The Annals of Statistics
1992, Vol. 20, No. 2, 675–694

EFFECT OF BIAS ESTIMATION ON COVERAGE ACCURACY OF BOOTSTRAP CONFIDENCE INTERVALS FOR A PROBABILITY DENSITY

BY PETER HALL

Australian National University

The bootstrap is a poor estimator of bias in problems of curve estimation, and so bias must be corrected by other means when the bootstrap is used to construct confidence intervals for a probability density. Bias may either be estimated explicitly, or allowed for by undersmoothing the curve estimator. Which of these two approaches is to be preferred? In the present paper we address this question from the viewpoint of coverage accuracy, assuming a given number of derivatives of the unknown density. We conclude that the simpler, undersmoothing method is more efficacious. Undersmoothing also has advantages from the standpoint of minimizing interval width. We derive formulae for bandwidths which are optimal in terms of coverage accuracy and also give formulae for the coverage error which results from using those bandwidths.

1. Introduction. The bootstrap can be a notoriously poor estimator of bias in problems of nonparametric curve estimation (e.g., [5]), and this fact should be taken into account when using bootstrap methods to construct confidence intervals. Bias is usually significant when a curve estimator is constructed so as to minimize mean squared error, but decreases as the amount of statistical smoothing is reduced. Therefore, one way of alleviating bias is to smooth the curve estimator less than would be appropriate for point estimation (as opposed to interval estimation). A statistician must either estimate bias explicitly, or allow for bias by undersmoothing the curve estimator. Which of these two methods is to be preferred? In this paper we address this problem in the case of nonparametric density estimation and from the viewpoint of coverage error. We use the percentile-t approach to confidence interval construction, and show that, in a sense which we shall make precise two paragraphs below, the undersmoothing method produces confidence intervals with greater coverage accuracy than those obtained by explicit bias correction.

To define the basic estimators, let $\mathscr{X} = \{X_1, \dots, X_n\}$ denote a random sample from the distribution with density f and write

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left\{\frac{x - X_i}{h}\right\} \tag{1.1}$$

for the density estimator based on kernel K and bandwidth h. Let $\hat{b}$ and $\hat{\sigma}^2$

Received January 1990; revised September 1991.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62E20.
Key words and phrases. Bias, bootstrap, confidence interval, coverage, smoothing.

denote estimates of the bias b and variance σ^2, respectively, of $\hat{f}$. Let u_α be the α-level quantile of the distribution of $S = (\hat{f} - E\hat{f})/\hat{\sigma}$, that is, the solution of the equation $P(S \le u_\alpha) = \alpha$; and let $\hat{u}_\alpha$ denote the bootstrap estimator of u_α. Then a one-sided α-level confidence interval for f has the form $(\hat{f} - \hat{\sigma}\hat{u}_\alpha - \hat{b}, \infty)$. By combining two such intervals, we obtain a two-sided interval $(\hat{f} - \hat{\sigma}\hat{u}_{(1+\alpha)/2} - \hat{b}, \hat{f} - \hat{\sigma}\hat{u}_{(1-\alpha)/2} - \hat{b})$.

We are now in a position to illustrate our main results. Assume that f has four derivatives and that K is a second-order kernel. Then the bias b of $\hat{f}$ is approximately proportional to $h^2 f''$, where h denotes bandwidth. If we estimate f'' using the second derivative of a second-order kernel density estimator with a new bandwidth h_1, say, then we need the full force of our assumption of four derivatives of f. It turns out that the minimum coverage error of a two-sided bias-corrected bootstrap confidence interval constructed in this manner is in general of size $n^{-12/17}$ and is obtained with h, h_1 of sizes $n^{-5/17}, n^{-1/17}$, respectively. An alternative way of solving the same problem, which also makes full use of the assumption of four derivatives, is to employ a fourth-order kernel (and bandwidth h) to estimate f and control bias by undersmoothing the estimator. In this approach it is optimal to take h to be of size $n^{-1/5}$, in which case the coverage error is of size $n^{-4/5}$, which is of smaller order than $n^{-12/17}$. (We should stress that we are using a *fourth*-order kernel here and so the familiar quantities $n^{-1/5}$ and $n^{-4/5}$ should not be interpreted in the usual manner for second-order kernels; compare, e.g., Silverman [14], pages 40 and 41.)

In the example discussed above and in other cases, both the minimum coverage error and the squared width of the confidence interval are (asymptotically) proportional to constant multiples of $(nh)^{-1}$, using either explicit bias estimation or the undersmoothing method. Therefore a method which is preferable on the grounds of minimizing coverage error, is also advantageous from the viewpoint of minimizing width.

A more obvious advantage of the undersmoothing method is that it requires choice of only one bandwidth. Furthermore, it does not involve explicit estimation of high-order density derivatives, a task which can be quite awkward.

The exact size of different bandwidths, such as $n^{-5/17}$ or $n^{-1/5}$, is not the most important conclusion to be drawn from this work. The main point is that methods based on undersmoothing can offer important improvements in coverage accuracy, both theoretically and empirically (see the simulation study summarized in Section 4), as well as providing narrower confidence intervals. These advantages have not, hitherto, been recognized by statisticians working on the problem of confidence intervals for a curve.

We shall give formulae for asymptotically optimal bandwidths and for asymptotic coverage error. In the case of a two-sided confidence interval constructed by explicit bias correction, employing second-order kernels and using bandwidth h in the estimation of f and bandwidth h_1 to estimate f'', the choice $h = Hn^{-5/17}$ and $h_1 = H_1 n^{-1/17}$ (for constants H and H_1) results in an asymptotic coverage error of $W(H, H_1)n^{-12/17}$. Here W is a function not depending on n, whose formula we give explicitly and which depends on f and

$f^{(4)}$. Depending on the choice of kernels and the value of $f^{(4)}$, it is sometimes possible to select H and H_1 so that $W(H, H_1) = 0$. There exist examples of quite reasonable second-order kernels where this is not possible and also examples where it is always possible. If the equation $W(H, H_1) = 0$ can be solved, then in theory, judicious choice of h and h_1 will result in a coverage error of $o(n^{-12/17})$. However, achieving this objective requires estimation of $f^{(4)}$ (in addition to estimation of f''). Such a procedure would be difficult to implement in many practical problems.

There are philosophical as well as practical difficulties in determining the appropriate number of derivatives to assume for curve estimation. However, it is common to ask that an unknown density have at least two derivatives, and simulations in Section 4 will treat this case.

Section 2 describes our methods for estimating density and bias and for constructing confidence intervals. Section 3 states our main theorem and discusses its consequences, which include the conclusions outlined above. Section 4 summarizes a simulation study which confirms the advantages of undersmoothing. Finally, Section 5 outlines the proof of the main theorem.

Our results are framed for confidence intervals, not confidence bands, since our method of proof does not extend to the case of bands. However, it would seem plausible that the advantages of undersmoothing would also be available for bands.

Although Edgeworth expansions for density estimators have been studied before [12], there appears to be no previous work on the effect of bias estimation on coverage accuracy of two-sided bootstrap confidence intervals. The influence of smoothing parameter choice on coverage accuracy has been discussed [7], but in a very different context. Härdle and Bowman [10], Härdle [8, 9] and Härdle and Marron [11] have analyzed properties of bootstrap confidence intervals in nonparametric regression, but were concerned only with the problem of consistency—that is, of ensuring that the coverage error converged to zero. They did not address the effect of bandwidth choice on coverage accuracy.

2. Constructing the estimators and intervals.

2.1. *Summary.* We begin in Section 2.2 be defining our estimators of density, bias and variance. Then in Section 2.3, we describe the percentile-t bootstrap method, and combine these ingredients to produce bias-corrected bootstrap confidence *intervals* (either one-sided or two-sided).

