ESTIMATION OF THE MODE*

HERMAN CHERNOFF

1. Introduction

The problems of estimating the density and the mode of a distribution are rather delicate and this delicacy is related to the fact that the density may be changed considerably over a short range without affecting the probability distribution substantially. On the other hand certain functions of the probability distribution which are related to the density are relatively easy to estimate. The key to the approach used by Rosenblatt, Whittle, and Parzen (see [2], [4], [5]) is that if $K(x)$ is a well behaved "kernel" function,

$$g(x)=\int K(x-y)dF(y) \tag{1.1}$$

is not sensitive to small changes in the probability distribution F. Note that if $K(x)$ is a probability density function, $g(x)$ represents the density of a "smoothened" version of F. Then, if F has a continuous density $f(x)$,

$$g_\delta(x)=\int \frac{1}{\delta}K\left(\frac{x-y}{\delta}\right)dF(y)\to f(x) \qquad \text{as } \delta\to 0\,, \tag{1.2}$$

and the problem of estimating $g(x)$ or $g_\delta(x)$ is related to that of estimating the density $f(x)$.

It seems natural to define $\tilde{x}$, *the mode with respect to a bounded measurable function* $K(x)$, as the value of x which maximizes $g(x)$. It is natural to estimate $\tilde{x}$ by the analogous function of the sample c.d.f. F_n, i.e., by $\hat{x}$, the value of x which maximizes

$$h(x)=\int K(x-y)dF_n(y)\,, \tag{1.3}$$

or by $\hat{x}_{s_n}$, which maximizes

* This work was supported in part by the Office of Naval Research Contract Nonr-225 (52) and National Science Foundation Grant GP 2220 at Stanford University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

$$h_{\delta_n}(x)=\int \frac{1}{\delta_n}K\left(\frac{x-y}{\delta_n}\right)dF_n(y)\,. \tag{1.4}$$

The estimate $\hat{x}$ may be called the analogue estimate of the mode with respect to K.

In [2] Parzen has stated conditions under which h_{δ_n} gives consistent and asymptotically normal estimates of $f(x)$ and under which $\hat{x}_{\delta_n}$ gives consistent and asymptotically normal estimates of the mode of the distribution. The latter results do *not* apply to the naive estimator of the mode which is the center of that interval of length $2a$ which contains the most observations. This estimator is of course the analogue estimate of the mode with respect to the kernel $K_a(x)=1/2a$ for $|x|\leqq a$ and 0 elsewhere.

This paper deals mainly with the application of the heat equation and Wiener-Lévy process to the derivation of the asymptotic distribution of the naive estimator. As a preliminary we informally outline a variation of Parzen's proof of the asymptotic normality of x under smoothness conditions on $K(x)$.

2. Asymptotic normality of the analogue estimate of the mode, under regularity conditions

We assume that $\tilde{x}$ is uniquely defined and that f and K are such that $g(x)$ has a negative second derivative $-c$ at $x=\tilde{x}$. This is the case if K, K' and K'' are bounded which we shall also assume. Then

$$0=g'(\tilde{x})=\int K'(\tilde{x}-y)dF(y)=E\{K'(\tilde{x}-X)\} \tag{2.1}$$

while

$$-c=g''(\tilde{x})=\int K''(\tilde{x}-y)dF(y)=E\{K''(\tilde{x}-X)\}\,. \tag{2.2}$$

The estimator $\hat{x}$ based on n independent observations $X_1, X_2, \cdots, X_n$ maximizes

$$\begin{aligned} H(x)-H(\tilde{x}) &= \int [K(x-y)-K(\tilde{x}-y)][dF_n(y)-dF(y)] \\ &\quad +\int [K(x-y)-K(\tilde{x}-y)]dF(y) \\ &\approx (x-\tilde{x})\int K'(\tilde{x}-y)\,[dF_n(y)-dF(y)]-\frac{c}{2}(x-\tilde{x})^2\,, \end{aligned} \tag{2.3}$$

and

$$\hat{x}-\tilde{x}\approx c^{-1}\int K'(\tilde{x}-y)[dF_n(y)-dF(y)]=c^{-1}\frac{1}{n}\sum_{i=1}^{n}K'(\tilde{x}-X_i)\,. \tag{2.4}$$

