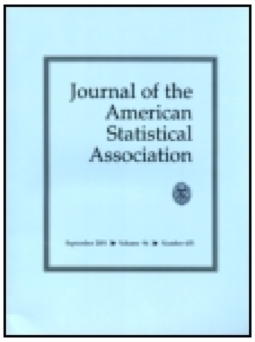

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands

Wolfgang Härdle & Adrian W. Bowman

To cite this article: Wolfgang Härdle & Adrian W. Bowman (1988) Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands, Journal of the American Statistical Association, 83:401, 102-110, DOI: 10.1080/01621459.1988.10478572

To link to this article: https://doi.org/10.1080/01621459.1988.10478572

Published online: 12 Mar 2012.

Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands

WOLFGANG HÄRDLE and ADRIAN W. BOWMAN*

The operation of the bootstrap in the context of nonparametric regression is considered. Bootstrap samples are taken from estimated residuals to study the distribution of a suitably recentered kernel estimator. The application of this principle to the problem of local adaptive choice of bandwidth and to the construction of confidence bands is investigated and compared with a direct method based on asymptotic means and variances. The technique of the bootstrap is to replace any occurrence of the unknown distribution in the definition of the statistical function of interest by the empirical distribution function of the observed errors. In a regression context these errors are not directly observed, although their role can be played by the residuals from the fitted model. In this article the fitted model is a kernel nonparametric regression estimator. Since nonparametric smoothing is involved, an additional difficulty is created by the bias incurred in smoothing. This bias, however, can be estimated in a consistent fashion. These considerations suggest the way in which the distribution of the nonparametric estimate about the true curve at some point of interest may be approximated by suitable recentering of the nonparametric estimates based on bootstrap samples. The bootstrap samples are constructed by adding to the observed estimate errors, which are randomly chosen without replacement from the collection of recentered and bias-corrected residuals from the original data. A theorem is proved to establish that the bootstrap distribution approximates the distribution of interest in terms of the Mallows metric. Two applications are considered. The first uses bootstrap sampling to approximate the mean squared error of the nonparametric estimate at some point of interest. This can then be minimized over the smoothing parameter to adapt the degree of smoothing applied at any point to the local behavior of the underlying curve. The second application uses the percentiles of the approximate distribution to construct confidence intervals for the curve at specific design points. In both of these cases the performance of the bootstrap is compared with a simple "plug-in" method based on direct estimation of the terms in an asymptotic expansion. The performances of the two methods are in general very similar. The bootstrap, however, has the slight advantage of not being as sensitive as the direct method to second derivatives near 0 in the local adaptive smoothing problem. In addition, in the construction of confidence intervals the bootstrap is able to reflect features such as skewness but falls slightly short of target confidence intervals as a result of inaccuracies in centering when the second derivative of the curve is high.

KEY WORDS: Regression smoothing; Resampling techniques.

1. INTRODUCTION

The bootstrap is a resampling technique whose aim is to gain information on the distribution of an estimator. In nonparametric regression there are several ways in which such information could be of considerable assistance. One application could be in choosing the parameter that controls the degree of smoothing that is applied to the data. Another area of interest is the construction of confidence intervals for the curve. Discussion of the first problem is usually directed toward methods that asymptotically minimize a global criterion such as mean integrated squared error. When estimating the regression function at a particular point, however, it would be helpful to tailor our choice of smoothing parameter to the features exhibited near that point. For example, near a peak a relatively small value of smoothing parameter is appropriate, whereas on an approximately linear section a larger value should be used.

The construction of confidence intervals extends the use of nonparametric smoothing beyond its role as a point estimator, often constructed with the sole purpose of giving visual information on the shape of the underlying regression curve. It would be very helpful to obtain, through confidence intervals, an impression of the variability of the estimator, providing a useful scale against which unusual features of the estimated curve may be assessed.

This article investigates the use of the bootstrap in providing approximations to a suitably centered distribution of kernel estimators of nonparametric regression curves. From the bootstrap distribution an estimate of local mean squared error is available, enabling a good choice of a local smoothing parameter to be made. Confidence bands for the true curve can also be derived from the bootstrap distribution. Both of these problems could be tackled in a more direct way by estimating the asymptotic means and variances of the estimators. Such an approach is simpler and can be very effective. One advantage of the bootstrap is that it does reflect the presence of nonstandard features such as skewness, although in the simulations of Section 3 the bootstrap proved to be slightly less effective than the direct method in attaining the target coverage probability of confidence bands.

The structure of the data is assumed to be of the following form.

Condition 1. $Y_i = m(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, where $E(\varepsilon_i) = 0$ and the design points x_i are equally spaced. For simplicity we assume that $x_i = (i - \frac{1}{2})/n$, where n is the total number of observations. Extensions to other patterns of design points are possible. m is a twice continuously differentiable function, and the errors ε_i are independent, with distribution F and constant variance σ^2.

* Wolfgang Härdle is Principal Researcher, Department of Economics, University of Bonn, D-5300 Bonn 1, West Germany. Adrian W. Bowman is Lecturer, Statistics Department, University Gardens, The University, Glasgow G12 8QW, Scotland. The authors are indebted to Steve Portnoy, Dennis Cox, and Mike Titterington for stimulating discussions and to an associate editor and referee for helpful suggestions. This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 and Sonderforschungsbereich 303.

Journal of the American Statistical Association
March 1988, Vol. 83, No. 401, Theory and Methods

We adopt the estimator of m originally proposed by Priestley and Chao (1972), namely

$$\hat{m}(x) = \hat{m}(x; h) = n^{-1}h^{-1} \sum_{i=1}^{n} K((x - x_i)/h)y_i,$$

and make the following assumptions on K.

Condition 2. The kernel function K is a symmetric probability density with bounded support that is Lipschitz continuous and has been parameterized so that $\int u^2 K(u)\, du = 1$.