2.2. *Estimators of density, bias and variance.* Let $\hat{f}$ denote the estimator defined at (1.1). We assume that K is an rth order kernel, which means that

$$\int y^i K(y)\,dy \begin{cases} = 1, & \text{if } i = 0, \\ = 0, & \text{if } 1 \le i \le r-1, \\ \ne 0, & \text{if } i = r. \end{cases}$$

For example, a symmetric density function is a second-order kernel. The bias of $\hat{f}$ is given by

$$
\begin{aligned}
b(x) &= E\hat{f}(x) - f(x) \\
&= \int K(y)\{f(x-hy) - f(x)\}\,dy = \kappa_r h^r f^{(r)}(x) + o(h^r)
\end{aligned}
\tag{2.1}
$$

as $h \to 0$, assuming $f^{(r)}$ is bounded and continuous. In this formula,

$$\kappa_r = (-1)^r (r!)^{-1} \int y^r K(y)\,dy.$$

To estimate bias one would typically estimate the dominant term in the expansion (2.1) and that would be achieved via a kernel estimator of $f^{(r)}$, say,

$$\tilde{f}^{(r)}(x) = \frac{1}{nh_1^{r+1}} \sum_{i=1}^{n} L^{(r)}\left\{\frac{x - X_i}{h_1}\right\},$$

where L is an sth order kernel having at least r derivatives. (Throughout this paper, excepting the last paragraph of Section 3.4, K and L denote kernels of orders r and s, respectively.) Thus, our estimator of bias would be

$$\hat{b}(x) = \kappa_r h^r \tilde{f}^{(r)}(x). \tag{2.2}$$

The explicit bias correction method uses this estimator of bias, whereas the undersmoothing method takes $\hat{b} = 0$ and relies on choosing h so small that b is negligible.

If $\hat{b}$ is defined as at (2.2), then the variance of $\hat{b}$ admits a simple asymptotic formula:

$$\operatorname{var}\{\hat{b}(x)\} \sim \left(nh_1^{2r+1}\right)^{-1} \kappa_r^2 \left(\int L^{(r)^2}\right) f(x),$$

assuming $h_1 \to 0$ and $nh_1^{2r+1} \to \infty$. The bias of $\hat{b}$ is more complex. It has two components, which derive from the bias of $\tilde{f}^{(r)}$ as an estimator of $f^{(r)}$ and from the error in $\kappa_r h^r f^{(r)}$ as an approximation of b, respectively. Our main theorem will be framed in terms of $E\hat{b} - b$ (see Section 3.2), but a subsequent, refined analysis will discuss the effects of the individual components (Section 3.3).

The variance of $\hat{f}$ equals

$$
\begin{aligned}
\sigma(x)^2 = {} & (nh^2)^{-1} \int K\{(x-y)/h\}^2 f(y)\,dy \\
& - n^{-1}\left[h^{-1} \int K\{(x-y)/h\} f(y)\right]^2,
\end{aligned}
$$

of which an estimator is

$$\hat{\sigma}(x)^2 = \frac{1}{nh}\left[\frac{1}{nh} \sum_{i=1}^{n} K\left\{\frac{x - X_i}{h}\right\}^2 - h\hat{f}(x)^2\right].$$

2.3. *Bootstrap confidence intervals.* We first describe the construction of bootstrap confidence intervals for $E\hat{f}$. Let $\{X_1^*, \dots, X_n^*\}$ denote a resample drawn randomly, with replacement, from $\mathscr{X} = \{X_1, \dots, X_n\}$. Put

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

$$\hat{\sigma}^*(x)^2 = \frac{1}{nh}\left[\frac{1}{nh} \sum_{i=1}^{n} K\left\{\frac{x - X_i^*}{h}\right\}^2 - h\hat{f}^*(x)^2\right],$$

$$S^*(x) = \frac{\hat{f}^*(x) - \hat{f}(x)}{\hat{\sigma}^*(x)}.$$

We shall usually drop the argument x and write $f, \hat{f}, \hat{\sigma}, \dots$ for $f(x)$, $\hat{f}(x), \hat{\sigma}(x), \dots$, respectively.

Note particularly that $E(\hat{f}^* | \mathscr{X}) = \hat{f}$. This identity helps explain why bias is a significant problem in curve estimation. In general applications of the bootstrap, where an unknown θ is estimated by $\hat{\theta}$, it would be common to regard $\tilde{b} = E(\hat{\theta}^* | \mathscr{X}) - \hat{\theta}$ as an estimator of $b = E(\hat{\theta}) - \theta$. The latter quantity equals bias, and while it is usually negligible in finite-dimensional problems [it equals $O(n^{-1})$, compared to an error-about-the-mean of size $n^{-1/2}$], it is usually nonnegligible in infinite-dimensional problems (here, bias and error-about-the-mean are often of the same order). On the other hand, $\tilde{b}$ vanishes identically in many curve estimation problems.

To construct a confidence interval for $E\hat{f}$, we would ideally wish to know the distribution of $S = (\hat{f} - E\hat{f})/\hat{\sigma}$, from which we would compute the quantile u_α defined by $P(S \le u_\alpha) = \alpha$. An ideal one-sided confidence interval for $E\hat{f}$ would then be $I_1 = (\hat{f} - \hat{\sigma} u_\alpha, \infty)$. The bootstrap estimate of u_α is $\hat{u}_\alpha$, defined by

$$P(S^* \le \hat{u}_\alpha | \mathscr{X}) = \alpha,$$

and the corresponding percentile-t bootstrap confidence interval ([1], [4], [13]) for $E\hat{f}$ is

$$\hat{I}_1 = \left(\hat{f} - \hat{\sigma}\hat{u}_\alpha, \infty\right).$$

These intervals may be corrected for bias and thereby converted into confidence intervals for f, by subtracting the bias estimator $\hat{b}$. The latter may either be given by (2.2) (in the case of explicit bias correction) or be identically zero (in the case of undersmoothing). The one-sided bias-corrected bootstrap interval is

$$\hat{I} = \left(\hat{f} - \hat{\sigma}\hat{u}_\alpha - \hat{b}, \infty\right).$$

Its two-sided counterpart is

$$\hat{J} = \left(\hat{f} - \hat{\sigma}\hat{u}_{(1+\alpha)/2} - \hat{b},\ \hat{f} - \hat{\sigma}\hat{u}_{(1-\alpha)/2} - \hat{b}\right).$$

Both the intervals $\hat{I}$ and $\hat{J}$ have nominal coverage probability α. The respective exact coverages are

$$\beta(\alpha) = P(f \in \hat{I}) = P\{S \le \hat{u}_\alpha + (\hat{b} - b)/\hat{\sigma}\},$$

$$\gamma(\alpha) = P(f \in \hat{J}) = \beta\{(1+\alpha)/2\} - \beta\{(1-\alpha)/2\},$$

and the respective coverage errors are $\beta(\alpha) - \alpha$, $\gamma(\alpha) - \alpha$. Note that $\beta(\alpha)$ and $\gamma(\alpha)$ depend on x. Our main theorem will provide an asymptotic formula for $\beta(\alpha)$, which in turn leads to a formula for $\gamma(\alpha)$.

3. Main results.

3.1. *Summary*. Section 3.2 states our main theorem. Section 3.3 examines properties of two-sided intervals when bias is estimated explicitly and Section 3.4 describes the case of bias correction by undersmoothing, comparing this approach with that in Section 3.3. In Section 3.5 we show that the interval type which has smallest coverage error also has smallest width. Section 3.6 discusses analogues of these results in the case of one-sided confidence intervals.

3.2. *Main theorem*. We assume the following regularity conditions. The kernels K and L are both bounded, and both vanish outside an interval $[c, d]$. There exist numbers $c = u_0 < u_1 < \cdots < u_m = d$ such that, for each $1 \le j \le m$, K' and $L^{(r+1)}$ are both bounded on (u_{j-1}, u_j) and each is either strictly positive or strictly negative there. The kernel K is of order r and L is of order s. The density f has $r + s$ continuous derivatives in a neighbourhood of x, and $f(x) > 0$. For some $\xi > 0$, the bandwidths $h = h(n)$ and $h_1 = h_1(n)$ satisfy

$$n^{\xi}\{h + (nh)^{-1} + (nh_1)^{-1} + (h/h_1)\} \to 0. \tag{3.1}$$

A word of explanation about these conditions is in order. The conditions on K are used to ensure a version of Cramér's continuity condition for vectors of kernel estimators [see Step (vi) of the proof in Section 5]. The standard normal kernel, which is not compactly supported, is also admissible although our proof will be framed for the case of compact support. When bias is not estimated explicitly, that is, when $\hat{b} \equiv 0$, it is permissible to take $s = 0$ in the regularity conditions and assume that f has $r + s = r$ derivatives. The assumption that $h \to 0$ at a faster rate than h_1, implied by the regularity condition (3.1), is quite reasonable since one would typically use a larger order of bandwidth for derivative estimation than for density estimation.

