be i.i.d. with c.d.f. F_n; let $\hat{F}_n$ be the empirical c.d.f. of the $X_{n,i}$, $1 \le i \le n$. Set*

$$\hat{f}_{n,h}(t) = \frac{1}{h}\int K\left(\frac{t-y}{h}\right) d\hat{F}_n(y)$$

and

$$f_{n,h}(t) = \frac{1}{h}\int K\left(\frac{t-y}{h}\right) dF_n(y).$$

Assume $nb_n^{2j+2}/[\log(n)a_n] \to \infty$ and $b_n > 0$. Assume $f_{n,h}(t)$ is bounded by a constant C as t and h vary, for large enough n. Then

(i) $$\sup_{\{h:\, b_n \le h \le a_n\}} \sup_t \left|\hat{f}_{n,h}^{(j)}(t) - f_{n,h}^{(j)}(t)\right| \to 0, \quad \text{in } F_n^n \text{ probability},$$

(ii) $$\sup_{\{h:\, b_n \le h \le a_n\}} \sup_t \left|\hat{f}_{n,h}^{(j)}(t) - f_{n,h}^{(j)}(t)\right| \to 0 \quad a.s. \text{ if } F_n = F.$$

In particular, if h_n is a fixed bandwidth sequence such that $nh_n^{2j+1}/\log(n) \to \infty$ and $F_n = F$, then

$$\sup_t \left|\hat{f}_{n,h_n}^{(j)}(t) - f_{n,h_n}^{(j)}(t)\right| \to 0 \quad a.s.$$

PROOF. Apply Theorem 5.1 with

$$\mathbf{F}_n = \left\{\frac{1}{M}K^{(j)}\left(\frac{t-\cdot}{h}\right),\ t \in \mathbf{R},\ b_n \le h \le a_n\right\}$$

and M the maximum of $\sup|K^{(j)}|$ and the integral of $|K^{(j)}|$. Then

$$\sup_{f \in \mathbf{F}_n} P_n f^2 \le \sup_{\{h:\, b_n \le h \le a_n\}} \sup_t \frac{1}{M}\int |K^{(j)}(z)| f_{n,h}(t - hz) h\, dz \le Ca_n.$$

Put $\delta_n^2 = Ca_n$ and $\alpha_n = b_n^{j+1}/Ca_n$, and the result follows. □

PROPOSITION 5.1. *Fix an integer $j \ge 0$ and let a_n and b_n be positive numbers such that $a_n \to 0$ as $n \to \infty$. Assume K (not necessarily a density) has an integrable jth derivative and such that, for every $\delta > 0$,*

$$\frac{1}{b_n^j}\int_{\{z:\, |z| > \delta/a_n\}} |K^{(j)}(z)|\, dz \to 0.$$

Suppose g_n and g are densities satisfying:

1. *g is j-times continuously differentiable in some neighborhood of x.*
2. *g_n and $g_n^{(j)}$ converge to g and $g^{(j)}$, respectively, uniformly in some neighborhood of x.*
3. *$\sup|g_n| < M$, $\sup|g| < M$, for some $M < \infty$.*

Define

$$g_{n,h}(t) = \frac{1}{h}\int K\left(\frac{t-y}{h}\right)g_n(y)\,dy$$

and let

$$\alpha_{n,h}(t) = g_{n,h}^{(j)}(t) - g^{(j)}(t)\int_{-\infty}^{\infty} K(y)\,dy.$$

Then there exists a $\delta_0 > 0$ such that

$$\lim_{n\to\infty} \sup_{\{h:\, b_n < h \le a_n\}} \sup_{\{t:\, |t-x|<\delta_0\}} \left|\alpha_{n,h}(t)\right| = 0.$$

Hence, if $x_n \to x$, then $g_{n,a_n}^{(j)}(x_n) \to g^{(j)}(x)\int_{-\infty}^{\infty}K(y)\,dy$.