With some care one can apply the above outline to prove rigorously that

$$\mathfrak{L}[\sqrt{n}\,(\hat{x}-\tilde{x})] \to \mathfrak{N}(0, \sigma^2) \tag{2.5}$$

where

$$\sigma^2=\frac{E\{[K'(\tilde{x}-X)]^2\}}{E^2[K''(\tilde{x}-X)]} \tag{2.6}$$

and where $\mathfrak{L}$ represents distribution law while $\mathfrak{N}(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2. The study of $\hat{x}_{\delta_n}$ with $\delta_n \to 0$ as $n \to \infty$, is a bit more delicate, but is also amenable to this treatment.

3. The naive estimator

If $K_a(x)=1/2a$ for $|x| \leqq a$ and 0 for $|x|>a$, $K'(x)$ is undefined at $x=\pm a$ and the conditions of Parzen's results fail to apply. In this case we shall assume that the density exists and is continuously differentiable. Furthermore we assume that $\tilde{x}_a$, the mode with respect to K_a and $\tilde{x}_0$, the true mode are uniquely defined.

For the time being we bypass consistency considerations and assume whatever conditions are required to assure that $\hat{x}_a$ is a consistent estimator of $\tilde{x}_a$. Incidentally, the naive estimator is not uniquely defined. Typically there are one or several intervals, any point of which is the center of an interval of length $2a$ which contains a maximal number of observations. While one could modify the definition to make the estimate unique, this will not be necessary, since these intervals typically all lie within a range which is small compared to the natural variability of $\hat{x}_a$.

In this section we shall indicate heuristically how the asymptotic distribution of the naive estimator is related to the distribution of $\hat{z}$ which maximizes $Z(z)-z^2$ where $Z(z)$ is a two-sided Wiener-Lévy process (Gaussian process with independent increments) through the origin with mean 0 and variance one per unit z.

First we note that for $\tilde{x}_a$ to be the mode with respect to K_a it is necessary that $\tilde{x}_a$ maximize $F(x+a)-F(x-a)$ and hence that

$$f(\tilde{x}_a+a)=f(\tilde{x}_a-a)\,. \tag{3.1}$$

The naive estimator $\hat{x}_a$ maximizes $h(x)$ and hence maximizes

$$h(x)-h(\tilde{x}_a)=[F_n(x+a)-F_n(x-a)]-[F_n(\tilde{x}_a+a)-F_n(\tilde{x}_a-a)]\,.$$

But

$$h(x)-h(\tilde{x}_a)=n^{-1/2}Y_n+u \tag{3.2}$$

where

$$n^{-1/2}Y_n=\{[F_n(x+a)-F(x+a)]-[F_n(\tilde{x}_a+a)-F(\tilde{x}_a+a)]\} - \{[F_n(x-a)-F(x-a)]-[F_n(\tilde{x}_a-a)-F(\tilde{x}_a-a)]\} \tag{3.3}$$