Assume that either $\hat{b}$ is given by formula (2.2) or $\hat{b} \equiv 0$. Define $\eta = (E\hat{b} - b)/\sigma$, a function of x which we assume converges to zero. This entails $(nh)^{1/2}h^r h_1^s \to 0$ if $\hat{b}$ is defined by (2.2) and $(nh)^{1/2}h^r \to 0$ if $\hat{b} \equiv 0$, which conditions are imposed in addition to (3.1). Let ϕ, Φ denote the standard normal density, distribution functions, respectively, write z_α for the α-level quantile of the standard normal distribution [so that $\Phi(z_\alpha) = \alpha$] and put $\lambda_r = \int K^r$,

$$
\begin{aligned}
&w_1 = f(x)^{-1}\tfrac{1}{6}\left(\tfrac{3}{2}\lambda_2^{-3}\lambda_3^2 - \lambda_1^{-2}\lambda_4\right)z_\alpha\left(2z_\alpha^2 + 1\right)\phi(z_\alpha),\\
&w_2 = \kappa_r\lambda_2^{-1}L^{(r)}(0)z_\alpha\phi(z_\alpha),\\
(3.2)\qquad &w_3 = -\tfrac{1}{2}z_\alpha\phi(z_\alpha),\\
&w_4 = \tfrac{1}{6}f(x)^{-1/2}\lambda_2^{-3/2}\lambda_3 z_\alpha\phi(z_\alpha),\\
&w_5 = \phi(z_\alpha).
\end{aligned}
$$

These weights do not depend on h, h_1 or n. Note that w_1, w_4 are functions of x and α, whereas w_2, w_3, w_5 are functions of α alone.

The coverage probability of the interval $\hat{I}$ is

$$\beta(\alpha) = P(f \in \hat{I}) = P\left(f > \hat{f} - \hat{\sigma}\hat{u}_\alpha - \hat{b}\right).$$

Our main theorem is an expansion of Edgeworth type for $\beta(\alpha)$.

THEOREM. *Under the stated conditions and for fixed x,*

$$
\begin{aligned}
\beta(\alpha) &= \alpha + (nh)^{-1}w_1(\alpha) + (h/h_1)^{r+1}w_2(\alpha)\\
(3.3)\qquad &\quad + \eta^2 w_3(\alpha) + \eta(nh)^{-1/2}w_4(\alpha)\\
&\quad + \eta w_5(\alpha) + o\left\{(nh)^{-1} + (h/h_1)^{r+1} + \eta^2\right\}
\end{aligned}
$$

uniformly in $0 < \alpha < 1$ *as* $n \to \infty$.

In the case $\hat{b} \equiv 0$, it is to be understood that we take $h_1 = \infty$ on the right-hand side of (3.3).

The analogue of (3.3) for two-sided confidence intervals is

$$
\begin{aligned}
\gamma(\alpha) = P(f \in \hat{J}) &= \alpha + 2\left\{(nh)^{-1}w_1(\alpha') + (h/h_1)^{r+1}w_2(\alpha')\right.\\
(3.4)\qquad &\qquad \left. + \eta^2 w_3(\alpha') + \eta(nh)^{-1/2}w_4(\alpha')\right\}\\
&\quad + o\left\{(nh)^{-1} + (h/h_1)^{r+1} + \eta^2\right\},
\end{aligned}
$$

where $\alpha' = (1+\alpha)/2$. This formula may be obtained directly from (3.3) on noting that (a) $\gamma(\alpha) = \beta(\alpha') - \beta(1-\alpha')$, (b) $z_{\alpha'} = -z_{1-\alpha'}$ and (c) of the weights defined at (3.2), $w_1, \dots, w_4$ are odd functions of z_α whereas w_5 is an even function of z_α.

The values of the weights w_i, in particular their signs, can be important in determining the size of coverage error, as we shall show in the next section. By way of example, if we take $K = L = \phi$, the standard normal kernel, then $r = s = 2$, $\lambda_2 = (2\pi^{1/2})^{-1}$, $\lambda_3 = (2\pi)^{-1}3^{-1/2}$, $\lambda_4 = (2\pi)^{-3/2}2^{-1}$, $L''(0) = -(2\pi)^{-1/2}$. It follows that for $\alpha > 1/2$, the weights w_1, w_4, w_5 are positive whereas w_2, w_3 are negative.

3.3. *Two-sided confidence intervals with explicit bias estimation.* Assume that we estimate b explicitly, using the estimator $\hat{b}$ defined at (2.2). Our analysis begins by considering the size of η in formula (3.4). Here and in Section 3.4, write w_i for $w_i(\alpha')$. Now

$$E\tilde{f}_1^{(r)}(x) = h_1^s f^{(r+s)}(x)(-1)^s(s!)^{-1}\int y^s L(y)\,dy + o(h_1^s)$$

and so

$$(3.5)\qquad \eta = (E\hat{b} - b)/\sigma = (nh)^{1/2}h^r h_1^s w_6 + \eta_1 + o\{(nh)^{1/2}h^r h_1^s\},$$

where $\eta_1 = (E\hat{b} - \kappa_r h^r f^{(r)})/\sigma$ and

$$(3.6)\qquad w_6 = w_6(x) = f^{(r+s)}(x)f(x)^{-1/2}\kappa_r\lambda_2^{-1/2}(-1)^s(s!)^{-1}\int y^s L(y)\,dy.$$

The quantity η_1 represents the error in $\kappa_r h^r f^{(r)}$ as an approximation to b, while the term $(nh)^{1/2}h^r h_1^s w_6$ in (3.5) represents the bias of $\tilde{f}^{(r)}$ in our formula for $\hat{b}$. Many practical applications of the bias estimation method would take K to be a symmetric second-order kernel and than η_1 would be of order $(nh)^{1/2}h^4$. In this setting, $r = 2$ and η_1 would make a negligible contribution to (3.5) if $h^2 = o(h_1^s)$. That would certainly be the case if $s = 2$; that is, if L were also a second-order kernel. In the analysis below we assume that η_1 is negligible, so that

$$(3.7)\qquad \eta = (nh)^{1/2}h^r h_1^s w_6 + o\{(nh)^{1/2}h^r h_1^s\}.$$

If this condition fails, then the explicit bias correction method is even less attractive relative to undersmoothing due to the appearance of additional, dominant error terms.

If η satisfies (3.7), then the coverage error of the two-sided confidence interval $\hat{\mathscr{J}}$ is, by (3.4),

$$\gamma(\alpha) - \alpha = 2\rho_1 + o\{(nh)^{-1} + (h/h_1)^{r+1} + nh^{2r+1}h_1^{2s}\},$$

where

$$\rho_1 = \rho_1(h, h_1) = (nh)^{-1}w_1 + (h/h_1)^{r+1}w_2 + nh^{2r+1}h_1^{2s}w_3w_6^2 + h^r h_1^s w_4 w_6,$$

$w_i = w_i(\alpha')$ for $1 \le i \le 4$ and w_6 is given by (3.6). To determine asymptotically optimal formulae for h and h_1, take

$$(3.8)\qquad h = Hn^{-(r+s+1)/\{(r+1)^2+s(r+2)\}},\qquad h_1 = H_1 n^{-1/\{(r+1)^2+s(r+2)\}},$$

where H, H_1 are constants. Then $\rho_1 = W_1 n^{-(r+1)(r+s)/\{(r+1)^2+s(r+2)\}}$, where

$$(3.9)\qquad \begin{aligned} W_1 &= W_1(H, H_1) \\ &= H^{-1}w_1 + (H/H_1)^{r+1}w_2 + H^{2r+1}H_1^{2s}w_3w_6^2 + H^r H_1^s w_4 w_6. \end{aligned}$$

The asymptotically optimal approach to bandwidth selection, in the sense of minimizing the absolute value of coverage error, is to define h, h_1 as in (3.8), with H, H_1 chosen to minimize $|W_1(H, H_1)|$. It will sometimes, but not always, be possible to select $H, H_1 > 0$ so that $W_1(H, H_1) = 0$. However, all the optimal bandwidths depend on the unknown density f, and so there are significant practical difficulties in taking these conclusions any further.

The values of h, h_1 which are optimal (from a mean-squared error viewpoint) for estimating $f, f^{(r)}$ are of orders $n^{-1/(2r+1)}, n^{-1/(2r+2s+1)}$, respectively; see, for example, Gasser, Müller and Mammitzsch [3] and Härdle, Marron and Wand [12]. However, if we define h, h_1 by (3.8), then $h \ll n^{-1/(2r+1)}$ and $h_1 \gg n^{-1/(2r+2s+1)}$. Therefore the prescription which minimizes coverage error results in undersmoothing for estimation of f and oversmoothing for estimation of $f^{(r)}$.

