

Nonparametric Regression with Infinite Order Flat-Top Kernels

Timothy L. McMurry
University of California, San Diego

Dimitris N. Politis
University of California, San Diego

April 24, 2003

Abstract

The problem of nonparametric regression is addressed, and a kernel smoothing estimator is proposed which has favorable asymptotic performance (bias, variance, and mean squared error). The proposed class of kernels is characterized by a Fourier transform which is flat near the origin and infinitely differentiable. This property allows the bias of the estimate to decrease at the maximal rate without harming the rate at which the variance decreases, thus leading to a faster rate of convergence.

1 Introduction

Suppose the data $(x_1, Y_1), \dots, (x_n, Y_n)$ are generated by a model of the form $Y_i = r(x_i) + \epsilon_i$, where the ϵ_i are uncorrelated random variables with mean 0 and variance σ^2 , and the x_i are non-random design points, with $0 < x_1 < \dots < x_n < 1$. The function $r : [0, 1] \rightarrow \mathbb{R}$ is assumed to be continuous on $[0, 1]$, and to possess a certain degree of smoothness on $(0, 1)$; r is unknown, and will be estimated from the data.

There are many approaches to estimating r , including orthogonal series, splines, local polynomials, and kernel smoothing. This paper will examine the asymptotic properties of the kernel type regression estimator proposed by Gasser and Müller [2] using a new type of kernel. The Gasser-Müller estimator is defined by

$$\hat{r}_h(x) = \frac{1}{h} \sum_{i=1}^n Y_i \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du, \quad (1)$$

where $s_0 = 0$, $s_n = 1$, and $s_i = (x_i + x_{i+1})/2$ for $i = 1, \dots, n-1$. In the above, K is the kernel function, and h is the bandwidth parameter. The estimator \hat{r} can be thought of as a weighted average of the data near x ; alternatively, (1) may be viewed as a convolution of the rough dataset with the smooth kernel function. The kernel is scaled via h ; the degree of scaling may depend on several factors, including the size of the dataset and the underlying function r .

Under some conditions on the asymptotic spacing of the design points (see Hart [5]), it is known that for any fixed $x \in (0, 1)$ as $n \rightarrow \infty$, $h \rightarrow 0$, and $nh \rightarrow \infty$,

$$\text{Var}(\hat{r}_h(x)) \sim C(x) \frac{\sigma^2}{nh} \int_{-\infty}^{\infty} K^2(z) dz,$$

where $C(x)$ is a constant that depends on the density of the design points near x .

If K has finite moments up to order q , and its moments up to order $q - 1$ are equal to zero, then K is said to be of order q . If r has k continuous derivatives, and K is of order q , then

$$\text{Bias}(\hat{r}_h(x)) = c_{K,r}(x)h^p + o(h^p),$$

where $p = \min\{q, k\}$ and $c_{K,r}(x)$ is a bounded function depending on K , r , and r 's derivatives.

If the underlying function r is sufficiently smooth, the bias of the estimator can be reduced to $O(h^k)$ by choosing a kernel of appropriately large order. This was recognized by Gasser and Müller [2], and the idea dates further back in other contexts. However, estimating the number of derivatives that a function possesses is an even more difficult task than estimating the function itself. Hence, it is never clear what order of kernel one should choose for a given problem. Once a kernel has been chosen, it is still possible that the underlying function is smooth enough that the order of the bias will be limited by the order of the kernel. To get around this problem, one can define a kernel that has “infinite order”; that is, a kernel which reduces the bias to $O(h^k)$, no matter how large k happens to be, see e.g. Devroye [1]. In the paper at hand, a class of infinite order kernels is proposed; these kernels are characterized by the fact that their Fourier transforms are functions which are even, infinitely differentiable, and constant in a neighborhood of the origin. Similar kernels have been proposed by Politis and Romano for use in spectral density estimation [9] and density estimation [10].

The remainder of this paper is organized as follows: Section 2 contains results on the asymptotic performance of our proposed estimator; Section 3 provides an example of a kernel whose Fourier transform satisfies the conditions stated above; Section 4 contains some simulation results; all technical proofs have been placed in Section 5.

2 Convergence Rates

As alluded to earlier, it is necessary to impose some restrictions on the asymptotic spacing of the design points. Roughly speaking, they are assumed to be generated by quantiles of a positive density function f defined on $[0, 1]$. This is stated more precisely in the following assumption.

Assumption 1 *The design points are given by*

$$x_i = Q\left(\frac{i - 1/2}{n}\right), \quad i = 1, \dots, n,$$

where

$$Q(u) = F^{-1}(u),$$

and

$$F(x) = \int_0^x f(t)dt.$$

The function f is assumed to be a Lipschitz continuous positive density function on $[0, 1]$.

Assumption 1 allows the design points to be spaced unevenly, while requiring that the maximum space between two adjacent design points decreases at a rate proportional to $1/n$.

Some assumptions about the model that generated the data will also be imposed.

Assumption 2 *The errors ϵ_i , $i = 1, \dots, n$ are uncorrelated with mean 0, and variance σ^2 .*

Assumption 3 *The unknown function r has at least one bounded continuous derivative on $(0, 1)$.*

We now give the following general definition.

Definition 1 A general infinite order 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

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

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

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

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 , but this dependence will not be explicitly denoted.

The following theorems investigate the performance of the Gasser-Müller estimator $\hat{r}_h(x)$ using the general flat-top kernel K from Definition 1.