and

$$u=[F(x+a)-F(\tilde{x}_a+a)]-[F(x-a)-F(\tilde{x}_a-a)]\,. \tag{3.4}$$

For small $y=x-\tilde{x}_a$

$$u=-\frac{1}{2}cy^2[1+o(1)] \tag{3.5}$$

with

$$c=f'(\tilde{x}_a-a)-f'(\tilde{x}_a+a)\,. \tag{3.6}$$

For $y\geqq 0$, y small compared to a and large compared to n^{-1}, $n^{1/2}Y_n$ represents the deviation from expectation of the number of observations between $\tilde{x}_a+a$ and $\tilde{x}_a+a+y$ minus the number of observations between $\tilde{x}_a-a$ and $\tilde{x}_a-a+y$. The numbers of observations in these small non-overlapping intervals are approximately uncorrelated, and approximately normally distributed with mean 0 and variance $nyf(\tilde{x}_a+a)=nyf(\tilde{x}_a-a)$. For $y^*>y$, $n^{1/2}[Y_n(y^*)-Y_n(y)]$ is the deviation from expectation of the number of observations between $\tilde{x}_a+a+y$ and $\tilde{x}_a+a+y^*$ minus the number between $\tilde{x}_a-a+y$ and $\tilde{x}_a-a+y^*$. Thus $Y_n(y)$ is approximately a Gaussian process with independent increments, mean 0 and variance $2f(\tilde{x}_a+a)$ per unit y. This heuristic argument also applies for $y<0$ making plausible the claim that Y_n behaves asymptotically like a two-sided Wiener-Lévy process Y and that $\hat{y}=\hat{x}_a-\tilde{x}_a$ is asymptotically distributed like the value of y which maximizes

$$n^{-1/2}Y(y)-\frac{1}{2}cy^2\,. \tag{3.7}$$

Let

$$y=rz\,, \tag{3.8}$$

$$Y(y)=[2f(\tilde{x}_a+a)r]^{1/2}Z(z)\,, \tag{3.9}$$

and

$$r=\left[\frac{8f(\tilde{x}_a+a)}{nc^2}\right]^{1/3}. \tag{3.10}$$

Then $\hat{z}=r^{-1}\hat{y}$ maximizes

$$Z(z)-z^2 \tag{3.11}$$

where Z is a two-sided Wiener-Lévy process with mean 0 and variance

one per unit z. This maximum occurs with probability one for z finite and thus $\hat{z}=O_p(1)$ and

$$\hat{y}=\hat{x}_a-\tilde{x}_a=O_p(n^{-1/3})\,. \tag{3.12}$$

In section 4 we derive an expression for the probability distribution of $\hat{z}$. It follows that this is the limiting distribution of

$$\frac{(\hat{x}_a-\tilde{x}_a)}{\left[\dfrac{8f(\tilde{x}_a+a)}{nc^2}\right]^{1/3}} \qquad \text{as } n\to\infty\,. \tag{3.13}$$

4. Distribution of $\hat{z}$

In this section we relate the distribution of $\hat{z}$ which maximizes $Z(z)-z^2$ to a solution of the heat equation. First let us define

$$u(x,t)=P\{Z(z)>z^2 \text{ for some } z>t|Z(t)=x\}\,, \tag{4.1}$$

where $Z(t)$ is a Wiener-Lévy process for $z>t$ originating at $Z(t)=x$. Then, for $x<t^2$,

$$u(x,t)=E\{u(x+\epsilon\sqrt{h}\,,\,t+h)\}+o(h) \qquad \text{as } h\to 0\,,$$

where ϵ is normally distributed with mean 0 and variance 1. The usual expansion argument yields

$$\frac{1}{2}u_{xx}=u_t \qquad \text{for } x<t^2 \tag{4.2a}$$

subject to the "boundary" conditions

$$u(x,t)=1 \qquad \text{for } x\geqq t^2 \tag{4.2b}$$

and

$$u(x,t)\to 0 \qquad \text{for } x\to-\infty\,. \tag{4.2c}$$

Now let

$$H_v(m)=P\{\max_{0\leq z\leq 1} Z(z)\leqq m|Z(0)=0\,,\ Z(1)=v\}\,. \tag{4.3}$$

Given $Z(t)=x$ and $Z(t+h)=x+\epsilon\sqrt{h}$, the maximum value of $Z(z)-z^2$ over the range $t<z<t+h$ is $-t^2+x+M\sqrt{h}+O(h)$ where M has the distribution $H_\epsilon(m)$. Given x, M, and ϵ, the probability that $Z(z)-z^2>-t^2+x+M\sqrt{h}+O(h)$ for some $z>t+h$ is the conditional probability that

$$Z(z)+t^2-x-M\sqrt{h}+O(h)>z^2 \qquad \text{for some } z>t+h$$

given

$$Z(t+h)+t^2-x-M\sqrt{h}+O(h)=t^2+(\epsilon-M)\sqrt{h}+O(h).$$

This probability is

$$P^+=u[t^2+(\epsilon-M)\sqrt{h}+O(h),\ t+h]. \tag{4.4}$$

Since $\epsilon-M<0$ with probability one, the arguments of the above expression lie in the domain where the heat equation holds, (for h small enough) and we may expand in terms of the boundary derivatives of the heat equation. Thus,