If we simply take h, h_1 to be given by (3.8), without specifying that the constants H, H_1 be chosen so that $W_1(H, H_1) = 0$, then the coverage error of the resulting confidence interval will be of size $n^{-(r+1)(r+s)/\{(r+1)^2+s(r+2)\}}$. For example, if $K = L = \phi$, then $r = s = 2$, formula (3.8) reduces to

$$h = Hn^{-5/17}, \qquad h_1 = H_1 n^{-1/17}$$

and the coverage error equals $W_1(H, H_1)n^{-12/17}$. If we note that the weights w_1, w_4 are positive whereas w_2, w_3 are negative (see Section 3.2), then we may deduce from (3.9) that in the case $K = L = \phi$, it is always possible to choose $H, H_1 > 0$ such that $W_1(H, H_1) = 0$.

We conclude this subsection with an example where $r = s = 2$ but the equation $W_1(H, H_1) = 0$ does not admit solutions $H, H_1 > 0$. In that event, the minimum coverage error of the confidence interval $\hat{J}$ is genuinely of size $n^{-12/17}$ and cannot be reduced below this level. To construct the example, take K to be a symmetric density for which $\lambda_2\lambda_3^{-2}\lambda_4 > 3/2$, let $L = \phi$ and assume that $f^{(4)}(x) < 0$. Then each of the constants w_1, w_2, w_3 and w_4w_6 is strictly negative and so $W_1(H, H_1) < 0$, no matter what the values of $H, H_1 > 0$. To construct a kernel K for which $\lambda_2\lambda_3^{-2}\lambda_4 > 3/2$, fix $p \geq 1$ and define

$$K_p(x) = \begin{cases} 1, & \text{for } |x| \leq 1, \\ |x|^{-1/2}, & \text{for } 1 < |x| \leq p, \\ 0, & \text{for } |x| > p, \end{cases}$$

$c = c(p) = \int K_p$ and $K = c^{-1}K_p$. Then $\lambda_2\lambda_3^{-2}\lambda_4 \to \infty$ as $p \to \infty$.

3.4. *Two-sided confidence intervals with bias corrected by undersmoothing.* Here we take $\hat{b} \equiv 0$, in which case

$$\eta = (E\hat{b} - b)/\sigma = -b/\sigma = (nh)^{1/2}h^r w_7 + o\{(nh)^{1/2}h^r\},$$

where $w_7 = -f^{(r)}f^{-1/2}\kappa_r\lambda_2^{-1/2}$. We may now deduce from (3.4) with $h_1 = \infty$, that the two-sided confidence interval $\hat{J}$ has coverage error

$$\gamma(\alpha) - \alpha = 2\rho_2 + o\{(nh)^{-1} + nh^{2r+1}\},$$

where

$$\rho_2 = \rho_2(h) = (nh)^{-1}w_1 + nh^{2r+1}w_3w_7^2 + h^r w_4 w_7$$

and $w_i = w_i(\alpha')$ for $i = 1, 3, 4$. The asymptotically optimal bandwidth is

$$h = Hn^{-1/(r+1)}, \tag{3.10}$$

where H is a constant. In this case, coverage error is asymptotic to $\rho_2(h) = W_2 n^{-r/(r+1)}$, with

$$W_2 = W_2(H) = H^{-1}w_1 + H^{2r+1}w_3 w_7^2 + H^r w_4 w_7.$$

To minimize the absolute value of coverage error we should select H in (3.10) to minimize $|W_2(H)|$. Now, the equation $W_2(H) = 0$ has solution $H = U^{1/(r+1)}$, where

$$U = \left(6\kappa_r \lambda_2 f^{(r)}\right)^{-1} \lambda_3 \left[\left\{36z^2 + 19 - 12\lambda_2\lambda_3^{-2}\lambda_4(2z^2+1)\right\}^{1/2} - 1\right]$$

and $z = z_{\alpha'}$. If $\lambda_2\lambda_3^{-2}\lambda_4 < 3/2$ (which is the case if $K = \phi$, for example), then U is real-valued and positive, and so it is possible to choose $H > 0$ such that $W_2(H) = 0$. However if $\lambda_2\lambda_3^{-2}\lambda_4 > 3/2$, then U is either complex-valued or negative and in such cases the equation $W_2(H) = 0$ does not admit a real positive solution. In summary, the asymptotically optimal value of h equals

$$H_0 n^{-r/(r+1)}, \tag{3.11}$$

where $H_0 > 0$ is the minimizer of

$$\Big| H^{-1}\left\{ f(x)^{-1} \tfrac{1}{6}\left(\tfrac{3}{2}\lambda_2^{-3}\lambda_3^2 - \lambda_2^{-2}\lambda_4\right)(2z^2+1)\right\}$$
$$- \tfrac{1}{2}H^{2r+1}\left\{ f^{(r)}(x)^2 f(x)^{-1}\kappa_r^2\lambda_2^{-1}\right\} - H^r\left\{\tfrac{1}{6} f^{(r)}(x) f(x)^{-1}\kappa_r\lambda_2^{-2}\lambda_3\right\}\Big|$$

and equals

$$\left(\left\{6\kappa_r\lambda_2 f^{(r)}(x)\right\}^{-1}\lambda_3\left[\left\{36z^2 + 19 - 12\lambda_2\lambda_3^{-2}\lambda_4(2z^2+1)\right\}^{1/2} - 1\right]\right)^{1/(r+1)}$$

if the latter is real and positive. (Again, $z = z_{\alpha'}$.)

The value of h which is optimal for estimating f, from the viewpoint of minimizing mean squared error, is of size $n^{-1/(2r+1)}$. If h is given by (3.10), then $h \ll n^{-1/(2r+1)}$, which confirms that coverage error is minimized by undersmoothing in the estimation of f.

For direct comparison with the case described in Section 3.3, we should make full use of our assumption that f has $r + s$ derivatives. This demands that k be a kernel of order $r + s$, rather than r, and so r should be replaced by $r + s$ in the argument above. In particular the optimal bandwidth is $h = Hn^{-1/(r+s+1)}$ and the minimum asymptotic coverage error is $W(H) = n^{-(r+s)/(r+s+1)}$. Since

$$(r+s)/(r+s+1) > (r+1)(r+s)/\left\{(r+1)^2 + s(r+2)\right\},$$

then an improvement in coverage accuracy may be gained from the undersmoothing method [unless we deliberately choose H and H_1 in (3.8) so that $W(H, H_1) = 0$]. When $r = s = 2$, the orders of coverage error are $n^{-12/17}$ in the case of explicit bias correction and $n^{-4/5}$ for the undersmoothing method.

3.5. *Width of two-sided confidence intervals.* The width of the two-sided confidence interval $\hat{J}$ is

$$\hat{\sigma}\big(\hat{u}_{(1+\alpha)/2} - \hat{u}_{(1-\alpha)/2}\big) \sim 2(nh)^{-1/2}(\lambda_2 f)^{1/2} z_{(1+\alpha)/2}.$$

Two features of this formula should be noted. First, the width of the interval does not depend on the manner in which bias is corrected, and second, the width is (asymptotically) proportional to $(nh)^{-1/2}$. Of the techniques developed in Sections 3.3 and 3.4, that is, explicit bias correction and undersmoothing, the latter gives the smaller value of $(nh)^{-1/2}$ and so is to be preferred on the counts of both coverage accuracy and interval width.

