REVSTAT – Statistical Journal
Volume 6, Number 2, June 2008, 123–150

MINIMALLY BIASED NONPARAMETRIC REGRESSION AND AUTOREGRESSION

Authors: Timothy L. McMurry
– Department of Mathematical Sciences, DePaul University,
Chicago, IL, USA
tmcmurry@depaul.edu

Dimitris N. Politis
– Department of Mathematics, University of California,
San Diego, La Jolla, CA, USA
politis@math.ucsd.edu

Received: September 2007 Revised: November 2007 Accepted: December 2007

Abstract:

- A nonparametric regression estimator is introduced which adapts to the smoothness of the unknown function being estimated. This property allows the new estimator to automatically achieve minimal bias over a large class of locally smooth functions without changing the rate at which the variance converges. Optimal convergence rates are shown to hold for both i.i.d. data and autoregressive processes satisfying strong mixing conditions.

Key-Words:

- *nonparametric regression; autoregression; Fourier transform.*

AMS Subject Classification:

- 62G08.

1. INTRODUCTION

Suppose the data $(X_1, Y_1), ..., (X_n, Y_n)$ are observations from a general real valued bivariate random process. The simplest example is when the data are generated by a model of the form $Y_i = r(X_i) + \epsilon_i$ where the ϵ_i are mean zero random errors satisfying some conditions; in general, the ϵ_i will not be independent. A second example of interest is nonparametric autoregression, where $Y_t = X_{t+1}$. The function r, the conditional mean of Y given X, is unknown and will be estimated from the data. There are many nonparametric approaches to estimating r, including various kernel methods proposed by Nadaraya [15] and Watson [25], Gasser and Müller [7, 8], and local polynomial estimators, Fan [6]. In each of these techniques $r(x)$ is in essence estimated through weighted local averaging on the data near x. The smoothness of the function r and properties of the weights used in this averaging determine the performance of the estimator. In this paper we propose a new class of kernels which allow the Nadaraya–Watson estimator to automatically achieve asymptotically optimal performance no matter how smooth r happens to be.

The Nadaraya–Watson estimator is defined to be

$$\hat{r}(x) := \frac{\sum_{i=1}^{n} Y_i K\big((X_i - x)/h\big)}{\sum_{i=1}^{n} K\big((X_i - x)/h\big)} \ . \tag{1.1}$$

The function $K(x)$ is the kernel; it is used to weight the observations. The denominator ensures the weights sum to 1. The parameter h is the bandwidth, or smoothing parameter. It balances a tradeoff between bias and variance. Small values of h concentrate the mass of the kernel near x, giving heavy weight to nearby observations and relatively little or no weight to more distant observations, resulting in a relatively unbiased but highly variable estimate. By contrast, large values of h average over many data points, resulting in an estimate with relatively low variance, but potentially large bias, as observations which are quite distant from x are included in the average. Since the number of data points included in the average is proportional to nh, each of these estimators has pointwise variance proportional to $1/(nh)$. For these reasons, we require that as $n \to \infty$, $h \to 0$ in such a way that $nh \to \infty$.

It is well known that the asymptotic bias of such nonparametric regression estimators is proportional to h^p, where p depends on the smoothness of r, the smoothness of the marginal density of the X_i, and the properties of the kernel, or in the case of local polynomials, the polynomial degree of the local fit. In this paper we show that through appropriate choice of kernel, the rate at which the bias converges to zero will only be limited by properties of the unknown function, and not the kernel.

Sections 2 and 3 contain some important definitions and background. The case where the pairs of data $(X_1,Y_1),...,(X_n,Y_n)$ are i.i.d. will be studied in Section 4; the case where the data satisfy strong mixing conditions will be presented in Section 5; a small simulation study is presented in Section 6. Technical proofs have been placed in Section 7.

2. INFINITE ORDER KERNELS

If the kernel K has finite moments up to order q and its first $q-1$ moments are 0, then K is said to be of *order* q. The most frequently used kernels are second order; common examples include the Epanechnikov kernel, $K_e(x) := (3/4)\,(1-x^2)\,1_{[-1,1]}(x)$, and the scaled normal density.

In general, if r is k times differentiable, with $k \geq 2$, the bias of a second order kernel estimate is $O(h^2)$. This rate of convergence can be improved up to $O(h^k)$ by choosing a kernel of order greater than or equal to k. However, the degree of smoothness in the underlying function is unknown and difficult to estimate, so it is difficult to know what order kernel to use.

In order to alleviate this difficulty we focus on a class of kernels that effectively have infinite order. These kernels automatically reduce the bias to $o(h^k)$ no matter how large k happens to be. As in Politis and Romano [18, 19, 20] and Politis [16, 17], we now state the following general definition.

Definition 2.1. A general flat-top kernel K is defined in terms of its Fourier transform λ, which in turn is defined as follows. Fix a constant $c > 0$. Let

$$\lambda(s) \;=\; \begin{cases} 1 & \text{if } |s| \leq c\,, \\ g\big(|s|\big) & \text{if } |s| > c\,, \end{cases} \tag{2.1}$$

where the function g is chosen to make $\lambda(s)$, $\lambda^2(s)$, and $s\,\lambda(s)$ integrable. The flat top kernel is now given by

$$K(x) \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty}\lambda(s)\,e^{-isx}\,ds\;, \tag{2.2}$$

i.e., the inverse Fourier transform of $\lambda(s)$.

Note that in the preceding definition, the choice of g is not unique. The function λ, and hence the kernel K, depend on the function g and the parameter c although this dependence will not be explicitly denoted.

Kernels satisfying this definition do not necessarily satisfy the moment conditions $\int z^k K(z)\,dz = 0$ for all integers k, as some of these integrals may not be defined in either the Riemann or Lebesgue sense. However, the Cauchy principal value of each of these integrals is zero, and in many cases this is sufficient for optimal asymptotic performance.

The simplest kernel satisfying Definition 2.1 is determined by

$$\lambda_D(s) \;=\; \begin{cases} 1 & \text{if } |s| \le 1\,, \\ 0 & \text{if } |s| > 1\;. \end{cases}$$

This is the example studied in the case of density estimation by Davis [2, 3], Devroye [4], and Ibragimov and Hasminksii [12], and it generates the Dirichlet kernel, $K(x) := \sin(x)/(\pi\, x)$. Both $\lambda(s)$ and the resulting kernel are shown in Figure 1. We can see that the tails of this kernel are very wiggly. This is problematic in two ways. First, the slow decay in the tails and the large negative oscillations increase $\int K^2(z)\, dz$, which will be shown to increase the variance of the estimate. Secondly, the large wiggles distant from 0 generate a finite sample bias because they allow observations which are relatively distant from x to have a substantial influence on the estimate at x. These difficulties make density estimators using this kernel relatively uncompetitive for all but extremely large sample sizes.

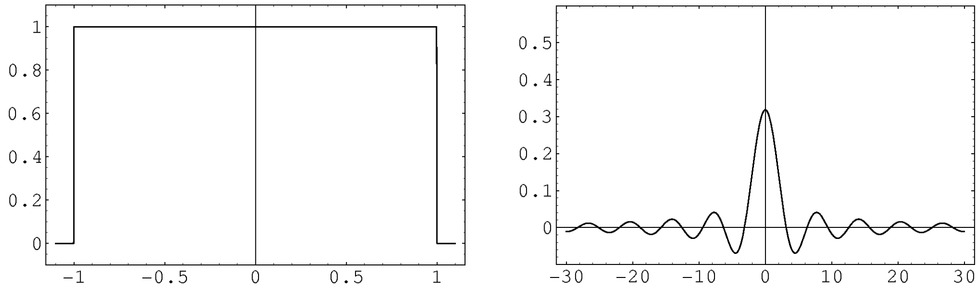

Figure 1: $\lambda(s)$ and the resulting Dirichlet kernel.

These problems can be substantially remedied by making the transition from 0 to 1 in the Fourier domain less abrupt. For example, Devroye and Gyorfi [5], Hall and Marron [11], and in the case of spectral density estimation, Politis and Romano [19], studied the kernel whose Fourier transform is given by

$$\lambda_{T,1/2}(s) \;=\; \begin{cases} 1 & \text{if } |s| \le 1/2\,, \\ 2\left(1-|s|\right) & \text{if } 1/2 < s \le 1\,, \\ 0 & \text{if } |s| > 1\;. \end{cases} \tag{2.3}$$

The corresponding kernel is

$$K(x) \;=\; \frac{2\left(\cos(x/2) - \cos(x)\right)}{\pi\, x^2}\;.$$

These are shown in Figure 2. Note the substantial improvement in the tails of the kernel.

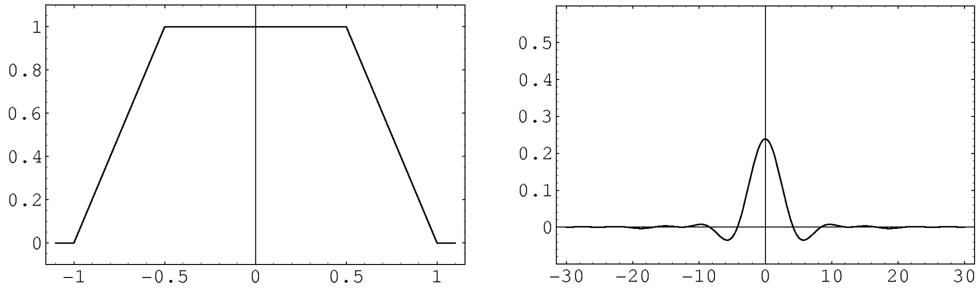

Figure 2: $\lambda(s)$ and the resulting improved kernel.