$$P^+=u(t^2, t)+(\epsilon-M)\sqrt{h}\,u_x(t^2, t)+O(h). \tag{4.5}$$

Note that P^+ represents the conditional probability that the maximum of $Z(z)-z^2$ over the range $z>t$ is attained for $z>t+h$, given $Z(t)=x$, $Z(t+h)=x+\epsilon\sqrt{h}$, and M. By symmetry, the conditional probability that the maximum of $Z(z)-z^2$ over the range $z<t+h$ is attained for $z<t$ given x, ϵ, and M is

$$P^-=u(t^2, -t)-M\sqrt{h}\,u_x(t^2, -t)+O(h). \tag{4.6}$$

Since increments are independent it follows that the conditional probability, given x, ϵ, and M, that $\hat{z}$ is between t and $t+h$ is

$$P=M(\epsilon-M)hu_x(t^2, t)u_x(t^2, -t)+o(h). \tag{4.7}$$

It follows that the density of $\hat{z}$ is

$$f^*(z)=K\alpha(z)\alpha(-z) \tag{4.8}$$

where

$$\alpha(z)=u_x(z^2, z) \tag{4.9}$$

and

$$K=E\{M(\epsilon-M)\}=\left[2\int_0^\infty \alpha(z)\alpha(-z)dz\right]^{-1}. \tag{4.10}$$

The standard reflection argument gives

$$P\{M>m\}=2[1-\Phi(m)]$$

where Φ is the normal c.d.f., and

$$E\{M^2\}=1.$$

For $\epsilon^*>0$,

$$\mathfrak{L}(M|\epsilon=\epsilon^*)=\mathfrak{L}(M+\epsilon^*|\epsilon=-\epsilon^*).$$

Hence

$$E(\epsilon M)=\frac{1}{2}E(\epsilon^2)=\frac{1}{2}$$

and

$$K=\frac{1}{2}.$$

We have proved

THEOREM 1. *The probability density function of $\hat{z}$, that value of z which maximizes $Z(z)-z^2$ where Z is a two-sided Wiener-Lévy process with mean 0 and variance 1 per unit z is given in terms of the solution of the heat equation (1) by*

$$f^*(z)=\frac{1}{2}u_x(z^2, z)u_x(z^2, -z). \tag{4.11}$$

5. Miscellaneous remarks

This section is devoted to the discussion of the rigorous proofs bypassed in section 3 and to some miscellaneous remarks.

THEOREM 1. *The estimator $\hat{x}_a$ is a consistent estimator of $\tilde{x}_a$.*

PROOF. Let

$$p_a=F(\tilde{x}_a+a)-F(\tilde{x}_a-a). \tag{5.1}$$

Given $\delta>0$, there is a $\lambda(a, \delta)>0$ such that

$$\sup_{|x-\tilde{x}_a|\geqq\delta} P\{x-a\leqq X\leqq x+a\}=p_a[1+3\lambda(a, \delta)]^{-1}<p_a. \tag{5.2}$$

Select $u_0=-\infty<u_1<u_2<\cdots<u_{M-1}<u_M=\infty$ so that $P\{u_i\leqq X\leqq u_{i+1}\}=p_a/r=M^{-1}$, where r is a real number large enough so that there is an integer m for which $m/r<(1+2\lambda)^{-1}$ but $(m-2)/r>(1+3\lambda)^{-1}$. For every point x outside $(\tilde{x}_a-\delta, \tilde{x}_a+\delta)$, the interval $(x-a, x+a)$ is a subinterval of at least one (u_i, u_{i+m}) each of which has probability less than $(1+2\lambda)^{-1}p_a$.