Theorem 1 Under Assumptions 1–3, the variance of the infinite order kernel estimator, $\hat{r}_h(x)$, as $n \rightarrow \infty$, $h \rightarrow 0$, and $n^2 h^3 \rightarrow \infty$ is given by

$$\text{Var}(\hat{r}_h(x)) = \frac{\sigma^2}{nh} \frac{1}{f(x)} \int_{-\infty}^{\infty} K^2(z) dz + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n^2 h^3}\right).$$

The variance is largest when $f(x)$ is small; this was to be expected, since small values of $f(x)$ correspond to a low density of design points. The variance is of the same order of magnitude as the variance of the traditional finite-order kernel Gasser-Müller estimator, but the bias is improved, as the following theorem shows.

Theorem 2 If r has k continuous derivatives, then under Assumptions 1–3, the bias of $\hat{r}_h(x)$ is

$$\text{Bias}(\hat{r}_h(x)) = E[\hat{r}_h(x)] - r(x) = O(1/n) + O(h^k).$$

If r is infinitely differentiable, the last term become $o(h^m)$ for all positive real m .

Thus, the infinite order kernel adapts to whatever degree of smoothness the underlying function r possesses, allowing the bias to converge at the optimal rate.

Corollary 3 Under the conditions of Theorems 1 and 2, the mean squared error of the infinite order kernel estimate at a point x is given by

$$\text{MSE}(\hat{r}_h(x)) = \frac{\sigma^2}{nh} \frac{1}{f(x)} \int_{-\infty}^{\infty} K^2(z) dz + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n^2 h^3}\right) + O(h^{2k}).$$

Under slightly stronger conditions, this estimator also has an asymptotic normal distribution about its mean. Since it is a biased estimate, it is not necessarily centered around the actual function r , unless undersmoothing occurs.

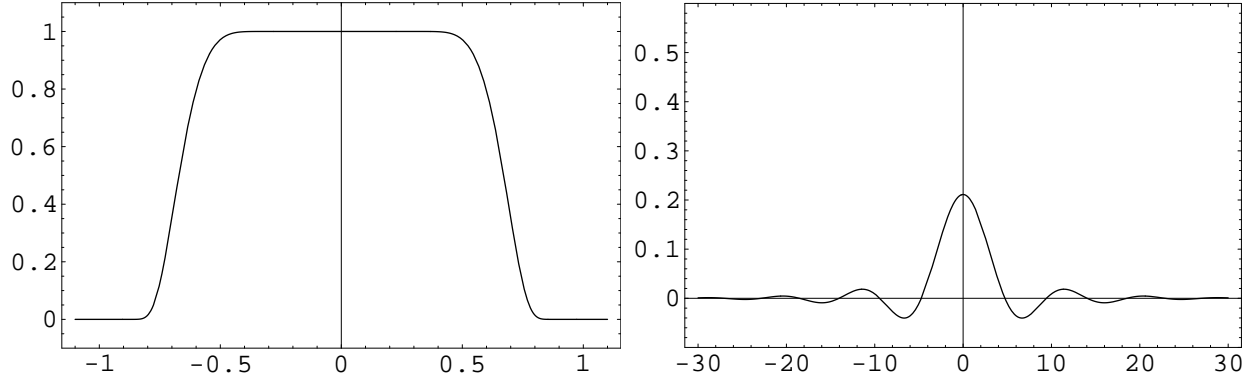


Figure 1: $\lambda(s)$ and the resulting kernel with $b = 1$ and $c = 0.05$.

Theorem 4 *Assume the conditions of Theorems 1 and 2. In addition, assume the data is generated by the model $Y_i = r(x_i) + \epsilon_i$, where $\epsilon_1, \dots, \epsilon_n$ are independent random variables with mean 0, variance σ^2 , and assume there exists a finite constant B such that the third moments of the errors satisfy $E|\epsilon_i|^3 < B$ for all i . Then as $n \rightarrow \infty$, $h \rightarrow 0$, and $nh^2 \rightarrow \infty$,*

$$\frac{\hat{r}(x) - E[\hat{r}(x)]}{\sqrt{\text{Var}(\hat{r}(x))}} \xrightarrow{\mathcal{D}} N(0, 1).$$

3 Example of a Flat-Top C^∞ Kernel

The problems caused by edge effects plague many nonparametric regression estimators, including the Gasser-Müller estimator. If the regression is performed using a standard finite order kernel with support $[-1, 1]$, then a finite sample bias is induced for x in the intervals $(0, h)$ and $(1 - h, 1)$ because the weights used to form the weighted averages in these regions do not add up to 1. Since infinite order kernels do not have compact support, these edge effects get spread out across the whole interval $[0, 1]$. For this reason, as discussed in the case of discontinuous density estimation in Politis [8], it is important that the tails of K decay as quickly as possible, to minimize the effect on the interior of $[0, 1]$. One way to accomplish this is to require that the Fourier transform of the kernel be very smooth; this ensures that the tails of K decay rapidly. For these reasons, attention focuses on the Fourier transform of the infinite order kernel. An example of such a kernel can be defined as follows.

Let b and c be constants satisfying $b > 0$ and $0 < c < 1$. Define $\lambda(s)$ by

$$\lambda(s) = \begin{cases} 1 & \text{if } |s| \leq c \\ \exp[-b \exp[-b/(|s| - c)^2]/(|s| - 1)^2] & \text{if } c < |s| < 1 \\ 0 & \text{if } |s| \geq 1. \end{cases} \quad (3)$$

As in Section 2, 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. Figures 1–3 show plots of λ (as defined above) and the resulting kernel K for $c = 0.05$ and several values of b .