Unfortunately, in the case of regression, it is often only reasonable to assume that the function being estimated is smooth over some interval rather than over its entire domain; if the marginal density of the X_i has compact support, then the endpoints often generate discontinuities. Since infinite order kernels do not have compact support, the effects caused by these breaks get spread across the whole region of interest, potentially worsening the rate of convergence. For this reason, as discussed in the case of discontinuous density estimation in Politis [16], it is important that the tails of K decay as quickly as possible, to minimize the effect on the interior of the interval. This can be ensured by requiring the Fourier transform of the kernel to be very smooth. If λ is infinitely differentiable, then the tails of $K(x)$ decay faster than x^{-m} for any positive m. In addition, $\lambda(s)$ as defined in equation (2.1) clearly has an infinite number of zero derivatives at $s = 0$. Together, these two conditions ensure that all moments of K are zero in the Lebesgue sense. For these reasons, for the remainder of this work, we will restrict ourselves to kernels satisfying the following stronger definition.

Definition 2.2. An infinitely differentiable flat-top kernel K is a flat-top kernel (as in Definition 2.1) with the added caveat that the function g is chosen to make $\lambda(s)$ infinitely differentiable for all s.

We now provide an example of such a kernel, which was first introduced in McMurry and Politis [14], where the case of fixed design regression was studied. Let b and c be constants satisfying $b > 0$ and $0 < c < 1$. Define $\lambda(s)$ by

$$(2.4) \qquad \lambda_{IO}(s) = \begin{cases} 1 & \text{if } |s| \le c\,, \\ \exp\Big[-b\exp\Big[-b/\big(|s|-c\big)^2\Big]/\big(|s|-1\big)^2\Big] & \text{if } c < |s| < 1\,, \\ 0 & \text{if } |s| \ge 1\,. \end{cases}$$

The parameter c determines the region over which the kernel is identically 1; the parameter b allows the shape of λ to be altered, making the transition from 0 to 1 less abrupt. Figure 3 show plots of λ (as defined above) and the resulting kernel K for $c = 0.05$ and $b = 1/4$.

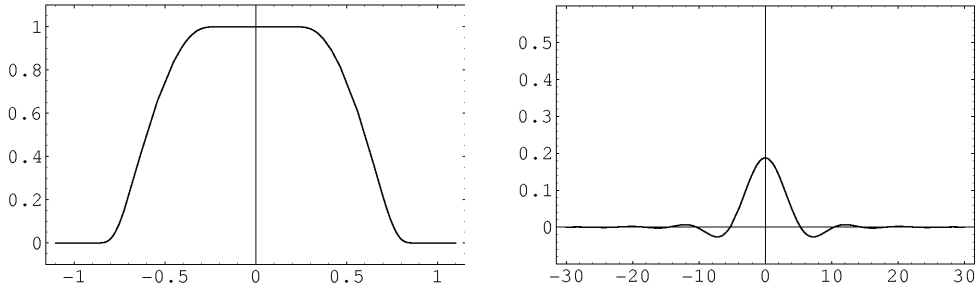

Figure 3: Smooth $\lambda(s)$ and the resulting kernel with $b = 1/4$ and $c = 0.05$.

The function $\exp\bigl[-b \exp[-b/(|s| - c)^2]/(|s| - 1)^2\bigr]$ was chosen because it connects the regions where λ is 0 and the region where λ is 1 in a manner such that $\lambda(s)$ is infinitely differentiable for all s, including where $|s| = c$, and $|s| = 1$.

3. BACKGROUND AND NOTATION

We examine the performance of the Nadaraya–Watson estimator when using infinite order kernels. The observed data is assumed to take the form of identically distributed pairs $(X_1, Y_1), ..., (X_n, Y_n)$, which satisfy $Y_i = r(X_i) + \epsilon_i$. Further restrictions will be necessary, but their discussion will be postponed for the moment. The Nadaraya–Watson estimator introduced in equation (1.1) can be written as

$$\hat{r}(x) := \frac{(1/n) \sum_{i=1}^{n} Y_i \, K_h(X_i - x)}{(1/n) \sum_{i=1}^{n} K_h(X_i - x)} ,$$

where

$$K_h(x) := (1/h) \, K(x/h) .$$

This estimator should be viewed in two different ways. As mentioned before, it is a weighted local average of the Y_i's, where the denominator normalizes the weights so they sum to 1. It is also an explicit estimator of conditional expectation. The denominator is the standard kernel estimate of the design density, the marginal density of the X_i's. The numerator is an approximation to $\int_{-\infty}^{\infty} y \, f(x, y) \, dy$, where $f(x, y)$ is the joint density of (X_i, Y_i). Put together, this is an approximation to $r(x) = E\bigl[Y|X = x\bigr]$.

In order to simplify notation, define

$$\hat{g}(x) := (1/n) \sum_{i=1}^{n} Y_i \, K_h(X_i - x)$$

and

$$\hat{f}(x) := (1/n) \sum_{i=1}^{n} K_h(X_i - x) \ ,$$

which are the finite sample approximations to

$$g(x) := \int_{-\infty}^{\infty} y \, f(x,y) \; dy$$

and

$$f(x) := \int_{-\infty}^{\infty} f(x,y) \; dy \ .$$

We note that $r(x) = g(x)/f(x)$.

4. THE INFINITE ORDER NADARAYA-WATSON ESTIMATOR FOR I.I.D. DATA

We first examine the behavior of the Nadaraya–Watson estimator when the observed pairs of data, $(X_1, Y_1), ..., (X_n, Y_n)$, are i.i.d. In order to understand the estimator as a whole, we begin with lemmas quantifying the asymptotic performance of the numerator and denominator, $\hat{g}(x)$ and $\hat{f}(x)$, as they approximate $g(x)$ and $f(x)$. In the process, it will be necessary to impose some assumptions, which will be introduced and discussed as needed. We first place some reasonable restrictions on the behavior of the bandwidth h as the sample size grows large and on the conditional distribution of the errors.

Assumption 1. As the sample size $n \to \infty$, the bandwidth $h \to 0$ in such a way that $nh \to \infty$.

Assumption 2. $E\big[\epsilon_i | X_i = x\big] = 0$, and $E\big[\epsilon_i^2 | X_i = x\big] := \sigma^2(x) < \infty$.

Under this assumption, $\hat{f}(x)$ and $\hat{g}(x)$ are infinite order estimators of $f(x)$ and $g(x)$; this is quantified in the following lemma.

Lemma 4.1. *If x is contained in an open interval on which $f(x)$ has p bounded continuous derivatives and $r(x)$ has q bounded continuous derivatives, then under Assumptions 1 and 2,*

$$E\big[\hat{f}(x)\big] - f(x) = o(h^p) \tag{4.1}$$

and

$$E\big[\hat{g}(x)\big] - g(x) \;=\; o(h^k) \;, \tag{4.2}$$

where $k = \min\{p, q\}$. If both $f(x)$ and $g(x)$ are infinitely differentiable, then each of these biases become $o(h^m)$ for all positive real m.

If we impose the additional assumptions that the observed pairs of data are i.i.d., and that f, g, and $\sigma^2(x)$ are reasonably well behaved, then the variance of $\hat{f}$ and $\hat{g}$ also behaves as expected.

Assumption 3. $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d.

The next assumption is necessary to ensure that the asymptotic approximations we use are valid, and to avoid division by zero.

Assumption 4. The point x is a continuity point of $\sigma^2(x)$, $f(x) > C$ for some $C > 0$, and r and f are each differentiable in a neighborhood of x.

Lemma 4.2. *Under Assumptions 1–4,*

$$\operatorname{var}\big[\hat{f}(x)\big] \;=\; \frac{f(x)}{nh}\int_{-\infty}^{\infty} K^2(z)\,dz \;+\; o\Big(\frac{1}{nh}\Big) + O\Big(\frac{1}{n}\Big)\,, \tag{4.3}$$

$$\operatorname{var}\big[\hat{g}(x)\big] \;=\; \frac{\Big(r^2(x) + \sigma^2(x)\Big)\,f(x)}{nh}\int_{-\infty}^{\infty} K^2(z)\,dz \;+\; o\Big(\frac{1}{nh}\Big) + O\Big(\frac{1}{n}\Big) \tag{4.4}$$

and

$$\operatorname{cov}\big[\hat{f}(x), \hat{g}(x)\big] \;=\; \frac{r(x)\,f(x)}{nh}\int_{-\infty}^{\infty} K^2(z)\,dz \;+\; o\Big(\frac{1}{nh}\Big) + O\Big(\frac{1}{n}\Big)\,. \tag{4.5}$$

Now that the behaviors of $\hat{f}$ and $\hat{g}$ are understood independently, the analysis will proceed by establishing their joint asymptotic normality. Once this has been shown, a Taylor series argument can be employed to show that $\hat{r}$ also has an asymptotic normal distribution with optimal bias and the standard variance. The joint asymptotic normality of $\hat{f}$ and $\hat{g}$ will be established via the Liapunov condition, which implies the Lindeberg–Feller central limit theorem. This requires a uniform bound on the $2+\delta$'th moments of the Y_i, for some $\delta > 0$.

Assumption 5. There exists a positive constants M and δ such that

$$E\Big[|Y_i|^{2+\delta}\,\big|X_i = x\Big] < M \;,$$

for all x.