Under Conditions 1 and 2, we have for any x in a subinterval $[\eta, 1 - n]$ $(\eta > 0)$,

$$E_F\hat{m}(x) = m(x) + \frac{h^2}{2} m''(x) + o(h^2)$$

$$\mathrm{var}_F(X) = n^{-1}h^{-1}\sigma^2 \int K^2(u)\, du + o(n^{-1}h^{-1}) \quad (1)$$

as $n \to \infty$, $h \downarrow 0$.

These asymptotic expressions indicate that an appropriate choice of the smoothing parameter h for estimation of $m(x)$ should be influenced by the local curvature of m, as expressed in the second derivative $m''(X)$. When $|m''(x)|$ is large, small values of h are required to keep the bias low, whereas when $|m''(x)|$ is small, large values of h are appropriate to deflate the variance. Local adaptive smoothing aims to balance these effects in a way that is appropriate for each particular location.

Section 2 discusses the general application of the bootstrap in the context of nonparametric regression. It is shown that the bootstrap works when an appropriate correction term is introduced. Section 3 discusses local adaptive smoothing, and Section 4 deals with confidence bands; Sections 3 and 4 give numerical examples and describe a small simulation study. Some brief discussion is given in Section 5.

2. THE BOOTSTRAP IN NONPARAMETRIC REGRESSION

The technique of the bootstrap is to replace any occurrence of the unknown distribution F in the definition of the statistical function of interest by the empirical distribution function F_n of $\{\varepsilon_i\}$. Since we cannot observe F_n, we need an initial estimate $\hat{m}$ of the regression function from which to estimate residuals $\hat{\varepsilon}_i = Y_i - \hat{m}(x_i)$. Special attention, however, must be paid to observations near the boundary of the interval [0, 1]. Since $\hat{m}$ has a slower rate of convergence near the boundary (Gasser and Müller 1979), it is advisable to use residuals only from an interior subinterval $[\eta, 1 - \eta]$ $(0 < \eta < \frac{1}{2})$, which contains the point x. The residuals need not necessarily have mean 0, so, to let the resampled residuals reflect the behavior of the true observation errors, they should first be recentered as

$$\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \frac{1}{[(1 - 2\eta)n]} \sum_i \hat{\varepsilon}_i,$$

where, to exclude boundary effects, $\eta n + 1 \leq i \leq (1 - \eta)n - 1$.

Bootstrap residuals ε_i^* are then created by sampling with replacement from $\{\tilde{\varepsilon}_i\}$, giving bootstrap observations $y_i^* = \hat{m}(x_i) + \varepsilon_i^*$. A bootstrap estimator m^* of m is then obtained by smoothing $\{Y_i^*\}$ rather than $\{Y_i\}$.

We define the bootstrap principle to hold if the distributions of $m^*(x)$ and $\hat{m}(x)$, when suitably normalized, become close as the sample size n increases. Specifically, we shall examine convergence of these distributions in the Mallows metric, following Bickel and Freedman (1981).

Since the variance of $\hat{m}(x)$ converges to 0 at the rate $n^{-1}h^{-1}$, as shown previously, we consider $\sqrt{nh}\{\hat{m}(x) - m(x)\}$. It is important, however, to note that $\hat{m}(x)$ is a biased estimator of $m(x)$ and that if h is chosen to balance this bias against the standard deviation of $\hat{m}$ then the variance and squared bias will have the same speed of convergence to 0. It is necessary, therefore, to ensure that this behavior is mirrored in the distribution of the bootstrap estimator $m^*(x)$.

The following approximate decomposition into a variance and bias part is helpful in understanding bootstrapping in this context:

$$\hat{m}(x) - m(x) \doteq n^{-1}h^{-1} \sum_{i=1}^{n} \times K((x - x_i)/h)\varepsilon_i + (h^2/2)m''(x). \quad (2)$$

In the bootstrap, any occurrence of ε_i is replaced by ε_i^* and we have

$$m^*(x; h, g) = n^{-1}h^{-1} \sum_{i=1}^{n} K((x - x_i)/h)(\hat{m}(x_i; g) + \varepsilon_i^*),$$

where the pilot bandwidth g is used to produce residuals $\hat{\varepsilon}_i = Y_i - \hat{m}(x_i; g)$. Note that in the definition of m^* there are two levels of smoothing involved. It is clearly helpful to have a good initial estimate $\hat{m}(x_i; g)$, giving reasonable residuals. Cross-validatory choice of g is a strong candidate, since Rice (1984) and Härdle and Marron (1985) showed that this produces estimators that asymptotically minimize the mean integrated squared error. In this article cross-validation will be used to choose the pilot bandwidth g.

The distribution of the bootstrap estimator is centered around its expectations (under the bootstrap distribution). This expectation is

$$n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} h^{-1}K\left[\frac{x - x_i}{h}\right] g^{-1}K\left[\frac{x_i - x_j}{g}\right] Y_j.$$

Using Conditions 1 and 2, it can be shown that the convolution term is approximated by the integral

$$K_1(v; h, g) = \int h^{-1}K(u/h)g^{-1}K\left[\frac{u - v}{g}\right] du.$$

Therefore, center m^* around

$$\hat{m}_c(x; h, g) = n^{-1} \sum_i K_1((x - x_i); h, g)Y_i$$

for reasons of computational efficiency. The kernel K_1 corresponds to the density of the sum of the two independent random variables hZ_1, gZ_2, where Z_1 and Z_2 have density K. K_1 can, therefore, be computed analytically.

The bias component in (2) may be estimated by employing a consistent estimator of $\hat{m}''(x)$. For example, a consistent kernel estimator is

$$\hat{m}''(x) = n^{-1}l^{-3} \sum_{i=1}^{n} K_{(2)} \left[\frac{x - x_i}{l}\right] Y_i,$$

where $K_{(2)}$ satisfies

$$\int K_{(2)}(u)\, du = \int uK_{(2)}(u)\, du = 0,$$
$$\int u^2K_{(2)}(u)\, du = 2$$

and $l \to 0$ and $nl^5 \to \infty$. To study the distribution of $(nh)^{1/2}(\hat{m}(x) - m(x))$, we will, therefore, use the following bootstrap approximation:

$$\sqrt{nh}\left(m^*(x; h, g) - \hat{m}_c(x; h, g) + \frac{h^2}{2} \hat{m}''(x)\right).$$

We make the following assumption on the smoothing parameters.