With probability approaching one the interval $(\tilde{x}_a-a, \tilde{x}_a+a)$ contains more than $np_a(1+\lambda)^{-1}$ observations. Thus to prove consistency it suffices to show that as $n\to\infty$,

$$\sum_{i=0}^{M-m} P\{(u_i, u_{i+m}) \text{ contains more than } np_a(1+\lambda)^{-1} \text{ observations}\}\to 0. \tag{5.3}$$

Since each of the intervals (u_i, u_{i+m}) has probability $p_a(1+2\lambda)^{-1}$, the result follows.

The case where $a=a_n\to 0$ as $n\to\infty$ requires an extension of this

argument based on the following lemma.

LEMMA 1. *If X has a binomial distribution with parameters n and p and $\epsilon>0$,*

$$P\{X>(1+\epsilon)np\}<\exp\{-np[1+\epsilon+\epsilon^2-e^\epsilon]\}. \tag{5.4}$$

PROOF. In general $P\{X>a\}\leqq E\{\exp[t(X-a)]\}$ for $t>0$. Hence

$$P\{X>(1+\epsilon)np\}<[pe^t+(1-p)]^n\exp[-t(1+\epsilon)np] \qquad \text{for } t>0.$$

Let $t=\epsilon$ and using the fact that $\log(1+x)<x$, we have

$$P\{X>(1+\epsilon)np\}<\exp\{n[p(e^\epsilon-1)-\epsilon(1+\epsilon)p]\}$$

from which the lemma follows.

THEOREM 2. *If $f(x)$ is bounded away from $f(\tilde{x}_o)$ outside every neighborhood of $\tilde{x}_o$, $a_n\to 0$, and $na_n+k\log a_n\to\infty$ for every $k>0$, then $\hat{x}_{a_n}$ is a consistent estimator of $\tilde{x}_o$.*

PROOF. We refer to the detailed proof above for fixed a. Under our assumptions, for fixed $\delta>0$, $\lambda=\lambda(a_n,\delta)$ is bounded away from zero as $a_n\to 0$ and $pa_n\approx f(\tilde{x}_0)a_n$. Then the appropriate r and m can be selected so that they are bounded and $M=O(a_n^{-1})$. Selecting an ϵ so that $1+\epsilon<(1+2\lambda)/(1+\lambda)$ and $1+\epsilon+\epsilon^2-e^\epsilon=\eta>0$, lemma 1 yields the fact that the sum in (5.3) is less than $M\exp[-np_{a_n}\eta/(1+2\lambda)]=o(1)$ according to our assumption. That $(\tilde{x}_0-a_n,\tilde{x}_0+a_n)$ contains more than $np_{a_n}(1+\lambda)^{-1}$ observations with probability approaching one is also easily seen. The consistency follows.

The reduction of the asymptotic distribution of $\hat{x}_a-\tilde{x}_a$ to that of $\hat{z}$ in section 4 involves a theorem which states that if (i) the distribution of W_n converges to the distribution of W and (ii) $g(W)$ is a function whose set of discontinuities has probability zero with respect to the distribution of W, then the distribution of $g(W_n)$ converges to that of $g(W)$. In our application W_n is a stochastic process related to Z and Y_n of section 3,

$$W(z)=Z(z)-z^2 \tag{5.5}$$

where Z is the two-sided Wiener-Lévy process of section 4, and $g(W)$ is that value of z for which W is maximized.

To determine continuity of g or to discuss convergence of distributions of stochastic processes, it is necessary to introduce a topology on the set of W. We use a topology related to that induced by the metric

$$\rho(W,W^*)=\sup_z[W(z)-W^*(z)](1+z^2)^{-1}. \tag{5.6}$$

A bare outline of a proof is presented here. The argument basically follows that used by Prokhorov (see [3], section 2.4) in illustrating his results with the Kolmogorov statistic.

