The Annals of Statistics
1995, Vol. 23, No. 6, 1921–1936

ON BANDWIDTH CHOICE IN NONPARAMETRIC REGRESSION WITH BOTH SHORT- AND LONG-RANGE DEPENDENT ERRORS

BY PETER HALL, SOUMENDRA NATH LAHIRI AND JÖRG POLZEHL

Australian National University, Iowa State University and Konrad-Zuse-Zentrum für Informationstechnik

We analyse methods based on the block bootstrap and leave-out cross-validation, for choosing the bandwidth in nonparametric regression when errors have an almost arbitrarily long range of dependence. A novel analytical device for modelling the dependence structure of errors is introduced. This allows a concise theoretical description of the way in which the range of dependence affects optimal bandwidth choice. It is shown that, provided block length or leave-out number, respectively, are chosen appropriately, both techniques produce first-order optimal bandwidths. Nevertheless, the block bootstrap has far better empirical properties, particularly under long-range dependence.

1. Introduction. In three seminal papers on nonparametric regression with short-range dependent data, Altman (1990), Chu and Marron (1991) and Hart (1991) addressed both the failure of cross-validation and the sort of remedy that might be appropriate to correct it. Chu and Marron considered a modified or "leave-k-out" form of cross-validation, and argued that, for processes exhibiting short-range dependence, this approach may produce asymptotically optimal performance if k is chosen to increase with sample size at an appropriate rate. On the other hand, the method of partitioned cross-validation was shown by Chu and Marron to be relatively unsuccessful in producing asymptotically optimal bandwidths.

In this note we take up the argument where it was left by Chu and Marron, and demonstrate that, even in the context of very-long-range dependent data, both modified cross-validation and a form of the block bootstrap produce asymptotically optimal bandwidths. We develop a simple asymptotic device that allows very long ranges of dependence to be modelled and analysed with relative ease. For example, it permits an elementary account of the way in which leave-out number (in cross-validation) or block size (for the block bootstrap) should depend on strength of dependence if first-order optimality of bandwidth choice is to be achieved. Thus, even in the context of short-range dependence and leave-k-out cross-validation we complement Chu and Marron's results by indicating the sort of leave-out numbers or block

Received May 1994; revised May 1995.

AMS 1991 *subject classifications*. Primary 62G07, 62G09; secondary 62M10.

Key words and phrases. Bandwidth choice, block bootstrap, correlated errors, cross-validation, curve estimation, kernel estimator, local linear smoothing, long-range dependence, mean squared error, nonparametric regression, resampling, short-range dependence.

sizes that are required. Furthermore, we address the block bootstrap approach to both local and global bandwidth choice. These theoretical results are described in Section 2, for which the technical details are outlined in Section 4. Section 3 summarizes the conclusions of a simulation study. That work makes it clear that while leave-k-out cross-validation has first-order theoretical properties similar to those of the block bootstrap, its empirical performance is very poor under long-range dependence. This is due to a marked tendency for cross-validation to select a bandwidth that is almost identical to the smallest one producing a well-defined cross-validation criterion. The problem becomes more pronounced as the range of dependence increases, with the result that leave-k-out cross-validation could not really be considered to perform satisfactorily with a variety of ranges of dependence. The block bootstrap is much more satisfactory.

A leave-k-out cross-validation method was also considered by Hart and Vieu (1990), in the context of density estimation. Hart and Wehrly (1986) studied bandwidth selection when measurements are repeated; Härdle and Vieu (1992) addressed leave-one-out cross-validation with mixing errors; Chiu (1989), Diggle and Hutchinson (1989), Hermann, Gasser and Kneip (1992) and Kohn, Ansley and Wong (1992) discussed other aspects of bandwidth choice for dependent data; and Hart (1994) introduced the method of time series cross-validation, appropriate when the dependence structure may be modelled parametrically. Surveys of the literature on nonparametric regression under dependence may be found in Györfi, Härdle, Sarda and Vieu (1989) and Härdle [(1990), Chapter 7]. The analysis of bootstrap methods for approximating error in curve estimation with independent data was initiated by Taylor (1989) and Faraway and Jhun (1990). The block bootstrap for dependent data was developed by Hall (1985), Carlstein (1986) and Künsch (1989).

2. Main results.

2.1. *Estimators and basic properties.* As in Altman (1990), Chu and Marron (1991) and Hart (1991) we suppose that the observed data $\mathscr{X} = \{Y_i,\ 1 \le i \le n\}$ are generated by the model $Y_i = m(x_i) + \varepsilon_i$, where $x_i = (i + c)/n$ for a constant c, m is a smooth function and $\{\varepsilon_i\}$ is a stationary sequence with zero mean. Let w_i denote a weight function. We take

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i \tag{2.1}$$

as our estimator of m. One candidate for w_i, producing the Nadaraya–Watson kernel estimator treated by Chu and Marron (1991), is

$$w_i(x) = K\{(x - x_i)/h\} \left[\sum_{j=1}^{n} K\{(x - x_j)/h\} \right]^{-1}, \tag{2.2}$$

where K is a kernel function and h is a bandwidth. Another, the local linear regression smoother proposed by Fan (1993), is

$$w_i(x) = v_i(x)\left\{\sum_{j=1}^{n} v_j(x) + n^{-2}\right\}^{-1}, \tag{2.3}$$

where

$$v_i(x) = K\left\{\frac{x - x_i}{h}\right\}\{s_2 - (x - x_i)s_1\} \text{ and } s_k = \sum_{j=1}^{n} K\left\{\frac{x - x_j}{h}\right\}(x - x_j)^k$$

for $k = 1, 2$.

Alternative choices of w_i include those proposed by Gasser and Müller (1979) or a simpler version of the Nadaraya–Watson prescription in which the denominator in (2.2) is replaced by nh. Our results have straightforward analogues in these cases. The mean squared error (MSE) and mean integrated squared error (MISE) of $\hat{m}$ are given by

$$\text{MSE}(x) = E\{\hat{m}(x) - m(x)\}^2, \qquad \text{MSE} = \int_{\mathscr{I}} E(\hat{m} - m)^2, \tag{2.4}$$

where $\mathscr{I} \subseteq (0, 1)$.