Condition 3. $\{h\}$ and $\{g\}$ are sequences of smoothing parameters that tend to 0 at the rate $n^{-1/5}$.

This is the rate entailed by choosing h to balance integrated squared bias against variance [see (1)] and was shown by Stone (1980) to be the optimal rate under our conditions on the regression function m.

At first sight the two levels of smoothing have some obvious similarities to twicing. Stuetzle and Mittal (1979), however, derived some asymptotic theory for twiced kernel estimators and showed that twicing is equivalent to using $2\hat{m} - \hat{m}_c$ as an estimator for m. Twicing is, therefore, different from bootstrapping.

The Mallows metric $d_2(F, G)$ between the distributions F and G is defined to be the infimum of $E\{(X - Y)^2\}^{1/2}$ over pairs of random variables X and Y having marginal distributions F and G, respectively. We shall adopt the convention that where random variables appear in the arguments of d_2 they represent the corresponding distributions.

Theorem 1. Under Conditions 1–3, the bootstrap principle holds in the following form:

$$d_2\left(\sqrt{nh}\{\hat{m}(x; h) - m(x)\},\right.$$
$$\left.\sqrt{nh}\left\{m^*(x; h, g) - \hat{m}_c(x; h, g) + \frac{h^2}{2} \hat{m}''(x)\right\}\right) \underset{p}{\to} 0,$$

as $n \to \infty$. Proof of this theorem is given in the Appendix.

This theorem shows that the bootstrap principle holds when resampling is carried out from the residuals $y_i - \hat{m}(x_i; g)$. Since an estimate of bias is already employed in the recentering of the distribution, we may also bias correct the residuals, so that resampling takes place from $(y_i - \hat{m}(x_i; g) + \frac{1}{2}g^2\hat{m}''(x))$. It is easy to see that the theory of Theorem 1 follows through without difficulty in this case; the bias component of the Mallows metric is the only part that is affected. This gives the following corollary.

Corollary 1. Theorem 1 holds when resampling is carried out from bias-corrected residuals.

The advantage of this is that the bootstrap distributions reflects the true error distribution more faithfully. Section 4 gives an example where the bootstrap is able to respond to skewness in the error distribution. Such a feature is less easily identified when the residuals are not corrected for bias. For the remainder of the article, bias correction of residuals will be assumed.

The mean squared error $\text{MSE}(x; h) = E_F(\hat{m}(x, h) - m(x))^2$ may be estimated from the bootstrap method by

$$\widehat{\text{MSE}}(x; h) = \int \left(m^*(x; h, g) - m_c(x; h, g) + \frac{h^2}{2} \hat{m}''(x; g)\right)^2 d\tilde{F}_n,$$

where $\tilde{F}_n$ denotes the empirical distribution function of $\{\tilde{\varepsilon}_i\}$. Denote by $\hat{h}$ the bandwidth that minimizes $\widehat{\text{MSE}}(x; h)$ over a range of smoothing parameters $H_n \subset (an^{-1/5}, bn^{-1/5})$ $(0 < a < b)$, with cardinality $\#H_n$. The following theorem can be proved by using methods similar to those of Rice (1984) and Härdle and Marron (1985).

Theorem 2. If the conditions of Theorem 1 are in force and if, for some $D > 0$, $\#H_n\, n^{-11/5} \leq D$, then $\hat{h}$ is asymptotically optimal in the sense that

$$\frac{\text{MSE}(x; \hat{h})}{\inf_{h \in H_n} \text{MSE}(x; h)} \underset{p}{\to} 1,$$

as $n \to \infty$.

3. LOCAL ADAPTIVE SMOOTHING

The bootstrap principle allows the estimation of mean squared error at specific estimation points x. To adapt the smoothing to local features this estimated mean squared error can be minimized over a range of smoothing parameters. A more direct estimator of mean squared error is obtained by plugging in estimates for the unknown quantities in expressions (1), namely

$$\frac{h^4}{4} \{\hat{m}''(x)\}^2 + n^{-1}h^{-1} \int K^2(u)\, du\, \hat{\sigma}^2,$$

which in turn provides what we will term a "plug-in" estimate of the optimal local smoothing parameter. An estimate of σ^2 is provided by

$$\hat{\sigma}^2 = \{(1 - 2\eta)n\}^{-1} \sum_i \left\{Y_i - \hat{m}(x_i) + \frac{h^2}{2} \hat{m}''(x)\right\}^2.$$

This shares with the bootstrap the need for an estimate of bias. The bootstrap attempts to remove some of the dependence on such asymptotic formulas by simulating from the data to provide an estimate of the variance part of the mean squared error. Alternative estimates σ^2 were dis-