The function $\exp[-b \exp[-b/(|s| - c)^2]/(|s| - 1)^2]$ 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 $|s| = 1$. Since $\lambda(s)$ is infinitely differentiable, the tails of $K(x)$, where K is defined by equation (2), decay faster than $|x|^{-m}$, for all positive finite m , as $|x| \rightarrow \infty$.

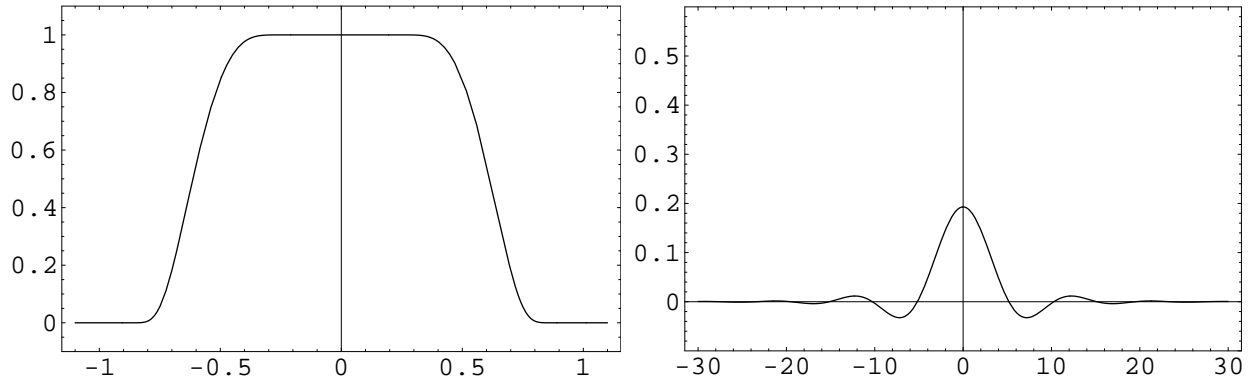


Figure 2: $\lambda(s)$ and the resulting kernel with $b = 1/2$ and $c = 0.05$.

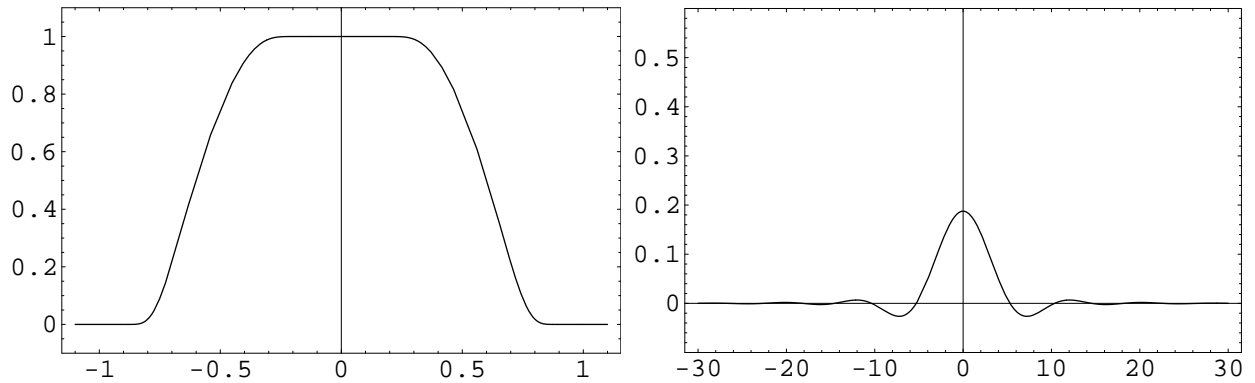


Figure 3: $\lambda(s)$ and the resulting kernel with $b = 1/4$ and $c = 0.05$.

4 Simulation Results

A small numerical study was undertaken to get a sense for the finite sample performance of this estimator. As in Gasser and Müller [3], analysis was performed on the function $r(x) = 2 - 2x + 3 \exp(-100(x - .5)^2)$, with $\sigma^2 = .4$. They note that if the underlying function is known, it is possible to calculate the exact bias and variance of the estimator at any point; this allows the integrated mean squared error (IMSE) to be estimated using a numerical integration technique, as in Messer and Goldstein [6]. A sample regression was performed for each of the following cases. In addition, the integrated bias and variance were estimated using Simpson's rule on 200 points in the interval $[0, 1]$. All programming was done in Mathematica. Before results are presented, a few words should be said about the choice of the parameters c and h .

The simulations were performed using the kernel defined by equation (3) with $c = 0.05$ and $b = 1$. Small values of c seem advisable in practice. Since $\lambda(s)$ was constructed to be very smooth, it ends up being very close to constant in a much wider region than $[-c, c]$; see Figure 1. Large values of c cause $\lambda(s)$ to become almost rectangular, which is undesirable as it corresponds to a kernel with very large side lobes, similar to those of the Dirichlet kernel.

The problem of optimal bandwidth choice is still in need of further study. However, as in Politis [7], we can make a practical recommendation regarding choice of the parameter h . Flat-top kernels perform a soft thresholding in the frequency domain. The data, viewed in this domain, should exhibit a large low frequency component resulting from the slowly varying underlying function r , and a much smaller high frequency component resulting from the errors. The bandwidth should be chosen so as to allow the low frequency component to pass undisturbed, while damping out the higher frequencies. With this in mind, we propose the following rule of thumb.