We model the dependence of the errors by taking $\{\varepsilon_i, 1 \le i \le n < \infty\}$ to be a triangular array, with the nth row having a joint distribution determined by defining $\varepsilon_i = Z(\lambda x_i)$, $1 \le i \le n$, where Z is a stationary stochastic process in the continuum. We assume that Z has zero mean and autocovariance γ, and take $\lambda = \lambda_n$ to be a sequence of positive numbers that would typically increase with n. Under this model, $E(\varepsilon_i \varepsilon_j) = \gamma\{\lambda(x_i - x_j)\}$, and λ may be interpreted as a measure of the strength of dependence of the process $\{\varepsilon_i\}$, with larger values of λ indicating weaker dependence. In particular, $\lambda = \infty$ corresponds to independence, and $\lambda/n \to \infty$ to asymptotic independence, in the sense that first-order asymptotic properties of $\hat{m}$ are identical to those under independence. We always assume that $\lambda \to \infty$ as $n \to \infty$. If λ does not diverge, then the amount of statistical information contained in any given sequence $\{Y_i\colon\ x_i \in (a, b)\}$, for any $a < b$, does not generally increase with increasing n. (For example, consider the case where the process Z is Gaussian.) The classical description of dependence among errors in nonparametric regression arises when $\lambda \equiv n$, the results in so-called "time-series errors" with $E(\varepsilon_i \varepsilon_j) = \gamma(i - j)$. This is the context studied by Altman (1990), Chu and Marron (1991) and Hart (1991).

Note particularly that under our model the sum of autocovariances, $s_n \equiv \sum_{i=1}^{n} E(\varepsilon_{i+1}\varepsilon_1)$, is not necessarily bounded. Indeed, if $\lambda/n \to 0$ and $\int\gamma \ne 0$, then s_n is asymptotic to a constant multiple of n/λ and so is unbounded, implying that the data exhibit long-range dependence.

As a prelude to describing asymptotic properties of mean squared error under our model, we assume that m has two bounded, continuous derivatives on the interval $[0, 1]$; that K satisfies the usual conditions of a second-order kernel (i.e., $\int y^i K(y)\,dy = 1$ if $i = 0$, 0 if $i = 1$ and 2κ, say, if $i = 2$) and is

compactly supported, Hölder continuous and of bounded variation, with $\int_{-x}^{\infty} K$ and $\int_{-\infty}^{x} K$ bounded away from zero for all $x \geq 0$; that γ is integrable, ultimately monotone and satisfies $\int \gamma \neq 0$; that $h = h(n) \to 0$ and $\min(\lambda, n)h \to \infty$ as $n \to \infty$; and that $\lambda/n \to C$, where $0 \leq C \leq \infty$. (The condition that K be compactly supported is imposed only for simplicity in technical arguments, and may be relaxed. In particular, our results are all valid if K is the standard normal density.) Define $R(K) = \int K^2$ and $\Lambda = \min(n, \lambda)$, and let $\beta = m'' - 1$ if the weights are given by (2.2), $\beta = m''$ if they are given by (2.3).

THEOREM 2.1. *If $C = 0$ or ∞, then*

$$E\{\hat{m}(x) - m(x)\}^2 = R(K)\left\{(nh)^{-1}\gamma(0) + (\lambda h)^{-1}\left(\int \gamma\right)\right\} + h^4\kappa^2\beta(x)^2 + o\{(\Lambda h)^{-1} + h^4\} \tag{2.5}$$

uniformly in $x \in (\delta, 1 - \delta)$ for each $\delta > 0$, and

$$\int_{\mathscr{I}} E\{\hat{m}(x) - m(x)\}^2\,dx = \int_{\mathscr{I}} \left[R(K)\left\{(nh)^{-1}\gamma(0) + (\lambda h)^{-1}\left(\int \gamma\right)\right\} + h^4\kappa^2\beta(x)^2\right] dx + o\{(\Lambda h)^{-1} + h^4\} \tag{2.6}$$

uniformly in measurable sets $\mathscr{I} \subseteq (\delta, 1 - \delta)$. If $0 < C < \infty$, then

$$E\{\hat{m}(x) - m(x)\}^2 = R(K)\left\{(nh)^{-1}\gamma(0) + (\lambda h)^{-1}\sum_{i \neq 0}\gamma(Ci)\right\} + h^4\kappa^2\beta(x)^2 + o\{(\Lambda h)^{-1} + h^4\} \tag{2.7}$$

uniformly in $x \in (\delta, 1 - \delta)$ for each $\delta > 0$, and

$$\int_{\mathscr{I}} E\{\hat{m}(x) - m(x)\}^2\,dx = \int_{\mathscr{I}} \left[R(K)\left\{(nh)^{-1}\gamma(0) + (\lambda h)^{-1}\sum_{i \neq 0}\gamma(Ci)\right\} + h^4\kappa^2\beta(x)^2\right] dx + o\{(\Lambda h)^{-1} + h^4\} \tag{2.8}$$

uniformly in measurable sets $\mathscr{I} \subseteq (\delta, 1 - \delta)$. If the weights at (2.3) *are employed, then* (2.6) *and* (2.8) *are available uniformly in all $\mathscr{I} \subseteq (0, 1)$.*

Locally and globally optimal bandwidth choices are obtained by minimizing the right-hand sides of (2.5)–(2.8). In all cases the optimal bandwidth is asymptotic to a constant multiple of $\Lambda^{-1/5}$, giving a convergence rate of $\Lambda^{-4/5}$ in terms of mean squared error. In the special case when $C = 0$, a version of (2.5) has been proved by Hart (1987) for the Gasser–Müller estimator of m.

2.2. *Bandwidth choice by block bootstrap.* Let $\hat{m}_1$ and $\hat{m}_2$ denote two estimators of m that are constructed according to the prescription at (2.1), but employing respective values h_1 and h_2 of the bandwidth h. In what follows, $\hat{m}_1$ will be used to compute centred residuals and $\hat{m}_2$ to generate bootstrap data.

Put $\hat{\varepsilon}_{1i} = Y_i - \hat{m}_1(x_i)$, $\bar{\hat{\varepsilon}}_{1\cdot} = N^{-1}\Sigma'\hat{\varepsilon}_{1i}$ and $\hat{\varepsilon}_i = \hat{\varepsilon}_{1i} - \bar{\hat{\varepsilon}}_{1\cdot}$, where Σ' denotes summation over all N design points x_i that lie within $\mathscr{I} \subseteq (0, 1)$. These are the centred residuals. Shortly we shall define the bootstrap errors ε_i^*. In terms of those, let

$$Y_i^* = \hat{m}_2(x_i) + \varepsilon_i^*, \qquad \hat{m}^*(x) = \sum_{i=1}^{n} w_i(x) Y_i^*,$$

where w_i is exactly as in (2.1). Our estimators of MSE and MISE are

$$\widehat{\mathrm{MSE}}(x) = E\Big[\{\hat{m}^*(x) - \hat{m}_2(x)\}^2 | \mathscr{X}\Big],$$

$$\widehat{\mathrm{MISE}} = N^{-1} \sum{}' E\Big[\{\hat{m}^*(x) - \hat{m}_2(x)\}^2 | \mathscr{X}\Big].$$

To minimize mean squared error, locally or globally, we select the smoothing parameter in the definition of w_i so as to minimize $\widehat{\mathrm{MSE}}$ or $\widehat{\mathrm{MISE}}$, respectively. It is of course not essential to use the same interval $\mathscr{I}$ to define $\widehat{\mathrm{MISE}}$ and the residuals $\hat{\varepsilon}_i$; we do so only to simplify notation and discussion.