PROOF. Choose δ_0 so g is j-times continuously differentiable in the neighborhood $N = [x - 2\delta_0, x + 2\delta_0]$ and

$$\sup_{x\in N}\left|g_n^{(k)}(x) - g^{(k)}(x)\right| \to 0, \quad \text{for } k = 0, j.$$

Then, given $\varepsilon > 0$, we can find $\delta > 0$ so that for n large enough: If $|t - y| < \delta$, $t \in N$, $y \in N$, then $|g_n^{(j)}(t) - g^{(j)}(y)| < \varepsilon$. If $t \in N$ and h is fixed, then

$$\alpha_{n,h}(t) = \frac{d}{dt^j}\frac{1}{h}\int K\left(\frac{t-y}{h}\right)\left[g_n(y) - g(t)\right]dy.$$

Let I_1 be this integral evaluated over the set $\{y\colon |y - t| < \delta\}$ and let I_2 be the integral over the complement of this set. Then $|I_1| \le \varepsilon$ and

$$I_2 = \frac{1}{h^{j+1}}\int_{\{y:\,|y-t|\ge\delta\}} K^{(j)}\left(\frac{t-y}{h}\right)g_n(y)\,dy - \int_{\{y:\,|y-t|\ge\delta\}} g^{(j)}(t)\frac{1}{h}K\left(\frac{t-y}{h}\right)dy$$

$$\le \frac{M}{b_n^j}\int_{\{z:\,|z|\ge\delta/a_n\}}\left|K^{(j)}(z)\right|dx + \left|g^{(j)}(t)\right|\int_{\{z:\,|z|\ge\delta/a_n\}} K(z)\,dz.$$

Using these bounds, the supremum of $|\alpha_{n,h}(t)|$ over the appropriate values of t tends to 0, as seen by first letting $n \to \infty$ and then $\varepsilon \to 0$. □

COROLLARY 5.2. *Let $X_1, X_2, \ldots$ be i.i.d. with c.d.f. F. Assume F has a density f with mode θ satisfying* (A1). *Assume K (not necessarily a probability density) is integrable, continuous and of bounded variation. Suppose*

$nh_n/\log(n) \to \infty$ and $h_n \to 0$. Then there exists $\delta_0 > 0$ such that

$$\text{(i)} \quad \sup_{\{t:\,|t-\theta|<\delta_0\}} \left| \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - X_1}{h_n}\right) - f(t)\int_{-\infty}^{\infty} K(y)\,dy \right| \to 0 \quad a.s.$$

Let $\hat{\theta}_n = M(\hat{f}_n)$ be the sample mode obtained from the first n observations. Then

$$\text{(ii)} \quad \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{\hat{\theta}_n - X_i}{h_n}\right) \to f(\theta)\int_{-\infty}^{\infty} K(y)\,dy \quad a.s.$$

PROOF. (i) Combine Proposition 5.1 with $j = 0$ and Corollary 5.1.
(ii) Theorem 1.1 yields $\hat{\theta}_n \to \theta$ a.s. Apply (i). □

REFERENCES

[1] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
[2] CHERNOFF, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31–41.
[3] DALENIUS, T. (1965). The mode—a neglected statistical parameter. *J. Roy. Statist. Soc. Ser. A* **128** 110–117.
[4] EDDY, W. (1980). Optimum kernel estimators of the mode. *Ann. Statist.* **8** 870–882.
[5] EDDY, W. (1982). The asymptotic distributions of kernel estimators of the mode. *Z. Wahrsch. verw. Gebiette* **59** 279–290.
[6] GRENANDER, U. (1965). Some direct estimates of the mode. *Ann. Math. Statist.* **36** 131–138.
[7] HASMINSKII, R. Z. (1979). Lower bounds for the risk of nonparametric estimates of the mode. In *Contributions to Statistics, Jaroslav Hájek Memorial Volume* (J. Jurečková, ed.) 91–97. Academia, Prague.
[8] IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, Berlin.
[9] MILLAR, P. W. (1983). The minimax principle in asymptotic statistical theory. *Ecole d'Eté de Probabilités de Saint Flour XI, 1981. Lecture Notes in Math.* **976** 75–265. Springer, Berlin.
[10] PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.
[11] POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
[12] ROMANO, J. (1986). On bootstrapping the joint distribution of the location and size of the mode. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.
[13] SAMANTA, M. (1973). Nonparametric estimation of the mode of a multivariate density. *South African Statist. J.* **7** 109–117.
[14] SILVERMAN, B. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6** 177–184.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305