Ann. Inst. Statist. Math.
Vol. 54, No. 2, 291–311 (2002)
©2002 The Institute of Statistical Mathematics

LOCAL POLYNOMIAL FITTING WITH LONG-MEMORY, SHORT-MEMORY AND ANTIPERSISTENT ERRORS

JAN BERAN AND YUANHUA FENG

Department of Mathematics and Statistics, University of Konstanz, Universitätsstr. 10, Postfach 5560, D-78457 Konstanz, Germany,
e-mail: jan.beran@uni-konstanz.de; yuanhua.feng@uni-konstanz.de

(Received May 12, 2000; revised October 30, 2000)

Abstract. Local polynomial smoothing for the trend function and its derivatives in nonparametric regression with long-memory, short-memory and antipersistent errors is considered. We show that in the case of antipersistence, the convergence rate of a nonparametric regression estimator is faster than for uncorrelated or short-range dependent errors. Moreover, it is shown that unified asymptotic formulas for the optimal bandwidth and the MSE hold for all of the three dependence structures. Also, results on estimation at the boundary are included. A bandwidth selector for nonparametric regression with different types of dependent errors is proposed. Its asymptotic property is investigated. The practical performance of the proposal is illustrated by simulated and real data examples.

Key words and phrases: Antipersistence, long-range dependence, local polynomial fitting, nonparametric regression, bandwidth selection.

1. Introduction

In this paper local polynomial fitting with long-range dependent, short-range dependent and antipersistent errors is investigated. A data-driven procedure for the practical use of this estimation is also proposed. Here, a stationary process X_i with autocovariances $\gamma(k) = \operatorname{cov}(X_i, X_{i+k})$ is said to have long-range dependence (or long memory), if the spectral density $f(\lambda) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \exp(ik\lambda)\gamma(k)$ has a pole at the origin of the form

$$f(\lambda) \sim c_f|\lambda|^{-\alpha} \qquad (\text{as } \lambda \to 0) \tag{1.1}$$

for constant $c_f > 0$ and $\alpha \in (0, 1)$, where "$\sim$" means that the ratio of the left and the right hand sides converges to one (see e.g. Mandelbrot (1983), Cox (1984), Künsch (1986), Hampel (1987) and Beran (1994) and references therein). In particular, this implies that, as $k \to \infty$, the autocovariances $\gamma(k)$ are proportional to $k^{\alpha-1}$ and hence their sum is infinite. If (1.1) holds with $\alpha = 0$, then the sum of all autocovariances is a positive constant and the process X_t is said to have short-memory. On the other hand, a stationary process is called antipersistent, if (1.1) holds for $\alpha \in (-1, 0)$. This implies that the sum of all autocovariances is zero.

The most important motivation to introduce the concept of antipersistence is to quantify the phenomenon of overdifferencing. That is, if a stochastic process is slightly nonstationary lying between a stationary long-memory process and a random walk, then its first difference will have antipersistence. There are some reports about antipersistent phenomena in practice. For instance, Beran *et al.* (1999) found that the first difference of the daily world copper price from Jan. 2, 1997 to Sep. 2, 1998 is significantly antipersistent together with significantly nonzero mean function. This means that the original series is the sum of a trend function and a nonstationary error process whose first difference is antipersistent. Karuppiah and Los (2000) pointed out, that the first difference of many intraday foreign exchange series, in particular those between Japanese Yen and US Dollar and German Mark and US Dollar, are clearly antipersistent.

Nonparametric regression with short- or long-range dependent errors has gained increasing attention in the literature. Kernel estimators are proposed for observations with short-range dependent errors (see e.g. Altman (1990), Herrmann *et al.* (1992)) and for observations with long-range dependent errors (see e.g. Hall and Hart (1990), Csörgö and Mielniczuk (1995) and Beran (1999)). Wang (1996) proposed nonparametric regression for data with long-range dependent errors via wavelet shrinkage. In this paper, the local polynomial fit introduced by Stone (1977) and Cleveland (1979) will be adapted to nonparametric regression with short- or long-range dependent as well as antipersistent errors. This approach is known to be an automatic kernel method (see Müller (1987) and Hastie and Loader (1993)) having many exciting statistical properties. For recent developments in this context see, e.g. Ruppert and Wand (1994), Wand and Jones (1995) and Fan and Gijbels (1995, 1996). Local polynomial fitting is adapted to the autoregressive context for modelling nonlinear time series under some mixing conditions e.g. by Masry (1996), Masry and Fan (1997) and Feng and Heiler (1998).

A crucial problem in the practical use of nonparametric regression is the data-driven bandwidth selection. Recent proposals for bandwidth selection in nonparametric regression with independent or short-range dependent errors may be found in, e.g. Chiu (1989), Gasser *et al.* (1991), Härdle *et al.* (1992), Herrmann *et al.* (1992), Fan and Gijbels (1995), Ruppert *et al.* (1995), Heiler and Feng (1998) and Feng and Heiler (1999). A bandwidth selector for nonparametric regression with long- memory errors is proposed by Ray and Tsay (1997). The proposals in Herrmann *et al.* (1992) and Ray and Tsay (1997) are adaptations of the iterative plug-in idea of Gasser *et al.* (1991).

The contributions of our paper are:

1. Asymptotic formulas for local polynomial estimators for the trend function and its derivatives are obtained under short memory, long memory and antipersistence. To our knowledge, *local polynomial fitting* and nonparametric estimation of *derivatives* has not been investigated in the literature for the cases of *long memory* and *antipersistence*. Note that antipersistence implies that the autocorrelations sum up to zero so that another technique has to be used than in the case of long-range dependence. We show, in particular, that *unified formulas* can be obtained that are valid for all three cases (short memory, long memory and antipersistence). Note that these three cases correspond to three very different types of dependence structures.

2. In contrast to previous literature on nonparametric regression with long memory, results are obtained not only for interior points but also for *boundary points*. The results hold for kernel estimators (with boundary correction) as well as local polynomial fits.

3. A unified iterative plug-in bandwidth selector is proposed for nonparametric regression with all of the three types of dependent structures with a so-called exponential inflation method (see Beran (1999) and Beran and Ocker (1999, 2001)). The asymptotic behaviour is briefly investigated. This proposal is applied to some simulated and practical examples. An extended simulation study will be given in a forthcoming paper (Beran and Feng (1999)).

The paper is organized as follows. The model and the local polynomial estimators for the mean function as well as its derivatives are introduced in Section 2. The asymptotic results are given in Section 3. The bandwidth selector is proposed in Section 4. Examples in Section 5 illustrate the practical use of the proposal. Section 6 contains some final remarks. Proofs of theorems are listed in the Appendix.

2. The model and the estimators

Consider the equidistant design nonparametric regression model

$$Y_i = g(t_i) + X_i, \tag{2.1}$$

where $t_i = (i - 0.5)/n$, $g : [0, 1] \to \mathbb{R}$ is a smooth function and X_i is a stationary process having the form

$$(1 - B)^{\delta} X_i = U_i, \tag{2.2}$$

where $\delta \in (-0.5, 0.5)$, B denotes the backshift operator such that $BY_i = Y_{i-1}$ and U_i is a stationary process with short memory so that $\sum_{k=-\infty}^{\infty} \mathrm{cov}(U_i, U_{i+k}) = C$ with $0 < C < \infty$. The parameter δ is called the fractional differencing parameter. The fractional difference $(1 - B)^{\delta}$, introduced by Granger and Joyeux (1980) and Hosking (1981), is defined by

$$(1 - B)^{\delta} = \sum_{k=0}^{\infty} b_k(\delta) B^k \tag{2.3}$$

with

$$b_k(\delta) = (-1)^k \frac{\Gamma(\delta + 1)}{\Gamma(k + 1)\Gamma(\delta - k + 1)}. \tag{2.4}$$

The spectral density of X_i in (2.1) has the form (1.1) with $\alpha = 2\delta$. Hence, X_i has long memory if $\delta > 0$. If $\delta = 0$, $X_i = U_i$ has short memory. And X_i is antipersistent, if $\delta < 0$. Model (2.1) is a special case of the SEMIFAR (semiparametric fractional autoregressive) model introduced by Beran (1999). See also Beran and Ocker (1999, 2001). In what follows we will consider local polynomial estimation of $g^{(\nu)}$, the ν-th derivative of g. The theorems in the next section give formulas for local polynomial estimators that are valid for the whole range $\delta \in (-0.5, 0.5)$.