First we refer to section 3 to define

$$Z_n(z)=[2f(\tilde{x}_a+a)r_n]^{-1/2}Y_n \tag{5.7}$$

where $x=\tilde{x}_a+y=x_a+r_nz$, $r_n=[8f(\tilde{x}_a+a)/nc^2]^{1/3}$. Then $\hat{z}_n=(\hat{x}_a-\tilde{x}_a)/r_n$ maximizes

$$W_n(z)=Z_n(z)-\theta(y)z^2$$

where (see equation (3.5))

$$\theta(y)=-\frac{2u}{cy^2}\to 1 \qquad \text{as } y\to 0\,.$$

The fact that $\theta(y)\to 0$ as $|y|\to\infty$ leads to some difficulty which we avoid by applying the consistency of $\hat{x}_a$ and modifying W_n. Let

$$W_n^*(z)=Z_n(z)-\theta_n^*(z)z^2 \tag{5.8}$$

where

$$\theta_n^*(z)=\begin{cases}\theta(s_n) & \text{for } y>s_n\\ \theta(-s_n) & \text{for } y<-s_n\\ \theta(y) & \text{for } -s_n\leqq y\leqq s_n\end{cases}$$

and where $\{s_n\}$ is a sequence of real numbers such that $s_n\to 0$, $s_n/r_n\to\infty$, and

$$P\{|\hat{x}_a-\tilde{x}_a|>s_n\}\to 1 \qquad \text{as } n\to\infty\,.$$

The consistency of $\hat{x}_a$ implies the existence of $\{s_n\}$ and that $g(W_n)$ has the same limiting distribution as that of $g(W_n^*)$ if the latter is non-degenerate. Note that

$$\sup_z|\theta_n^*(z)-1|\to 0 \qquad \text{as } n\to\infty\,. \tag{5.9}$$

The first main step is to show that the limiting distribution of W_n^* is equal to the distribution of W. The second is to establish that $g(W)$ is continuous on a set of W of probability one.

The heuristic demonstration in section 3 is basically adequate to show that for every k-tuple $(z_1, z_2, \cdots, z_k)$, the joint distribution of $(W^*(z_1), W^*(z_2), \cdots, W^*(z_k))$ converges to that of $(W(z_1), W(z_2), \cdots, W(z_k))$. In addition to this a compactness condition must be satisfied. To make it easier to refer to results in [3] where the processes studied are restricted to $[0, 1]$ and continuous in the limit, we let

$$V_n(t)=Z_n(z)(1+z^2)^{-1} \tag{5.10}$$

$$V(t)=Z(z)(1+z^2)^{-1} \tag{5.11}$$

$$t=z(1+z^2)^{-1/2}\,. \tag{5.12}$$

Thus $V_n(t)\to 0$ and $V(t)\to 0$ as $t\to -1$ or 1 while $z\to\pm\infty$. Again, for every k-tuple $t_1, t_2, \cdots, t_k$, $-1\leqq t_i\leqq 1$, we have the convergence in distribution of $(V_n(t_1), V_n(t_2), \cdots, V_n(t_k))$. Note that the metric (5.6) corresponds to maximum deviation in the V functions.

Prokhorov's illustration is applicable to V_n with minor modifications. Our problem permits some simplification because $Z_n(0)=V_n(0)=0$ and hence it suffices to deal with $0\leqq t\leqq 1$. Furthermore Z_n remains constant for $y\geqq 2a$ (and also for $y\leqq -2a$). On the other hand our problem is more complicated in several respects. First Z_n is not a Markov process. Analysis is simplified if we decompose Z_n into $Z_n=Z_n^+-Z_n^-$ where

$$\begin{aligned} Z_n^+&=n^{2/3}\{[F_n(x+a)-F_n(\tilde{x}_a+a)]-[F(x+a)-F(\tilde{x}_a+a)]\}\\ Z_n^-&=n^{2/3}\{[F_n(x-a)-F_n(\tilde{x}_a-a)]-[F(x-a)-F(\tilde{x}_a-a)]\}\,. \end{aligned} \tag{5.13}$$