The final assumption forces the conditional variance of the errors to be bounded above and below for all x. The bound from below is assumed for technical simplicity.

Assumption 6. There exist strictly positive constants b and B such that $b < \sigma^2(x) < B$ for all x.

Lemma 4.3. *Under Assumptions 1–6, for all real c_1 and c_2 (not both zero),*

$$\sqrt{nh}\left[c_1\Big(\hat{f}(x) - E\big[\hat{f}(x)\big]\Big) + c_2\Big(\hat{g}(x) - E\big[\hat{g}(x)\big]\Big)\right] \xrightarrow{\mathcal{D}} N\big(0, \theta(x)\big)\,. \tag{4.6}$$

where $\theta(x) := \Big(c_1^2 + 2\,c_1 c_2\, r(x) + c_2^2\big[r^2(x) + \sigma^2(x)\big]\Big)\, f(x) \int_{-\infty}^{\infty} K^2(z)\,dz$. *This implies the joint asymptotic normality of $\hat{f}$ and $\hat{g}$.*

The consequence of the preceding Lemma is the asymptotic normality of our estimator.

Theorem 4.1. *If x is contained in an open interval on which $f(x)$ has p bounded continuous derivatives and $r(x)$ has q bounded continuous derivatives, then under Assumptions 1–6,*

$$\sqrt{nh}\,\Big(\hat{r}(x) - r(x) + o(h^k)\Big) \xrightarrow{\mathcal{D}} N\bigg(0\,,\, \frac{\sigma^2(x)}{f(x)} \int_{-\infty}^{\infty} K^2(z)\,dz\bigg)\,, \tag{4.7}$$

where $k = \min\{p, q\}$.

Remark 4.1. Letting h proportional to $n^{-1/(2k+1)}$, the mean square optimal rate, we get $\hat{r}(x) = r(x) + O_p\big(n^{-k/(2k+1)}\big)$, and $\sqrt{nh}\big(\hat{r}(x) - r(x)\big) \xrightarrow{\mathcal{D}} N\Big(0\,,\, \frac{\sigma^2(x)}{f(x)} \int_{-\infty}^{\infty} K^2(z)\,dz\Big)$, which demonstrates the higher order accuracy provided by infinite order kernels.

5. DEPENDENT DATA AND NONPARAMETRIC AUTOREGRESSION

It is desirable to weaken the condition that $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d. In particular we wish to be able to estimate nonparametric autoregression, where X_t may be an unknown function of X_{t-1}. Mathematically, autoregressive processes are assumed to satisfy a model of the form, $X_t = r(X_{t-1}) + \sigma(X_{t-1})\,\epsilon_t$, where r and σ are unknown, and the ϵ_t are mean zero errors; further restrictions similar to those in Section 4 will be imposed as necessary. The problem of interest is the estimation of $r(x) := E\big[X_t|X_{t-1} = x\big]$; this can be done by pairing consecutive observations, $(X_1, X_2), (X_2, X_3), ..., (X_{n-1}, X_n)$, and then performing a standard nonparametric regression.

Since nothing is gained by restricting ourselves to the case of autoregression, we will study the infinite order kernel estimator in the case where $(X_1, Y_1), ..., (X_n, Y_n)$ satisfy the same type of asymptotic dependence conditions that we wish the autoregressive process to satisfy. The results in this more general situation will then imply the desired result in the specific case of interest.

Any meaningful analysis will require conditions that ensure some sort of asymptotic independence; that is, random samples at times which are very distant from each other should behave as if they are independent. We will focus on the case of α-mixing because it is the weakest of the most commonly studied conditions.

Definition 5.1. Let $\mathcal{F}_l^m$ be the σ-field generated by $U_l, U_{l+1}, ..., U_m$. A stationary time series $\{U_n\}_{n\in\mathbb{Z}}$ is said to be α-mixing (or strong-mixing), if

$$\sup_{k\in\mathbb{Z}} \sup_{A\in\mathcal{F}_{-\infty}^k, B\in\mathcal{F}_{k+i}^\infty} \Big|P(A)\,P(B) - P(AB)\Big| := \alpha(i) \;\rightarrow\; 0 \ ,$$

as $i \to \infty$. The $\alpha(i)$'s are called the α-mixing coefficients.

The analysis proceeds as in the i.i.d. case. We begin by proving, under some conditions, the joint asymptotic normality of $\hat{f}(x)$ and $\hat{g}(x)$. Once this has been established, we will be able to use the same argument used in the proof of Theorem 4.1 to show the asymptotic normality of $\hat{g}/\hat{f}$. Similar results for local polynomial regression and for Nadaraya–Watson estimators with finite order kernels have been obtained by Masry and Fan [13] and Robinson [21] respectively. Although their arguments are similar, the central limit theorem we prove here will be more closely related to that of Masry and Fan [13]. Let c_1 and c_2 be real numbers (not both zero). Define

$$Z_i \ := \ c_1\Big[K_h(X_i - x) - E\big[K_h(X_i - x)\big]\Big] + c_2\Big[Y_i\,K_h(X_i - x) - E\big[Y_i\,K_h(X_i - x)\big]\Big]$$

and

$$Q_n := \frac{1}{n}\sum_{i=1}^{n} Z_i \ .$$

We will establish asymptotic normality for $\sqrt{nh}\,Q_n$. In order to do so, we impose further assumptions on the marginal distributions of (X_1, Y_1), and on the α-mixing coefficients associated with the time series defined by $U_i = (X_i, Y_i)$.

Assumption 7. There exist finite positive bounds M_1, M_2, and M_3, such that

(i) $f_i(u,v) \leq M_1$, where $f_i(u,v)$ is the joint density of (X_1, X_i);

(ii) $E\big[Y_1^2 + Y_i^2 | X_1,\, X_i\big] \leq M_2$;

(iii) There exists $\delta > 2$ and $\beta > 1 - 2/\delta$ such that $E\big[|Y_1|^\delta \, |X\big] \leq M_3$ and $\sum_{i=1}^{\infty} i^\beta [\alpha(i)]^{1-2/\delta} < \infty$.

Lemma 5.1. *Let x be a continuity point of conditional mean and variance functions, $r(\cdot)$ and $\sigma^2(\cdot)$. In addition, suppose that the marginal density of the X_i, $f(\cdot)$, and the product of $r(\cdot)$ and $f(\cdot)$, $g(\cdot)$, have k bounded continuous derivatives in a neighborhood of x, where $k \geq 1$. Under Assumptions 2, 4, 6, and 7, we have the following convergences as $n \to \infty$, $h \to 0$, and $nh \to \infty$:*

(**a**) $h \operatorname{var}[Z_1] \to \theta(x)$,

(**b**) $h \sum_{i=1}^{n-1} \big|\operatorname{cov}[Z_1, Z_{i+1}]\big| \to 0$,

(**c**) $nh \operatorname{var}[Q_n] \to \theta(x)$,

where, as before, $\theta(x) := \Big(c_1^2 + 2\, c_1 c_2\, r(x) + c_2^2 \big[r^2(x) + \sigma^2(x)\big]\Big)\, f(x) \int_{-\infty}^{\infty} K^2(z)\, dz.$

In order to establish the central limit theorem, we need one final condition on the α-mixing coefficients.

Assumption 8. There exists a sequence of positive integers satisfying $s_n \to \infty$ and $s_n = o(\sqrt{nh})$ such that $\sqrt{n/h}\; \alpha(s_n) \to 0$.

Remark 5.1. Assumption 8 is a technical assumption that may dictate some particular rates at which $h \to 0$; nevertheless, Assumption 8 is weak enough to allow for a wide range of useful rates. To elaborate, Assumption 7 (iii) requires that the mixing coefficients decay at a polynomial rate depending on δ. In particular, it requires that there exist $C > 0$, $\epsilon > 0$, and $n_0 > 0$ such that for all $n > n_0$, $\alpha(n) < Cn^{-(\delta/(\delta-2)+1+\epsilon)}$. For example, if $\delta = 3$, then the mixing coefficients need to decay slightly faster than n^{-4}. Assuming that the function being estimated is at least twice differentiable, h will optimally decrease at a rate equal to or slower than $n^{-1/5}$. This means that $\sqrt{n/h} \leq Cn^{3/5}$ for some constant C. Similarly, $\sqrt{nh} \geq Cn^{2/5}$. Under the strongest moment assumptions, $\alpha(n)$ is required to decay faster than n^{-2}. If we put these together, $\sqrt{n/h}\; \alpha(s_n) \leq Cn^{3/5} s_n^{-2}$. From this expression it is easily seen that as long as s_n grows faster than $n^{3/10}$, the second requirement of Assumption 8 will be satisfied. By the preceding argument the first requirement is satisfied if $s_n = o(n^{2/5})$. Since these conditions can be met simultaneously, Assumption 8 generally imposes no additional restrictions.

Lemma 5.2. *Under Assumptions 2, 4, and 6–8, we have as* $n \to \infty$, $h \to 0$, *and* $nh \to \infty$,

$$\sqrt{nh}\, Q_n \xrightarrow{\mathcal{D}} N\big(0, \theta(x)\big)\,.$$

The immediate consequence of this result is the following asymptotic normality for $\hat{r}$.

Theorem 5.1. *If x is contained in an open interval on which $f(x)$ has p bounded continuous derivatives and $r(x)$ has q bounded continuous derivatives, then under Assumptions 1, 2, 4, and 6–8,*

$$\sqrt{nh}\,\Big(\hat{r}(x) - r(x) + o(h^k)\Big) \;\xrightarrow{\mathcal{D}}\; N\bigg(0\,,\, \frac{\sigma^2(x)}{f(x)} \int_{-\infty}^{\infty} K^2(z)\,dz\bigg)\,, \tag{5.1}$$

where $k = \min\{p, q\}$.