Assume that g is at least $(p + 1)$-times differentiable at a point t_0. Then $g(t)$ can be approximated locally by a polynomial of order p:

$$g(t) = g(t_0) + g'(t_0)(t - t_0) + \cdots + g^{(p)}(t_0)(t - t_0)^p/p! + R_p \tag{2.5}$$

for t in a neighbourhood of t_0, where R_p is a remainder term. Let K be a symmetric density (a kernel of order two without boundary correction) having compact support $[-1, 1]$. Giving n observations $Y_1, \ldots, Y_n$, we can obtain an estimator of $g^{(\nu)}$ $(\nu \le p)$ by solving the locally weighted least squares problem

$$Q = \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (t_i - t_0)^j \right\}^2 K\left(\frac{t_i - t_0}{h}\right) \Rightarrow \min, \tag{2.6}$$

where h is the bandwidth and K is called the weight function. Let $\hat\beta = (\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_p)^{\mathrm{T}}$, then it is clear from (2.5) that $\nu!\hat\beta_\nu$ estimates $g^{(\nu)}(t_0)$, $\nu = 0, 1, \ldots, p$. Let

$$\boldsymbol{X} = \begin{bmatrix} 1 & t_1 - t_0 & \cdots & (t_1 - t_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n - t_0 & \cdots & (t_n - t_0)^p \end{bmatrix}.$$

Let $\boldsymbol{e}_j$, $j = 1, \ldots, p+1$, denote the j-th $(p{+}1)\times 1$ unit vector, and let $\boldsymbol{K}$ denote the diagonal matrix with

$$k_i = K\left(\frac{t_i - t_0}{h}\right)$$

as its i-th diagonal entry. Finally, let $\boldsymbol{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Then $\hat g^{(\nu)}(t_0)$ can be written as

$$\begin{aligned} \hat g^{(\nu)}(t_0) &= \nu!\boldsymbol{e}_{\nu+1}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} \\ &=: \{\boldsymbol{w}^\nu(t_0)\}^{\mathrm{T}}\boldsymbol{y}, \end{aligned} \tag{2.7}$$

where $\{\boldsymbol{w}^\nu(t_0)\}^{\mathrm{T}} = \nu!\boldsymbol{e}_{\nu+1}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{K}$ is called the weighting system. We see that $\hat g^{(\nu)}(t_0)$ is a linear smoother with the weighting system $\boldsymbol{w}^\nu(t_0) = (w_1^\nu, \ldots, w_n^\nu)^{\mathrm{T}}$, where $w_i^\nu \neq 0$ only if $|t_i - t_0| \leq h$. The weighting system does not depend on the dependence structure of the errors. For any interior point $t_0 \in [h, 1-h]$ the non-zero part of $\boldsymbol{w}^\nu(t_0)$ is the same, i.e. $\hat g^{(\nu)}$ works as a moving average in the interior. Furthermore, $\boldsymbol{w}^\nu(t_0)$ satisfies:

$$\sum_{i=1}^n w_i^\nu (t_i - t_0)^\nu = \nu! \quad \text{and} \quad \sum_{i=1}^n w_i^\nu (t_i - t_0)^j = 0 \quad \text{for } j = 0, \ldots, p,\ j \neq \nu. \tag{2.8}$$

It is the property (2.8) that makes $\hat g^{(\nu)}$ exactly unbiased if g is a polynomial of order not larger than p. The property (2.8) also shows the main difference between a local polynomial estimator and a kernel estimator, since for a kernel estimator property (2.8) only holds approximately.

3. Asymptotic results

3.1 *Assumptions*

In this section we discuss the asymptotic properties of the estimators proposed in the last section. It is well known that a local polynomial estimator is asymptotically equivalent to a certain kernel estimator, called an (asymptotically) equivalent kernel estimator of the local polynomial estimator. Hence, the asymptotic properties of a local polynomial estimator are the same as those of the equivalent kernel estimator. Although it is shown in Ruppert and Wand (1994) that one can analyze local polynomial fitting directly as a weighted least squares estimator rather than as an approximate kernel estimator, the asymptotic results given in this section will be proved by means of the asymptotically equivalent kernel function.

From here on we will mainly consider the case with $p-\nu$ odd. This implies $p-\nu \geq 1$. Let $k = p+1$. We have $k \geq \nu+2$ and $k-\nu$ even. A brief discussion of the case with $p-\nu$ even will be given latter. For the asymptotic results given below we need the following assumptions:

A1. The weight function $K(u)$ is a symmetric density (i.e. a kernel of order two) with compact support $[-1,1]$ having the polynomial form

$$K(u)=\sum_{l=0}^{r}\alpha_l u^{2l}\mathbb{I}_{[-1,1]}(u)$$

(see e.g. Gasser and Müller (1979)).

A2. g is an at least k times continuously differentiable function on $[0,1]$ with $k\geq \nu+2$ and $k-\nu$ even.

A3. The bandwidth satisfies: $h\rightarrow 0$, $(nh)^{1-2\delta}h^{2\nu}\rightarrow\infty$ as $n\rightarrow\infty$.

A4. A local polynomial fit of order $p=k-1$ is used.

Under the assumptions A1 and A4 it can be shown that $\hat{g}^{(\nu)}$ converges to $g^{(\nu)}$ at the same rate in the interior as well as at the boundary.

3.2 *Asymptotic bias and variance*

In the following, formulas for asymptotic bias and variance of $\hat{g}^{(\nu)}$ will be given at a point t in the interior ($t\in[h,1-h]$), at the left boundary ($t\in[0,h)$) as well as at the right boundary ($t\in(1-h,1]$). The discussion will only be carried out for $t\in[h,1-h]$ and the left boundary $t\in[0,h)$. The formula for the right boundary $(1-h,1]$ is analogous to the left boundary. Note that asymptotically any fixed point $t\in(0,1)$ will not be a boundary point, since $h\rightarrow 0$ as $n\rightarrow\infty$. A standard definition of a left boundary point is $t=ch$ with $0\leq c<1$. For each $t\in[h,1-h]$ we define $c=1$. For each $c\in[0,1]$, the truncated kernel $K_c(u)$ is defined by

$$K_c(u)=\left(\int_{-c}^{1}K(x)dx\right)^{-1}K(u)\mathbb{I}_{[-c,1]}(u). \tag{3.1}$$

Two special cases of $K_c(u)$ are $K_1(u)=K(u)$ for the estimation at an interior point and $K_0(u)=2K(u)\mathbb{I}_{[0,1]}$ for the estimation at the left end point. Let $\mu_{j,c}=\int_{-c}^{1}u^jK_c(u)du$ be the j-th moment of K_c. And define

$$\boldsymbol{N}_{p,c}=\begin{bmatrix}1 & \mu_{1,c} & \cdots & \mu_{p,c}\\ \mu_{1,c} & \mu_{2,c} & \cdots & \mu_{p+1,c}\\ \vdots & \vdots & \ddots & \vdots\\ \mu_{p,c} & \mu_{p+1,c} & \cdots & \mu_{2p,c}\end{bmatrix}. \tag{3.2}$$

Denote by $\mu_j=\mu_{j,1}=\int_{-1}^{1}u^jK(u)du$ the j-th moment of $K(u)$. Let $\boldsymbol{N}_p=\boldsymbol{N}_{p,1}$ is the matrix as defined in (3.2) with $\mu_{j,c}$ being replaced by μ_j. An important difference between $\boldsymbol{N}_{p,c}$ and $\boldsymbol{N}_p$ is that $\mu_j=0$ for j odd, but in general, any $\mu_{j,c}$ is non-zero, if $c\neq 1$. A k-th order kernel $K_{(\nu,k,c)}$ can be defined as follows. For $i,j=1,\dots,k$, let $(\alpha_{i,j,c})=\boldsymbol{N}_{p,c}^{-1}$ and define

$$K_{(\nu,k,c)}(u)=\nu!Q_{(\nu,k,c)}(u)K_c(u), \tag{3.3}$$

where

$$Q_{(\nu,k,c)}(u)=\sum_{j=1}^{k}\alpha_{\nu+1,j,c}u^{(j-1)}.$$

It is easily established that the function defined in (3.3) satisfies

$$\int_{-c}^{1} u^j K_{(\nu,k,c)}(u)du = \begin{cases} 0, & j = 0,\dots,\nu-1,\nu+1,\dots,k-1, \\ \nu!, & j=\nu, \\ \beta_{(\nu,k,c)}, & j=k, \end{cases} \tag{3.4}$$

where $\beta_{(\nu,k,c)}$ is a non-zero constant. Hence, $K_{(\nu,k,c)}$ is a k-th order kernel (for $c=1$) or boundary kernel (for $c<1$) for estimating the ν-th derivative, which we will call an "equivalent kernel". Denote $K_{(\nu,k,1)}$ by $K_{(\nu,k)}$. $K_{(\nu,k)}$ is the same as defined by Gasser *et al.* (1985) up to a $(-1)^\nu$ sign. Like K, $K_{(\nu,k,c)}$ is a polynomial kernel, too.

Let $\tilde{\boldsymbol{w}}_c^\nu(t) = (\tilde{w}_{1,c}^\nu,\dots,\tilde{w}_{n,c}^\nu)^{\mathrm{T}}$ denote the weighting system of a kernel estimator with $K_{(\nu,k,c)}$. Then we obtain a kernel estimator of $g^{(\nu)}$:

$$\tilde{g}^{(\nu)}(t) = \sum_{i=1}^{n} \tilde{w}_{i,c}^\nu Y_i, \tag{3.5}$$

where

$$\tilde{w}_{i,c}^\nu = \frac{1}{nh^{\nu+1}} K_{(\nu,k,c)}\left(\frac{t_i - t}{h}\right). \tag{3.6}$$

$\tilde{g}(\nu)$ is a kernel estimator with boundary correction. Observe that, here $K_{(\nu,k,c)}$ changes over $c\in[0,1]$. In the case of equidistant design the definition (3.5) is asymptotically equivalent to that given by Gasser and Müller (1984), provided that corresponding boundary kernels are used. For $\hat{g}^{(\nu)}$ and $\tilde{g}^{(\mu)}$ we have:

LEMMA 1. *Under the assumptions* A1 *to* A4 *the p-th order local polynomial estimator $\hat{g}^{(\nu)}$ at any point $t\in[0,1]$ is asymptotically equivalent to the k-th order kernel estimator $\tilde{g}^{(\nu)}$ defined in* (3.5) *in the sense that*

$$\lim_{n\to\infty} \sup_{1\le i\le n} \left| \frac{w_i^\nu}{\tilde{w}_{i,c}^\nu} - 1 \right| = 0, \quad \text{defining } \frac{0}{0} = 1. \tag{3.7}$$

Lemma 1 is given by Müller (1987) for an interior point. This result is extended to the boundary area by Feng (1999). Lemma 1 shows that a local polynomial estimator with $p-\nu$ odd has automatic boundary correction.