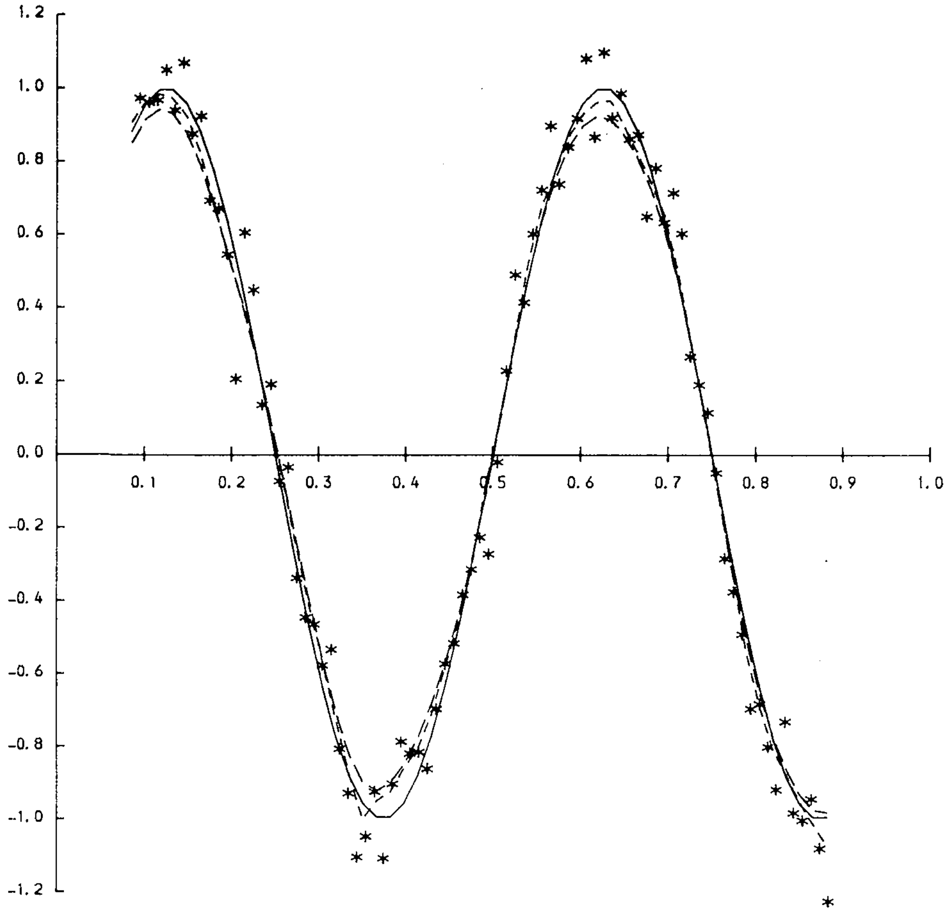

Figure 1. Data Simulated From the Curve $m(x) = \sin(4\pi x)$, With $N(0, (.1)^2)$ Error Distribution. The curve is shown by a solid line, global smoothing is shown by a dashed line, and local adaptive smoothing is shown by a fine dashed line.

cussed by Rice (1984) and Gasser, Sroka, and Jennen-Steinmetz (1986).

Figure 1 displays some data simulated by adding normally distributed errors, with standard deviation .1, to the curve $m(x) = \sin(4\pi x)$ evaluated at $x = (i - \frac{1}{2})/100$ ($i = 1, \ldots, 100$). To avoid problems with edge effects, the curves and data have been plotted only over an interior region of (0, 1). Cross-validation was used to select a good global smoothing parameter ($g = .03$; sum of squares based on an interior region to avoid edge effects). The resulting estimate of the regression function shows the problems caused by bias at the peaks and troughs, where $|m''(x)|$ is high.

Estimation of derivatives and appropriate smoothing parameters was discussed by Gasser and Müller (1984), who showed that a larger smoothing parameter will be required to obtain a good estimator of $m''(x)$. The asymptotic formulas given by these authors suggest that, for sample size 100, a simple but reasonable smoothing parameter for estimation of $m''(x)$ is obtained by approximately doubling the cross-validatory one, g. To use a level of smoothing that deviates greatly from this rule of thumb would require the assumption that $m''(x)$ is extremely smooth or extremely rough, since higher-order derivatives enter the asymptotic formulas with only a very small power. Here we will use $2g$. [Notice that estimation of $m''(x)$ requires the smoothing parameter to converge to 0 at a slower rate than in estimation of $m(x)$. The proposal to use $2g$ is tailored to sample size 100 and, in general, an approximate formula such as $1.5g\, n^{1/10}$ might be used.]

Figure 2 plots the local smoothing parameters obtained by minimizing the bootstrap estimate of mean squared error over a grid of smoothing parameters near g. For comparison, the asymptotically optimal local smoothing parameters are also plotted, and it can be seen that an appropriate pattern of local smoothing has been achieved. Again, to avoid edge effects, residuals near 0 and near 1 were not included in the bootstrap sampling and the curve was evaluated over the corresponding interior region. Comparison with the "plug-in" local smoothing parameters also reveals very little difference. Since the bootstrap is estimating only the variance of $\hat{m}(x)$, its performance is not markedly superior to the direct method. The estimate of the regression curve producd by the local parameters is also displayed in Figure 1, where it can be seen that this estimate is considerably nearer the true curve at most of the peaks and troughs. Since the two estimates based on local smoothing are virtually indistinguishable, only the bootstrap one has been plotted. A normal kernel with truncated support was used in this example because it has the helpful property that the convolution kernel $K_1(u; h, g)$ is well approximated by a normal density, with