Let $w_{b,c}$ be the half width of the region over which $\lambda(s)$ is close to flat. For example, when $b = 1$ and $c = .05$, $\lambda(s)$ appears to be approximately constant for $|s| < .4$ (see Figure 1), so choose $w_{1,.05} = .4$. Define the sample Fourier transform $\phi_n(s)$ by

$$\phi_n(s) = \sum_{j=1}^n \int_{S_{j-1}}^{S_j} Y_j \exp(isx) dx,$$

and let $\rho_n(s) = |\phi_n(s)/\phi_n(0)|$. If a plot of $\rho_n(s)$ reveals that there is a constant \hat{B} such that $\rho_n(s)$ is negligible for $|s| > \hat{B}$, and nonnegligible for $|s| < \hat{B}$, choose $h = w_{b,c}/\hat{B}$.

Case I: On the first run, efforts were made to control edge effects by placing 200 evenly spaced design points in the interval $(-5, 1.5)$, while performing the regression only on the smaller interval $[0, 1]$. For comparison, the same regression was performed using the Epanechnikov kernel, $K_e(x) = (3/4)(1 - x^2)1_{[-1,1]}(x)$, at the asymptotically optimal bandwidth. Note that finding the asymptotically optimal bandwidth requires knowledge of the underlying function r ; this knowledge puts the Epanechnikov kernel estimator at an advantage relative to the infinite order kernel estimator, which used a data driven bandwidth selection process. The results are summarized in Table 1. The flat-top kernel estimator has a larger integrated variance, but it substantially improves the bias, yielding a smaller IMSE. Thus, our small simulation confirms the validity of our Theorems 1–3 even for a moderate sample size of $n = 200$. See Figure 4 for the scatterplot and the two smoothers.

Case II: On the second run, the same function was used with 200 design points randomly spaced, from a uniform distribution on $[0, 1]$. No attempts were made to control for edge effects. The smoothed scatterplot is shown in Figure 5, and the results are summarized in Table 1.

Case III: For the final regression, the data from Case I was used, except the 100 design points outside of $[0, 1]$ were thrown away. To control for edge effects, a version of the reflection technique proposed by Hall and Wehrly [4] was used. The regression was performed at $x = 0$, and $x = 1$,

Case	Flat Top Kernel			Epanechnikov Kernel		
	IVAR	ISB	IMSE	IVAR	ISB	IMSE
I	3.74×10^{-2}	1.40×10^{-3}	3.88×10^{-2}	3.38×10^{-2}	7.24×10^{-3}	4.10×10^{-2}
II	4.02×10^{-2}	1.31×10^{-2}	5.33×10^{-2}	2.85×10^{-2}	1.94×10^{-2}	4.79×10^{-2}
III		1.53×10^{-3}			7.30×10^{-3}	

Table 1: Integrated squared bias, variance, and mean squared error for the designs studied in Cases I–III.

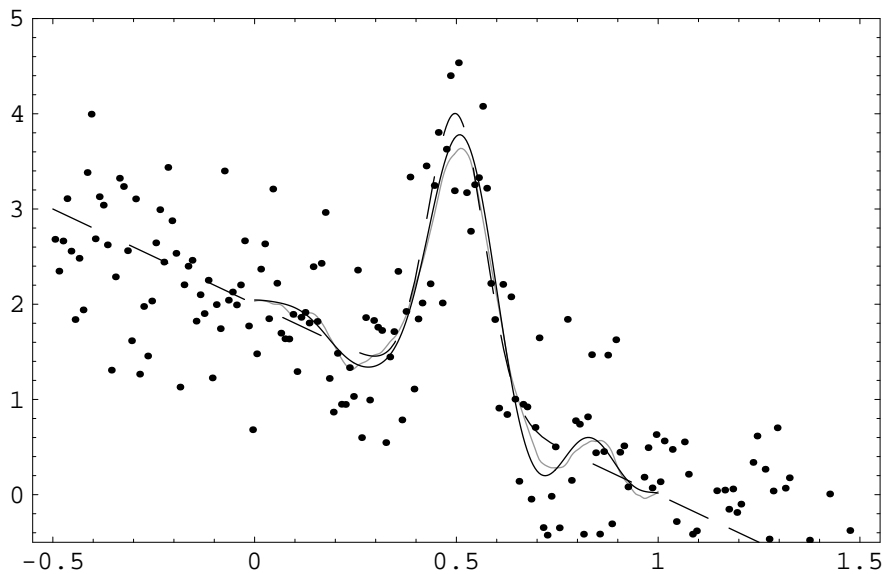


Figure 4: Smoothed scatterplot, Case I. The dashed line is the underlying function, the dots are the data, the gray line is the Epanechnikov kernel estimator, and the black line is the infinite order kernel estimator.