Define $x_i = (t_i - t)/h$, $y_j = (t_j - t)/h$. Let $n_0 = [nt]+1$, $b = [nh]$, $b_c = [nch]$, where $[\cdot]$ denotes the integer part. Using the notation

$$V_n(h,c) = (nh)^{-1-2\delta} \sum_{i,j=n_0-b_c}^{n_0+b} K_{(\nu,p,c)}(x_i) K_{(\nu,p,c)}(y_j)\gamma(i-j), \tag{3.8}$$

we obtain

THEOREM 1. *Let the assumptions* A1 *to* A4 *hold. Then for $\delta\in(-0.5,0.5)$, we have*

i) *Bias*:

$$E[\hat{g}^{(\nu)} - g^{(\nu)}] = h^{(k-\nu)} \frac{g^{(k)}(t)\beta_{(\nu,k,c)}}{k!} + o(h^{(k-\nu)}); \tag{3.9}$$

ii):

$$\lim_{n\to\infty} V_n(h,c) = V(c), \tag{3.10}$$

where $0 < V(c) < \infty$ *is a constant*;

iii) *Variance*:

$$(nh)^{1-2\delta} h^{2\nu} \operatorname{var}[\hat{g}^{(\nu)}(t)] = V(c) + o(1). \tag{3.11}$$

For $p-\nu$ even, let $k = p+2$, if t is an interior point, and $k = p+1$, if t is a boundary point. Then Theorem 1 also holds. This means that the bias term is of a lower order at a boundary point than at an interior point.

Remark 1. For bandwidth selection with the plug-in method one has to calculate the value of $V := V(1)$, i.e. $V(c)$ with $c = 1$ in the interior. Simple explicit formulas for V can be given as follows:

$$V = 2\pi c_f \int_{-1}^{1} K^2_{(\nu,k)}(x)dx \tag{3.12}$$

for $\delta = 0$ and

$$V = 2c_f\Gamma(1-2\delta)\sin\pi\delta \int_{-1}^{1}\int_{-1}^{1} K_{(\nu,k)}(x)K_{(\nu,k)}(y)|x-y|^{(2\delta-1)}dxdy \tag{3.13}$$

for $\delta > 0$ (Hall and Hart (1990)), where $c_f = (2\pi)^{-1}C$. The explicit form of V for $\delta < 0$ is more complex, since the integral $\int_{-1}^{1} K_{(\nu,k)}(y)|x-y|^{2\delta-1}dy$ does not exist. However, at any point x the kernel $K_{(\nu,k)}(y)$ may be written as $K_{(\nu,k)}(y) = \sum_{l=0}^{r} \beta_l(x)(x-y)^l =: K_0(x)+K_1(x-y)$, where r is an integer, $K_0(x) = \beta_0(x)$ and $K_1(x-y) = \sum_{l=1}^{r} \beta_l(x)(x-y)^l$. Note that, in the case of antipersistence we have $\sum_{k=-\infty}^{\infty} \gamma(k) = 0$. For $\delta < 0$,

$$\begin{aligned} V &= 2c_f\Gamma(1-2\delta)\sin(\pi\delta)\int_{-1}^{1} K_{(\nu,k)}(x) \\ &\quad \times \left\{\int_{-1}^{1} K_1(x-y)|x-y|^{2\delta-1}dy - K_0(x)\int_{|y|>1}|x-y|^{2\delta-1}dy\right\}dx. \end{aligned} \tag{3.14}$$

If g is estimated by a local linear fit with the uniform kernel as the weight function, then we have, in the interior, $K_{(0,1)}(x) = K(x) = \mathbb{1}_{\{|x|\le 1\}}/2$. In this case we have $K_0(x) = \mathbb{1}_{\{|x|\le 1\}}/2$ and $K_1 \equiv 0$. The formulas (3.12), (3.13) and (3.14) give the same result

$$V = \frac{2^{2\delta} c_f \Gamma(1-2\delta)\sin(\pi\delta)}{\delta(2\delta+1)} \tag{3.15}$$

with $V(0) = \lim_{\delta\to 0} V(\delta) = \pi c_f$ (see corollary 1 in Beran 1999).

3.3 *The MISE*

Let

$$I(g^{(k)}) = \int_0^1 [g^{(k)}(t)]^2 dt. \tag{3.16}$$

Denote by $\beta_{(\nu,k)}$ the kernel constant of $K_{(\nu,k)}$. The following result holds:

THEOREM 2. *Under the assumptions* A1 *to* A4 *and for* $\delta \in (-0.5, 0.5)$, *we have*

i) *The MISE (mean integrated squared error) of* $\hat{g}^{(\nu)}$ *is given by*

$$\begin{aligned}(3.17)\qquad &\int_0^1 E\{[\hat{g}^{(\nu)}(t)-g^{(\nu)}(t)]^2\}dt\\ &=\mathrm{MISE}_{\mathrm{asympt}}(n,h)+o(\max(h^{2(k-\nu)},(nh)^{2\delta-1}h^{-2\nu}))\\ &=h^{2(k-\nu)}\frac{I(g^{(k)})\beta^2_{(\nu,k)}}{k!}+(nh)^{2\delta-1}h^{-2\nu}V\\ &\quad+o(\max(h^{2(k-\nu)},(nh)^{2\delta-1}h^{-2\nu}));\end{aligned}$$

ii) *The asymptotically optimal bandwidth that minimizes the asymptotic MISE is given by*

$$(3.18)\qquad h_{\mathrm{A}}=C_{\mathrm{opt}}n^{(2\delta-1)/(2k+1-2\delta)},$$

where

$$(3.19)\qquad C_{\mathrm{opt}}=\left[\frac{2\nu+1-2\delta}{2(k-\nu)}\frac{[k!]^2V}{I(g^{(k)})\beta^2_{(\nu,k)}}\right]^{1/(2k+1-2\delta)},$$

where it is assumed that $I(g^{(k)})>0$.

Note that by inserting h_{A} in (3.17), Theorem 2 implies that the optimal MISE is of the order

$$(3.20)\qquad \int_0^1 E\{[\hat{g}^{(\nu)}(t)-g^{(\nu)}(t)]^2\}dt=O(n^{2(2\delta-1)(k-\nu)/(2k+1-2\delta)}).$$

The following remarks clarify the results given above.

Remark 2. Theorem 2 does not hold for $p-\nu$ even. For $p-\nu$ even, the integrated squared bias over the interior is of the order $O(h^{2(p+2-\nu)})$, but that over [0, 1] is of the order $O(h^{2(p+1-\nu)+1})$. That is, the bias in the boundary region dominates the bias in the interior and it will cause a slower convergence rate of the MISE. This is the so- called boundary effect for nonparametric regression estimators. In particular, for a kernel estimator for g, i.e. a local polynomial one with $p=0$ and $\nu=0$, the integrated squared bias over the interior is of the order $O(h^4)$, and that over [0, 1] is of the order $O(h^3)$ (see Gasser and Müller (1979)). For $p-\nu$ odd, the integrated squared bias over the interior and over [0, 1] is of the same order $O(h^{2(k-\nu)})$. In this case we have $\mathrm{MISE}([0,1])=\mathrm{MISE}([h,1-h])(1+O(h))$. Hence, in local polynomial fitting $p-\nu$ is often taken to be odd in order that the MISE can be calculated over the whole support $[0,1]$ of g. Similar to the terminology in Gasser and Müller (1984), we call a $(\nu+1)$-th local polynomial estimator $\hat{g}^{(\nu)}$ with $k=\nu+2$ a standard local polynomial estimator. So, standard local polynomial estimators are those with lowest polynomial order such that $p-\nu$ is odd. The first three standard local polynomial estimators are the local linear estimate for g, the local quadratic estimate for g' and the local cubic estimate for g''.

Remark 3. Theorems 1 and 2 extend previous results (see Altman (1990), Hall and Hart (1990), Herrmann *et al.* (1992), Csörgö and Mielniczuk (1995) and Beran (1999)) in several ways by including $\delta<0$, estimation of derivatives, estimation with higher order kernels and the pointwise asymptotic behaviour in the boundary region.

These results hold for kernel estimators with boundary correction. For kernel estimators without boundary correction Theorem 2 holds with the whole interval $[0,1]$ replaced by $[\Delta, 1-\Delta]$, where Δ is a small positive constant.

Remark 4. In the case when the error process is antipersistent, one obtains higher rates of convergence than for iid errors. For the most important case with $k = 2$ and $\nu = 0$, (3.20) yields the well known rate of convergence $O(n^{-4/5})$ for the MISE in nonparametric regression, if $\delta = 0$ (short memory, for example iid). However, consider for instance $\delta = -0.25$ or -0.4 respectively, then for the same k and ν, the rate of convergence is $O(n^{-12/11})$ and $O(n^{-36/29})$ respectively. Also note that, as $\delta \to -0.5$ and $k \to \infty$, the rate in (3.20) tends to $O(n^{-2})$.