3.6. *One-sided confidence intervals.* The problem of one-sided confidence intervals may be analyzed by paralleling arguments in Sections 3.3 and 3.4 for either of the two bias correction methods. We shall give only an outline here. The starting point for analysis is the following formula for coverage error, derived directly from (3.3),

$$\begin{aligned} \beta(\alpha) - \alpha &= P(f \in \hat{I}) \\ &= (nh)^{-1}w_1 + (h/h_1)^{r+1}w_2 + \eta w_5 \\ &\quad + o\{(nh)^{-1} + (h/h_1)^{r+1} + \eta\}, \end{aligned} \tag{3.12}$$

where $w_i = w_i(\alpha)$ for $i = 1, 2, 5$. If bias is estimated by $\hat{b}$, defined at (2.2), and if η satisfies (3.7), then (3.12) simplifies to

$$\beta(\alpha) - \alpha = 2\rho_3 + o\{(nh)^{-1} + (h/h_1)^{r+1} + n^{1/2}h^{r+(1/2)}h_1^s\},$$

where

$$\rho_3 = \rho_3(h, h_1) = (nh)^{-1}w_1 + (h/h_1)^{r+1}w_2 + n^{1/2}h^{r+(1/2)}h_1^s w_5 w_6.$$

It may be shown that the optimal choices of h and h_1 are

$$\begin{aligned} h &= Hn^{-\{3(r+1)+2s\}/\{(r+1)(2r+3)+2(r+2)s\}}, \\ h_1 &= H_1 n^{-(r+3)/\{(r+1)(2r+3)+2(r+2)s\}} \end{aligned}$$

for constants $H, H_1 > 0$; compare (3.8). Asymptotic coverage error is then

$$\rho_3(h, h_1) = W_3(H, H_1)n^{-2(r+1)(r+s)/\{(r+1)(2r+3)+2(r+2)s\}},$$

where

$$W_3(H, H_1) = H^{-1}w_1 + H^{r+1}H_1^{-(r+1)}w_2 + H^{r+(1/2)}H_1^s w_5 w_6.$$

Once again, by careful choice of H and H_1 it is sometimes, but not always, possible to ensure that $W_3(H, H_1) = 0$.

If the estimator $\hat{b}$ is taken to be zero, then the optimal choice of h is $Hn^{-3/(2r+3)}$ for a constant $H > 0$ and the asymptotic coverage error equals $W_4(H)n^{-2r/(2r+3)}$, where

$$W_4(H) = H^{-1}w_1 + H^{r+1}H_1^{-(r+1)}w_2 + H^{r+(1/2)}w_5 w_7.$$

Sometimes, but not always, we may choose H so that $W(H) = 0$. Since

$$\frac{2(r+s)}{2r+2s+3} > \frac{2(r+1)(r+s)}{(r+1)(2r+3)+2(r+2)s},$$

then, if $W_3(H, H_1)$ and $W_4(H)$ are nonzero, the undersmoothing method produces lower coverage error than explicit bias correction, for the same smoothness assumptions.

4. Numerical results.

4.1. *Summary.* In this section we report details of a simulation study which addresses performance of the two different methods of bias correction. The study has two parts: First, an examination of the undersmoothing method when a second-order kernel estimator is employed to estimate f; and second, a comparison of the undersmoothing method and explicit bias correction when four derivatives are assumed of f. Throughout, the intervals were two-sided, bootstrap approximations were based on $B = 299$ resamples, and approximations to coverage probability, etc., were based on 300 independent samples.

4.2. *Undersmoothing a second-order kernel density estimator.* Let K denote Epanečnikov's kernel, $K(u) = (3/4)5^{-1/2}\{1 - (u^2/5)\}$ for $|u| \le 5^{1/2}$, $K(u) = 0$ otherwise. When the true density f is that of an $N(0, \tau^2)$ distribution, the bandwidth which asymptotically minimizes mean integrated squared error is $1.05\tau n^{-1/5}$. This suggests the empirical version, $1.05\hat{\tau} n^{-1/5}$, where $\hat{\tau}^2$ denotes sample variance. To assess the effect of undersmoothing we took $h = 1.05c\hat{\tau} n^{-1/5}$, where $0 < c \le 1$. Table 1 lists the coverage p (as a percentage), mean width w and width standard deviation s of nominal 95% confidence intervals, for various sample sizes, distributions and values of c.

It will be seen that coverage accuracy tends to improve at first, and then decline, with increasing amount of undersmoothing. The inaccuracy of coverage is generally greater for $x = 0$ than for $x = 0.75$ or 1.50, since bias is more pronounced for $x = 0$. The case where $c = 1$, $x = 0$ and the sampling distribution is the normal mixture $(1/2)N(0,1) + (1/2)N(3,1)$, produces extremely poor coverage accuracy. However, accuracy improves by the time c has decreased to 0.3. The pattern in Table 1, of coverage accuracy at first improving and then declining as h decreases, is readily apparent for larger sample sizes. Of course, the optimal value of c decreases as n increases.

On comparing the values of w and s for small values of c, it is clear that substantial undersmoothing is not a practical proposition. The values of c which give good coverage accuracy for given x, n and distribution type are not easy to determine empirically, and as yet we have no good practical advice to offer.

TABLE 1
True coverage probabilities p (as percentages), interval widths w and width standard deviations s, estimated by simulation, of nominal 95% confidence for different values of the bandwidth $h = 1.05c\hat{\tau}n^{-1/5}$, *where* $0 < c \leq 1$. *Further details are given in Section* 4.2

		$n = 50$			$n = 100$		
c		$x = 0.00$	$x = 0.75$	$x = 1.50$	$x = 0.00$	$x = 0.75$	$x = 1.50$
				(i) $N(0, 1)$ distribution			
1.0	p	78.7	93.0	93.7	80.3	94.7	91.3
	w	0.170	0.170	0.152	0.134	0.132	0.111
	s	0.0207	0.0216	0.0253	0.0126	0.0121	0.0112
0.75	p	92.7	94.0	93.3	96.7	96.3	96.3
	w	0.227	0.222	0.201	0.179	0.168	0.135
	s	0.0306	0.0333	0.0966	0.0174	0.0175	0.0164
0.5	p	96.3	97.0	93.3	94.7	94.7	96.3
	w	0.328	0.312	0.284	0.244	0.228	0.183
	s	0.0456	0.0692	0.111	0.0266	0.0254	0.0339
0.3	p	97.7	98.3	88.0	96.7	95.7	94.7
	w	0.490	0.475	0.559	0.359	0.327	0.305
	s	0.0733	0.0872	1.01	0.0447	0.0430	0.272
0.2	p	97.3	94.7	80.0	96.7	97.7	93.3
	w	0.775	0.907	1.038	0.481	0.445	0.552
	s	0.337	1.333	3.829	0.0701	0.0758	0.890
0.1	p	87.7	78.0	60.7	95.3	88.7	71.0
	w	2.15	9.19	1.06	1.02	1.12	0.918
	s	5.54	5.82	6.04	0.982	2.66	3.18
				(ii) $\frac{1}{2}N(0, 1) + \frac{1}{2}N(3, 1)$ distribution			
1.0	p	4.00	43.3	86.7	80.3	94.7	91.3
	w	0.188	0.120	0.115	0.950	0.095	0.088
	s	0.0160	0.0167	0.0156	0.0107	0.0109	0.0094
0.75	p	33.7	77.0	90.7	17.3	67.3	82.0
	w	0.164	0.222	0.201	0.179	0.168	0.135
	s	0.0306	0.164	0.147	0.126	0.124	0.109
0.5	p	67.7	93.0	94.0	55.3	86.7	93.3
	w	0.237	0.231	0.214	0.181	0.171	0.146
	s	0.0344	0.0323	0.0881	0.0206	0.0188	0.0187
0.3	p	94.0	97.7	91.3	78.7	93.7	94.3
	w	0.357	0.343	0.443	0.260	0.245	0.209
	s	0.0582	0.0584	0.970	0.0315	0.0327	0.0308
0.2	p	96.3	98.7	90.7	93.0	97.0	95.0
	w	0.808	0.631	1.011	0.344	0.342	0.483
	s	4.89	1.20	5.55	0.0465	0.273	2.91
0.1	p	87.7	84.0	69.3	96.7	96.7	84.0
	w	1.53	1.51	1.36	0.225	1.74	0.569
	s	3.94	7.76	4.65	0.270	16.6	0.844

4.3. *Comparison of undersmoothing and explicit bias correction.* First we took K to be the Epanečnikov kernel for which $\kappa_r = 1/2$ and $r = 2$ in the bias formula (2.1). Under the assumption that f has four derivatives, we should take $s = 2$ in the discussion in Section 2. Therefore, we estimated $f^{(2)}$ by using the $(2,4)$th order optimal kernel suggested by Gasser, Müller and Mammitzsch [3]; this kernel is given by $L^{(2)}(u) = (105/16)(6u^2 - 5u^4 - 1)$ for $|u| \le 1$ and $L^{(2)}(u) = 0$ otherwise. We took $h = 1.05\hat{\tau}n^{-1/5}$ and $h_1 = 2.70\hat{\tau}n^{-1/9}$, which are asymptotically optimal if the data are normal. With these conventions, the bias estimator $\hat{b}(x)$ is given by (2.2).