Next we describe a block bootstrap algorithm for generating the ε_i^*'s. Write n_1, n_2 for integers such that $x_i \in \mathscr{I}$ if and only if $i \in \{j\colon n_1 \le j \le n_2\}$. Let $l \le n_2 - n_1 + 1$ denote block length and let b denote the integer satisfying $(b-1)l < n \le bl$. (We shall argue in Theorem 2.2 that l should be of smaller order than nh_1 and of larger order than n/λ.) Write $\mathscr{B}_i = (\hat{\varepsilon}_i, \ldots, \hat{\varepsilon}_{i+l-1})$ for the block of centred residuals that starts from position i, where $n_1 \le i \le n_2 - l + 1$. Resample randomly, with replacement, b times from the sequence of all $n_2 - n_1 - l + 2$ such blocks, obtaining the sequence $\mathscr{B}_j^*$, $1 \le j \le b$, say. Put the elements of these blocks into a string of length bl, and let ε_i^* denote the ith element of the string.

We are now in a position to address performance of the block bootstrap. Take $\mathscr{I}$ to be a subset of $(\delta, 1-\delta)$ for some $\delta > 0$, to eliminate edge effects. In addition to the conditions of Theorem 2.1, assume that m'' is Hölder continuous with exponent $\alpha \in (0, 1)$. Suppose that the process Z used to define the errors ε_i is stationary with all moments finite, and satisfies the Rosenblatt mixing condition

$$\sup\{|P(A \cap B) - P(A)P(B)|\colon A \in \mathscr{F}_{-\infty}^{a}, B \in \mathscr{F}_{a+t}^{\infty}, -\infty < a < \infty\} \le C_1 \exp(-C_2 t)$$

for all $t > 0$ and for constants $C_1, C_2 > 0$, where $\mathscr{F}_a^b$ denotes the σ-field generated by $\{Z(x)\colon x \in [a, b]\}$. To simplify the formulation of our next result we assume that h_1 is of the same order as the optimal bandwidth, and in fact take $h_1 = \xi\Lambda^{-1/5}$ for some $\xi > 0$; and suppose that h_2 satisfies $C_3\Lambda^{-(1/5)+\delta} \le h_2 = o(\Lambda^{-(1/5)+(\alpha/10)})$ for some C_3, $\delta > 0$. This reflects the fact that $\hat{m}_1$

should ideally use a bandwidth of similar order to the optimal one, whereas $\hat{m}_2$ should employ a bandwidth of larger order, in order to address the problem of implicitly estimating the second derivative of m when approximating the squared bias contribution to mean squared error. (In practice, h_1 at least would be computed using an iterative scheme.) Finally, we assume that

$$n(\lambda l)^{-1} + l(nh_1)^{-1} + n^{\delta}\lambda^{-1} \to 0 \tag{2.9}$$

as $n \to \infty$. The first two portions of this condition assert maximum and minimum orders of magnitude, respectively, for the block length l. Note that $l = nh_1^a$ satisfies both the requirements whenever $1 < a < 5$. The last part is a technical assumption and rules out extraordinary long-range dependence.

THEOREM 2.2. *Under the above conditions*,

$$\widehat{\mathrm{MSE}}(x) - \mathrm{MSE}(x) = o_p(\Lambda^{-4/5}), \qquad \widehat{\mathrm{MISE}} - \mathrm{MISE} = o_p(\Lambda^{-4/5})$$

uniformly in $x \in (\delta, 1-\delta)$ *and* $h \in H_n = [C_4\Lambda^{-1/5}, C_5\Lambda^{-1/5}]$ *for any* $\delta > 0$ *and any* $0 < C_4 < C_5 < \infty$.

In the event that the weights w_i are chosen by the prescription at (2.3) we may take $\delta = 0$, both in the definition of $\mathscr{I}$ and the statement of the theorem.

It follows from Theorems 2.1 and 2.2 that the ratio of the bandwidth that minimizes $\widehat{\mathrm{MSE}}(x)$ (respectively, $\widehat{\mathrm{MISE}}$) over H_n to that which minimizes $\mathrm{MSE}(x)$ (respectively, MISE) converges to 1 in probability.

2.3. *Bandwidth choice by leave-out cross-validation.* Let $\hat{m}_j$ represent the version of the estimator $\hat{m}$, defined at (2.1), in which the sum over $1 \le i \le n$ is replaced by a sum over those values in this range that satisfy $|i - j| > l$, where l is an integer. Put

$$\widehat{\mathrm{MISE}} = N^{-1} \sum{}' \{\hat{m}_j(x_j) - Y_j\}^2.$$

Then choosing h to minimize $\widehat{\mathrm{MISE}}$ amounts to using leave-k-out cross-validation with $k = 2l + 1$. The integer k, or equivalently l, plays a role similar to block length in the block bootstrap.

Our main result in this section is an analogue of Theorem 2.2 and holds under identical conditions except that we change (2.9) to

$$n(\lambda l)^{-1} + l\left(nh_1^3\right)^{-1} + n^{\delta}\lambda^{-1} \to 0 \tag{2.10}$$

as $n \to \infty$.

THEOREM 2.3. *Under the above conditions*,

$$\widehat{\mathrm{MISE}} - \mathrm{MISE} - N^{-1} \sum{}' \varepsilon_i^2 = o_p(\Lambda^{-4/5})$$

uniformly in $h \in H_n = [C_4\Lambda^{-1/5}, C_5\Lambda^{-1/5}]$ *for any* $\delta > 0$ *and any* $0 < C_4 < C_5 < \infty$.

It follows that if l is chosen appropriately to satisfy (2.10), then the ratio of the bandwidth that minimizes $\widetilde{\text{MISE}}$ over H_n to that which minimizes MISE converges to 1 in probability.