Then an analysis similar to Prokhorov's, applied to the Markov process (Z_n^+, Z_n^-), shows that the distribution of (Z_n^+, Z_n^-) converges for z in any bounded interval $[0, z_0]$, (t bounded away from 1). Second, it is required to account for the neighborhood of $z=\infty$ ($t=1$). To do so, it suffices to show that

$$P\{\sup_{t\leq t'\leq 1}|V_n(t')|\geqq\epsilon\}\to 0 \tag{5.14}$$

uniformly in n as $t\to 1$. This may be accomplished by noting that $(Z_n^+)^2-n^{1/3}[F(x+a)-F(\tilde{x}_a+a)]$ is a lower semi-martingale, and applying semi-martingale inequalities to $[Z_n^+(z)/1+2^{2r}z_0^2]^2$ in the intervals $(2^r z_0, 2^{r+1}z_0)$, $r=1, 2, \cdots$, (see [1]). In this way we conclude that the limiting distribution of V_n is the distribution of V. This result applies with respect to the topology called d convergence or Skorokhod convergence [3] on the class of functions with discontinuities of the first kind. This topology is slightly weaker than that of the metric of maximum deviation. Finally $\theta_n^*(z)z^2$ is a deterministic continuous function which converges to z^2 and W_n^* converges in distribution to W.

The second major step consists of showing that the set of discontinuities W of the function g has probability zero with respect to the limiting distribution. With probability one, the two-sided Wiener-Lévy process yields a continuous realization $W(z)$. It follows that V is continuous and since Skorokhod convergence to a *continuous* function V is equivalent to uniform convergence, it suffices to show that W is a point of continuity with respect to the ρ metric (5.6). With probability one,

(i) $Z(z)=O[2|z|\log\log|z|]^{1/2}$ as $|z|\to\infty$ and hence $W(z)\approx -z^2$, and (ii) for $\delta>0$, $W(z)$ is bounded away from its maximum outside $(g(W)-\epsilon, g(W)+\epsilon)$. It follows easily that with probability one, W is a point of continuity of g. This concludes our outlined demonstration.

We conclude our paper with a heuristic discussion of the results of sections 2 and 3 in the cases where $\delta_n\to 0$ and $a_n\to 0$. Suppose that the distribution has mode $\tilde{x}_0=0$ with density $f(x)=c_0+c_2x^2+c_3x^3+o(x^3)$ for $x\to 0$. If K and f are symmetric $\tilde{x}_\delta=\tilde{x}_0$ for δ sufficiently small and there are no bias effects. It is more interesting to consider the case where K is symmetric but $c_2<0$ and $c_3\neq 0$. Then, it can be shown that $\tilde{x}_{\delta_n}\sim\delta_n^2$. On the other hand, the asymptotic variance of the estimate is of the order of magnitude of $(n\delta^3)^{-1}$. Taking $\delta_n\sim n^{-1/7}$ indicates the possibility of obtaining an estimator $\hat{x}_{\delta_n}$ such that $\hat{x}_{\delta_n}-\tilde{x}_0=O_p(n^{-2/7})$.

A similar analysis can be applied to the naive estimator with $a_n\to 0$. Then $\tilde{x}_{a_n}\sim a_n^2$ and the variance is of the order of $n^{-2/3}a_n^{-4/3}$. Setting $a_n\sim n^{-1/8}$ would give $\hat{x}_{a_n}-\tilde{x}_0=O_p(n^{-1/4})$.

Finally suppose that $f(x)$ does not have a derivative at the mode. For example we could have $f(x)=c_0+c_1x+o(x)$ for $x>0$, $x\to 0$ and $f(x)=c_0+c_1^*x+o(x)$ for $x<0$, $x\to 0$. It seems evident that the approach of this paper should be applicable to this case.

STANFORD UNIVERSITY

REFERENCES

[1] J. L. Doob, *Stochastic Processes*, Wiley, New York, 1953.

[2] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, 33 (1962), 1065-1076.

[3] Yu. V. Prokhorov, "Convergence of random processes and limit theorems in probability theory," *Theory of Prob. and Applic.* (translation of *Teor. Veroyatnost. i Primenen.*), 1 (1956), 157-214.

[4] M. Rosenblatt, "Remarks on some non-parametric estimators of a density function," *Ann. Math. Statist.*, 27 (1956), 832-837.

[5] P. Whittle, "On the smoothing of probability density functions," *J. R. Statist. Soc. (B)*, 20 (1958), 334-343.