6. SIMULATIONS

An extensive simulation study was undertaken to investigate the performance of the proposed estimator. For each combination of regression function, design density, error variance, and sample size, 100 data sets were created and smoothed using the infinite order and local linear estimators. Finally the integrated square error was estimated using Simpson's rule. Bandwidths for the infinite order estimator were selected by the rule of thumb suggested in [14] and developed further in [17]. Bandwidths for the local linear estimator were selected using the direct plug-in method suggested by Ruppert, Sheather, and Wand [22] and implemented in the R package KernSmooth [24].

The first regression function was taken to be $r(x) = x + 4\exp(-2x^2)/\sqrt{2\pi}$, which includes sections of almost linear behavior and an exponential bump with more curvature. Design densities were uniform on $[-2,2]$ and $N(0,1)$, and the integrated square error is over the interval $[-2,2]$. The resulting integrated square errors for one simulation are shown in Figure 4. A scatterplot along with the two smoothings is shown in Figure 5.

The second regression function was taken to be $r(x) = \sin(4\pi x)$ with uniform design density on $[0,1]$. The integrated square error was calculated on both the entire interval $[0,1]$ and over the interval $[0.15, 0.85]$ to exclude edge effects.

It is clear from the simulations that the two estimators have different strengths. The infinite order estimator is clearly superior in the interior of the data set, when the error variance is large, and when the sample size is moderate to large. Since the local linear estimator automatically adapts to the edges of the design and the infinite order estimator does not, it is unsurprising that the local linear estimator is superior in these regions.

Table 1: Comparison of infinite order and local linear estimators.

Function	Design	n	σ	Median Integrated Square Error	
				Infinite Order	Local Linear
Exponential	Normal	100	0.3	0.0870	0.0631
			0.5	0.1312	0.1627
			0.7	0.2101	0.2521
Exponential	Normal	200	0.3	0.0337	0.0324
			0.5	0.0633	0.0686
			0.7	0.1091	0.1375
Exponential	Normal	1000	0.3	0.0065	0.0077
			0.5	0.0132	0.0166
			0.7	0.0204	0.0274
Exponential	Uniform	100	0.3	0.0520	0.0384
			0.5	0.0954	0.1613
			0.7	0.0813	0.1481
Exponential	Uniform	200	0.3	0.0251	0.0190
			0.5	0.0481	0.0474
			0.7	0.0731	0.0823
Exponential	Uniform	1000	0.3	0.0066	0.0051
			0.5	0.0110	0.0112
			0.7	0.0175	0.0189
Sin	Uniform	100	0.3	0.0065	0.0052
	(edges excluded)		0.5	0.0135	0.0139
			0.7	0.0232	0.0228
Sin	Uniform	100	0.3	0.0120	0.0077
	(edges included)		0.5	0.0221	0.0191
			0.7	0.0412	0.0333
Sin	Uniform	200	0.3	0.0032	0.0031
	(edges excluded)		0.5	0.0069	0.0068
			0.7	0.0108	0.0126
Sin	Uniform	200	0.3	0.0066	0.0042
	(edges included)		0.5	0.0115	0.0091
			0.7	0.0191	0.0183
Sin	Uniform	1000	0.3	0.0007	0.0008
	(edges excluded)		0.5	0.0013	0.0019
			0.7	0.0021	0.0030
Sin	Uniform	1000	0.3	0.0029	0.0011
	(edges included)		0.5	0.0037	0.0026
			0.7	0.0050	0.0040

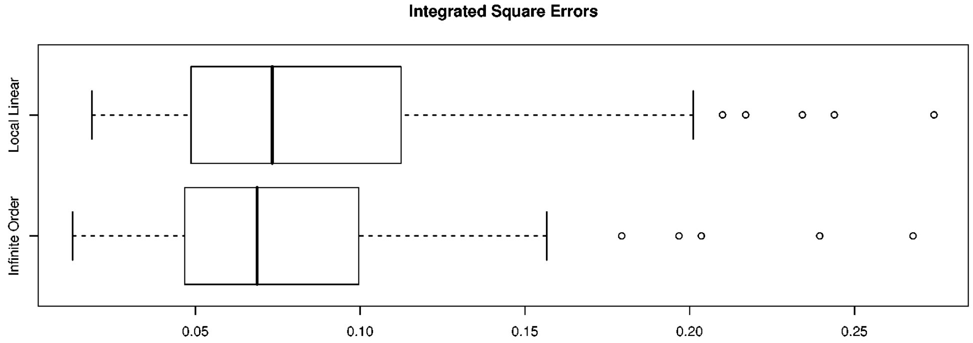

Figure 4: Comparison of integrated square errors for 100 simulations of the exponential function with $n = 200$ and $\sigma = 0.5$.

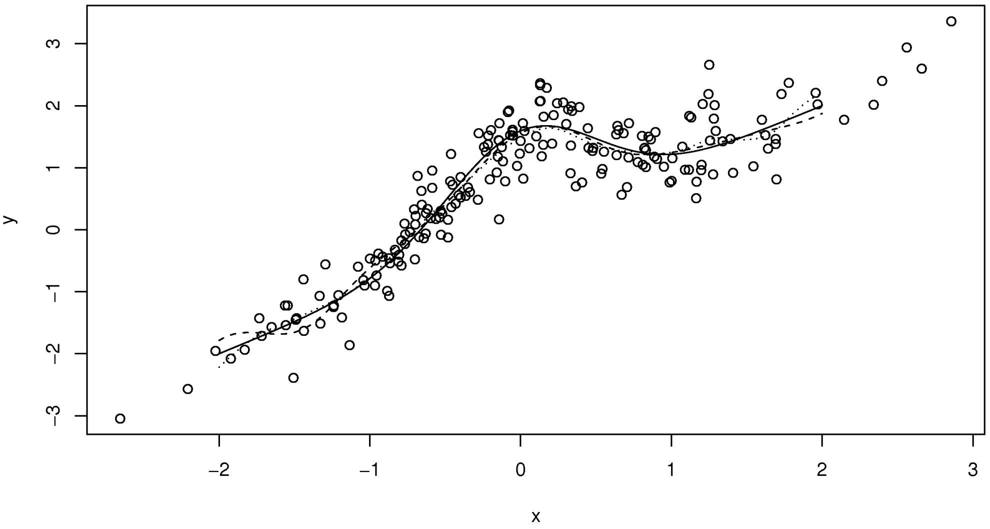

Figure 5: A sample regression. The solid line is the true function, the dashed line is the infinite order estimate, and the dotted line is the local linear estimate.

7. TECHNICAL PROOFS

Proof of Lemma 4.1: The proof of (4.2) is almost identical to, but slightly more complicated than the proof of (4.1). For this reason, only (4.2) will be shown. The proof is similar in spirit to the proof of Theorem 2 in McMurry and Politis [14], except this time it requires both r and f to be smooth. It will be proved using a different, although more standard, technique. This method of proof is somewhat less elegant than the Fourier transform method used previously, but it has the advantage of showing that the same convergence rates hold even if f and g are only smooth on an open interval containing x. This proof technique does have a slight disadvantage. If f and g are smooth over all $\mathbb{R}$, then the Fourier transform technique can be employed to show the same convergence rates hold even if $\lambda(s)$, the Fourier transform of K, is not smooth. By conditioning on X_i,

$$\begin{aligned} E\big[\hat{g}(x)\big] - g(x) &= E\Big[E\big[Y_i\, K_h(X_i - x)\,\big|\, X_i\big]\Big] - g(x) \\ &= E\Big[r(X_i)\, K_h(X_i - x)\Big] - g(x) \\ &= \int_{-\infty}^{\infty} r(u)\, f(u)\, K_h(u-x)\, du \;-\; g(x)\ . \end{aligned}$$

Suppose (rf) has k bounded continuous derivatives on an interval (a,b) containing x. Should (rf) be smooth over all $\mathbb{R}$, then the proof can be simplified by taking $(a,b) = (-\infty,\infty)$:

$$\begin{aligned} E\big[\hat{g}(x)\big] - g(x) \;=\; & \int_a^b r(u)\, f(u)\, K_h(u-x)\, du + \int_{-\infty}^{a} r(u)\, f(u)\, K_h(u-x)\, du \\ & + \int_b^{\infty} r(u)\, f(u)\, K_h(u-x)\, du \;-\; g(x)\ . \end{aligned}$$