Remark 5. The results in this paper can easily be extended to a parametric regression model with time series errors. If this is done, we can see that the traditional $O(n^{-1})$ lower bound for the rate of convergence of the MISE under the iid assumption no longer holds for strongly dependent errors. If the errors are just short range dependent, the rate is still correct but with some change in the constant. In the case of long-memory errors, the rate is strongly reduced. An interesting fact is that the rate of convergence is higher than $O(n^{-1})$, if the errors are antipersistent, and this rate tends to $O(n^{-2})$ as $\delta \to -0.5$. Another, more extreme example is the simple linear regression model

$$Y_t = \beta_0 + \beta_1 t_i + X_i$$

with $t_i = i/n$ and an error process $X_i = \epsilon_i - \epsilon_{i-1}$, where ϵ_i are iid mean zero random variables with $\mathrm{var}(\epsilon) = \sigma_\epsilon^2 < \infty$. Note that in this case $\delta = -1$ so that Theorem 2 is not directly applicable. It is however easy to show that in this case an $O(n^{-2})$ convergence rate is achieved.

Remark 6. Suppose that the degree of smoothness at a point t is given, i.e. there is a k_0 such that $g^{(k_0)}$ is continuous but $g^{(k_0+1)}$ not. Clearly, $\hat{g}^{(\nu)}$ is only meaningful for $\nu \leq k_0$. The rate of convergence of the MSE is derived under the assumption that $k \leq k_0$ and $\nu \leq k_0 - 2$. For $k \leq k_0$, the theoretical rate of convergence is faster for larger values of k. For $k > k_0$, Theorem 1 does not hold, but it may be conjectured that using $k > k_0$ does not improve the rate of convergence as compared to $k = k_0$. An exact derivation of this result is an open problem.

4. Bandwidth selection

Based on the results in the last section a plug-in bandwidth selector may be developed. The proposal here follows the iterative plug-in idea of Gasser *et al.* (1991). However, it differs from the existing iterative plug-in bandwidth selectors in the following points: 1. The bandwidth selector proposed here is a unified procedure for nonparametric regression with different dependence structures (antipersistence, short memory, long memory); 2. The starting bandwidth is different; 3. The bandwidth h_k for estimating $g^{(k)}$ is obtained from h (i.e. the bandwidth used for $\hat{g}$) with an exponential inflation method. In contrast, Gasser *et al.* (1991), Herrmann *et al.* (1992) and Ray and Tsay (1997) use a multiplied inflation method.

Denote by h_{M} the optimal bandwidth that minimizes the MISE of $\hat{g}$. Plug-in estimators for h_{M} use the formula (3.18) with the unknown constants δ, V as well as $I(g^{(k)})$

being replaced by some consistent estimators. If the error process X_i is modeled parametrically, then both, δ and V, can be estimated with the rate of convergence $n^{-1/2}$ (see Beran (1995, 1999)). In Ray and Tsay (1997) δ and V are estimated nonparametrically. In the following, the finite sample effect of estimating δ and V is not discussed. This problem will be considered in detail elsewhere.

A natural estimator of $I(g^{(k)})$ is

$$\hat{I}(g^{(k)}) = n^{-1}\sum_{i=1}^{n}[\hat{g}^{(k)}(t_i)]^2 \tag{4.1}$$

with a bandwidth h_k. In the following proposition, the following modified versions of assumptions A2 to A4 are used:

A4′. A local polynomial of order $p_1 = p + 2m$ $(m \in N)$ with $p = k - 1$ is used.

A2′. Let $l = p_1 + 1$. Then g is at least k' times continuously differentiable with $k' = \max(l, 2k)$;

A3′. $h_k \to 0$, $(nh_k)^{1-2\delta}h_k^{2k} \to \infty$ as $n \to \infty$.

Let $L_{(k,l)}$ be the asymptotically equivalent kernel for estimating $g^{(k)}$. The accuracy of $\hat{I}(g^{(k)})$ is quantified by the following proposition.

PROPOSITION 1. *Let* A1, A2′ *to* A4′ *hold. Assume that* U_i *in* (2.2) *are normal. Then*

$$E[\hat{I}(g^{(k)}) - I(g^{(k)})] \doteq 2h_k^{(l-k)}\frac{\beta_{(k,l)}}{l!}\int_0^1 g^{(k)}(t)g^{(l)}(t)dt + (nh_k)^{2\delta-1}h_k^{-2k}V \tag{4.2}$$

and

$$\mathrm{var}[\hat{I}(g^{(k)})] \doteq o[(nh_k)^{(4\delta-2)}h_k^{-4k}] + O(n^{2\delta-1}). \tag{4.3}$$

The normality of U_i are assumed for simplifying the proof. This results may be extended to nonnormal innovations. Similar results for independent data may be found in Gasser *et al.* (1991) and Ruppert *et al.* (1995).

The mean squared error (MSE) of $\hat{I}(g^{(k)})$ is dominated by the squared bias, i.e.

$$\mathrm{MSE}\{\hat{I}(g^{(k)})\} \doteq \left\{2h_k^{(l-k)}\frac{\beta_{(k,l)}}{l!}\int_0^1 g^{(k)}(t)g^{(l)}(t)dt + (nh_k)^{2\delta-1}h_k^{-2k}V\right\}^2.$$

The optimal bandwidth for estimating $I(g^{(k)})$ which minimizes the MSE is of the order $O(n^{(2\delta-1)/(k+l+1-2\delta)})$. Note that, this is not the same as the optimal bandwidth for estimating $g^{(k)}$ itself. For a bandwidth $h_k = O(n^{(2\delta-1)/(k+l+1-2\delta)})$ we have $\mathrm{MSE}(\hat{I}(g^{(k)})) = O(n^{2(l-k)(2\delta-1)/(k+l+1-2\delta)})$. When $\int g^{(k)}(x)g^{(l)}(x)dx < 0$, the results can be slightly improved, supplied that the constant in h_k is properly estimated. However, this will not be discussed here. In the most important special case with $k = 2$, $l = 4$, the optimal choice is $h = O(n^{(2\delta-1)/(7-2\delta)})$ which results in $\mathrm{MSE}(\hat{I}(g'')) = O(n^{4(2\delta-1)/(7-2\delta)})$.

Based on Proposition 1 we propose the following algorithm for selecting bandwidth in nonparametric regression with unknown dependent structures. The proposal is made for general k and l. However, what we keep in mind are the two practical cases: $k = 2, 4$ with $l = k + 2$. The algorithm is defined as follows:

i) Start with the bandwidth $h_0 = n^{-\beta}$ with $\beta = (2k + 1)/(k + l + 1)$;

ii) For $i = 1, 2, \ldots$ estimate g using h_{i-1} and let $r_i = y_i - \hat{g}(t_i)$. Estimate δ and V from r_i with an appropriate method;

iii) Set $h_{k,i} = h_{i-1}^{\alpha}$ with $\alpha = (2k+1-2\hat{\delta})/(k+l+1-2\hat{\delta})$ and set

$$h_i = \left(\frac{1-2\hat{\delta}}{2k\beta_{(\nu,k)}^2} \frac{[k!]^2\hat{V}}{\hat{I}(g^{(k)}(t;h_{k,i}))} \right)^{1/(2k+1-2\hat{\delta})} \cdot n^{(2\hat{\delta}-1)/(2k+1-2\hat{\delta})}, \tag{4.4}$$

vi) Increase i by 1 and repeat steps ii) and iii) untill convergence is reached at some i^0 and set $\hat{h} = h_{i^0}$.

The constants α and β may be chosen differently for different purposes. For instance, $\beta = (2k+1)/(k+l+1)$ may be used in order that h_0 is not too large and assumption A3 holds. This means that $\hat{g}$ at the first iteration is already consistent and hence $\hat{\delta}$ and $\hat{V}$ can be estimated by the method introduced in Beran (1999). Gasser *et al.* (1991) show that the starting bandwidth should not be too large. This is the reason, why they propose to use the starting bandwidth $h_0' = n^{-\beta}$ with $\beta = 1$. For long-memory data $h_0' = n^{-1}$ is too small to be used at the beginning. Hence, in Ray and Tsay (1997) the method of Herrmann *et al.* (1992) is used as a pilot method at the first step by assuming short memory. Compared with h_{A}, h_0 is also very small. For example, for $k = 2$ and $l = 4$, $h_0 = o(h_{\mathrm{A}}^2)$ for all $\delta \in (-0.5, 0.5)$. h_0 is determined by the fact that, if $h_0' = n^{-1}$ is used, under the assumption $\delta = 0$, then $h_1' = O(n^{-5/7})$.

The formula $h_{k,i} = (h_{i-1})^{\alpha}$ may be called an exponential inflation method. The factor $\alpha = (2k+1-2\hat{\delta})/(k+l+1-2\hat{\delta})$ is chosen in order that the MSE of $\hat{I}(g^{(k)})$ is of the optimal order, when convergence is reached. The choice of α in the inflation method used by Gasser *et al.* (1991), Herrmann *et al.* (1992) and Ray and Tsay (1997), i.e. $h_{k,i} = h_{i-1}n^{\alpha}$ with $\alpha = (1-2\hat{\delta})/(2(2k+1-2\hat{\delta}))$, is motivated by minimizing the variance in $\hat{I}(g^{(k)})$. For our inflation method, the order of the variance of $\hat{I}(g^{(k)})$ is minimized by $\alpha = 1/2$ for all δ and k.