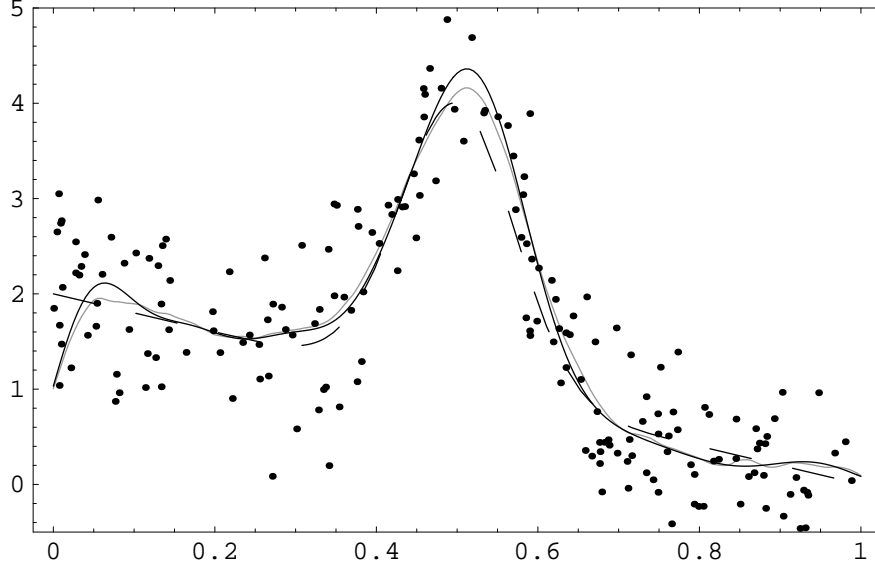


Figure 5: Smoothed scatterplot, Case II. The dashed line is the underlying function, the dots are the data, the gray line is the Epanechnikov kernel estimator, and the black line is the infinite order kernel estimator.

and the data was then reflected through the points $(0, 2\hat{r}(0))$, and $(1, 2\hat{r}(1))$. The regression was then performed on the interval $[0, 1]$ using the expanded data set. This approach is suboptimal in terms of asymptotic IMSE, since the bias at a boundary point is at best $O(h^2)$, but it is a substantial improvement over doing nothing at all. In addition, it makes the regression invariant under vertical translations of the data. The integrated variance was not calculated due to a slightly more complicated dependence structure. The smoothed scatterplot is shown in Figure 6.

Asymptotic Normality: Lastly, a small simulation was undertaken to verify the asymptotic normality result. On each iteration, 100 design points were evenly spaced on $[0, 1]$, and a new data set was generated from the model used in the preceding simulations. The residuals were then calculated at $x = .5$. The process was repeated 100 times, and the results are shown in the form of a QQ-plot in Figure 7.

5 Technical Proofs

PROOF OF THEOREM 1:

$$\text{Var}(\hat{r}_h(x)) = \text{Var} \left(\frac{1}{h} \sum_{i=1}^n Y_i \int_{s_{i-1}}^{s_i} K \left(\frac{x-u}{h} \right) du \right) \quad (4)$$

$$= \sigma^2 \sum_{i=1}^n \left(\frac{1}{h} \int_{s_{i-1}}^{s_i} K \left(\frac{x-u}{h} \right) du \right)^2. \quad (5)$$

By the intermediate value theorem,

$$= \frac{\sigma^2}{h^2} \sum_{i=1}^n (s_i - s_{i-1})^2 K^2 \left(\frac{x - x_i^*}{h} \right)$$

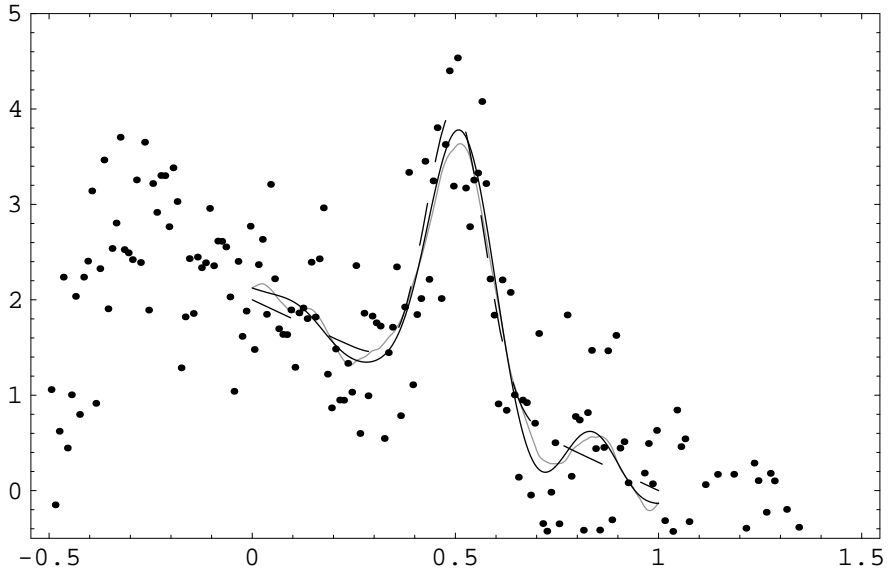


Figure 6: Smoothed scatterplot, Case III. The dashed line is the underlying function, the dots are the data, the gray line is the Epanechnikov kernel estimator, and the black line is the infinite order kernel estimator.

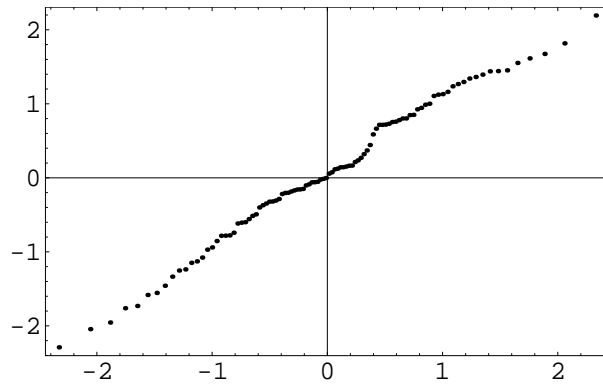


Figure 7: QQ-plot of the standardized residuals of $\hat{r}_h(.5)$ against the quantiles of the standard normal distribution.