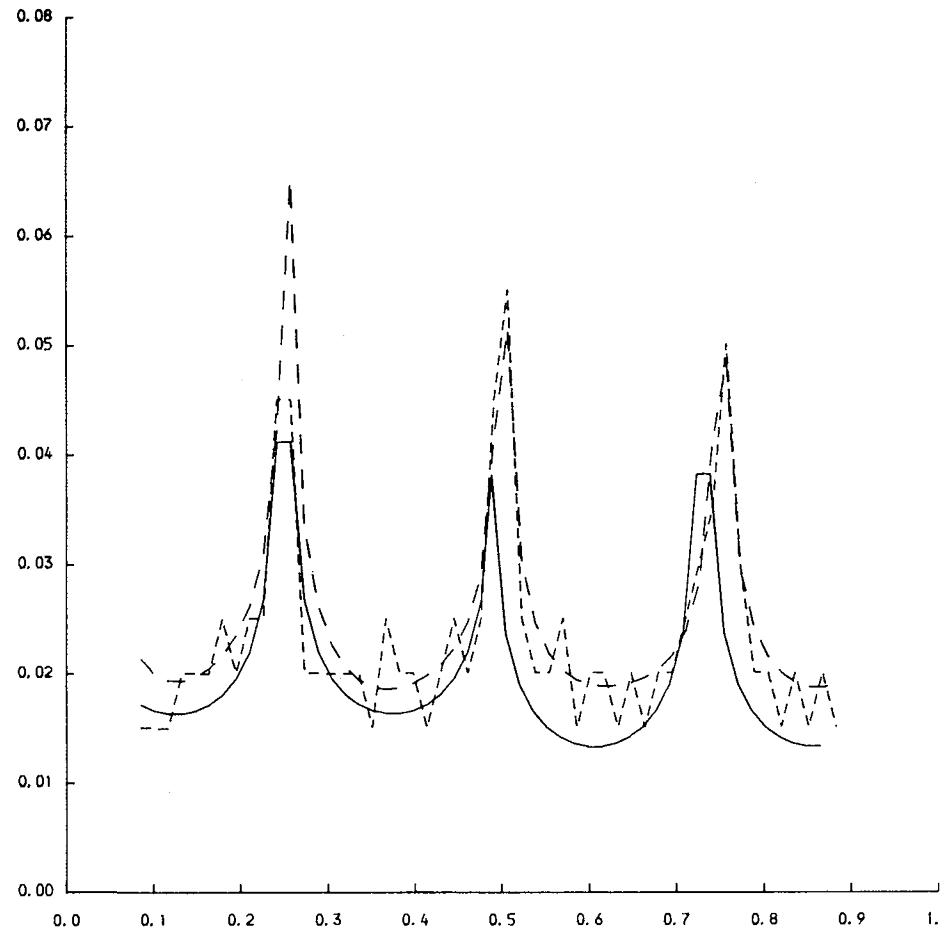

Figure 2. Local Smoothing Parameters for the Simulated Data of Figure 1. Asymptotically optimal is shown by a solid line, direct estimation is shown by a dashed line, and bootstrap is shown by a fine dashed line.

variance $h^2 + g^2$. In addition, the kernel $K_{(2)}$ used in the estimation of m'' can be taken to be the second derivative of the normal kernel.

To quantify the comparisons, 10 simulations were carried out and the squared errors of each estimated curve were averaged over the simulations and over the design points. The results for global smoothing, local smoothing by bootstrapping, and local smoothing by direct estimation were .000997, .000582, and .000581, respectively. This confirms the improved performance of local adaptive smoothing over global smoothing, and the similar results of the bootstrap and direct methods.

The asymptotically optimal local smoothing parameter contains the factor $m''(x)^{-2}$ and so takes very large values when m'' is close to 0. The curve on Figure 2 is truncated because the grid over which the parameters have been calculated does not contain the points .25, .5, or .75, where m'' is exactly 0. One disadvantage of the "plug-in" method of local smoothing, compared with the bootstrap, is that the factor $\hat{m}''(x)^{-2}$ is present, causing oversensitivity at locations where m'' is near 0, as Figure 2 shows.

4. CONFIDENCE BANDS

In addition to estimation of mean squared error, the bootstrap principle allows the construction of pointwise confidence bands for the true regression curve, since Theorem 1 shows that the bootstrap distribution approximates the distribution of $(nh)^{1/2}(\hat{m} - m)$. A confidence interval for the curve at a specific point x may be obtained by bootstrap sampling and calculation of the empirical quantiles. This contrasts with a direct approach based on asymptotic normality with estimated bias and variance.

In this section, global smoothing based on the cross-validatory bandwidth is used, both for the original data and for the bootstrap samples. The use of a global bandwidth allows results to be pooled across different points on the curve. The potential advantages of local adaptive smoothing in the context of confidence intervals are not clear, since a simple bias correction can be added to the globally smoothed curve, as suggested by the material of Section 2.

Figures 3 and 4 display the bias-corrected estimate with nominal 95% pointwise confidence intervals at 32 estimation points for the example described in Section 3. For clarity, the figures are drawn on recentered scales by subtracting the true regression curve. Figure 3 shows that when the error distribution is normal there is very little difference between the intervals produced by the bootstrap and the direct method, as would be expected from the discussion of mean squared error estimation in Section 2. The empirical confidence levels (coverage relative frequencies, averaged over design points) of the two methods