The main reason for using the exponential inflation method is to reduce the number of iterations i^0 required for obtaining convergence of the bandwidth. Note that, i^0 depends strongly on δ. In fact, $i^0 \to \infty$ as $\delta \to 0.5$ (see Ray and Tsay (1997)). However, i^0 also depends on the inflation method. In general, our exponential inflation method needs a smaller number of iterations than the method by Ray and Tsay (1997). For example, set $k = 2$, $\alpha = 1/2$ and $\beta = 5/7$. Then the relative rate of convergence of the estimated bandwidth is the same as in Gasser *et al.* (1991), Herrmann *et al.* (1992) or Ray and Tsay (1997), respectively. However, the number of iterations is equal to $i^0 = 4$ for $\delta = 0$, as opposed to 11 for the method by Gasser *et al.* (1991) and Herrmann *et al.* (1992). For $\delta = 0.4$, we have $i^0 = 6$ whereas for the proposal by Ray and Tsay (1997), i^0 is equal to 43.

The asymptotic properties of $\hat{h}$ is given by theorem 3.

THEOREM 3. *Let* A1 *and* A2′ *hold. Let g and $g^{(k)}$ be estimated with local polynomials of orders k and l respectively. Assume U_i in* (2.2) *are normal and the error in $\hat{\delta}$ and $\hat{V}$ is negligible, then $\hat{h}$ is a consistent estimator of h_{M} such that*

$$\hat{h} = h_{\mathrm{M}}\{1 + O[(h_{\mathrm{A}} - h_{\mathrm{M}})/h_{\mathrm{M}}] + O_p(n^{(l-k)(2\delta-1)/(k+l+1-2\delta)})\}. \tag{4.5}$$

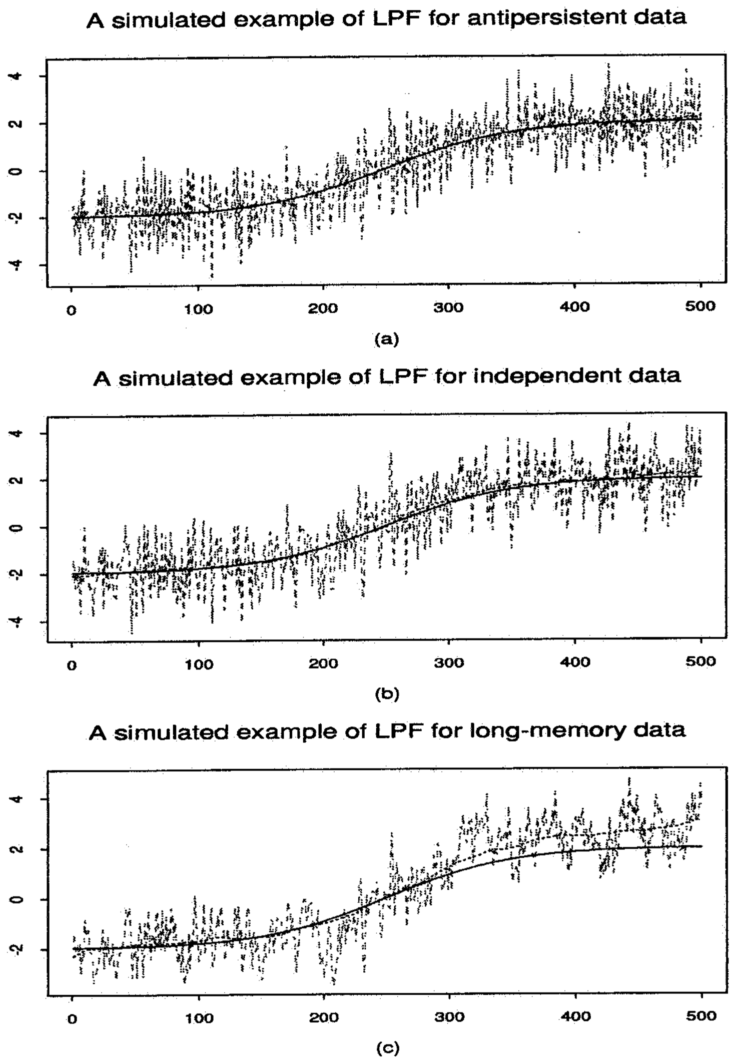

Fig. 1. The data (long dashes), the trend (solid line) and $\hat{g}$ (short dashes) for the three simulated examples with $\delta = -0.4$ (a), $\delta = 0$ (b) and $\delta = 0.4$ (c), respectively.

The difference between $h_{\rm A}$ and $h_{\rm M}$ provides a natural bound for the rate of convergence of a plug-in bandwidth selector. The exact form of $h_{\rm A} - h_{\rm M}$ will not be discussed here. Results on this topic may be found in Gasser *et al.* (1991) for independent data and in Ray and Tsay (1997) for data with long-memory. For independent data with $k = 2$, $l = 4$, we have $(h_{\rm A} - h_{\rm M})/h_{\rm M} = o_p\{\hat{I}(g^{(k)}) - I(g^{(k)})\}$. Now, the rate of convergence of the proposed bandwidth selector is $n^{-2/7}$. This is the same as for the direct plug-in estimator proposed by Ruppert *et al.* (1995) and it is higher than that of the proposal of Gasser *et al.* (1991), where the rate of convergence is dominated by a bias term of the order $O(n^{-1/5})$ with a root n consistent variance term.

Table 1. Estimates of δ and h_A for the simulated examples.

Example	δ	$\hat{\delta}$	95%-CI for δ	h_A	$\hat{h}$
Fig. 1a	-0.4	-0.336	$[-0.405, -0.267]$	0.075	0.076
Fig. 1b	0	0.048	$[-0.020, 0.117]$	0.106	0.110
Fig. 1c	0.4	0.417	$[0.349, 0.486]$	0.173	0.166

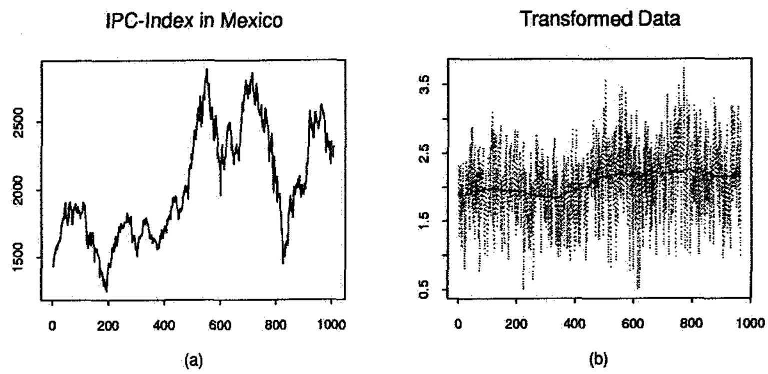

Fig. 2. The IPC stock market index in Mexico from Jan. 2, 1992 to Nov. 10, 1995 (a) and transformed data r_t with the estimated trend (b).

5. Application

Examples in this section were computed with the data-driven program for SEMIFAR model written by Beran in S-Plus (see Beran and Feng (1999)), with the kernel smoothing being replaced by a function for local polynomial fitting. For simplicity, unweighted standard local fits for estimating g and g'', respectively, are used. Here the optimal bandwidth is defined on the support $[\Delta, 1-\Delta]$ with $\Delta = 0.1$.

5.1 *Simulated examples*

In order to show how the estimation and the selected bandwidth depend on the dependence level, three simulated examples are shown in figures 1a to 1c. The data are generated with $\delta = -0.4, 0, 0.4$, respectively, following the model:

$$Y_i = 2\tanh(5(t_i - 0.5)) + X_i \tag{5.1}$$

with

$$(1-B)^{\delta} X_i = \epsilon_i, \tag{5.2}$$

where ϵ_i are iid normal distributed random variables, which are normalized according to δ such that the variance of X_i is equal to one. The estimates and confidence intervals given by the SEMIFAR program are summarized in Table 1.

Figure 1 shows that the accuracy of the estimation depends strongly on the dependence level in the data. When δ assumes a large positive value, a large bandwidth will be selected in order to reduce the variance of the estimation. This causes a larger bias. In this case $\hat{g}$ may be far from g, if n is not large enough. Another explanation for this phenomenon is the local trend caused by long memory (Beran (1994)).

5.2 *A practical example*

Beran and Ocker (2001) propose to analyze volatility in financial time series by fitting a SEMIFAR model to the power transformed absolute differences $r_t = |Y_t - Y_{t-1}|^{1/4}$. These turn out to be almost normal. See Ding *et al.* (1993) and Ding and Granger (1996) for a similar proposal. Beran and Ocker (2001) show that long-memory phenomenon in r_t only occurs for relatively small stock markets. As an application of the proposal in this paper, a local polynomial fit is used to analyze r_t for the IPC stock market index in Mexico from Jan. 2, 1992 to Nov. 10, 1995 (Fig. 2a). The volatility series r_t is displayed in Fig. 2b together with the estimated trend. The detrended "residual" series X_t has significant long memory with $\hat{\delta} = 0.121$ and a 95%-confidence interval of $[0.072, 0.171]$.