where $x_i^* \in [s_{i-1}, s_i]$. This can be broken up as

$$\text{Var}(\hat{r}_h(x)) = C_{n,h} + E_{n,h},$$

where,

$$C_{n,h} = \frac{\sigma^2}{h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \left(Q \left(\frac{i+1/2}{n} \right) - Q \left(\frac{i-1/2}{n} \right) \right),$$

and

$$E_{n,h} = \frac{\sigma^2}{h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \times \left(s_i - s_{i-1} - \left[Q \left(\frac{i+1/2}{n} \right) - Q \left(\frac{i-1/2}{n} \right) \right] \right).$$

For the moment, concentrate on $C_{n,h}$. By the mean value theorem,

$$Q \left(\frac{i+1/2}{n} \right) - Q \left(\frac{i-1/2}{n} \right) = Q'(u_i) \frac{1}{n},$$

for some $u_i \in [(i-1/2)/n, (i+1/2)/n]$. Since, $Q'(u) = 1/f(Q(u))$,

$$Q \left(\frac{i+1/2}{n} \right) - Q \left(\frac{i-1/2}{n} \right) = \frac{1}{f(Q(u_i))} \frac{1}{n} \tag{6}$$

$$= \frac{1}{f(\tilde{x}_i)} \frac{1}{n}, \tag{7}$$

where $\tilde{x}_i \in [x_i, x_{i+1}]$. So,

$$\begin{aligned} C_{n,h} &= \frac{\sigma^2}{nh^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \frac{1}{f(\tilde{x}_i)} \\ &= \frac{\sigma^2}{nh^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \frac{1}{f(x_i^*)} + R_{n,h}, \end{aligned}$$

where,

$$R_{n,h} = \frac{\sigma^2}{nh^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \left(\frac{1}{f(\tilde{x}_i)} - \frac{1}{f(x_i^*)} \right).$$

By the intermediate value theorem, we can write,

$$\begin{aligned} \int_0^1 \frac{1}{f(t)} K^2 \left(\frac{x-t}{h} \right) dt &= \sum_{i=1}^n \int_{s_{i-1}}^{s_i} \frac{1}{f(t)} K^2 \left(\frac{x-t}{h} \right) dt \\ &= \sum_{i=1}^n (s_i - s_{i-1}) \frac{1}{f(x'_i)} K^2 \left(\frac{x - x'_i}{h} \right), \end{aligned}$$

where $x'_i \in [s_{i-1}, s_i]$. If we stick this into the equation for $C_{n,h}$, we get,

$$C_{n,h} = \frac{\sigma^2}{nh^2} \int_0^1 \frac{1}{f(t)} K^2 \left(\frac{x-t}{h} \right) dt + R_{n,h} + R_{n,h}^*$$

where,

$$R_{n,h}^* = \frac{\sigma^2}{nh^2} \sum_{i=1}^n (s_i - s_{i-1}) \left(\frac{1}{f(x_i^*)} K^2 \left(\frac{x - x_i^*}{h} \right) - \frac{1}{f(x'_i)} K^2 \left(\frac{x - x'_i}{h} \right) \right).$$

If we make the change of variable $z = (x - t)/h$, our expression for $C_{n,h}$ becomes

$$C_{n,h} = \frac{\sigma^2}{nh} \int_{(x-1)/h}^{x/h} \frac{1}{f(x-hz)} K^2(z) dz + R_{n,h} + R_{n,h}^* \quad (8)$$

Let us examine the integral term in this expression. The fourier transform of K is infinitely differentiable, so the tails of K decay faster than z^{-k} for any positive integer k . Since f is positive, bounded away from zero, and Lipschitz continuous, $1/f$ is also Lipschitz continuous. Therefore,

$$\begin{aligned} & \left| \frac{\sigma^2}{nh} \frac{1}{f(x)} \int_{-\infty}^{\infty} K^2(z) dz - \frac{\sigma^2}{nh} \int_{(x-1)/h}^{x/h} \frac{1}{f(x-hz)} K^2(z) dz \right| \\ & \leq \frac{\sigma^2}{nh} \left(\int_{(x-1)/h}^{x/h} \left| \frac{1}{f(x)} - \frac{1}{f(x-hz)} \right| K^2(z) dz \right. \\ & \quad \left. + \left| \int_{x/h}^{\infty} \frac{1}{f(x)} K^2(z) dz \right| + \left| \int_{-\infty}^{(x-1)/h} \frac{1}{f(x)} K^2(z) dz \right| \right) \\ & \leq \frac{\sigma^2}{nh} \int_{(x-1)/h}^{x/h} m|hz| K^2(z) dz + o(h^k/n) \\ & = O(1/n) + o(h^k/n), \end{aligned}$$

for some positive constant m and all positive integers k . Hence,

$$C_{n,h} = \frac{\sigma^2}{nh} \frac{1}{f(x)} \int_{-\infty}^{\infty} K^2(z) dz + O(1/n) + o(h^k/n) + R_{n,h} + R_{n,h}^*.$$

Now we need to show that the remainder terms are asymptotically negligible with respect to $C_{n,h}$. We will begin with $R_{n,h}$.

$$\begin{aligned} |R_{n,h}| & \leq \frac{\sigma^2}{nh^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \left| \frac{1}{f(\tilde{x}_i)} - \frac{1}{f(x_i^*)} \right| \\ & \leq m \frac{\sigma^2}{nh^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) |\tilde{x}_i - x_i^*|. \end{aligned}$$