3. Simulations. We investigated numerical aspects of the block bootstrap and leave-out cross-validation in a simulation study, the results of which are summarized here. The mean function m and autocovariance γ were chosen as $m(x) = \cos(4\pi x)$ and $\gamma(t) = \sigma^2 e^{-(1/5)|t|}$ with $\sigma^2 = 0.04$. We took K to be the triweight kernel $K(u) = (35/32)(1 - u^2)^3$, and employed the w_i sequence defined at (2.3). The process Z was assumed Gaussian with zero mean, and its values were simulated using a Fourier-based algorithm developed by Wood and Chan (1993). We used the procedures described in Sections 2.2 and 2.3 to estimate the bandwidth that minimizes MISE, defined at (2.4) with $\mathscr{I} = (0, 1)$.

To implement the block bootstrap, we calculated h_1 by iteration, selecting first a plausible bandwidth $h_{11} = 0.4n^{-1/5}$, then using it to calculate an estimate h_{12} of the optimal bandwidth by minimizing the block bootstrap estimate of MSE(x) or MISE, then replacing h_{11} by h_{12} and repeating the operation, and so on until convergence was achieved. Of course, the bandwidth to which these iterations converge is equal to that which we seek.

To ensure the right rate of decay for h_2 under both short- and long-range dependence, we chose h_2 to depend on h_1 in the form $h_2 = C * h_1^{5/9}$. This gives a bandwidth of the correct size $\Lambda^{-1/9}$ for minimizing $\int_{\mathscr{I}} E(\hat{m}'' - m'')^2$, the mean integrated squared error (MISE) of $\hat{m}''$, as an estimate of m''. The constant C, obtained from asymptotic theory under independence [see Gasser, Engel and Seifert (1993)], is of the form $C = (5\int y^2 K(y)\,dy / \int y^2 K''(y)\,dy * R(K'')/R(K) * R(m'')/R(m''''))^{1/9}$. The terms $R(m'')$ and $R(m'''')$ were, respectively, estimated by $R(\hat{m}'')$ and $R(\hat{m}'''')$, using bandwidths h_2 for the first and $h_3 = h_1^{5/13}$ in case of the second term.

For comparison we implemented the leave-$(2l + 1)$-out cross-validation $\widetilde{\text{MISE}}$ using local linear regression weights. More precisely, we minimized

$$\widetilde{\text{MISE}} = n^{-1} \sum_{j=1}^{n} \{\hat{m}_j(x_j) - Y_j\}^2$$

with

$$\hat{m}_j(x_j) = \sum_{|i-j|>l} \omega_{ij}(x_j) Y_i,$$

where

$$\omega_{ij}(x) = v_{ij}(x) \left\{ \sum_{|k-j|>l} v_{kj}(x) + n^{-2} \right\}^{-1},$$

$$v_{ij}(x) = K\{(x - x_i)/h\}\{s_{2j} - (x - x_i)s_{1j}\}$$

and

$$s_{kj} = \sum_{|i-j|>l} K\{(x_j - x_i)/h\}(x_j - x_i)^k.$$

We considered the same situations as in the case of the block bootstrap.

Our simulations covered a very wide variety of parameter settings and sample sizes. One feature of the results is that when cross-validation produces a plausible estimate of the MISE curve, it tends to be flatter than its counterpart obtained using the block bootstrap, and so is harder to minimize numerically. In consequence, if the minimum does not occur near the smallest possible bandwidth, then its value is strongly influenced by sampling fluctuations, with the result that in such cases cross-validation produces more variable bandwidths than does the block bootstrap. Moreover, as the range of dependence increases, cross-validation shows a marked tendency to select the smallest bandwidth that is consistent with the method being well defined, leading to chronic undersmoothing. This problem is not observed with the block bootstrap.

Naturally these effects are influenced by choice of l, for either method. However, apart from relatively extreme cases such as $l = 1$, the estimate of MISE provided by the block bootstrap tends to be nicely curved, and so is easily minimized, for a wide range of choices of l. With both methods the position of the minimum of the estimate of MISE tends at first to increase with l. In the case of cross-validation with very long-range dependent data, however, this tendency is little more than an artifact of the property that the smallest bandwidth for which the criterion is well defined increases with l; it is equal to $(l + 1)/n$. When using the block bootstrap the minimizing bandwidth tends to decrease with l after reaching a maximum. We did not observe this property clearly in the case of cross-validation. For both methods the variance of the minimizing bandwidth tends to increase with l, except for those cases of cross-validation where the minimizing bandwidth is very near to the smallest value for which the criterion is well defined. As suggested by the theory in Section 2, the value of l that is required for good performance using either the block bootstrap or cross-validation tends to increase with range of dependence, and our simulation study bears this out.

For the sake of brevity we illustrate only one set of results from the simulation study, corresponding to $n = 1600$ and the two extreme cases $\lambda = 1600$ (short-range dependence) and $\lambda = 200$ (long-range dependence). We chose a large sample size principally because there the inferior properties of cross-validation are starker. For smaller sample sizes those problems can perhaps be put down to insufficient data, but when they are so obvious with $n = 1600$ we feel that our conclusions are quite convincing. All the features described above are evident for smaller sample sizes.

Figures 1 ($\lambda = 1600$) and 2 ($\lambda = 200$) depict the case of the block bootstrap, whereas Figures 3 ($\lambda = 1600$) and 4 ($\lambda = 200$) address cross-validation. Each treats several values of l. The left-hand panel of each figure illustrates the average (over 20 out of 100 samples in the case of Figures 1 and 2, and 200 samples for Figures 3 and 4) of the estimate of MISE. We used only 20 of the

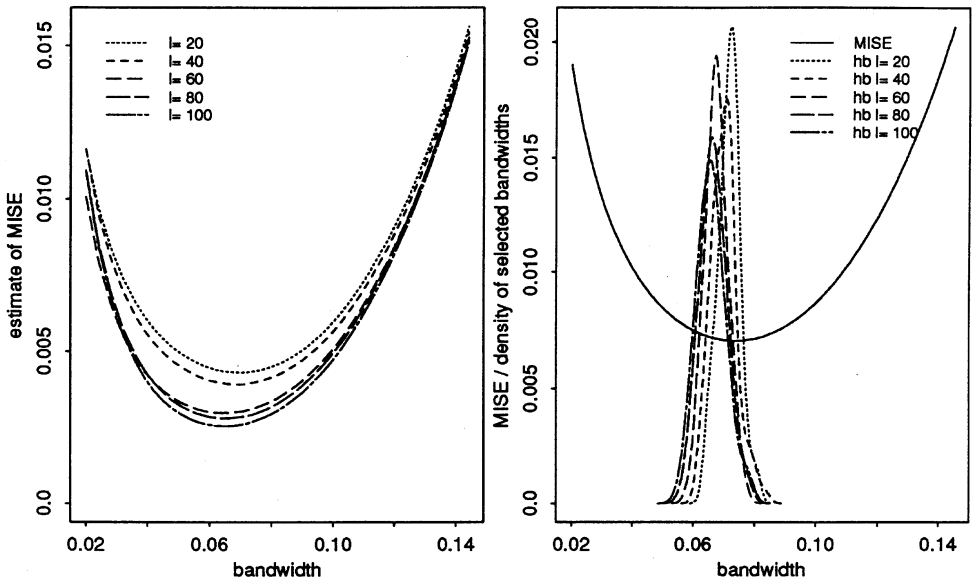

FIG. 1. *Block bootstrap for short-range dependent data. The parameter values* $n = 1600$ *and* $\lambda = 1600$ *were employed, and* $B = 20$ *replications were conducted. The left-hand panel depicts the estimate of MISE averaged over these replications. The right-hand panel illustrates a kernel estimate of the density of the estimated bandwidths. A MISE curve is added to provide information about the mean performance of the selected bandwidths. Curves for* $l = 20, 40, 60, 80$ *and* 100 *are illustrated.*