6. Final remarks

It is clear that the bias of a nonparametric regression estimator for observations with dependent errors is the same as that for uncorrelated errors if the estimator is a linear smoother and the error process is stationary. However, the variance of a linear smoother depends on how the errors are correlated. If the errors have short memory with $\delta = 0$, then only the constant V is influenced. In particular, V is larger than $\gamma(0) \int K^2_{(\nu,k)}(u)du$, when the dependence structure is dominated by positive correlations. In this case, the optimal bandwidth is larger than in the case of independent errors (Herrmann *et al.* (1992)). If the error process has long memory, i.e. $\delta > 0$, then not only the constant V but also the order of the variance is changed. The variance converges at a slower rate to zero than in the case of short memory. On the other hand, if $\delta < 0$, the data are mostly negatively correlated and $\sum_{k=-\infty}^{\infty} \gamma(k) = 0$. In this case, the variance of $\hat{g}^{(\nu)}$ converges to zero at a higher rate than under independence or short-range dependence.

For the practical application of the results in Section 3, the iterative plug-in method can be adapted for bandwidth selection with unknown dependent structures. However, as shown in Section 4, the rate of convergence of a plug-in method is determined by the difference between $h_{\rm A}$ and $h_{\rm M}$. An interesting question is, if other bandwidth selection rules for nonparametric regression with independent data, which aim at a direct estimation of the MISE (e.g. Fan and Gijbels (1995), Feng and Heiler (1999)), can be adapted to bandwidth selection in nonparametric regression with long memory. These methods have some advantages over the plug-in method. For instance, for independent data, the double smoothing bandwidth selector by Feng and Heiler (1999) does not require the estimation of $g^{(k)}$, and the rate of convergence is faster than the best possible convergence rate of a plug-in bandwidth selector based on $h_{\rm A}$. This and other open questions on bandwidth selection in nonparametric regression with long- memory will be discussed elsewhere.

Acknowledgements

This paper was supported in part by the Center of Finance and Econometrics at the University of Konstanz and by an NSF (SBIR, phase 2) grant to MathSoft, Inc. We would like to thank Dr. Dirk Ocker for providing us the data of the IPC index. We are also grateful to the associate editor and two referees for their helpful comments, which lead improve the quality of the paper.

Appendix: Proofs of the theorems

PROOF OF THEOREM 1. The proof will only be carried out in detail for an interior point t. Define $x_i = (t_i - t)/h$, $y_j = (t_j - t)/h$, $n_0 = [nt] + 1$ and $b = [nh]$ as before.

(i): The proof of the pointwise bias is standard.

(iii): (iii) holds if (ii) holds, since

$$\begin{aligned}
&(nh)^{1-2\delta} h^{2\nu} \operatorname{var}(\hat{g}^{(\nu)}(t)) \\
&= (nh)^{1-2\delta} h^{2\nu} \sum_{i,j=n_0-b}^{n_0+b} w_i^\nu w_j^\nu \gamma(i-j) \\
&= (nh)^{1-2\delta} h^{2\nu} \sum_{i,j=n_0-b}^{n_0+b} \tilde{w}_i^\nu \tilde{w}_j^\nu \gamma(i-j)[1+o(1)] \\
&= (nh)^{-1-2\delta} \sum_{i,j=n_0-b}^{n_0+b} K_{(\nu,k)}(x_i) K_{(\nu,k)}(y_j) \gamma(i-j)[1+o(1)] \\
&= V_n(h)[1+o(1)]
\end{aligned}$$

(ii): The results will be proved separately for $\delta = 0$, $0 < \delta < 0.5$ and $-0.5 < \delta < 0$. The formulas (3.12), (3.13) and (3.14) will be obtained immediately.

a) $\delta = 0$. Observe that, for $\delta = 0$, $\sum_{k=-\infty}^{\infty} \gamma(k) = C$ and

$$V_n(h) = (nh)^{-1} \sum_{i=n_0-b}^{n_0+b} K_{(\nu,k)}(x_i) \sum_{j=n_0-b}^{n_0+b} K_{(\nu,k)}(y_j)\gamma(i-j).$$

Let b' be an integer such that $b' = o(b)$, $b' \to \infty$ as $n \to \infty$ (e.g. $b' = [\sqrt{b}]$). For $n_0 - b + b' < i < n_0 + b - b'$ we have

$$\sum_{j=n_0-b}^{n_0+b} K_{(\nu,k)}(y_j)\gamma(i-j) = \sum_{|i-j|<b'} K_{(\nu,k)}(y_j)\gamma(i-j) + \sum_{|i-j|\geq b'} K_{(\nu,k)}(y_j)\gamma(i-j).$$

For $|i-j| < b'$ we have $|x_i - y_j| = o(1)$ and hence $K(y_j) = K(x_i)[1+o(1)]$, because $b' = o(b)$ and $K_{(\nu,k)}$ is Lipschitz-continuous. Then we obtain

$$\sum_{|i-j|<b'} K_{(\nu,k)}(y_j)\gamma(i-j) = CK_{(\nu,k)}(x_i)[1+o(1)],$$

because $\sum_{|i-j|<b'} \gamma(i-j) = C + o(1)$. Furthermore,

$$\sum_{|i-j|\geq b'} K_{(\nu,k)}(y_j)\gamma(i-j) = o(1),$$

because $K_{(\nu,k)}$ is bounded and $\sum_{|i-j|\geq b'} \gamma(i-j) = o(1)$.

For $i = n_0 - b$ or $i = n_0 + b$ we have, analogously,

$$\begin{aligned}
\sum K_{(\nu,k)}(y_j)\gamma(i-j) &= [(C+\gamma(0))/2] K_{(\nu,k)}(x_i)(1+o(1)) \\
&= O(K_{(\nu,k)}(x_i)).
\end{aligned}$$

It is clear that $\sum K_{(\nu,k)}(y_j)\gamma(i-j) = O(K_{(\nu,k)}(x_i))$ holds for all i such that $n_0 - b < i \le n_0 - b + b'$ or $n_0 + b - b' < i \le n_0 + b$. Since $b' = o(b)$, we have

$$V_n(h) = (nh)^{-1} \sum_{i=n_0-b}^{n_0+b} K^2_{(\nu,k)}(x_i)[C + o(1)] \tag{A.1}$$
$$= C \int_{-1}^{1} K^2_{(\nu,k)}(x)dx + o(1).$$

b) $\delta > 0$. In this case we have $\gamma(k) \sim c_\gamma |k|^{2\delta-1}$ with $c_\gamma = 2c_f\Gamma(1-2\delta)\sin\pi\delta > 0$ (see Beran (1994), pp. 61–63).

$$V_n(h) = (nh)^{-1-2\delta} \sum_{i,j=n_0-b}^{n_0+b} K_{(\nu,k)}(x_i)K_{(\nu,k)}(y_j)\gamma(i-j) \tag{A.2}$$
$$\doteq c_\gamma (nh)^{-1-2\delta} \sum_{\substack{i,j=n_0-b \\ i\neq j}}^{n_0+b} K_{(\nu,k)}(x_i)K_{(\nu,k)}(y_j)|i-j|^{2\delta-1}$$
$$= c_\gamma (nh)^{-2} \sum_{\substack{i,j=n_0-b \\ i\neq j}}^{n_0+b} K_{(\nu,k)}(x_i)K_{(\nu,k)}(y_j)|x_i-y_j|^{2\delta-1}$$
$$\doteq c_\gamma \int_{-1}^{1}\int_{-1}^{1} K_{(\nu,k)}(x)K_{(\nu,k)}(y)|x-y|^{2\delta-1}dxdy$$

c). The proof for $\delta < 0$ is based on the decomposition of the equivalent kernel and the property $\sum_{k=-\infty}^{\infty} \gamma(k) = 0$, where $\gamma(k) \sim c_\gamma |k|^{2\delta-1}$ for large k with $c_\gamma = 2c_f\Gamma(1-2\delta)\sin(\pi\delta) < 0$ in the case of antipersistence (see Beran (1994)). For given i we have $\sum_{j=n_0-b}^{n_0+b} \gamma(i-j) = -\sum_{|j-n_0|>b} \gamma(i-j)$. Recall that the equivalent kernel has the form $K_{(\nu,k)}(x) = \sum_{l=0}^{r} \alpha_l x^l \mathbf{1}_{\{|x|\le 1\}}$. At a point x_i, $K_{(\nu,k)}(y)$ can be rewritten as $K_{(\nu,k)}(y) = \sum_{l=0}^{r} \beta_l(x_i)(x_i-y)^l =: K_0(x_i) + K_1(x_i-y)$, where $K_0(x_i) = \beta_0(x_i)$, $K_1(x_i-y) = \sum_{l=1}^{r} \beta_l(x_i)(x_i-y)^l$.