Since the tails of $K(x)$ decay faster than x^{-m} for all positive m, the two error terms are $o(h^m)$ for all positive m. At this point we perform a Taylor series expansion of the product $(rf)(z)$ around x:

$$\begin{aligned} & E\big[\hat{g}(x)\big] - g(x) \;= \\ & \quad = \int_a^b r(u)\, f(u)\, K_h(u-x)\, du \;-\; g(x) \;+\; o(h^m) \\ & \quad = \int_{(a-x)/h}^{(b-x)/h} r(x+hv)\, f(x+hv)\, K(v)\, dv \;-\; g(x) \;+\; o(h^m) \\ & \quad = \int_{(a-x)/h}^{(b-x)/h} \left[(rf)(x) + hv\,(rf)'(x) + \cdots + \frac{(hv)^k}{k!}\,(rf)^{(k)}\,(x+\xi)\right] K(v)\ dv \\ & \qquad - \;g(x) \;+\; o(h^m)\ , \end{aligned}$$

where ξ is between x and $x + hv$. Since K integrates to one, its moments are zero, and since $g(x) = r(x)\, f(x)$,

$$\begin{aligned} & E\big[\hat{g}(x)\big] - g(x) \;= \\ & \quad = \int_{-\infty}^{\infty} \left[(rf)(x) + hv\,(rf)'(x) + \cdots + \frac{(hv)^{k-1}}{(k-1)!}\,(rf)^{(k-1)}(x)\right] K(v)\ dv \\ & \qquad + \int_{(a-x)/h}^{(b-x)/h} \frac{(hv)^k}{k!}\,(rf)^{(k)}(x+\xi)\, K(v)\ dv \;-\; g(x) \;+\; o(h^m) \\ & \quad = \int_{(a-x)/h}^{(b-x)/h} \frac{(hv)^k}{k!}\,(rf)^{(k)}(x+\xi)\, K(v)\ dv \;+\; o(h^m)\ . \end{aligned}$$

Since (rf) has bounded continuous derivatives on (a,b), we can apply the dominated convergence theorem, yielding

$$\lim_{h\to 0}\int_{(a-x)/h}^{(b-x)/h}(rf)^{(k)}(x+\xi)\,v^k\,K(v)\,dv \;=\; \int_{-\infty}^{\infty}(rf)^{(k)}(x)\,v^k\,K(v)\,dv \;=\; 0\;.$$

Therefore,

$$E\big[\hat{g}(x)\big]-g(x) \;=\; o(h^k)\;. \qquad \square$$

Proof of Lemma 4.2: As $f(x)$ can be viewed as a special case of $g(x)$ with $Y_i=1$ for all i, (4.3) will follow immediately from the proof of (4.4). The proof of (4.5) is almost identical to the proof of (4.4), so it is omitted.

By conditioning on X_1, and by Lemma 4.1,

$$\begin{aligned}\operatorname{var}\big[\hat{g}(x)\big] &= \frac{1}{n}\operatorname{var}\big[Y_1K_h(X_1-x)\big]\\ &= \frac{1}{n}\left[\int_{-\infty}^{\infty}\Big(r^2(u)+\sigma^2(u)\Big)f(u)\,K_h^2(u-x)\,du\;-\;r^2(x)\;+\;o(h^k)\right]\\ &= \frac{1}{nh}\int_{-\infty}^{\infty}\Big(r^2(x+hz)+\sigma^2(x+hz)\Big)f(x+hz)\,K^2(z)\,dz\;+\;O\Big(\frac{1}{n}\Big)\,,\end{aligned}$$

since x is a continuity point of r, f, and σ^2, the dominated convergence theorem yields

$$\operatorname{var}\big[\hat{g}(x)\big] \;=\; \frac{\Big(r^2(x)+\sigma^2(x)\Big)f(x)}{nh}\int_{-\infty}^{\infty}K^2(z)\,dz\;+\;o\Big(\frac{1}{nh}\Big)\;+\;O\Big(\frac{1}{n}\Big)\,. \qquad \square$$

Proof of Lemma 4.3: The proof proceeds by verifying that the Liapunov condition holds, which is sufficient for the Lindeberg–Feller central limit theorem. The result is trivial if $c_1=c_2=0$, so assume that at least one of these constants is nonzero. Let C denote a positive constant.

$$\begin{aligned}&\frac{\sum_{i=1}^{n}E\Big[\big|(c_1+c_2Y_i)\,K_h(X_i-x)\big|^{2+\delta}\Big]}{\operatorname{var}\Big[c_1\sum_{i=1}^{n}K_h(X_i-x)\;+\;c_2\sum_{i=1}^{n}Y_iK_h(X_i-x)\Big]^{(2+\delta)/2}}\;\leq\\ &\qquad\leq\;\frac{Cnh^{1+\delta}}{n^{1+\delta/2}\Big[\theta(x)+o(1/h)+O(1)\Big]^{(2+\delta)/2}}\;,\end{aligned}$$

where the inequality follows from the proof of Lemma 4.2. After multiplying the numerator and denominator by $h^{(2+\delta)/2}$, it is clear that this quantity goes to zero as n goes to infinity. Therefore, the Liapunov condition is satisfied, and the Lemma follows immediately. $\square$

Proof of Theorem 4.1: We begin with a lemma which ensures that for large enough n, $\hat{f}(x) \geq c > 0$ for some constant c, as long as $f(x) > 0$.

Lemma 7.1. *Suppose $f(x) > c$ for some $c > 0$. Also suppose h decreases slowly enough that $n^{-2/7} + \delta = o(h)$ for some $\delta > 0$. Then for all $\epsilon > 0$, $P\big[|f(x) - \hat{f}(x)| > \epsilon \text{ i.o.}\big] = 0$.*

It should also be noted that much stronger results hold. Under additional conditions, rates of uniform almost sure convergence over compact sets can be established. See Bosq [1] or Györfi *et al.* [9].

Proof of Lemma 7.1: We make use of the following Bernstein inequality, which is Theorem 1.3 part (2) in Bosq [1].

Lemma 7.2. *Let $\{W_t\}_{t\in\mathbb{Z}}$ be a mean zero real valued random process such that $\sup_{1\leq t\leq n} \|W_t\|_\infty \leq b$, and let $S_n = \sum_{t=1}^n W_t$. Then for each integer $q \in \lfloor 1, n/2 \rfloor$ and each $\epsilon > 0$,*

$$\mathrm{P}\big[|S_n| > n\epsilon\big] \;\leq\; 4\exp\!\left(-\frac{\epsilon^2}{8\,v^2(q)}\,q\right) + 22\left(1 + \frac{4\,b}{\epsilon}\right)^{1/2} q\,\alpha\!\left(\left\lfloor \frac{n}{2\,q} \right\rfloor\right),$$

where $v^2(q) = \frac{2}{p^2}\,\sigma^2(q) + \frac{b\epsilon}{2}$, $p = \frac{n}{2q}$, and

$$\begin{aligned}\sigma^2(q) \;=\; \max_{0\leq j\leq 2q-1} E\Big[&\Big(\lfloor jp \rfloor + 1 - jp\Big) X_{\lfloor jp \rfloor + 1} + X_{\lfloor jp \rfloor + 2} + \cdots \\ &+ X_{\lfloor (j+1)p \rfloor} + \Big[(j+1)p - \big\lfloor (j+1)p \big\rfloor\Big] X_{\lfloor (j+1)p+1 \rfloor}\Big].\end{aligned}$$

The Borel–Cantelli lemma will be used to show

$$\mathrm{P}\Big[\big|\hat{f}_n(x) - E\hat{f}_n(x)\big| > \epsilon \text{ i.o.}\Big] \;=\; 0\,.$$

Since $E\hat{f}_n(x) \to f(x)$, this will establish the desired result.

Let $W_i = K_h(X_i - x) - E\big[K_h(X_i - x)\big]$. Then $\|W_i\|_\infty \leq \hat{K}/h$ for all i, where $\hat{K} = 2\sup_{x\in\mathbb{R}} K(x)$. In addition, it can easily be seen that $\sigma^2(q) \leq \frac{(p+1)^2}{n^2h^2}\,\hat{K}^2$.

Therefore,

$$\begin{aligned} p_n \;&:=\; \mathrm{P}\left[\frac{1}{n}\left|\sum_{i=1}^n W_i\right| > \epsilon\right] \\ &\leq\; 4\,\exp\!\left(-\frac{C_1\,\epsilon^2}{\frac{(p+1)^2}{n^2\,h^2\,p^2} + \frac{1}{h}}\,q\right) + C_2\left(1 + \frac{C_3}{h\epsilon}\right)^{1/2} q\,\alpha\!\left(\left\lfloor \frac{n}{2\,q} \right\rfloor\right). \end{aligned}$$

We need $\sum_n p_n < \infty$. In order for the first term in the sum to be fininte, it is necessary that $n^2 h^2 q/(n^2 h) \to \infty$ at a rate equal to or faster than n^{δ_1} for some $\delta_1 > 0$; this requires q grow at least as fast as $n^{\delta_1} h^{-1}$. On the other hand, the second term requires $\sum_n (q/\sqrt{h})\, \alpha(\lfloor n/2q \rfloor) < \infty$. As noted in the Remark 5.1, it is sufficient to choose q such that $\sum_n (q/\sqrt{h})(q/n)^2 < \infty$. The latter condition can be satisfied if $q^3 h^{-1/2}$ grows at a rate $n^{1-\delta_2}$, for some $\delta_2 > 0$. Equivalently, it suffices for q to grow at a rate equal to or slower than $n^{(1/3)-\delta_2} h^{1/6}$. It can easily be seen that these requirements can be met simultaneously as long as h satisfies $n^{-2/7+\delta} = o(h)$, for some $\delta > 0$; this includes all optimal rates. This result could be further strengthened by imposing additional assumptions. For example, in the case where $X_1, ..., X_n$ are i.i.d., the mixing coefficients are 0, and hence summable. In this situation, the only restriction on h is that it decrease slightly slower than $1/n$. In the case of a mixing process, the possible range of rates for h could be expanded if one were to assume that the joint density of (X_1, X_i) is differentiable with partial derivatives uniformly bounded in i. □