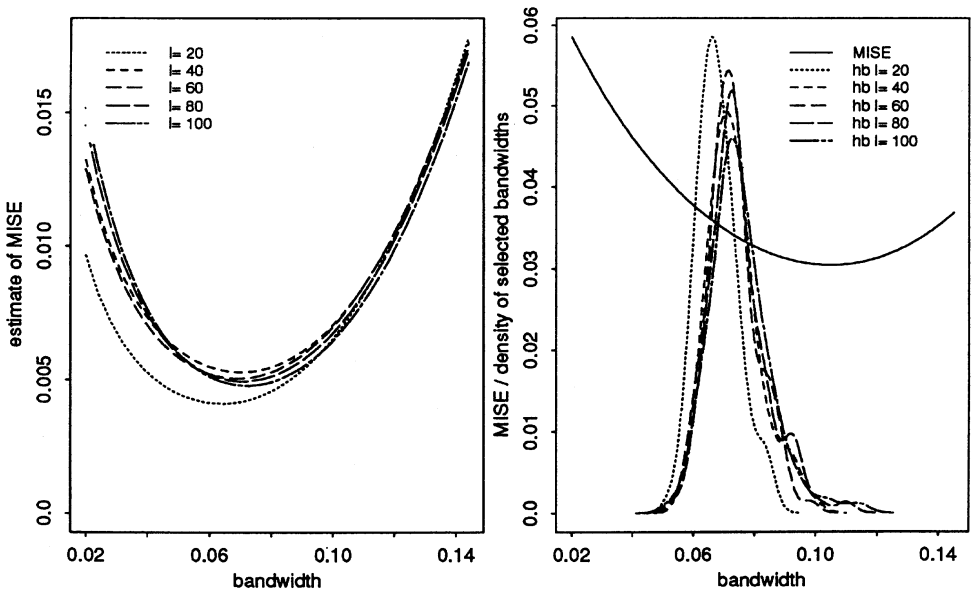

FIG. 2. *Block bootstrap for long-range dependent data; parameter specifications are as for Figure* 1, *except that now* $\lambda = 200$.

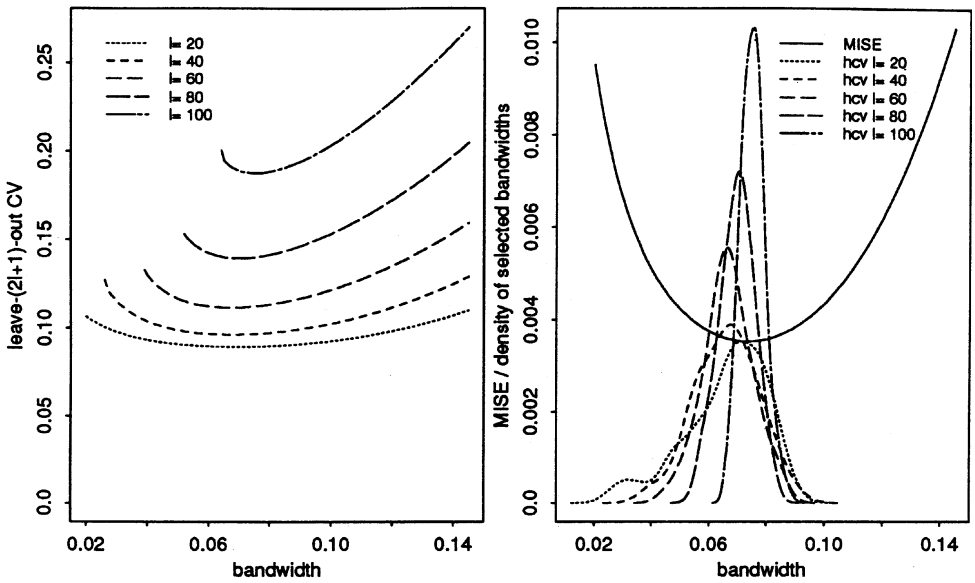

FIG. 3. *Cross-validation for short-range dependent data*; *parameter specifications are as for Figure* 1, *except that now* $B = 200$.

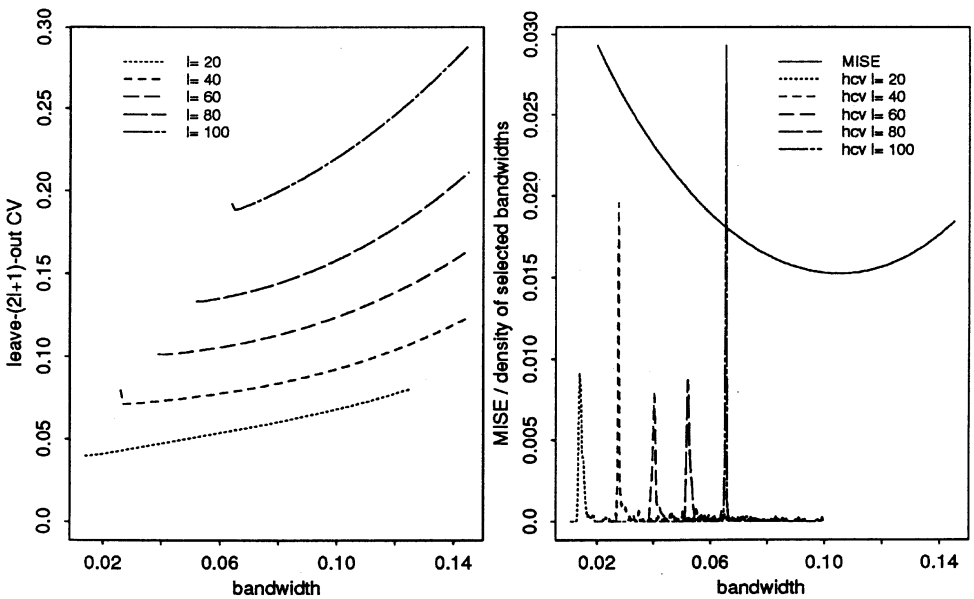

FIG. 4. *Cross-validation for short-range dependent data*; *parameter specifications are as for Figure* 3, *except that now* $\lambda = 200$.