By the same argument as the one used to establish equation (7), there exists $u_i \in [x_{i+1}, x_{i-1}]$, and a constant M such that

$$\begin{aligned} |\tilde{x}_i - x_i^*| & \leq |x_{i+1} - x_{i-1}| \\ & = Q \left(\frac{i+1/2}{n} \right) - Q \left(\frac{i-3/2}{n} \right) \\ & = \frac{1}{2n} \frac{1}{f(u_i)} \\ & \leq \frac{M}{n}. \end{aligned}$$

So,

$$\begin{aligned}
|R_{n,h}| &\leq M_1 \frac{\sigma^2}{n^2 h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \\
&\leq M_1 \frac{\sigma^2}{n^2 h^2} \max K^2(x) \sum_{i=1}^n (s_i - s_{i-1}) \\
&= O\left(\frac{1}{n^2 h^2}\right).
\end{aligned}$$

Using a similar technique the convergence rate for $R_{n,h}^*$ can be calculated. K has a bounded derivative, so it is Lipschitz continuous. Since the product of bounded Lipschitz continuous functions is Lipschitz continuous,

$$\begin{aligned}
|R_{n,h}^*| &\leq \frac{\sigma^2}{nh^2} \sum_{i=1}^n (s_i - s_{i-1}) \left| \frac{1}{f(x_i^*)} K^2 \left(\frac{x - x_i^*}{h} \right) - \frac{1}{f(x'_i)} K^2 \left(\frac{x - x'_i}{h} \right) \right| \\
&= O\left(\frac{1}{n^2 h^3}\right).
\end{aligned}$$

Finally,

$$\begin{aligned}
|E_{n,h}| &= \left| \frac{\sigma^2}{h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \right. \\
&\quad \left. \times \left(s_i - s_{i-1} - \left[Q \left(\frac{i+1/2}{n} \right) - Q \left(\frac{i-1/2}{n} \right) \right] \right) \right| \\
&= \left| \frac{\sigma^2}{h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) (s_i - s_{i-1} - [x_{i+1} - x_i]) \right| \\
&= \left| \frac{\sigma^2}{h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \right. \\
&\quad \left. \times \left(\frac{x_i + x_{i+1}}{2} - \frac{x_i + x_{i-1}}{2} - [x_{i+1} - x_i] \right) \right| \\
&= \left| \frac{\sigma^2}{h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \left(\frac{1}{2}(x_i - x_{i-1}) - \frac{1}{2}(x_{i+1} - x_i) \right) \right| \\
&= \left| \frac{\sigma^2}{2h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \left(\frac{1}{n} \frac{1}{f(v_{i-1})} - \frac{1}{n} \frac{1}{f(v_i)} \right) \right|,
\end{aligned}$$

where $v_{i-1} \in [x_{i-1}, x_i]$ and $v_i \in [x_i, x_{i+1}]$. So,

$$\begin{aligned}
|E_{n,h}| &\leq m \frac{\sigma^2}{2nh^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) |v_i - v_{i-1}| \\
&\leq M_1 \frac{\sigma^2}{2n^2 h^2} \sum_{i=1}^n (s_i - s_{i-1}) K^2 \left(\frac{x - x_i^*}{h} \right) \\
&= O\left(\frac{1}{n^2 h^2}\right).
\end{aligned}$$

□

PROOF OF THEOREM 2: For technical reasons, it will also be necessary to assume that r decays to 0 outside the interval $[0, 1]$ with as much smoothness as is possessed by r inside $(0, 1)$, and that this decay is rapid enough that r is integrable over the whole real line. This is a harmless assumption since it can just be thought of as an extension of r beyond the region of interest and into a region where its behavior will have no impact on our regression estimates.

$$\begin{aligned} \text{Bias}(\hat{r}_h(x)) &= \frac{1}{h} \sum_{i=1}^n r(x_i) \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du - r(x) \\ &= \frac{1}{h} \int_0^1 r(u) K\left(\frac{x-u}{h}\right) du \\ &\quad + \frac{1}{h} \sum_{i=1}^n \int_{s_{i-1}}^{s_i} (r(x_i) - r(u)) K\left(\frac{x-u}{h}\right) du - r(x). \end{aligned}$$

Since r has a bounded continuous derivative,

$$\begin{aligned} |r(x_i) - r(u)| &= |r'(x_i^*)(x_i - u)| \\ &\leq M|x_i - u| \\ &= O\left(\frac{1}{n}\right), \end{aligned}$$

where x_i^* is between x_i and u , and $|r'(\cdot)| \leq M$. So,

$$\begin{aligned} |\text{Bias}(\hat{r}_h(x))| &\leq \left| \frac{1}{h} \int_0^1 r(u) K\left(\frac{x-u}{h}\right) du - r(x) \right| + O\left(\frac{1}{n}\right) \\ &\leq \left| \frac{1}{h} \int_{-\infty}^{\infty} r(u) K\left(\frac{x-u}{h}\right) du - r(x) \right| \\ &\quad + \left| \frac{1}{h} \int_{-\infty}^0 r(u) K\left(\frac{x-u}{h}\right) du \right| + \left| \frac{1}{h} \int_1^{\infty} r(u) K\left(\frac{x-u}{h}\right) du \right| \\ &\quad + O\left(\frac{1}{n}\right). \end{aligned}$$

Since the Fourier transform of K is infinitely differentiable, the tails of K decay faster than the inverse of any polynomial. Therefore, as a function of h , the two integral remainder terms converge to zero faster than the inverse of any polynomial in h .