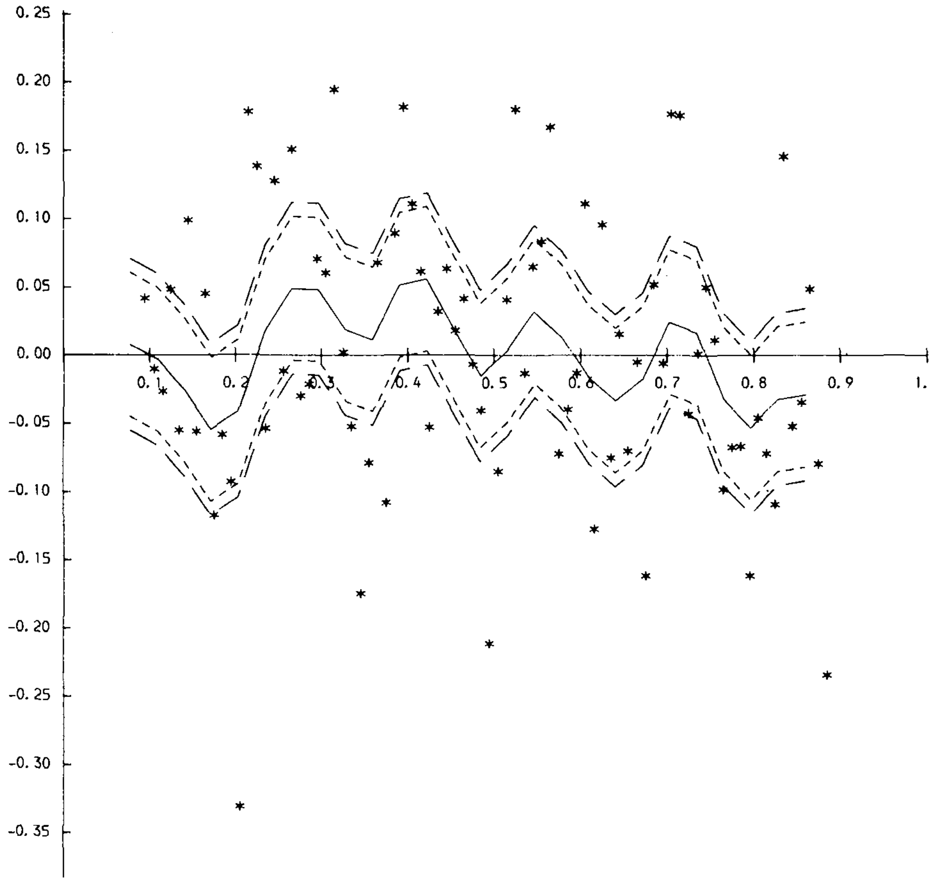

Figure 3. Pointwise Confidence Bounds, Recentered by Subtraction of the True Curve m(x) = sin(4πx), With Normal Errors, Scaled in Each Case to Have Mean 0 and Standard Deviation .1. Point estimate is shown by a full line, directly estimated bands are shown by a dashed line, and bootstrap estimation is shown by a fine dashed line.

are 87% for the bootstrap and 92% for the direct method. This is based on 100 simulations of original data followed in each case by 100 bootstrap simulations.

Since the direct method is based only on an estimate of error variance, we may expect the bootstrap to perform better in the presence of skewness. This is investigated in Figure 4, where the error distribution in the simulations is exponential, shifted to have mean 0 and scaled to have standard deviation .1. Here the bootstrap reflects the variation of the estimate about its mean more satisfactorily than the direct method and the asymmetry is clearly apparent. (Notice that positive skewness of the bootstrap distribution about its mean leads to negative skewness of the confidence interval.) The empirical confidence levels, however, remain at 86% for the bootstrap and 92% for the direct method.

The slightly low empirical confidence levels are due to the imperfect estimation of bias, as can be seen by the fact that the target confidence levels are very nearly achieved when the true second derivative is employed in the bias-correction term. The differences between the bootstrap and the direct methods can be explained by the fact that the bootstrap correctly reflects the variation of the estimate about its mean, and so the confidence intervals have a smaller width than the direct method. In both cases, however, the intervals are slightly incorrectly centered because of the bias estimation. In the direct method this also leads to a slightly inflated estimate of variance, which counteracts the incorrect centering and achieves a confidence level close to the target one.

A potential advantage of the bootstrap is that it can be applied to the construction of uniform confidence bands. Bootstrap sampling can be used to approximate the distribution of $\sup_x|\hat{m}(x) - m(x)|$, which in general is not amenable to theoretical treatment without the further assumption of normality, as in Knafl, Sacks, and Ylvisaker (1985) and Hall and Titterington (1986). This could be done in practice by examining the estimated regression curve over a very fine grid. With the methods described previously, however, the achieved confidence levels are unacceptably low as a result of the slight incorrect centering already discussed. The simultaneous bands are particularly sensitive to this effect because it needs only one point of the true curve to lie outside the confidence bands for coverage to fail. Practical use of this approach for simultaneous bands awaits a more satisfactory estimate of second derivatives.

5. DISCUSSION

The theory of Section 2 shows that the bootstrap is successful, in an asymptotic sense, in estimating features of the distribution of a nonparametric regression estimator. The numerical results of Sections 3 and 4 show that bootstrapping does not always perform better than a simple

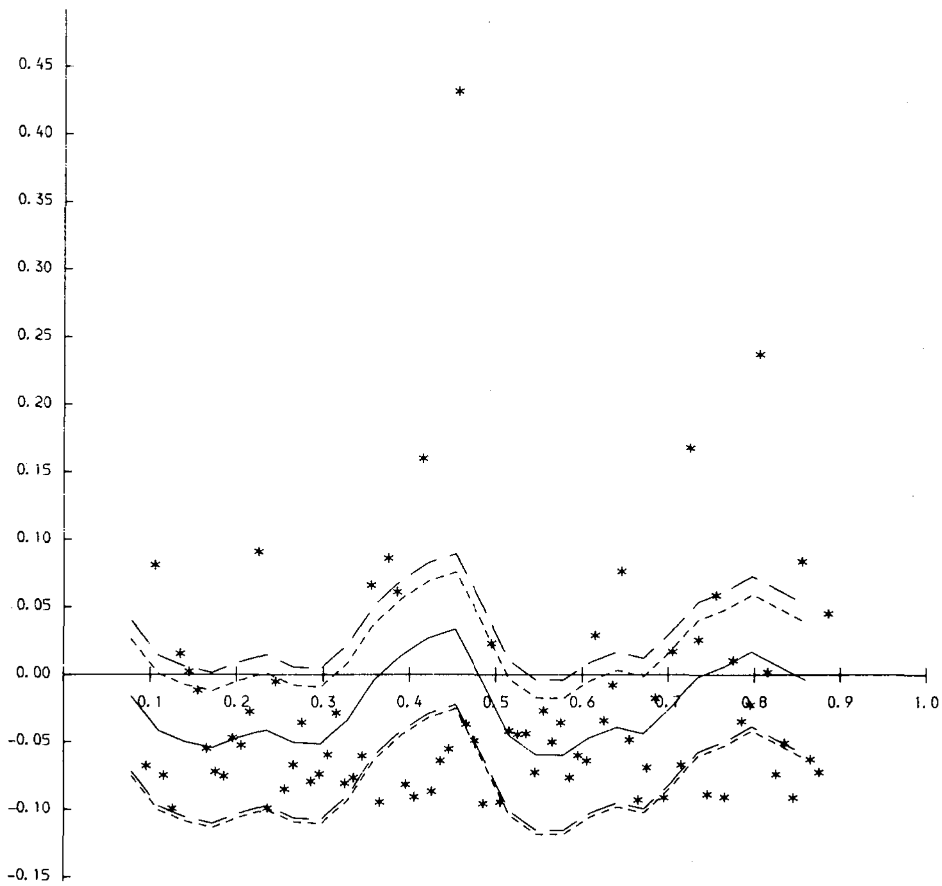

Figure 4. Pointwise Confidence Bounds, Recentered by Subtraction of the True Curve m(x) = sin(4πx), With Exponential Errors, Scaled in Each Case to Have Mean 0 and Standard Deviation .1. Point estimate is shown by a full line, directly estimated bands are shown by a dashed line, and bootstrap estimation is shown by a fine dashed line.

plug-in estimator. The root problem is estimation of bias and, in particular, estimation of the second derivative of the regression curve. The example used in the article is a particularly exacting one since the second derivative is large in several places along the curve. The methods described in the article will work more successfully for smoother curves.