replicates in the case of the block bootstrap because of the computational costs of the method. The right-hand panels of the figures depict estimates of the densities of the distributions of those bandwidths that minimize the estimators of MISE, computed using the same data as the respective left-hand panels (but with 100 samples in the case of the block bootstrap). In Figure 4, cross-validation is selecting the extreme bandwidth $h = (l + 1)/n$ in almost all the simulations. In the case of the block bootstrap and for the values of l treated, $l = 20$ performs well when $\lambda = 1600$, and any l between 40 and 100 seems appropriate when $\lambda = 200$. For cross-validation, $l = 40$ or $l = 60$ is appropriate when $\lambda = 1600$, but the method does not appear to work well for any plausible l when $\lambda = 200$.

4. Proofs. The proofs are given only in outline.

PROOF OF THEOREM 2.1. Mean squared errors are of course made up of squared bias and variance components, of which the first may be derived by arguments identical to those in the case of independent random variables. Indeed, $E\hat{m}(x) - m(x) = h^2\kappa\beta(x) + o(h^2)$ uniformly in $x \in (\delta, 1 - \delta)$ for each $\delta > 0$, and if the weights w_i are given by (2.2), then this result holds for $\delta = 0$. Therefore it suffices to confine attention to deriving that version of Theorem 2.1 which is obtained by replacing $E\{\hat{m}(x) - m(x)\}^2$, β and $o\{(\Lambda h)^{-1} + h^4\}$ by $\operatorname{var} \hat{m}(x)$, 0 and $o\{(\Lambda h)^{-1}\}$, respectively, in each appearance made by the former in the statement of Theorem 2.1. Let the corresponding versions of (2.5)–(2.8) be (2.5′)–(2.8′), say. Derivation of (2.5′) and (2.7′) is not particularly awkward, and so is not discussed further here. Similarly it may be proved that $\operatorname{var} \hat{m}(x) = O\{(\Lambda h)^{-1}\}$ uniformly in $x \in (0, 1)$, and so (2.6′) and (2.8′) follow from (2.5′) and (2.7′), respectively. □

PROOF OF THEOREM 2.2. Let $C, C_1, C_2, \ldots$ denote generic positive constants, not depending on n, h or x. Write E^* and var^* for expectation and variance, respectively, conditional on the data $\mathscr{X}$. Note that

$$\widehat{\operatorname{MSE}}(x) = V_n(x) + \{B_{1n}(x) + B_{2n}(x)\}^2,$$

where $V_n(x) = \sum_{i=1}^{b} \operatorname{var}^*\{\sum_{j=1}^{l} w_{(i-1)l+j}(w)\varepsilon^*_{(i-1)l+j}\}$, $B_{1n}(x) = \sum_{i=1}^{n} w_i(x)\hat{g}_2(x_i) - \hat{g}_2(x)$ and $B_{2n}(x) = \sum_{i=1}^{b}\sum_{j=1}^{l} w_{(i-1)l+j}(x)E^*(\varepsilon^*_j)$. We shall establish Theorem 2.2 by proving that

$$\sup{}^{\dagger}|B_{1n}(x) + B_{2n}(x) - \{E\hat{g}(x) - g(x)\}| = o_p(\Lambda^{-2/5}), \tag{4.1}$$

$$\sup{}^{\dagger}|V_n(x) - \operatorname{var} \hat{g}(x)| = o_p(\Lambda^{-4/5}), \tag{4.2}$$

where $\sup^{\dagger}$ denotes the supremum taken over all $x \in (\delta, 1 - \delta)$ and $h \in H_n$.

Using the Cauchy–Schwarz inequality and the fact that the $\hat{\varepsilon}_i$'s are centred, we may prove that

$$B_{2n}(x)^2 \le C\Bigg(\sum_{i=1}^{b} \sum_{j=1}^{l} N_1^{-2}(nh)^{-1}\Big(\sum\nolimits^{(j)} \varepsilon_k \Big)^2 + \Big[\sum\nolimits^{(\cdot)} \{\hat{g}_1(x_k) - g(x_k)\}^2 \Big] N_1^{-2} l + \big(\bar{\hat{\varepsilon}}_{1\cdot}\big)^2 N_1^{-2} l^2 \Bigg)$$

uniformly in $x \in (\delta, 1-\delta)$ and $h \in H_n$, where $N_1 = N - l + 1$, $\Sigma^{(j)}$ denotes summation over all $k \in \{n_1, n_1+1, \ldots, n_2\} \setminus \{n_1 + j - 1, n_1 + j, \ldots, n_1 + N_1 + j - 2\}$ and $\Sigma^{(\cdot)}$ denotes summation over the union of indices under $\Sigma^{(j)}$ for $1 \le j \le l$. Therefore,

$$E\big\{\sup{}^{\dagger} B_{2n}(x)^2\big\} \le C\big\{(nl)^{-2}\Lambda^{-4/5} + (nh_1)^{-1} l \lambda^{-1}\big\}. \tag{4.3}$$

Let w_{2j} denote the version of w_j which arises when h in the latter is replaced by h_2, and put $w_{3j}(x) = \sum_{i=1}^{n} w_i(x) w_{2i}(x_i) - w_{2j}(x)$. In this notation, $B_{1n}(x) - EB_{1n}(x) = \sum_{j=1}^{n} w_{3j}(x)\varepsilon_j$, $|w_{3j}(x)| \le C(nh_2)^{-1}$ and $\sum_{j=1}^{n} |w_{3j}(x)| \le C$. Hence, applying Corollary A.2 of Hall and Heyde [(1980), page 278], we may deduce that for each integer $r \ge 1$,

$$\begin{aligned} E\{B_{1n}(x) - EB_{1n}(x)\}^{2r} &\le C_1(r) \sum_{m=1}^{2r} \sum_{m,\alpha} \sum_{1 \le j_1 < \cdots < j_m \le n} \Big| E \prod_{p=1}^{m} \big\{ w_{3j_p}(x) \varepsilon_{j_p} \big\}^{\alpha_p} \Big| \\ &\le C_2(r) \Bigg\{ (nh_2)^{-r} + \sum_{m=r+1}^{2r} \sum_{m,\alpha} \sum_{1 \le j_1 < \cdots < j_m \le n} \\ &\qquad \times \prod_{p=1}^{m} |w_{3j_p}(x)|^{\alpha_p} \exp\Big(-C_3 \lambda |x_{j_q} - - x_{j_s}| \Big) \Bigg\}, \end{aligned} \tag{4.4}$$