Next, we took K to be the optimal fourth order kernel from [3], defined by $K(u) = (15/32)(7u^4 - 10u^2 + 3)$ if $|u| \le 1$ and $K(u) = 0$ otherwise. The asymptotically optimal bandwidth for normal data is $5.05\tau n^{-1/9}$, which suggests taking $h = 5.05c\hat{\tau}n^{-1/9}$ where $0 < c \le 1$. This we did.

These approaches to the confidence interval problem both require four derivatives of f. Their numerical performance is summarized in panels (b) and (c), respectively, of Table 2. The poor performance of the first method and of the second method for c close to 1 is immediately apparent. However, the second method performs substantially better for smaller c's. Similar patterns of performance occur for larger sample sizes.

5. Proof of the theorem. The proof which we shall give of the main theorem is valid uniformly in $\alpha \in (\xi, 1 - \xi)$ for any $0 < \xi < 1/2$. It is readily extended to $\alpha \in (n^{-1}, 1 - n^{-1})$, on noting that if $z_\alpha = \Phi^{-1}(\alpha)$ is the α-level normal quantile, then $\sup_{n^{-1} \le \alpha \le 1 - n^{-1}} |z_\alpha| = O\{(\log n)^{1/2}\}$.

To make the proof as clear as possible we shall present it in a sequence of six steps, of which the last two outline technical arguments behind rigorous theoretical justification of earlier steps.

STEP (i): Simplified formula for $\beta(\alpha)$. Let X and X^* denote generic versions of X_i and X_i^*, respectively, and define

$$Y_k = K\{(x - X)/h\}^k - E\big[K\{(x - X)/h\}^k\big],$$

$$Y_k^* = K\{(x - X^*)/h\}^k - E\big[K\{(x - X^*)/h\}^k \big| \mathscr{X}\big],$$

$$\mu_{ij} = h^{-1}E\big(Y_1^i Y_2^j\big), \qquad \hat{\mu}_{ij} = h^{-1}E\big(Y_1^{*i} Y_2^{*j} \big| \mathscr{X}\big)$$

for $i, j = 0, 1, 2$. Then $\sigma^2 = (nh)^{-1}\mu_{20}$, $\hat{\sigma}^2 = (nh)^{-1}\hat{\mu}_{20}$. Define $\delta = \hat{u}_\alpha - u_\alpha$, $\varepsilon = (\hat{b} - b)/\hat{\sigma}$. In this notation,

$$\beta(\alpha) = P\big(f \in \hat{I}\big) = P(S \le u_\alpha + \delta + \varepsilon). \tag{5.1}$$

Next we develop an approximation to δ. Put $\Delta_{ij} = \hat{\mu}_{ij} - \mu_{ij}$,

$$\Delta_j = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left\{\frac{x - X_i}{h}\right\}^j - EK\left\{\frac{x - X}{h}\right\}^j\right].$$

TABLE 2

The coverage probabilities p (as percentages), interval widths w and width standard deviations s, estimated by simulation, of nominal 95% confidence intervals constructed by the methods described in Section 4.3. Panel (a) represents the method in the first paragraph of Section 4.3 and panel (b) represents the method in the second paragraph

		$n = 50$			$n = 100$		
c		$x = 0.00$	$x = 0.75$	$x = 1.50$	$x = 0.00$	$x = 0.75$	$x = 1.50$
				(i) $N(0, 1)$ distribution			
				(a)			
	p	87.0	90.0	91.7	89.3	91.3	88.7
	w	0.168	0.170	0.149	0.135	0.133	0.122
	s	0.0217	0.0215	0.0197	0.0135	0.0128	0.0111
				(b)			
1.0	p	16.0	76.7	57.7	13.3	77.0	53.3
	w	0.077	0.092	0.103	0.060	0.071	0.077
	s	0.0102	0.0154	0.0135	0.0065	0.0078	0.0067
0.75	p	71.0	96.0	89.3	71.7	95.3	89.3
	w	0.128	0.141	0.136	0.099	0.106	0.100
	s	0.0161	0.0198	0.0147	0.0096	0.0117	0.0088
0.5	p	94.7	94.3	95.3	95.7	96.3	95.3
	w	0.218	0.219	0.186	0.163	0.160	0.129
	s	0.0264	0.0252	0.0217	0.0154	0.0245	0.0131
0.3	p	94.0	98.3	94.3	95.0	94.3	96.3
	w	0.349	0.330	0.289	0.255	0.236	0.182
	s	0.0499	0.0460	1.127	0.0274	0.0261	0.0255
0.2	p	96.3	96.3	96.3	94.7	95.7	97.7
	w	0.487	0.455	4.223	0.340	0.310	0.250
	s	0.0700	0.0692	63.867	0.0386	0.0382	0.0659
0.1	p	97.7	94.3	86.7	98.3	98.3	92.3
	w	0.867	12.79	1.710	0.530	0.501	0.532
	s	0.197	5.71	5.23	0.0786	0.0752	0.627
				(ii) $\frac{1}{2}N(0, 1) + \frac{1}{2}N(3, 1)$ distribution			
				(a)			
	p	13.7	63.7	82.7	9.67	47.0	77.0
	w	0.118	0.120	0.115	0.094	0.095	0.088
	s	0.0163	0.0164	0.0143	0.0104	0.0104	0.0083
				(b)			
1.0	p	0.00	3.67	38.0	0.00	0.67	13.0
	w	0.056	0.061	0.069	0.043	0.046	0.052
	s	0.0084	0.0111	0.0122	0.0052	0.0062	0.0068
0.75	p	2.33	29.7	66.7	0.67	18.0	52.0
	w	0.091	0.096	0.101	0.070	0.073	0.075
	s	0.0141	0.0154	0.0159	0.0077	0.0089	0.0081
0.5	p	26.0	75.3	88.0	12.3	63.0	86.7
	w	0.153	0.154	0.145	0.115	0.115	0.166
	s	0.0212	0.0227	0.0176	0.0125	0.0123	0.0115
0.3	p	78.3	92.3	94.3	60.7	85.7	90.0
	w	0.253	0.242	0.214	0.184	0.176	0.150
	s	0.0384	0.0348	0.0330	0.0199	0.0206	0.0178
0.2	p	90.0	96.7	94.0	80.0	89.0	94.7
	w	0.353	0.333	0.315	0.248	0.230	0.192
	s	0.0543	0.0497	0.0860	0.0286	0.0289	0.0260
0.1	p	98.7	99.3	92.0	91.3	94.7	95.0
	w	0.632	0.967	0.901	0.397	0.370	0.341
	s	0.147	6.71	1.55	0.0557	0.0584	0.0855

Note that $\Delta_{ij} = \Delta_{i+2j} + o_p\{(nh)^{-1/2}\} = O_p\{(nh)^{-1/2}\}$ and that

$$(5.2) \qquad \delta = -(nh)^{-1/2}\{\hat{q}_1(z_\alpha) - q_1(z_\alpha)\} + O_p\{(nh)^{-1}\},$$

where

$$(5.3) \qquad q_1(y) = \tfrac{1}{2}\mu_{20}^{-3/2}\mu_{11} - \tfrac{1}{6}\mu_{20}^{-3/2}(\mu_{30} - 3\mu_{11})(y^2 - 1)$$

and $\hat{q}_1(y)$ has the same formula except that μ_{ij} is replaced by $\hat{\mu}_{ij}$ [6]. We may prove by Taylor expansion that

$$D_1 = \hat{\mu}_{20}^{-3/2}\hat{\mu}_{11} - \mu_{20}^{-3/2}\mu_{11} = \mu_{20}^{-3/2}\left(\Delta_3 - \tfrac{3}{2}\mu_{20}^{-1}\mu_{11}\Delta_2\right) + o_p\{(nh)^{-1/2}\},$$

$$D_2 = \hat{\mu}_{20}^{-3/2}\hat{\mu}_{30} - \mu_{20}^{-3/2}\mu_{30} = \mu_{20}^{-3/2}\left(\Delta_3 - \tfrac{3}{2}\mu_{20}^{-1}\mu_{30}\Delta_2\right) + o_P\{(nh)^{-1/2}\}.$$

Hence, noting that $\mu_{30} = \mu_{11} + o(1)$,

$$\hat{q}_1(y) - q_1(y) = \mu_{20}^{-3/2}\left(\Delta_3 - \tfrac{3}{2}\mu_{20}^{-1}\mu_{11}\Delta_2\right)\tfrac{1}{6}(2y^2 + 1) + o_p\{(nh)^{-1/2}\}.$$

and so by (5.2),

$$(5.4) \qquad \delta = (nh)^{-1/2}(c_1\Delta_2 + c_2\Delta_3) + o_p\{(nh)^{-1}\},$$

where $c_1 = (2z_\alpha^2 + 1)\mu_{20}^{-5/2}\mu_{11}/4$, $c_2 = -(2z_\alpha^2 + 1)\mu_{20}^{-3/2}/6$.