In summary, the performances of the bootstrap and direct methods in the two problems discussed in the article may be compared by first observing that the direct method is simpler but based on asymptotic normality, whereas the bootstrap requires more computational effort in an attempt to reflect nonnormal features of the underlying distribution. In the local adaptive smoothing problem the bootstrap has a slight advantage of not being as sensitive to second derivatives near 0. In the construction of confidence intervals the bootstrap is able to reflect features such as skewness but falls slightly short of target confidence levels as a result of inaccuracies in centering when the second derivative of the curve is high.

APPENDIX: PROOF OF THEOREM 1

An application of Bickel and Freedman's (1981) lemma 8.8, using the modulus norm on the real line, allows the Mallows metric to be split up into a variance part

$$V_n = d_2(\sqrt{nh}\{\hat{m}(x; h) - E_F\hat{m}(x; h)\}, \sqrt{nh}\{m^*(x; h, g) - \mathrm{E}^*m^*(x; h, g)\}),$$

where E^* denotes expectation with respect to the bootstrap sampling, and a squared bias part

$$nh|b_n(x) - b_n^*(x)|^2,$$

where

$$b_n(x) = E_F\hat{m}(x; h) - m(x)$$

and

$$b_n^*(x) = \mathrm{E}^*m^*(x; h, g) - \hat{m}_c(x; h, g) + \frac{h^2}{2}\hat{m}''(x).$$

The variance part may be handled by fairly straightforward application of some of the results of Bickel and Freedman. Their lemma 8.9 shows that

$$\sup_{h\in H_n} d_2(\sqrt{nh}\{\hat{m}(x; h) - E\hat{m}(x; h)\}, \sqrt{nh}\{m^*(x; h, g) - \mathrm{E}^*m^*(x; h, g)\})$$

is bounded above by

$$\sup_{h\in H_n} nh \sum_i \alpha_i(x; h)^2 \cdot d_2(y_i - m(x_i), y_i^* - \hat{m}(x_i; g))^2, \quad \text{(A.1)}$$

where $\alpha_i(x; h)$ denotes $n^{-1}h^{-1}K((x - x_i)/h)$.

Let F denote the distribution function of the errors ε, let F_n denote the empirical distribution function (edf) of $\{\varepsilon_i\}$ where i is such that $x_i \geqslant \eta$ or $x_i \leqslant 1 - \eta$, let $\hat{F}_n$ denote the edf of the uncentered residuals $\{\hat{\varepsilon}_i\}$, and let $\tilde{F}_n$ denote the edf of the centered residuals $\{\tilde{\varepsilon}_i\}$. The bound (A.1) may be denoted by

$$\sup_{h\in H_n} nh \sum_i \alpha_i(x; h)^2 \cdot d_2(F, \tilde{F}_n)^2,$$

if we adopt the convention on d_2 made previously. It suffices to show that $d_g(F, \tilde{F}_n)$ converges in probability to 0. Since d_2 is a metric, we have that

$$d_2(F, \tilde{F}_n)^2 \leq 2(d_2(F, F_n)^2 + d_2(F_n, \tilde{F}_n)^2).$$

$d_2(F, F_n) \underset{p}{\rightarrow} 0$ by lemma 8.4 of Bickel and Freedman.

The general result for random variables U and V,

$$d_2(U, V)^2 = d_2(U, V - E(V))^2 + E^2(U - V) - E^2(U),$$

can be proved by a slight amendment of the proof of Bickel and Freedman's lemma 8.8. An application of this to $d_2(F_n, \tilde{F}_n)^2$, with $U \sim F_n$ and $V \sim \tilde{F}_n$, yields

$$d_2(F_n, \tilde{F}_n)^2 = d_2(F_n, \hat{F}_n)^2 - \left\{((1 - 2\eta)n)^{-1} \sum_i (\hat{\varepsilon}_i - \varepsilon_i)\right\}^2 + \left\{((1 - 2\eta)n)^{-1} \sum_i \varepsilon_i\right\}^2$$

and hence

$$E_F d_2(F_n, \tilde{F}_n)^2 \leq E_F d_2(F_n, \hat{F}_n)^2 + \frac{\sigma^2}{(n - 2\eta n)}.$$

By definition of the Mallows metric $d_2(F_n, \hat{F}_n)^2$, we may consider the joint distribution of $\{\varepsilon_i\}$ and $\{\hat{\varepsilon}_i\}$, which puts probability $(n - 2\eta n)^{-1}$ at each $\{\varepsilon_i, \hat{\varepsilon}_i\}$ to establish that

$$E_F d_2(F_n, \hat{F}_n)^2 \leq E_F\left\{(n - 2\eta n)^{-1} \sum_i (\varepsilon_i - \hat{\varepsilon}_i)^2\right\} = (n - 2\eta n)^{-1} \sum_i \mathrm{MSE}(x_i),$$

where $\mathrm{MSE}(x_i)$ denotes the mean squared error of $\hat{m}(x_i; g)$.

The convergence result is now established by combining these inequalities and observing that $\sum_i \alpha_i(x; h)^2$ converges uniformly over $h \in H_n$ to 0 with speed $n^{-1}h^{-1}$ (Priestley and Chao 1972).