This leaves just the convolution to be dealt with. For notational convenience, let $K_h(x) := (1/h)K(x/h)$, and let \check{g} denote the Fourier transform of a function g .

$$\begin{aligned} \left| \int_{-\infty}^{\infty} r(u) K_h(x-u) du - r(x) \right| &= \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} \check{r}(s) \check{K}_h(s) e^{-isx} ds \right. \\ &\quad \left. - \frac{1}{2\pi} \int_{-\infty}^{\infty} \check{r}(s) e^{-isx} ds \right| \\ &= \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} (\lambda(hs) - 1) \check{r}(s) e^{-isx} ds \right|. \end{aligned}$$

Since $\lambda(hs) = 1$ when $|hs| \leq c$, or equivalently when $|s| \leq c/h$,

$$\begin{aligned}
\left| \frac{1}{2\pi} \int_{-\infty}^{\infty} (\lambda(hs) - 1) \check{r}(s) e^{-isx} ds \right| &= \frac{1}{2\pi} \left| \int_{-\infty}^{-c/h} (\lambda(hs) - 1) \check{r}(s) e^{-isx} ds \right. \\
&\quad \left. + \int_{c/h}^{\infty} (\lambda(hs) - 1) \check{r}(s) e^{-isx} ds \right| \\
&\leq \frac{1}{2\pi} \int_{|s| > c/h} |\check{r}(s)| ds \\
&= \frac{1}{2\pi} \int_{|s| > c/h} \frac{s^k}{s^k} |\check{r}(s)| ds \\
&\leq \frac{1}{2\pi} \left(\frac{h}{c} \right)^k \int_{|s| > c/h} s^k |\check{r}(s)| ds \\
&= O(h^k).
\end{aligned}$$

□

PROOF OF THEOREM 4: The proof proceeds by verifying that the Liapunov condition holds, which is sufficient for the Lindeberg-Feller central limit theorem.

Let $w_i = (1/h) \int_{s_{i-1}}^{s_i} K((x-u)/h) du$. Then $(\hat{r}(x) - E[\hat{r}(x)]) = \sum_{i=1}^n w_i \epsilon_i$. By the Liapunov condition, it is sufficient to show

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E|w_i \epsilon_i|^3}{[\sum_{i=1}^n \text{Var}(w_i \epsilon_i)]^{3/2}} = 0.$$

If the denominator is multiplied by $(nh)^{3/2}$ it will converge to a nonzero constant. By equation (8) and Theorem 1,

$$\begin{aligned}
\text{Var}(\hat{r}(x)) &= \sum_{i=1}^n \text{Var}(w_i \epsilon_i) \\
&= \frac{\sigma^2}{nh} \int_{(x-1)/h}^{x/h} \frac{1}{f(x-hz)} K^2(z) dz + O\left(\frac{1}{n^2 h^3}\right).
\end{aligned}$$

By dominated convergence,

$$\lim_{n \rightarrow \infty} nh \sum_{i=1}^n \text{Var}(w_i \epsilon_i) = \sigma^2 \int_{-\infty}^{\infty} \frac{1}{f(x)} K^2(z) dz > 0.$$

Hence, it will suffice to show $(nh)^{3/2} \sum_{i=1}^n E|w_i \epsilon_i|^3 \rightarrow 0$.

$$\begin{aligned}
\sum_{i=1}^n |w_i|^3 E|\epsilon_i|^3 &\leq \frac{B}{h} \max_{1 \leq i \leq n} (s_i - s_{i-1}) \max_{u \in \mathbf{R}} |K(u)| \sum_{i=1}^n w_i^2 \\
&\leq \frac{C}{nh} \sum_{i=1}^n w_i^2.
\end{aligned}$$

This is $O(1/(n^2 h^2))$ by Theorem 1 and equation (5). Therefore,

$$(nh)^{3/2} \sum_{i=1}^n |w_i|^3 E|\epsilon_i|^3 \rightarrow 0.$$

□

References

- [1] Luc Devroye. A note on the usefulness of superkernels in density estimates. *Annals of Statistics*, 20(4):2037–2056, 1992.
- [2] Th. Gasser and H.-G. Müller. Kernel estimation of regression functions. In Th. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, number 757 in Springer Lecture Notes in Mathematics, pages 23–68. Springer-Verlag, Berlin, 1979.
- [3] Th. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11:171–185, 1984.
- [4] Peter Hall and Thomas E. Wehrly. A geometrical method for removing edge effects from kernel type nonparametric regression estimators. *Journal of the American Statistical Association*, 86(415):665–672, 1991.
- [5] Jeffrey D. Hart. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York, 1997.
- [6] Karen Messer and Larry Goldstein. A new class of kernels for nonparametric curve estimation. *Annals of Statistics*, 21(1):179–195, 1993.
- [7] Dimitris N. Politis. Adaptive bandwidth choice. Submitted, 2001.
- [8] Dimitris N. Politis. On nonparametric function estimation. In Ch. A. Charalambides, Markos V. Koutras, and N. Balakrishnan, editors, *Probability and Statistical Models with Applications*, pages 469–483. Chapman & Hall/CRC, Washington, D.C., 2001.
- [9] Dimitris N. Politis and Joseph P. Romano. Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis*, 16(1):67–103, 1995.
- [10] Dimitris N. Politis and Joseph P. Romano. Multivariate density estimation with general flat-top kernels of infinite order. *Journal of Multivariate Analysis*, 68:1–25, 1999.