Now we develop an approximation to ε. Observe that $(nh)^{-1/2}\hat{\sigma}^{-1} = \mu_{20}^{-1/2}(1 - \mu_{20}^{-1}\Delta_{20}/2)^{-1/2} + O_p\{(nh)^{-1}\}$ and with $\mu_j = h^{-1}E[K^j\{(x - X)/h\}]$, $\Delta_{20} = \Delta_2 - 2h\mu_1\Delta_1 + O_p\{(nh)^{-1}\}$. Therefore, since $(nh)^{1/2}(\hat{b} - b) \to 0$ in probability,

$$\varepsilon = (nh)^{1/2}(\hat{b} - b)\mu_{20}^{-1/2}\{1 - \tfrac{1}{2}\mu_{20}^{-1}(\Delta_2 - 2h\mu_1\Delta_1)\} + o_p\{(nh)^{-1}\}.$$

Put $\eta = (nh)^{1/2}(E\hat{b} - b)\mu_{20}^{-1/2}$, $\Delta = (nh)^{1/2}(\hat{b} - E\hat{b})\mu_{20}^{-1/2}$, $c_3 = \mu_{20}^{-1}\mu_1$, $c_4 = -\mu_{20}^{-1}/2$. Then

$$(5.5) \quad \varepsilon = \eta + \Delta + \eta(c_3h\Delta_1 + c_4\Delta_2) + o_p\{(nh)^{-1}\} + O_p\{(nh)^{-1/2}(E\Delta^2)^{1/2}\}.$$

Combining (5.1), (5.4) and (5.5) we deduce that

$$(5.6) \qquad \beta(\alpha) = P\left[S \le v_\alpha + R + o_p\{(nh)^{-1}\} + O_p(g)\right],$$

where $v_\alpha = u_\alpha + \eta$, $g = (nh)^{-1/2}(E\Delta^2)^{1/2}$ and

$$(5.7) \qquad R = (nh)^{-1/2}(c_1\Delta_2 + c_2\Delta_3) + \eta(c_3h\Delta_1 + c_4\Delta_2) + \Delta.$$

The delta method for Edgeworth expansion argues that for any $\xi > 0$, (5.6) can be written as

$$(5.8) \qquad \beta(\alpha) = P(S \le v_\alpha + R) + o\{(nh)^{-1}\} + O(gn^\xi);$$

see Step (v) for justification.

STEP (ii): Method for expanding $P(S \le y + R)$. We expand the difference between the distributions of $S - R$ and S. The difference between the characteristic functions equals

$$\psi(t) = E[\exp\{it(S - R)\} - \exp(itS)]$$
$$(5.9) \qquad = -itE(Re^{itS}) + o\{(nh)^{-1}\} + O\{E(\Delta^2)\}$$

uniformly in $|t| \le t_0$ for any $t_0 > 0$.

Define $S' = (\hat{f} - E\hat{f})/\sigma$. Then $S = S' + O_p\{(nh)^{-1/2}\}$ and it follows from (5.9) that

$$(5.10) \qquad \psi(t) = -itE(Re^{itS'}) + o\{(nh)^{-1}\} + O\{g + E(\Delta^2)\}.$$

STEP (iii): Formula for $E(Re^{itS'})$. Define

$$A(y) = \sigma^{-1}(nh)^{-1}[K\{(x-y)/h\} - EK\{(x-X)/h\}]$$

and let B denote any other real-valued function such that $E\{B(X)\} = 0$. Put $T = \sum_{j \le n} B(X_j)$. Then

$$E(Te^{itS'}) = nE[B(X_1)\exp\{itA(X)\}]E\left[\exp\left\{it\sum_{j=2}^{n} A(X_j)\right\}\right],$$

$$E[B(X)\exp\{itA(X)\}] = itE\{A(X)B(X)\} + \tfrac{1}{2}(it)^2E\{A(X)^2B(X)\}$$
$$+ O\left[\{EA(X)^6\}^{1/2}\{EB(X)^2\}^{1/2}\right],$$

$$\left[\{EA(X)^6\}^{1/2}\{EB(X)^2\}^{1/2}\right] = O\{n^{-1}(nh)^{-1}(ET^2)^{1/2}\},$$

$$E\left[\exp\left\{it\sum_{j=2}^{n} A(X_j)\right\}\right] = e^{-t^2/2} + O\{(nh)^{-1/2}\}.$$

Therefore

$$E(Te^{itS'}) = n\left[itE\{A(X)B(X)\} + \tfrac{1}{2}(it)^2E\{A(X)^2B(X)\}\right]e^{-t^2/2}$$
$$(5.11) \qquad + O\{(nh)^{-1}(ET^2)^{1/2}\}$$
$$+ O\left[n(nh)^{-1/2}\{E|A(X)B(X)| + E|A(X)^2B(X)|\}\right].$$

Noting formula (5.7) for R and the fact that each of $\Delta_1, \Delta_2, \Delta_3, \Delta$ is a sum of independent random variables with zero mean, we see that (5.11) may be used to develop an expression for $E(Re^{itS'})$. Put

$$A_j(y) = K\{(x-y)/h\}^j - EK\{(x-X)/h\}^j, \qquad j \ge 1,$$
$$A_0(y) = L^{(r)}\{(x-y)/h_1\} - EL^{(r)}\{(x-X)/h_1\},$$

$a_0 = (nh)^{-1/2}\sigma^{-1}$, $a_1 = h^{-1}E\{A_1(X)A_0(X)\}$. If we take $T = \Delta_j$, then $B(y) = (nh)^{-1}A_j(y)$, whence by (5.11),

$$E\big(\Delta_j e^{itS'}\big) = (nh)^{-1/2}a_0\mu_{j+1}ite^{-t^2/2} + o\big\{(nh)^{-1/2}\big\}.$$

In the case $T = \Delta$,

$$E(\Delta e^{itS'}) = (h/h_1)^{r+1}\kappa_r\mu_{20}^{-1/2}a_0a_1ite^{-t^2/2} + o\big\{(h/h_1)^{r+1} + (nh)^{-1}\big\}.$$

Therefore by (5.7),

$$\begin{aligned} E(Re^{itS'}) = \big\{&(nh)^{-1}a_0(c_1\mu_3 + c_2\mu_4) + (nh)^{-1/2}\eta a_0(c_3\mu_2h + c_4\mu_3) \\ &+(h/h_1)^{r+1}\kappa_r\mu_{20}^{-1/2}a_0a_1\big\}ite^{-t^2/2} \\ &+ o\big\{(h/h_1)^{r+1} + (nh)^{-1}\big\}. \end{aligned} \tag{5.12}$$

STEP (iv): Formula for $P(S \le y + R)$. Combining (5.10) and (5.12) and noting that $E(\Delta^2) = O\{(h/h_1)^{2r+1}\} = o\{(h/h_1)^{r+1}\}$, we see that

$$\begin{aligned} \psi(t) = -\big\{&(nh)^{-1}a_0(c_1\mu_3 + c_2\mu_4) - (nh)^{-1/2}\eta a_0(c_3\mu_2h + c_4\mu_3) \\ &+(h/h_1)^{r+1}\kappa_r\mu_{20}^{-1/2}a_0a_1\big\}(it)^2e^{-t^2/2} + o\big\{(h/h_1)^{r+1} + (nh)^{-1}\big\}, \end{aligned}$$

uniformly in $|t| \le t_0$ for any $t_0 > 0$. Noting that $\mu_{11} = \mu_3 + o(1)$ we see that formal inversion of this Fourier–Stieltjes transform yields

$$\begin{aligned} &P(S \le y + R) - P(S \le y) \\ &\quad= \Big[(nh)^{-1}a_0\big\{\tfrac{1}{4}(2z_\alpha^2 + 1)\mu_{20}^{-5/2}\mu_3^2 - \tfrac{1}{6}(2z_\alpha^2 + 1)\mu_{20}^{-3/2}\mu_4\big\} \\ &\qquad+(nh)^{-1/2}\eta a_0\big(\mu_{20}^{-1}\mu_1\mu_2h - \tfrac{1}{2}\mu_{20}^{-1}\mu_3\big) \\ &\qquad+(h/h_1)^{r+1}\kappa_r\mu_{20}^{-1/2}a_0a_1\Big]y\phi(y) + o\big\{(h/h_1)^{r+1} + (nh)^{-1}\big\}. \end{aligned} \tag{5.13}$$