Observing that t is an interior point we have

$$V_n = (nh)^{-1-2\delta} \sum_{i=n_0-b}^{n_0+b} K_{(\nu,k)}(x_i) \sum_{j=n_0-b}^{n_0+b} K_{(\nu,k)}(y_j)\gamma(i-j).$$

$$\sum_{j=n_0-b}^{n_0+b} K_{(\nu,k)}(y_j)\gamma(i-j)$$
$$= \sum_{j=n_0-b}^{n_0+b} K_1(x_i-y_j)\gamma(i-j) - K_0(x_i) \sum_{|j-n_0|>b} \gamma(i-j)$$
$$\doteq c_\gamma \left\{ \sum_{\substack{j=n_0-b \\ j\neq i}}^{n_0+b} K_1(x_i-y_j)|i-j|^{2\delta-1} - K_0(x_i) \sum_{|j-n_0|>b} |i-j|^{2\delta-1} \right\}$$

$$= c_\gamma (nh)^{2\delta-1} \left\{ \sum_{\substack{j=n_0-b \\ j\neq i}}^{n_0+b} K_1(x_i - y_j)|x_i - y_j|^{2\delta-1} - K_0(x_i) \sum_{|j-n_0|>b} |x_i - y_j|^{2\delta-1} \right\}$$

$$\doteq c_\gamma (nh)^{2\delta} \left\{ \int_{-1}^{1} K_1(x_i - y)|x_i - y|^{2\delta-1} dy - K_0(x_i) \int_{|y|>1} |x_i - y|^{2\delta-1} dy \right\}$$

Finally, we obtain

$$\begin{aligned} V_n &\doteq c_\gamma (nh)^{-1} \sum_{i=n_0-b}^{n_0+b} K_{(\nu,k)}(x_i) \left\{ \int_{-1}^{1} K_1(x_i - y)|x_i - y|^{2\delta-1} dy \right. \\ &\qquad \left. - K_0(x_i) \int_{|y|>1} |x_i - y|^{2\delta-1} dy \right\} \\ &\doteq c_\gamma \int_{-1}^{1} K_{(\nu,k)}(x) \left\{ \int_{-1}^{1} K_1(x - y)|x - y|^{2\delta-1} dy \right. \\ &\qquad \left. - K_0(x) \int_{|y|>1} |x - y|^{2\delta-1} dy \right\} dx. \end{aligned}$$

The proof at a boundary point $t = ch$ with $0 \le c < 1$ is similar to that given above and is hence omitted. Especially, note that the proof in "(ii): c)" can also be carried out for $c = 0$. This concludes the proof. □

PROOF OF THEOREM 2. Theorem 2 follows from Theorem 1, since in the case that $p - \nu$ is odd the MISE on the boundary area is asymptotically negligible. □

PROOF OF PROPOSITION 1. From here on $I(g^{(k)})$ will be abbreviated as I.

a) Bias in $\hat{I}$, denoted by $B(\hat{I})$:

$$\begin{aligned} B(\hat{I}) &= n^{-1} \sum_{i=1}^{n} B\{[\hat{g}^{(k)}(t_i)]^2\} \\ &= n^{-1} \sum_{i=1}^{n} \{E[(\hat{g}^{(k)}(t_i))^2] - [g^{(k)}(t_i)]^2\}. \end{aligned}$$

From Theorem 1, at an interior point t_i we have

$$\begin{aligned} E\{[\hat{g}^{(k)}(t_i)]^2\} &= [E\{\hat{g}^{(k)}(t_i)\}]^2 + \mathrm{var}\{\hat{g}^{(k)}(t_i)\} \\ &\doteq [g^{(k)}(t_i)]^2 + 2h_k^{(l-k)} \frac{\beta_{(k,l)}}{l!} g^{(k)}(t_i) g^{(l)}(t_i) + (nh_k)^{2\delta-1} h_k^{-2k} V. \end{aligned}$$

Since $l-k$ is even, the part of $B[\hat{I}]$ due to the boundary region is asymptotically negligible. We have

$$\begin{aligned} B(\hat{I}) &\doteq n^{-1} \sum_{t_i \in [h_k, 1-h_k]} \left\{ 2h_k^{(l-k)} \frac{\beta_{(k,l)}}{l!} g^{(k)}(t_i) g^{(l)}(t_i) + (nh_k)^{2\delta-1} h_k^{-2k} V \right\} \\ &\doteq 2h_k^{(l-k)} \frac{\beta_{(k,l)}}{l!} \int_0^1 g^{(k)}(t) g^{(l)}(t) dt + (nh_k)^{2\delta-1} h_k^{-2k} V. \end{aligned}$$

b) Variance in $\hat{I}$:

The formula of var$[\hat{I}]$ in this paper aims at the relationship between it and $B(\hat{I})$. A more explicit formula of var$[\hat{I}]$ in the case when X_t are iid, may be found in Ruppert *et al.* (1995) (see also Gasser *et al.* (1991) for results with $k = 2$).

$$
\begin{aligned}
\text{(A.3)}\quad \operatorname{var}[\hat{I}] &= n^{-2}\operatorname{var}\left\{\sum_{i=1}^{n}(\hat{g}^{(k)}(t_i))^2\right\}\\
&= n^{-2}\operatorname{var}\left\{\sum_{i=1}^{n}[(\hat{g}^{(k)}(t_i)-E(\hat{g}^{(k)}(t_i)))^2+2\hat{g}^{(k)}(t_i)E(\hat{g}^{(k)}(t_i))]\right\}\\
&\doteq O\left(n^{-2}\operatorname{var}\left[\sum_{i=1}^{n}(\hat{g}^{(k)}(t_i)-E(\hat{g}^{(k)}(t_i)))^2\right]\right)\\
&\quad +O\left(n^{-2}\operatorname{var}\left[\sum_{i=1}^{n}\hat{g}^{(k)}(t_i)E(\hat{g}^{(k)}(t_i))\right]\right)=V_1+V_2,\qquad \text{say.}
\end{aligned}
$$

We consider at first V_1. Define $\xi_i = \hat{g}^{(k)}(t_i) - E(\hat{g}^{(k)}(t_i))$. We have

$$
\text{(A.4)}\qquad n^{-2}\operatorname{var}\left[\sum_{i=1}^{n}\xi_i^2\right] = n^{-2}\sum_{i=1}^{n}\sum_{j=i}^{n}\operatorname{cov}(\xi_i^2,\xi_j^2) = o(\operatorname{var}[\xi_{i_0}^2])
$$

with t_{i_0} being an interior point. (A.4) is due to the fact that var$[\xi_i^2]$ is of the same order at each point and $\operatorname{cov}(\xi_i^2,\xi_j^2) = o[\operatorname{var}(\xi_{i_0}^2)]$ as $|i-j| \to \infty$. Under the normal assumption,

$$
\begin{aligned}
\text{(A.5)}\quad \operatorname{var}(\xi_{i_0}^2) &= \operatorname{var}[(\hat{g}^{(k)}(t_{i_0})-E(\hat{g}^{(k)}(t_{i_0})))^2]\\
&= E[(\hat{g}^{(k)}(t_{i_0})-E(\hat{g}^{(k)}(t_{i_0})))^4]-E^2[(\hat{g}^{(k)}(t_{i_0})-E(\hat{g}^{(k)}(t_{i_0})))^2]\\
&= 3\operatorname{var}^2[\hat{g}^{(k)}(t_{i_0})]-\operatorname{var}^2[\hat{g}^{(k)}(t_{i_0})]\\
&= 2\operatorname{var}^2[\hat{g}^{(k)}(t_{i_0})].
\end{aligned}
$$

By combining (A.4) and (A.5) we obtain $V_1 = o(\operatorname{var}^2[\hat{g}^{(k)}(t_{i_0})]) = o((nh_k)^{4\delta-2}h_k^{-4k})$.

Let $b_k = [nh_k]$. And let $\tilde{w}_i$, $i = 1, 2, \ldots, 2b_k + 1$, be the possible nonzero kernel weights for $\hat{g}^{(k)}(t_i)$ at an interior point (see Section 3), which are symmetric about $i = b_k + 1$. Observing that the influence of the points with $i \le 2b_k$ or $i > n - 2b_k$ on V_2 is asymptotically negligible, if $h_k \to 0$.

$$
\begin{aligned}
V_2 &\doteq O\left(n^{-2}\operatorname{var}\left[\sum_{i=2b_k+1}^{n-2b_k}\hat{g}^{(k)}(t_i)E(\hat{g}^{(k)}(t_i))\right]\right)\\
&\doteq O\left(n^{-2}\operatorname{var}\left\{\sum_{i=2b_k+1}^{n-2b_k}E(\hat{g}^{(k)}(t_i))\left[\sum_{j=i-b_k}^{i+b_k}\tilde{w}_{j-i+b_k+1}Y_j\right]\right\}\right)\\
&= O\left(n^{-2}\operatorname{var}\left\{\sum_{i=2b_k+1}^{n-2b_k}\left[\sum_{j=i-b_k}^{i+b_k}\tilde{w}_{j-i+b_k+1}E(\hat{g}^{(k)}(t_{i+b_k+1-j}))\right]Y_i\right\}\right).
\end{aligned}
$$

Observing that $\sum_{j=i-b_k}^{i+b_k}\tilde{w}_{j-i+b_k+1}E(\hat{g}^{(k)}(t_{i+b_k+1-j})) \doteq g^{(2k)}(t_i)$ and $g^{(2k)}$ is bounded, we have

$$V_2 \doteq O\left(n^{-2}\operatorname{var}\left[\sum_{i=2b_k+1}^{n-2b_k} g^{(2k)}(t_i)Y_i\right]\right) \tag{A.6}$$
$$\doteq O\left(n^{-2}\operatorname{var}\left[\sum_{i=1}^{n} Y_i\right]\right) = O(n^{2\delta-1}).$$

Proposition 1 is proved. □

A SKETCHED PROOF OF THEOREM 3. In the following it is assumed that the error in $\hat{\delta}$ and $\hat{V}$ is asymptotically negligible. Hence the unknown parameters δ and V will be directly used in the proof.