We now return to the proof of the main result. By the preceding lemma, for large enough n, we can assume that $\hat{f}_n(x) > c/2$. So, we can apply the intermediate value theorem to see,

$$\begin{aligned} \hat{r}(x) - r(x) &= \frac{\hat{g}(x)}{\hat{f}(x)} - \frac{g(x)}{f(x)} \\ &= \hat{g}(x) \left(\frac{1}{f(x)} - \frac{1}{\xi_n^2} \Big(\hat{f}(x) - f(x) \Big) \right) - \frac{g(x)}{f(x)} \ , \end{aligned}$$

where $|\xi_n - f(x)| \le |\hat{f}(x) - f(x)|$. This can be further simplified to

$$\begin{aligned} \hat{r}(x) - r(x) &= \frac{1}{f(x)} \Big(\hat{g}(x) - g(x) \Big) - \frac{\hat{g}(x)}{\xi_n^2} \Big(\hat{f}(x) - f(x) \Big) \\ &= \frac{1}{f(x)} \Big(\hat{g}(x) - E\big[\hat{g}(x)\big] \Big) - \frac{\hat{g}(x)}{\xi_n^2} \Big(\hat{f}(x) - E\big[\hat{f}(x)\big] \Big) + o(h^k) \ . \end{aligned}$$

By Lemmas 4.1 and 4.2, $\hat{g}(x)$ and $\hat{f}(x)$ converge in probability to $g(x)$ and $f(x)$ respectively. Therefore, ξ_n also converges to $f(x)$ in probability. By Slutsky's theorem, and Lemma 4.3,

$$\sqrt{nh} \Big(\hat{r}(x) - r(x) + o(h^k) \Big) \ \xrightarrow{\mathcal{D}} \ N\left(0, \, \frac{\sigma^2(x)}{f(x)} \int_{-\infty}^{\infty} K^2(z)\, dz \right) ,$$

the desired result. □

Proof of Lemma 5.1: The proof of part (a) follows from similar results for the i.i.d. case:

$$\begin{aligned} \operatorname{var}\big[Z_1\big] &= E\Big[(c_1 + c_2 Y_1)^2 K_h^2(X_1 - x) \Big] - \Big(c_1 f(x) + c_2\, r(x) f(x) + O(h^k) \Big)^2 \\ &= E\left[\Big(c_1^2 + 2\, c_1 c_2\, r(X_1) + c_2^2 \big[r^2(X_1) + \sigma^2(X_1) \big] \Big) K_h^2(X_1 - x) \right] + O(1) \\ &= \theta(x)/h + o(1/h) + O(1) \ . \end{aligned}$$

The proof of part (b) is more challenging. Let d_n be a sequence of integers such that $d_n \to \infty$ and $d_n h \to 0$. Define

$$J_1 := \sum_{i=1}^{d_n-1} \big|\mathrm{cov}[Z_1, Z_{i+1}]\big| \,,$$

and

$$J_2 := \sum_{i=d_n}^{n-1} \big|\mathrm{cov}[Z_1, Z_{i+1}]\big| \,.$$

We wish to show $J_1 = o(1/h)$ and $J_2 = o(1/h)$. We begin with J_1. By conditioning on (X_1, X_i),

$$\begin{aligned}
&\big|\mathrm{cov}[Z_1, Z_i]\big| \le \\
&\quad\le \Big|E\Big[(c_1 + c_2 Y_1)\, K_h(X_1 - x)\, (c_1 + c_2 Y_i)\, K_h(X_i - x)\Big]\Big| + O(1) \\
&\quad\le E\Big[\big|K_h(X_1 - x)\, K_h(X_i - x)\big|\, E\Big[\big|(c_1 + c_2 Y_1)\, (c_1 + c_2 Y_i)\big| \Big| X_1, X_i\Big]\Big] + O(1) \\
&\quad\le E\Big[\big|K_h(X_1 - x)\, K_h(X_i - x)\big| \\
&\qquad \times \Big(E\Big[(c_1 + c_2 Y_1)^2 \,\big|\, X_1, X_i\Big]\, E\Big[(c_1 + c_2 Y_i)^2 \,\big|\, X_1, X_i\Big]\Big)^{1/2}\Big] + O(1) \\
&\quad\le C\, E\Big[\big|K_h(X_1 - x)\, K_h(X_i - x)\big|\Big] + O(1) \\
&\quad\le C \left(\int_{-\infty}^{\infty} \big|K_h(u - x)\big|\, du\right)^2 + O(1) \,.
\end{aligned}$$

Since the $O(1)$ term is the same for all i, $\big|\mathrm{cov}[Z_1, Z_i]\big| < C$ for some positive C. Therefore $J_1 = o(1/h)$. For the second term, J_2, we employ Davydov's Lemma (see Hall and Heyde [10]), which tells us

$$\big|\mathrm{cov}[Z_1, Z_{i+1}]\big| \le 8\, \big[\alpha(i)\big]^{1-2/\delta}\, \big[E|Z_1|^\delta\big]^{2/\delta} .$$

We now need to put a bound on $E|Z_i|^\delta$.

$$\begin{aligned}
E|Z_i|^\delta &= E\Big[\Big| c_1\Big[K_h(X_i - x) - E\big[K_h(X_i - x)\big]\Big] \\
&\qquad + c_2\Big[Y_i\, K_h(X_i - x) - E\big[Y_i\, K_h(X_i - x)\big]\Big]\Big|^\delta\Big] \\
&\le 2\, E\Big[\Big| c_1\Big[K_h(X_i - x) - E\big[K_h(X_i - x)\big]\Big]\Big|^\delta\Big] \\
&\quad + 2\, E\Big[\Big| c_2\Big[Y_i\, K_h(X_i - x) - E\big[Y_i\, K_h(X_i - x)\big]\Big]\Big|^\delta\Big] .
\end{aligned} \tag{7.1}$$

These two terms behave similarly, so it is sufficient to examine only the second. Let C denote a generic positive constant which may take on different values:

$$\begin{aligned} 2\,E\left[\left|c_2\Big[Y_i\,K_h(X_i-x) - E\big[Y_i\,K_h(X_i-x)\big]\Big]\right|^{\delta}\right] &\le \\ &\le\; 4\,E\Big[\big|c_2 Y_i\,K_h(X_i-x)\big|^{\delta}\Big] + 4\Big|E\big[Y_i\,K_h(X_i-x)\big]\Big|^{\delta} \\ &\le\; C\,E\Big[\big|K_h(X_i-x)\big|^{\delta}\,E\big[|Y_i|^{\delta}\big|X_i\big]\Big] + C \\ &\le\; C\,h^{1-\delta} + C\;. \end{aligned}$$

An identical result holds for the first term in equation (7.1). Putting these two terms together yields

$$\begin{aligned} \big[E\,|Z_1|^{\delta}\big]^{2/\delta} &\le\; \big[C\,h^{1-\delta} + C\big]^{2/\delta} \\ &\le\; C\,h^{2/\delta-2} + C\;. \end{aligned}$$

Returning to J_2,

$$\begin{aligned} J_2 &\le\; \sum_{i=d_n}^{\infty} 8\,\big[\alpha(i)\big]^{1-2/\delta}\,\big[E\,|Z_1|^{\delta}\big]^{2/\delta} \\ &\le\; \sum_{i=d_n}^{\infty} \big[\alpha(i)\big]^{1-2/\delta}\big(Ch^{2/\delta-2} + C\big)\;. \end{aligned}$$

By Assumption 7, $\sum_{i=d_n}^{\infty}[\alpha(i)]^{1-2/\delta} \to 0$ as $d_n \to \infty$. So,

$$\begin{aligned} J_2 &\le\; C\,h^{2/\delta-2} \sum_{i=d_n}^{\infty} \big[\alpha(i)\big]^{1-2/\delta} + o(1) \\ &\le\; C\,h^{2/\delta-2}\,d_n^{-\beta} \sum_{i=d_n}^{\infty} i^{\beta}\big[\alpha(i)\big]^{1-2/\delta} + o(1)\;. \end{aligned}$$

By choosing d_n such that $h^{2/\delta-1}d_n^{-\beta} \to 1$, we see $J_2 = o(1/h)$ and $hd_n \to 0$, which ensures the convergence of J_1. This completes the proof of (b). The proof of (c) is an immediate consequence of (a) and (b). □

Proof of Lemma 5.2: The proof employs a small-block large-block argument. The set $\{1,...,n\}$ is partitioned into $2k+1$ alternating large and small subsets. Let r_n be the size of the large blocks and s_n be the size of the small blocks. Then $k_n = \lfloor n/(r_n + s_n)\rfloor$. For $0 \le j \le k-1$, define

$$U_j \;:=\; \sqrt{h} \sum_{i=j(r+s)+1}^{j(r+s)+r} Z_i\;,$$

$$V_j \;:=\; \sqrt{h} \sum_{i=j(r+s)+r+1}^{(j+1)(r+s)} Z_i$$

and

$$W_j := \sqrt{h} \sum_{i=k(r+s)+1}^{n} Z_i \, .$$

We see immediately that U_j sums the Z_i over the blocks of size r, V_j sums the Z_i over the blocks of size s, and W_j accounts for the remaining terms that do not fit evenly into the first $2k$ blocks.

The idea of the proof is to show that the small blocks separate the large blocks by enough to make them asymptotically independent while being small enough that they don't make a substantial contribution to the limiting distribution. The Lindeberg condition can then be checked for the separated large blocks.