To deal with the bias part, denote by $\alpha_i(x; h)$ the weight $n^{-1}h^{-1}K((x - x_i)/h)$ and by $\beta_i(x; h, g)$ the weight $n^{-1}K_1(x - x_i; h, g)$. Then

$$\hat{m}(x; h) = \sum_i \alpha_i(x; h) Y_i$$

and

$$b_n(x) = \sum_i \alpha_i(x; h)m(x_i) - m(x).$$

The bootstrap bias is

$$\begin{aligned} b_n^*(x) &= \mathrm{E}^* \left\{\sum_i \alpha_i(x; h) Y_i^*\right\} - \hat{m}_c(x; h, g) + \frac{h^2}{2} \hat{m}''(x) \\ &= \sum_i \alpha_i(x; h)\hat{m}(x_i; g) - \hat{m}_c(x; h, g) + \frac{h^2}{2} \hat{m}''(x) \\ &= \sum_i \alpha_i(x; h) \sum_j \alpha_j(x_i; g) Y_j - \sum_i \beta_i(x; h, g) Y_i + \frac{h^2}{2} \hat{m}''(x). \end{aligned}$$

By writing $Y_i = m(x_i) + \varepsilon_i$ and combining the bias components, we have

$$\begin{aligned} b_n(x) - b_n^*(x) &= \sum_i \alpha_i(x; h)m(x_i) - m(x) - \sum_i \alpha_i(x; h) \sum_j \alpha_j(x_i; g)m(x_j) \\ &\quad + \sum_i \beta_i(x; h, g)m(x_i) - \frac{h^2}{2} \hat{m}''(x) \\ &\quad - \sum_i \alpha_i(x; h) \sum_j \alpha_j(x_i; g)\varepsilon_j + \sum_i \beta_i(x; h, g)\varepsilon_i. \end{aligned}$$

Consider first the ε_i terms. These may be gathered together as

$$T_{1n} = - \sum_i \left\{\sum_j \alpha_i(x; h)\alpha_j(x_i; g) - \beta_i(x; h, g)\right\} \varepsilon_i,$$

which has mean 0 and variance

$$n^{-2} \sum_i \left\{n^{-1} \sum_j h^{-1}K((x - x_j)/h)g^{-1}K((x_i - x_j)/g) - K_1((x - x_i); h, g)\right\}^2 \sigma^2. \quad \text{(A.2)}$$

Since $K_1(x - x_i; h, g) = \int h^{-1}K((x - y)/h)g^{-1}K((x_i - y)/g)\, dy$, we may use the mean–value theorem and the Lipschitz continuity of the kernel to show that an upper bound for (A.2) is provided by $C_1 n^{-1}\{n^{-1}g^{-1}(h^{-1} + g^{-1})\}^2$ for some constant C_1. This shows that (A.2) is $o(n^{-1})$ under Condition 3.

Consider now the other terms of $b_n(x) - b_n^*(x)$, which may be grouped together as

$$\begin{aligned} T_{2n} &= \sum_i \alpha_i(x; g)m(x_i) - m(x) - \sum \alpha_i(x; h) \sum \alpha_j(x_i; g)m(x_j) \\ &\quad + \sum \alpha_i(x; h)m(x_i) + \sum_j \beta_j(x; h, g)m(x_j) \\ &\quad - \sum_i \alpha_i(x; g)m(x_i) - \frac{h^2}{2} \hat{m}''(x). \quad \text{(A.3)} \end{aligned}$$

The first two terms of (A.3) may be written as $(g^2/2)\, m''(x) + o(g^2)$, and the second two terms may be written as $-\sum_i \alpha_i(x; h) (g^2/2)\, m''(x_i) + o(g^2)$, where $o(g^2)$ is uniform in i since m'' is uniformly continuous. Boundary problems cannot occur since x was assumed to be between η and $1 - \eta$ and the kernel K has compact support. So, for n large enough, no design point with $x_i \geq 1 - \eta/2$ or $x_i \leq \eta/2$ enters the approximations of the terms in formula (A.3).

The first four terms of (A.3) can be replaced by the following integral:

$$-g^2/2 \int h^{-1}K((x - y)/h)\{m''(y) - m''(x)\}\, dy + o(g^2).$$

Splitting this integral and using continuity of m'' as in Parzen (1962), it can be shown that this term is $o(g^2)$.

Since the variance associated with the density $K_1(\cdot; h, g)$ is $(h^2 - g^2)$, the remaining terms of (A.3) may be written as

$$\frac{(h^2 + g^2)}{2} \cdot m''(x) - \frac{g^2}{2} m''(x) - \frac{h^2}{2} \hat{m}''(x) + o_p(h^2 + g^2) = \frac{h^2}{2} \{m''(x) - \hat{m}''(x)\} + o_p(h^2 + g^2).$$

Since $\hat{m}''(x)$ is a consistent estimator of $m''(x)$, these terms are $o_p(h^2 + g^2)$.

Collecting together all of the terms of the squared bias part of the Mallows metric, we now have that this converges in probability to 0. This general result allows us to use the bootstrap to investigate the distribution of any quantities of interest.

[Received June 1985. Revised April 1987.]

REFERENCES

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217.

Gasser, T., and Müller, H.-G. (1979), "Kernel Estimation of Regression Functions," in *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Springer Lecture Notes, 757, 23–68.

——— (1984), "Estimating Regression Functions and Their Derivatives by the Kernel Method," *Scandinavian Journal of Statistics*, 11, 171–185.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986), "Residual Variance and Residual Pattern in Nonlinear Regression," *Biometrika*, 73, 625–634.

Hall, P., and Titterington, D. M. (1986), "On Confidence Bands in Nonparametric Density Estimation and Regression," unpublished manuscript.

Härdle, W., and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics*, 13, 1465–1481.

Knafl, G., Sacks, J., and Ylvisaker, D. (1985), "Confidence Bands for Regression Functions," *Journal of the American Statistical Association*, 80, 683–691.

Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, 33, 1065–1076.

Priestley, M. B., and Chao, M. J. (1972), "Non-parametric Function Fitting," *Journal of the Royal Statistical Society*, Ser. B, 34, 385–392.

Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 12, 1215–1230.

Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

Stuetzle, W., and Mittal, Y. (1979), "Some Comments on the Asymptotic Behavior of Robust Smoothers," in *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Springer Lecture Notes, 757, 191–195.