The iterative plug-in algorithm is motivated by fixpoint search. Here a bandwidth $h'_k = O(n^{(2\delta-1)/(2k+1-2\delta)})$ plays an important role. Observe that, on one hand we have $h'_k = O(h_{\mathrm{M}})$. On the other hand, h'_k is a bandwidth such that $\hat{g}^{(k)} = O_p(g^{(k)})$ and $\hat{I} = O_p(I)$. But now, both of $\hat{g}^{(k)}$ and $\hat{I}$ are not yet consistent. Assuming that the starting bandwidth satisfies $h_0 = o(h'_k)$, e.g. $h_0 = n^{-(2k+1)/(k+l+1)}$ as proposed in this paper, an iterative plug-in bandwidth selection procedure may be divided into the following three steps according to the relationship between $h_{k,i}$ and h'_k.

Step 1. When $h_{k,i} = o_p(h'_k)$, $\hat{I} = O_p(n^{2\delta-1}h^{-(2k+1+\delta)})$ is infinite in probability. In this case one obtains $h_i = O_p(h_{k,i})$. This implies that $h_{i-1} = o(h_i)$, i.e. the here bandwidth is of larger order.

Step 2. When $h_{k,i} = O_p(h'_k)$, $\hat{I} = O_p(I)$. Now, $h_i = O_p(h_{\mathrm{M}})$. But, in general, the constant is not yet consistent.

Step 3. When $h'_k = o_p(h_{k,i})$, $\hat{I}$ is consistent and one obtains a consistent bandwidth selector h_i.

Let i^0 be defined as in Section 4. Let i' be the number of iterations in Step 1, and i^* the number of iterations in Step 2 ($i^* = 0$ or 1). Then $i^0 = i' + i^* + k + 1$ where $k+1$ is the number of iterations in Step 3. The selected bandwidth h_i is consistent in the first and all subsequent iterations in step 3. However, the next k iterations (after the first one in Step 3) still improve the variability of the estimator $\hat{h}$ (Gasser *et al.* (1991)). For given δ, k and l, the required number of iterations depends on α, β and the inflation method. Assume that $k = 2$, $l = 4$. By the proposal of Gasser *et al.* (1991), $i' = 7$, $i^* = 1$ and $i^0 = 11$. The purpose of Gasser *et al.* (1991) is to minimize the variance in $\hat{I}$. When the MSE of $\hat{I}$ is to be minimized, more iterations are required. By the proposal in this paper for minimizing the MSE of $\hat{I}$, if $\delta = 0$, we have $i' = 3$, $i^* = 0$ and $i^0 = 6$. For $\delta = 0.4$, we have $i' = 6$, $i^* = 0$ and $i^0 = 9$ (note that $i^0 = 43$ by the proposal of Ray and Tsay (1997) for the case with $\delta = 0.4$). In both cases, Step 2 does not appear.

The asymptotic results and hence the rate of convergence does not depend on h_0. Theoretically, any h_0 such that $h_{k,1} \to 0$, $nh_{k,1} \to \infty$ as $n \to \infty$ could be used as the starting bandwidth. The asymptotic results depend only on the inflation method and α. The rate of convergence for our proposal can be calculated as follows.

We have

$$(\hat{h} - h_{\mathrm{M}})/h_{\mathrm{M}} = (\hat{h} - h_{\mathrm{A}})/h_{\mathrm{M}} + (\hat{h}_{\mathrm{A}} - h_{\mathrm{M}})/h_{\mathrm{M}} \tag{A.7}$$
$$\doteq (\hat{h} - h_{\mathrm{A}})/h_{\mathrm{A}} + (\hat{h}_{\mathrm{A}} - h_{\mathrm{M}})/h_{\mathrm{M}}.$$

Taylor's expansion of $\hat{h}$ gives

$$\hat{h} - h_{\mathrm{A}} \doteq -\frac{1}{2k+1-2\delta} h_{\mathrm{A}} I^{-1}(\hat{I} - I).$$

It follows that

$$(\hat{h} - h_{\mathrm{A}})/h_{\mathrm{A}} \doteq -\frac{1}{2k+1-2\delta} I^{-1}(\hat{I} - I).$$

Let $h_{k,i^0} = O(n^{(2\delta-1)/(k+l+1-2\delta)})$, then we obtain

$$(\hat{h} - h_{\mathrm{A}})/h_{\mathrm{A}} \doteq O_p(n^{(l-k)(2\delta-1)/(k+l+1-2\delta)}). \tag{A.8}$$

Theorem 3 is proved by combining (A.7) and (A.8). □

References

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

Beran, J. (1994). *Statistics for Long-memory Processes*, Chapman & Hall, New York.

Beran, J. (1995). Maximum likelihood of estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models, *J. Roy. Statist. Soc. Ser. B*, **57**, 659–672.

Beran, J. (1999). SEMIFAR models—A semiparametric framework for modelling trends, long range dependence and nonstationarity, Discussion paper No. 99/16, Center of Finance and Econometrics, University of Konstanz.

Beran, J. and Feng, Y. (1999). Data-driven estimation of semiparametric fractional autoregressive models, Preprint, University of Konstanz.

Beran, J. and Ocker, D. (1999). SEMIFAR forecasts, with applications to foreign exchange rates, *J. Statist. Plann. Inference*, **80**, 137–153.

Beran, J. and Ocker, D. (2001). Volatility of stock market indices—An analysis based on SEMIFAR models, *J. Bus. Econom. Statist.*, **19**, 103–116.

Beran, J., Feng, Y., Franke, G., Hess, D. and Ocker, D. (1999). SEMIFAR model, with application to commodities, exchange rates and the volatility of stock market indices, Preprint, University of Konstanz.

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *J. Amer. Statist. Assoc.*, **74**, 829–836.

Cox, D. R. (1984). Long-range dependence: A review, *Statistics: An Appraisal. Proceedings 50th Anniversary Conference* (eds. H. A. David and H. T. David), 55–74, The Iowa State University Press.

Csörgö, S. and Mielniczuk, J. (1995). Nonparametric regression under long-range dependent normal errors, *Ann. Statist.*, **23**, 1000–1014.

Ding, Z. and Granger, C. W. J. (1996). Modeling volatility persitence of speculative returns: A new approach, *J. Econometrics*, **73**, 185–215.

Ding, Z., Granger, C. W. J. and Engle, R. F. (1993). A long memory property of stock market returns and a new model, *Journal of Empirical Finance*, **1**, 83–106.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation, *J. Roy. Statist. Soc. Ser. B*, **57**, 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*, Chapman & Hall, London.

Feng, Y. (1999). *Kernel- and Locally Weighted Regression—With Applications to Time Series Decomposition*, Verlag für Wissenschaft und Forschung, Berlin.

Feng, Y. and Heiler, S. (1998). Locally weighted autoregression, *Econometrics in Theory and Practice* (eds. R. Galata and H. Küchenhoff), 101–117, Physica, Heidelberg.

Feng, Y. and Heiler, S. (1999). Selecting bandwidth for nonparametric regression based on double smoothing rule, Preprint, University of Konstanz.

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions, *Smoothing Techniques for Curve Estimation* (eds. T. Gasser and M. Rosenblatt), 23–68, Springer, Heidelberg.

Gasser, T. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method, *Scand. J. Statist.*, **11**, 171–185.

Gasser, T., Müller, H. G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation, *J. Roy. Statist. Soc. Ser. B*, **47**, 238–252.

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing, *J. Amer. Statist. Assoc.*, **86**, 643–652.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing, *J. Time Ser. Anal.*, **1**, 15–30.

Hall, P. and Hart, J. D. (1990). Nonparametric regression with long- range dependence, *Stochastic Process. Appl.*, **36**, 339–351.

Hampel, F. R. (1987). Data analysis and self-similar processes, *Proceeding of the 46th Session of ISI, Tokyo*, Book 4, 235–254.

Härdle, W., Hall, P. and Marron, J. S. (1992). Regression smoothing parameters that are not far from their optimum, *J. Amer. Statist. Assoc.*, **87**, 227–233.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry (with discussion), *Statist. Sci.*, **8**, 120–143.

Heiler, S. and Feng, Y. (1998). A simple root n bandwidth selector for nonparametric regression, *J. Nonparametr. Statist.*, **9**, 1–21.

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

Hosking, J. R. M. (1981). Fractional differencing, *Biometrika*, **68**, 165–176.

Karuppiah, J. and Los, C. A. (2000). Wavelet multiresolution analysis of high-frequency FX rates, Summer 1997, Working Paper, Adelaide University.

Künsch, H. (1986). Statistical aspects of self-similar processes, *Proc. First World Congress of the Bernoulli Society, Tashkent*, Vol. 1, 67–74.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*, Freeman, New York.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates, *J. Time Ser. Anal.*, **17**, 571–599.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes, *Scand. J. Statst.*, **24**, 165–179.

Müller, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting, *J. Amer. Statist. Assoc.*, **82**, 231–238.

Ray, B. K. and Tsay, R. S. (1997). Bandwidth selection for kernel regression with long-range dependence, *Biometrika*, **84**, 791–802.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *Ann. Statist.*, **22**, 1346–1370.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression, *J. Amer. Statist. Assoc.*, **90**, 1257–1270.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion), *Ann. Statist.*, **5**, 595–620.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman & Hall, London.

Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data, *Ann. Statist.*, **24** 466–484.