where $\Sigma_{m,\alpha}$ extends over all $\alpha_1 \ge 1, \ldots, \alpha_m \ge 1$ such that $\alpha_1 + \cdots + \alpha_m = 2r$, and q, s are integers such that $|j_q - j_s| = \max\{|j_p - j_{p-1}| \wedge |j_p - j_{p+1}|\colon 1 \le p \le m,\ p \in \mathscr{L}_m^{(\alpha)}\}$, with $\mathscr{L}_m^{(\alpha)} = \{1 \le p \le m\colon \alpha_p = 1\}$. Note that $\#\mathscr{L}_m^{\alpha} \ge 2(m - r) \ge 2$ for all $m \ge r + 1$. Then by (4.4),

$$\begin{aligned} E\{B_{1n}(x) - EB_{1n}(x)\}^{2r} &\le C_2(r)(nh_2)^{-r} + C_4(r) \sum_{m=r+1}^{2r} (nh_2)^{-r} \\ &\qquad \times \sum_{k=1}^{n-1} k^{m-r} \exp(-C_3 \lambda k/n) \\ &\le C_5(r)\big\{(nh_2)^{-r} + (n/\lambda)(\lambda h_2)^{-r}\big\}. \end{aligned} \tag{4.5}$$

More simply, we may show that

$$\sup_{\substack{x \in (\delta, 1-\delta) \\ h \in H_n}} |EB_{1n}(x) - \{E\hat{g}(x) - g(x)\}| = O\big(h_2^2 h_1^{\alpha}\big) = o(\Lambda^{-2/5}),$$

which in conjunction with (4.3) and (4.5) establishes (4.1).

The remainder of our proof is devoted to deriving (4.2). Write Σ'' for summation over $n_1 \le k \le n_2 - l + 1$, and observe that we may write $V_n = V_{1n} - V_{2n}$, where $V_{1n}(x) = N_1^{-1}\Sigma_{i=1}^{b}\Sigma_k''\{\Sigma_{j=1}^{l} w_{(i-1)l+j}(x)\hat{\varepsilon}_{j+k-1}\}^2$ and $V_{2n}(x) = N_1^{-2}\Sigma_{i=1}^{b}\{\Sigma_k''\Sigma_{j=1}^{l} w_{(i-1)l+j}(x)\hat{\varepsilon}_{k+j-1}\}^2$. Arguments similar to those employed to derive (4.3) may be used to show that $E\{\sup^{\dagger} V_{2n}(x)\} \le C\Lambda^{-4/5}h_1^2$. Also, writing V_{11n} for the random variable obtained by replacing each $\hat{\varepsilon}_i$ in the definition of $V_{1n}(x)$ by the respective ε_i, and putting $V_{12n} = V_{1n} - V_{11n}$ and

$$V_{121n}(x) = N_1^{-1}\sum_{i=1}^{b}\sum_{k}''\left[\sum_{j=1}^{l} w_{(i-1)l+j}(x)\left\{g(x_{j+k-1}) - \hat{g}(x_{j+k-1}) - \bar{\hat{\varepsilon}}_{1\cdot}\right\}\right]^2,$$

we may prove by the Cauchy–Schwarz inequality that $V_{12n} \le V_{121n} + 2(V_{11n}V_{121n})^{1/2}$. Furthermore, it is readily shown that $E\{\sup^{\dagger} V_{121n}(x)\} \le C\Lambda^{-4/5}l(nh)^{-1}$. The desired result (4.2) will follow these bounds if we prove that, for some $\eta > 0$ and all integers $r > 1$,

$$\sup^{\dagger}|EV_{11n}(x) - \operatorname{var}\hat{g}(x)| = o(\Lambda^{-4/5}), \tag{4.6}$$

$$\sup^{\dagger} E\{V_{11n}(x) - EV_{11n}(x)\}^{2r} = O(\Lambda^{-8r/5}n^{-\eta r}). \tag{4.7}$$

To derive (4.6), observe that

$$\begin{aligned}
&|EV_{11n}(x) - \operatorname{var}\hat{g}(x)| \\
&\quad\le 2\left|\sum_{j_1=1}^{l-1}\gamma(\lambda j_1/n)\left\{\sum_{j_2=1}^{l-j_1}\sum_{i=1}^{b} w_{(i-1)l+j_2}(x)w_{(i-1)l+j_1+j_2}(x)\right.\right. \\
&\qquad\qquad\left.\left. - \sum_{j_2=1}^{n-j_1} w_{j_2}(x)w_{j_1+j_2}(x)\right\}\right| \\
&\qquad + 2\left|\sum_{j_1=l}^{n-1}\gamma(\lambda j_1/n)\sum_{j_2=1}^{n-j_1} w_{j2}(x)w_{j_1+j_2}(x)\right|.
\end{aligned}$$

The fact that $\lambda l/n \to \infty$ may be used to prove that the second term on the right-hand side equals $o(\Lambda^{-4/5})$. To bound the first term, let $l_1 = l_1(n)$ denote integers such that $l_1/l \to 0$ and $\lambda l_1/n \to \infty$. Using the compactness of the support of K we may show that the first term is bounded above by

$$\begin{aligned}
&2\sum_{j_1=1}^{l_1}|\gamma(\lambda j_1/n)|\sum_{j_1=l-l_1}^{l}\sum_{i=1}^{b}|w_{(i-1)l+j_2}(x)w_{(i-1)l+j_1+j_2}(x)| \\
&\qquad + 2\sum_{j_1=l_1}^{l}|\gamma(\lambda j_1/n)|\sum_{i=1}^{n} w_i(x)^2 \\
&\qquad \le C_1(nh_1)^{-1}\{(l_1/l) + \exp(-C_2\lambda l_1/n)\} = o(\Lambda^{-4/5}),
\end{aligned}$$

uniformly in $x \in (\delta, 1-\delta)$ and $h \in H_n$, completing the proof of (4.6).