Justification will be given in Step (vi). Taking $y = v_\alpha = u_\alpha + \eta = z_\alpha + o(1)$ and noting that $\mu_{20} = \mu_2 + o(1)$, $\mu_r = \lambda_r f + o(1)$, $a_0 = (\lambda_2 f)^{-1/2} + o(1)$ and $a_1 = L^{(r)}(0)f + o(1)$, we see that

$$\begin{aligned} &P(S \le u_\alpha + \eta + R) - P(S \le u_\alpha + \eta) \\ &\quad= \Big\{(nhf)^{-1}\tfrac{1}{6}\big(\tfrac{3}{2}\lambda_2^{-3}\lambda_3^2 - \lambda_2^{-2}\lambda_4\big)\big(2z_\alpha^2 + 1\big) \\ &\qquad-(nh)^{-1/2}\eta\tfrac{1}{2}f^{-1/2}\lambda_2^{-3/2}\lambda_3 + (h/h_1)^{r+1}\kappa_r\lambda_2^{-1}L^{(r)}(0)\Big\}z_\alpha\phi(z_\alpha) \\ &\qquad+ o\big\{(h/h_1)^{r+1} + (nh)^{-1} + \eta^2\big\}. \end{aligned} \tag{5.14}$$

Finally, since

$$P(S \le y) = \Phi(y) + (nh)^{-1/2}q_1(y)\phi(y) + (nh)^{-1}q_2(y)\phi(y) + o\big\{(nh)^{-1}\big\},$$

where q_1 is given by (5.3) and q_2 is a polynomial with bounded coefficients, then

$$P(S \le y + \eta) - P(S \le y) = \eta\phi(y) - \tfrac{1}{2}\eta^2 y\phi(y) + (nh)^{-1/2}\eta f^{-1/2}\lambda_2^{-3/2}\lambda_3 \times \tfrac{1}{6}y(3 - 2y^2)\phi(y) + o\{\eta^2 + (nh)^{-1}\}.$$

Taking $y = u_\alpha - (nh)^{-1/2}q_2(z_\alpha) + O\{(nh)^{-1}\}$, we see that

$$\begin{aligned} &P(S \le u_\alpha + \eta) - \alpha \\ &\quad = \eta\phi(z_\alpha) - \tfrac{1}{2}\eta^2 z_\alpha\phi(z_\alpha) + (nh)^{-1/2}\eta\tfrac{2}{3}f^{-1/2}\lambda_2^{-3/2}\lambda_3 z_\alpha\phi(z_\sigma) \\ &\qquad + o\{\eta^2 + (nh)^{-1}\}. \end{aligned} \tag{5.15}$$

The theorem follows by combining (5.8), (5.14) and (5.15).

STEP (v): Justification of passage from (5.6) to (5.8). Since the functions K and $L^{(r)}$ are bounded, then we may readily apply Bernstein's inequality to derive large deviation formulae for quantities such as $\hat{f}$ and $\hat{b}$. For example, we may prove that for any $u > 0$,

$$P\{|\hat{f}(x) - E\hat{f}(x)| > u(nh)^{-1/2}(\log n)^{1/2}\} = O(n^{-l(u)}),$$

where $l(u) \to \infty$ as $u \to \infty$. Therefore $P\{|\hat{f}(x) - E\hat{f}(x)| > (nh)^{-1/2}\log n\} = O(n^{-\lambda})$ for all $\lambda > 0$. Arguing thus, and using condition (5.6), we may show that for some $\xi > 0$, the random variable V which is represented by the term $o_p\{(nh)^{-1}\} + O_p(g)$ in (5.6) has the property

$$P\Big[|V| > \{(nh)^{-1}n^{-\xi} + g\}\log n\Big] = O(n^{-\lambda})$$

for all $\lambda > 0$. Put $\zeta_1 = \{(nh)^{-1}n^{-\xi} + g\}\log n$ and

$$\zeta_2 = \sup_{+,-} |P(S \le v_\alpha + R \pm \zeta_1) - P(S \le v_\alpha + R)|.$$

Then

$$\beta(\alpha) = P(S \le v_\alpha + R + V) = P(S \le v_\alpha + R) + O(\zeta_2 + n^{-\lambda}). \tag{5.16}$$

In Steps (ii)–(iv), we derived an expansion of $P(S \le y + R)$ for general y; that result gives $P(S \le v_\alpha + y + R) - P(S \le v_\alpha + y) = O(|y|)$. Therefore $\zeta_2 = O(\zeta_1)$, and so by (5.16), $\beta(\alpha) = P(S \le v_\alpha + R) + O(\zeta_1 + n^{-\lambda})$, from which follows (5.8).

STEP (vi): Justification of (5.13). The key property of $S - R$ which facilitates the Edgeworth expansion in (5.13) is that it is a smooth function of sums of independent and identically distributed random variables. Indeed, if we define $X_{ji} = (nh)^{-1}A_j(X_i)$ for $1 \le j \le 3$, $X_{0i} = (nh)^{1/2}h^r A_0(X_i)$ and

$$\begin{aligned} l(\xi_0, \dots, \xi_3) = {} & (nh)^{1/2}\xi_1\{\mu_2 + \xi_2 - h(\mu_1 + \xi_1)^2\}^{-1/2} \\ & -(nh)^{-1/2}(c_1\xi_1 + c_2\xi_3) - \eta(c_3 h\xi_1 + c_4\xi_2) - \kappa_r\mu_{20}^{-1/2}\xi_0, \end{aligned}$$

then

$$S - R = l(\Sigma X_{0i}, \ldots, \Sigma X_{3i}). \tag{5.17}$$

The methods in [6] may be used to develop an Edgeworth expansion of arbitrary length for the distribution of $Z = (\Sigma X_{0i}, \ldots, \Sigma X_{3i})$, standardized for scale. The key feature of the argument is a version of Cramér's condition, which is provided by an analogue of Lemma 4.1 of [6]. That result uses the piecewise monotonicity properties assumed of K and $L^{(r)}$. The resulting expansion of the distribution of Z is a straightforward analogue of Proposition 4.1 in [6] and is available uniformly over a large class of Borel sets. From that point it is a simple matter to obtain an Edgeworth expansion for the distribution of $S - R$, noting (5.17) and using the approach of Bhattacharya and Ghosh [2]. This establishes existence of the expansion.

Acknowledgments. I am grateful to three referees and an Associate Editor for their helpful comments, which have led to this substantially improved version of the paper.

REFERENCES

[1] BERAN, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–466.

[2] BHATTACHARYA, R. N. and GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451.

[3] GASSER, T., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

[4] HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.* **16** 237–249.

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

[6] HALL, P. (1991). Edgeworth expansions for nonparametric density-estimators with applications. *Statistics* **22** 215–232.

[7] HALL, P. and SHEATHER, S. J. (1988). On the distribution of a studentized quantile. *J. Roy. Statist. Soc. Ser. B* **50** 381–391.

[8] HÄRDLE, W. (1989). Resampling for inference from curves. *Proc. 47th Session Internat. Statist. Inst.*

[9] HÄRDLE, W. (1989). *Applied Nonparametric Regression*. Cambridge Univ. Press.

[10] HÄRDLE, W. and BOWMAN, A. W. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102–110.

[11] HÄRDLE, W. and MARRON, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19** 778–796.

[12] HÄRDLE, W., MARRON, J. S. and WAND, M. P. (1990). Bandwidth choice for density derivatives. *J. Roy. Statist. Soc. Ser. B* **52** 223–232.

[13] HINKLEY, D. V. and WEI, B.-C. (1984). Improvement of jackknife confidence methods. *Biometrika* **71** 331–339.

[14] SILVERMAN, B. W. (1988). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

CENTRE FOR MATHEMATICS AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4,
CANBERRA, ACT 2601
AUSTRALIA