To formalize this, we write

$$\begin{aligned} \sqrt{nh}\, Q_n &= \frac{1}{\sqrt{n}} \left[\sum_{j=0}^{k-1} U_j + \sum_{j=0}^{k-1} V_j + W_j \right] \\ &:= \frac{1}{\sqrt{n}} \Big[Q_n' + Q_n'' + Q_n''' \Big] \, . \end{aligned}$$

We will establish the following identities:

$$\frac{1}{n} E\big[(Q_n'')^2\big] \to 0 \, , \tag{7.2}$$

$$\frac{1}{n} E\big[(Q_n''')^2\big] \to 0 \, , \tag{7.3}$$

$$\left| E \exp(it\, Q_n') - \prod_{j=1}^{k} E\big[\exp(it\, U_j)\big] \right| \to 0 \, , \tag{7.4}$$

$$\frac{1}{n} \sum_{i=1}^{k} E\big[U_i^2\big] \to \theta^2(x) \, , \tag{7.5}$$

and

$$\frac{1}{n} \sum_{j=0}^{k-1} E\Big[U_j^2 \, 1_{[|U_j| \geq \epsilon\, \theta(x) \sqrt{n}]}\Big] \to 0 \, , \tag{7.6}$$

for all $\epsilon > 0$.

Once these have been established, by a Taylor Series expansion, we will have

$$\begin{aligned} &\Big| E\Big[\exp\big(it\sqrt{nh}\, Q_n\big)\Big] - \exp\big(-t^2\theta^2(x)/2\big)\Big| = \\ &\qquad = \left| E\left[\exp\left(it\, \frac{1}{\sqrt{n}}\, [Q_n' + Q_n'' + Q_n''']\right)\right] - \exp\big(-t^2\theta^2(x)/2\big)\right| \\ &\qquad \leq \left| E\left[\exp\left(it\, \frac{1}{\sqrt{n}}\, Q_n'\right)\right] - \exp\big(-t^2\theta^2(x)/2\big)\right| \\ &\qquad\quad + E\left[\left|\frac{2t}{\sqrt{n}}\, Q_n''\right|\right] + E\left[\left|\frac{2t}{\sqrt{n}}\, Q_n'''\right|\right] \, . \end{aligned} \tag{7.7}$$

The final two terms will converge to 0 by the Cauchy–Schwarz inequality and equations (7.2) and (7.3). Equations (7.4), (7.5), and (7.6) are enough to verify the conditions of the Lindeberg–Feller central limit theorem, which will establish the desired result. The proof will be complicated somewhat because (7.6) will be established first for bounded random variables, and then the bound will be allowed to tend to infinity.

We begin by choosing block sizes. By Assumption 8, there exists a sequence q_n such that $q_n \to \infty$ and $q_n s_n = o\big(\sqrt{nh}\big)$, and $q_n \sqrt{n/h}\, \alpha(s_n) \to 0$. Define the large block size r_n by

$$r_n := \lfloor \sqrt{nh}/q_n \rfloor \ .$$

From this definition, we see

$$\frac{s_n}{r_n} \le \frac{s_n}{\sqrt{nh}/q_n - 1} = \frac{q_n s_n/\sqrt{nh}}{1 - q_n/\sqrt{nh}} \to 0 \ , \tag{7.8}$$

$$\begin{aligned} \frac{n}{r_n}\,\alpha(s_n) &= \frac{n}{\lfloor \sqrt{nh}/q_n \rfloor}\,\alpha(s_n) \le \frac{n}{\sqrt{nh}/q_n - 1}\,\alpha(s_n) \\ &= \Big(\sqrt{n/h}\, q_n + o(1)\Big)\,\alpha(s_n) \to 0 \end{aligned} \tag{7.9}$$

and

$$\frac{r_n}{\sqrt{nh}} = \frac{\lfloor \sqrt{nh}/q_n \rfloor}{\sqrt{nh}} \le \frac{\sqrt{nh}/q_n + 1}{\sqrt{nh}} = \frac{1}{q_n} + \frac{1}{\sqrt{nh}} \to 0 \ . \tag{7.10}$$

We begin by verifying equations (7.2) and (7.3):

$$E\big[(Q''_n)^2\big] = \sum_{i=0}^{k-1} \mathrm{var}[V_j] + \sum_{i \ne j} \mathrm{cov}[V_i, V_j] \ , \tag{7.11}$$

where, by Lemma 5.1,

$$\begin{aligned} \mathrm{var}[V_i] &= s\,h\,\mathrm{var}[Z_1] + 2\,s\,h \sum_{i=1}^{s-1} (1 - j/s)\,\mathrm{cov}[Z_1, Z_{1+i}] \\ &= s\,\big[\theta(x) + o(1)\big] \ . \end{aligned}$$

By equation (7.8),

$$\begin{aligned} \sum_{i=0}^{k-1} \mathrm{var}[V_j] &\le k_n s_n \big[\theta(x) + o(1)\big] \\ &\le \frac{n\, s_n}{s_n + r_n}\,\big[\theta(x) + o(1)\big] \\ &= o(n) \ . \end{aligned}$$

The second term of (7.11) can be treated as follows:

$$\sum_{i \ne j} \mathrm{cov}[V_i, V_j] = h \sum_{i \ne j}^{k-1} \sum_{l=1}^{s} \sum_{m=1}^{s} \mathrm{cov}\big[Z_{i(r+s)+r+m}\,, Z_{j(r+s)+r+l}\big] \ .$$

Since $i \neq j$, the difference between the indices, $\big|i(r{+}s){+}r{+}m{-}(j(r{+}s){+}r{+}l)\big| \geq r$, so

$$\left|\sum_{i\neq j} \mathrm{cov}[V_i, V_j]\right| \;\leq\; 2\,h \sum_{l=1}^{n-r} \sum_{m=l+r}^{n} \big|\mathrm{cov}[Z_l, Z_m]\big|$$
$$=\; 2\,h \sum_{l=1}^{n-r} \sum_{j=r}^{n-l} \big|\mathrm{cov}[Z_l, Z_{l+j}]\big|$$
$$\leq\; 2\,n\,h \sum_{j=r}^{n-1} \big|\mathrm{cov}[Z_1, Z_{j+1}]\big| \;=\; o(n)\ .$$

This establishes (7.2). We now turn our attention to (7.3):

$$\frac{1}{n}\,E\big[(Q_n''')^2\big] \;=\; \frac{h}{n}\Big(n - k(r+s)\Big)\,\mathrm{var}[Z_1] \;+\; 2\,h \sum_{i=2}^{n-k(r+s)} \mathrm{cov}[Z_1, Z_i]$$
$$\leq\; \frac{r_n + s_n}{n}\,\theta^2(x) + o(1) \;\to\; 0\ .$$

To prove (7.4), we use a lemma of Volkonskii and Rozanov [23], which is stated in Lemma 7.3, following this proof:

$$\left|E\,\exp(it\,Q_n') - \prod_{j=1}^{k} E\big[\exp(it\,U_j)\big]\right| \;\leq\; 16\,(k-1)\,\alpha(s_n{+}1)$$
$$=\; 16\,\frac{n}{r_n}\,\alpha(s_n{+}1) + o(1) \;\to\; 0\ ,$$

by equation (7.9).

We now turn our attention to equation (7.5):

$$\frac{1}{n}\sum_{i=1}^{k} E[U_i^2] \;=\; \frac{k_n}{n}\,\mathrm{var}[U_1]$$
$$=\; \frac{k_n\,r_n}{n}\,\big(\theta^2(x) + o(1)\big)$$
$$=\; \frac{r_n}{r_n + s_n}\,\theta^2(x) + o(1) \;\to\; \theta^2(x)\ .$$

Finally, we verify equation (7.6). We begin establishing the result for truncated random variables, and then subsequently letting the truncation point go to infinity. Define

$$Z_i^L \;:=\; (c_1 + c_2\,Y_i)\;1_{[|Y_i|\leq L]}\,K_h(X_i{-}\,x) \;-\; E\Big[(c_1 + c_2\,Y_i)\;1_{[|Y_i|\leq L]}\,K_h(X_i{-}\,x)\Big]\ ,$$
$$Q_n^L \;:=\; \frac{1}{n}\sum_{i=1}^{n} Z_i^L\ ,$$
$$\tilde{Q}_n^L \;:=\; \frac{1}{n}\sum_{i=1}^{n}\big(Z_i - Z_i^L\big)\ ,$$

and

$$U_j^L \;:=\; \sqrt{h} \sum_{i=j(r+s)+1}^{j(r+s)+r} Z_i^L .$$

We first need to estimate the asymptotic variance of Z_1^L. Assume conditions strong enough that for all c_1 and c_2, and for all $L > L_0$,

$$E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i| \le L]}\,\big|\,X_1 = x\Big]$$

is continuous as a function of x:

$$\left| E\Big[(c_1 + c_2\,Y_i)\, 1_{[|Y_i|\le L]}\, K_h(X_i - x)\Big]\right| \;\le\; \Big(|c_1| + |c_2\,L|\Big)\, E\Big[\big|K_h(X_1 - x)\big|\Big] \;\le\; C\ ,$$

and

$$\begin{aligned}
h\,E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i|\le L]}\, K_h^2(X_i - x)\Big] \;=& \\
&= \; h\,E\Big[K_h^2(X_1 - x)\; E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i|\le L]}\,\big|\,X_1\Big]\Big] \\
&= \; h \int_{-\infty}^{\infty} K_h^2(u - x)\; E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i|\le L]}\,\big|\,X_1 = u\Big]\, f(u)\; du \\
&= \; \int_{-\infty}^{\infty} K^2(v)\; E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i|\le L]}\,\big|\,X_1 = x + hv\Big]\, f(x + hv)\; dv \\
&= \; E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i|\le L]}\,\big|\,X_1 = x\Big]\; f(x) \int_{-\infty}^{\infty} K^2(v)\, dv \;+\; o(1)\ .
\end{aligned}$$