To prove (4.7), put $w(x; j_1, j_2) = a_{j1}\sum_{i=1}^{b} w_{(i-1)l+j_2}(x)w_{(i-1)l+j_1+j_2}(x)$, where $a_{j_1} = 1$ or 2 according as $j_1 = 0$ or $j_1 \geq 1$, and put

$$W_k(x) = \sum_{j_1=0}^{l-1} \sum_{j_2=1}^{l-j_1} w(x; j_1, j_2)\varepsilon_{k+j_2-1}\varepsilon_{k+j_1+j_2-1}.$$

In this notation, $V_{11n} = N_1^{-1}\sum_k'' W_k$. Write $M - 1$ for the integer part of $N_1/8l$. If $1 \leq m \leq 2M$, let B_m denote the sum of W_k over $4l(m-1) < k \leq (4lm) \wedge (n_1 + N_1)$. Using an argument similar to that employed to derive (4.5), we may show that, for each integer $r \geq 1$,

$$\begin{aligned} E(B_m^{2r}) &\leq \sum_{k=1}^{4r} \sum_{k}^{\#} \left| \prod_{p=1}^{2r} \sum_{j_2=1}^{l-j_{1p}} w(x; j_{1p}, j_2) E\left(\prod_{p=1}^{2r} \varepsilon_{k_{1p}}\varepsilon_{j_{1p}} \right) \right| \\ &\leq C_1(r)(nh_1)^{-2r}\left[l^{2r} + \sum_{k=2r+1}^{4r} l^{2r}\left\{ \sum_{j=1}^{l-1} j^{m-2r}\exp(-C\lambda j/n) \right\} \right] \\ &\leq C_2(r)(1 + \lambda^{-1}n)\zeta^{2r} \end{aligned}$$

uniformly in $x \in (\delta, 1-\delta)$ and $h \in H_n$, where $\sum_k^{\#}$ denotes summation over integers $j_{11}, \ldots, j_{1,2r}$, $k_{11}, \ldots, k_{1,2r}$ satisfying $0 \leq j_{1p} \leq l-1$ and $4l(m-1) < k_{1p} \leq 4lm$ and such that there are precisely k distinct indices among them, and where $\zeta = l\{(nh_1)^{-1} + (\lambda h_1)^{-1}\}$. For any $1 \leq m$, $m + k \leq M$, the minimum separation between the indices of ε_i's in B_{2m} and $B_{2(m+k)}$, respectively, is $(2k-1)4l - l + 1$. Hence, by Corollary A.2 of Hall and Heyde [(1980), page 278] and parallel to the argument used earlier to derive (4.5), we may show that, for each integer $r \geq 1$,

$$\begin{aligned} &E\{V_{11n}(x) - EV_{11n}(x)\}^{2r} \\ &\quad \leq C_1(r, \eta')N_1^{-2r}(1 + \lambda^{-1}n) \\ &\qquad \times \left[M^r\zeta^{4r} + \left\{ (\zeta^4)^{(1+\eta')/4}(\zeta^{4(2r-1)})^{1/4} \right\}^{(2+\eta')/(2+2\eta')} \right] \\ &\qquad \times \left[\sum_{p=r+1}^{2r} M^r \sum_{j=1}^{M/2} j^{k-r}\exp\{-C_2(\eta')\lambda lj/n\} \right] \\ &\quad \leq C_3(r, \eta')(1 + \lambda^{-1}n)\zeta^{(1-\eta)2r}(nl)^{-r}\left[1 + \int_{\lambda l/n}^{\infty} y^r\exp\{-C_2(\eta')y\}\,dy \right] \\ &\quad \leq C_4(r, \eta')(1 + \lambda^{-1}n)\Lambda^{-8r/5}(l/n)^r\zeta^{2r\eta} \end{aligned}$$

uniformly in $x \in (\delta, 1-\delta)$ and $h \in H_n$ for each $\eta' > 0$, where $\eta = \eta(\eta') \downarrow 0$ as $\eta' \downarrow 0$. This proves (4.7). □

The proof of Theorem 2.3 is similar to that of Theorem 2.2 and so is not given here.

Acknowledgments. The authors would like to thank the referees and the Associate Editor for constructive criticisms that improved the earlier version of the paper.

REFERENCES

ALTMAN, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85** 749–759.

CARLSTEIN, E. (1986). The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *Ann. Statist.* **14** 1171–1179.

CHIU, S.-T. (1989). Bandwidth selection for kernel estimation with correlated noise. *Statist. Probab. Lett.* **8** 347–354.

CHU, C. K. and MARRON, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.* **19** 1906–1918.

DIGGLE, P. J. and HUTCHINSON, M. F. (1989). On spline smoothing with autocorrelated errors. *Austral. J. Statist.* **31** 166–182.

FAN, J. (1993). Local linear smoothers and their minimax efficiency. *Ann. Statist.* **21** 196–216.

FARAWAY, J. J. and JHUN, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85** 1119–1122.

GASSER, T and MÜLLER, H. G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation Lecture Notes in Math.* **757** 23–68. Springer, Berlin.

GASSER, T., ENGLE, J. and SEIFERT, B. (1993). Nonparametric function estimation. In *Handbook of Statistics* (C. R. Rao, ed.) **9** 423–465. North-Holland, Amsterdam.

GYÖRFI, L., HÄRDLE, W., SARDA, P. and VIEU, P. (1989). *Nonparametric Curve Estimation from Time Series. Lecture Notes in Statist.* **60**. Springer, New York.

HALL, P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl.* **20** 231–246.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HÄRDLE, W. and VIEU, P. (1992). Kernel regression smoothing of time series. *J. Time Ser. Anal.* **13** 209–232.

HART, J. D. (1987). Kernel smoothing when the observations are correlated. Technical report, Dept. Statistics, Texas A & M Univ.

HART, J. D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc. Ser. B* **53** 173–187.

HART, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *J. Roy. Statist. Soc. Ser. B* **56** 529–542.

HART, J. D. and VIEU, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.* **18** 873–890.

HART, J. D. and WEHRLY, T. E. (1986). Kernel regression using repeated measurements data. *J. Amer. Statist. Assoc.* **81** 1080–1088.

HERMANN, E., GASSER, T. and KNEIP, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika* **79** 783–795.

KOHN, R., ANSLEY, C. G. and WONG, C.-M. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika* **79** 335–346.

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241.

TAYLOR, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **76** 705–712.

WOOD, A. T. A. and CHAN, G. (1993). Simulation of stationary Gaussian processes in $[0, 1]^d$. Research Report CMA-SR06-93, Centre for Mathematics and its Applications, Australian National Univ.

P. HALL
CENTRE FOR MATHEMATICS AND
ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
G.P.O. BOX 4
CANBERRA, ACT 0200
AUSTRALIA

S. N. LAHIRI
DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
AMES, IOWA 80011

J. POLZEHL
KONRAD-ZUSE-ZENTRUM FÜR
INFORMATIONSTECHNIK
HEILBRONNES STRASSE 10
BERLIN D-10711
GERMANY