Putting these two together,

$$h\,\mathrm{var}[Z_1^L] \;=\; E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i|\le L]}\,\big|\,X_1 = x\Big]\; f(x) \int_{-\infty}^{\infty} K^2(v)\, dv \;+\; o(1)\ .$$

For the sake of notational simplicity, we now define

$$(\theta^L)^2\,(x) \;:=\; E\Big[(c_1 + c_2\,Y_i)^2\, 1_{[|Y_i|\le L]}\,\big|\,X_1 = x\Big]\; f(x) \int_{-\infty}^{\infty} K^2(v)\, dv\ .$$

Returning now to the proof of (7.6), since K and Y_i^L are bounded, hZ_i^L is also bounded. Equivalently, for some D,

$$\sqrt{h}\; Z_i^L \;\le\; D/\sqrt{h}\ .$$

Therefore, by equation (7.10), $\max_{0\le j\le k-1} U_j^L/\sqrt{n} \;\le\; (Dr_n)/\sqrt{nh} \to 0$. For large enough n, the set $\big\{|U_j^L| \ge \theta^L(x)\,\epsilon\,\sqrt{n}\big\}$ becomes empty. Therefore, by the same argument as used to establish (7.7),

$$\sqrt{nh}\; Q_n^L \;\xrightarrow{\mathcal{D}}\; N\big(0, (\theta^L)^2(x)\big)\ . \tag{7.12}$$

We are now prepared to put these pieces together to finish the proof of the theorem:

$$
\begin{aligned}
&\Big|E\Big[\exp\big(it\sqrt{nh}\,Q_n\big)\Big]-\exp\big(-t^2\theta^2(x)/2\big)\Big| =\\
&\quad= \Big|E\Big[\exp\Big(it\sqrt{nh}\,[Q_n^L+\tilde{Q}_n^L]\Big)\Big]-\exp\big(-t^2(\theta^L)^2(x)/2\big)\\
&\qquad+\exp\big(-t^2(\theta^L)^2(x)/2\big)-\exp\big(-t^2\theta^2(x)/2\big)\Big|\\
&\quad= \Big|E\Big[\exp\big(it\sqrt{nh}\,Q_n^L\big)+\exp\big(it\sqrt{nh}\,Q_n^L\big)\Big(\exp\big[it\sqrt{nh}\,\tilde{Q}_n^L\big]-1\Big)\Big]\\
&\qquad-\exp\big(-t^2(\theta^L)^2(x)/2\big)+\exp\big(-t^2(\theta^L)^2(x)/2\big)-\exp\big(-t^2\theta^2(x)/2\big)\Big|\\
&\quad\le \Big|E\Big[\exp\big(it\sqrt{nh}\,Q_n^L\big)\Big]-\exp\big(-t^2(\theta^L)^2(x)/2\big)\Big|\\
&\qquad+\Big|E\Big[\exp\big(it\sqrt{nh}\,Q_n^L\big)\Big(\exp\big[it\sqrt{nh}\,\tilde{Q}_n^L\big]-1\Big)\Big]\Big|\\
&\qquad+\Big|\exp\big(-t^2(\theta^L)^2(x)/2\big)-\exp\big(-t^2\theta^2(x)/2\big)\Big|\\
&\quad\le \Big|E\Big[\exp\big(it\sqrt{nh}\,Q_n^L\big)\Big]-\exp\big(-t^2(\theta^L)^2(x)/2\big)\Big|\\
&\qquad+E\Big[\Big|\exp\big[it\sqrt{nh}\,\tilde{Q}_n^L\big]-1\Big|\Big]+\Big|\exp\big(-t^2(\theta^L)^2(x)/2\big)-\exp\big(-t^2\theta^2(x)/2\big)\Big|\,.
\end{aligned}
$$

We analyze each term separately, first letting $n\to\infty$, and then letting $L\to\infty$. For fixed t, the first term goes to zero by equation (7.12). The third term goes to 0 by dominated convergence, since $(\theta^L)^2(x)\to\theta^2(x)$ as $L\to\infty$. Only the second term remains to be analyzed. By a Taylor series expansion,

$$\Big|\exp\big[it\sqrt{nh}\,\tilde{Q}_n^L\big]-1\Big| \;\le\; 2\,\Big|t\sqrt{nh}\,\tilde{Q}_n^L\Big|\,.$$

By the Cauchy–Schwarz inequality, the bound will converge to 0 if it can be shown that $nh\,\mathrm{var}\big[(\tilde{Q}_N^L)\big]\to 0$. As $\tilde{Q}_n^L$ satisfies the same dependence assumptions as Q_n, the calculations of Lemma 5.1(c) apply. So, it is sufficient to show that $(\theta^L)^2(x)\to 0$ as $L\to\infty$. This follows immediately by dominated convergence. □

Lemma 7.3 (Volkonskii and Rozanov [23]). *Let $V_1,\dots,V_N$ be strong mixing random variables, which are measurable with respect to the σ-algebras $\mathcal{F}_{i_1}^{j_1},\dots,\mathcal{F}_{i_N}^{j_N}$ respectively, with $1\le i_1<j_1<i_2<\dots<j_N\le n$, $i_{l+1}-j_l\ge w\ge 1$, and $|V_l|\le 1$, for $l=1,\dots,N$. Then*

$$\left|E\left[\prod_{j=1}^{N}V_j\right]-\prod_{j=1}^{N}E[V_j]\right| \;\le\; 16\,(L-1)\,\alpha(w)\;,$$

where $\alpha(w)$ is the strong mixing coefficient.

REFERENCES

[1] Bosq, D. (1998). *Nonparametric statistics for stochastic processes*, "Lecture Notes in Statistics", vol. 110, Springer-Verlag, New York, second edition.

[2] Davis, K. (1975). Mean square error properties of density estimates, *Annals of Statistics*, **3**(4), 1025–1030.

[3] Davis, K. (1977). Mean integrated square error properties of density estimates, *Annals of Statistics*, **5**(3), 530–535.

[4] Devroye, L. (1992). A note on the usefulness of superkernels in density estimates, *Annals of Statistics*, **20**(4), 2037–2056.

[5] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L^1 View*, Wiley, New York.

[6] Fan, J. (1992). Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004.

[7] Gasser, Th. and Müller, H.-G. (1979). *Kernel estimation of regression functions*. In "Smoothing Techniques for Curve Estimation" (Th. Gasser and M. Rosenblatt, Eds.), Springer Lecture Notes in Mathematics, number 757, Springer-Verlag, Berlin, 23–68.

[8] Gasser, Th. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method, *Scandinavian Journal of Statistics*, **11**, 171–185.

[9] Györfi, L.; Härdle, W.; Sarda, P. and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*, "Lecture Notes in Statistics", vol. 60, Springer-Verlag, New York.

[10] Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and its Applications*, Academic Press, New York.

[11] Hall, P. and Marron, J.S. (1988). Choice of kernel order in density estimation, *Annals of Statistics*, **16**(1), 161–173.

[12] Ibragimov, I.A. and Khasminksii, R.Z. (1983). Estimation of distribution density belonging to a class of entire functions, *Theor. Probab. Appl.*, **27**(3), 551–562.

[13] Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes, *Scandinavian Journal of Statistics*, **24**, 165–179.

[14] McMurry, T. and Politis, D.N. (2004). Nonparametric regression with infinite order flat-top kernels, *Journal of Nonparametric Statistics*, **16**(3–4), 549–562.

[15] Nadaraya, E.A. (1964). On estimating regression, *Theory Probab. Appl.*, **9**, 141–142.

[16] Politis, D.N. (2001). *On nonparametric function estimation with infinite order flat-top kernels*. In "Probability and Statistical Models with Applications" (Ch.A. Charalambides, Markos V. Koutras, and N. Balakrishnan, Eds.), Chapman & Hall/CRC, Boca Raton, 469–483.

[17] Politis, D.N. (2003). Adaptive bandwidth choice, *Journal of Nonparametric Statistics*, **25**(4–5), 517–533.

[18] Politis, D.N. and Romano, J.P. (1993). On a family of smoothing kernels of infinite order. In "Computing Science and Statistics, Proceedings of the 25th Symposium on the Interface" (M. Tarter and M. Lock, Eds.), The Interface Foundation of North America, 141–145.

[19] Politis, D.N. and Romano, J.P. (1995). Bias-corrected nonparametric spectral estimation, *Journal of Time Series Analysis*, **16**(1), 67–103.

[20] Politis, D.N. and Romano, J.P. (1999). Multivariate density estimation with general flat-top kernels of infinite order, *Journal of Multivariate Analysis*, **68**, 1–25.

[21] Robinson, P.M. (1983). Nonparametric estimators for time series, *Journal of Time Series Analysis*, **4**(3), 185–207.

[22] Ruppert, D.; Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, **90**, 1257–1270.

[23] Volkonski, V.A. and Rozanov, Yu.A. (1959). Some limit theorems for random functions, *Theory of Probability and its Applications*, **4**, 178–197.

[24] Wand, M. and Ripley, B. (2005). *KernSmooth: Functions for kernel smoothing for Wand and Jones (1995)*, S original by Matt Wand, R port by Brian Ripley, R package version 2.22-15.

[25] Watson, G.S. (1964). Smooth regression analysis, *Sankhya Ser. A*, **26**, 359–372.