THE INSTITUTE
OF STATISTICS

THE CONSOLIDATED UNIVERSITY
OF NORTH CAROLINA

ADAPTATION TO HIGH SPATIAL INHOMOGENEITY BASED ON WAVELETS AND ON LOCAL LINEAR SMOOTHING

by

Jianqing Fan Peter Hall
Michael Martin Prakash Patil

August 1993

Mimeo Series #2307

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

MIMEO SERIES #2307

J. Fan P.Hall
M. Martin P. Patil

ADAPTATION TO HIGH SPATIAL INHOMOGENEITY BASED ON WAVELETS AND ON LOCAL LINEAR SMOOTHING

NAME DATE

ADAPTATION TO HIGH SPATIAL INHOMOGENEITY BASED ON WAVELETS AND ON LOCAL LINEAR SMOOTHING *

JIANQING FAN[1] PETER HALL[2]
MICHAEL MARTIN[3] PRAKASH PATIL[2]

August 24, 1993

ABSTRACT. We develop mathematical models for functions with aberrant, high-frequency episodes, and describe the ability of wavelet-based estimators to capture those features. Our results have a genuinely local character, in that they describe pointwise asymptotic properties of curve estimators. Previous accounts of the performance of wavelet methods have been based on global rates of convergence uniformly over very large function classes; in marked contrast, we establish local convergence rates for single function. We allow those functions to depend on sample size, so that we may describe the extent to which sample size influences the type of short, sharp aberrations that may be accurately recovered from noisy observations. It is shown that wavelet methods based on thresholding, and employing a relatively arbitrary level of primary resolution, capture high-frequency episodes with an accuracy that is within a logarithmic factor of being optimal. We point out that this factor derives from the estimators being somewhat oversmoothed, with systematic errors of larger order than their stochastic errors. That undersmoothing is, in turn, a consequence of inadequate choice of primary resolution. In principle, this difficulty may be overcome by adjusting the primary resolution level in an adaptive way, but that is not a practically appealing proposition, not least because of its computational complexity. By way of contrast, methods based on more traditional smoothing approaches can be applied locally to obtain estimators that outperform wavelet methods in terms of pointwise convergence rates. In particular, we show that estimators based on local linear smoothing attain the optimal convergence rates, even in the presence of unusually high frequencies in the curve. Moreover, these local smoothing methods are straightforward, in both conception and execution.

AMS (1991) SUBJECT CLASSIFICATION. Primary 62G07, Secondary 62G20.

SHORT TITLE. Local adaptivity of curve estimator.

[1] Department of Statistics, University of North Carolina at Chapel Hill.
[2] Centre for Mathematics and its Applications, Australian National University.
[3] Department of Statistics, Stanford University.

1 Introduction

Wavelet transforms are a device for representing functions $f(x)$ in a way that is local in both the argument x and the roughness of f. Consequently, when wavelet methods are applied to produce statistical estimates of a function from noisy data, they provide levels of smoothing which automatically adapt to local variations in the roughness of f as x changes. This *local* adaptability has traditionally been expressed in terms of the *global* performance of wavelet curve estimates over very large function classes. In particular, such an approach is taken in the seminal work of Donoho (1992), Donoho and Johnstone (1992a,b), Donoho, Johnstone, Kerkyacharian and Picard (1993) and Kerkyacharian and Picard (1993a,b), who introduced wavelet methods to statistics. In the present paper we develop theory that explicitly describes the performance of wavelet methods in a local setting. In particular, we develop upper bounds to pointwise rates of convergence. These upper bounds enable us to show explicitly how wavelet estimators react to local episodes in a curve, and how they adapt to them in a nearly-optimal way. Our distinctly local approach to mathematical modeling differs markedly from the global viewpoint adopted by earlier authors, who have relied on uniform global convergence rates to convey information about the issues that we are addressing here.

From the viewpoint of mathematical modeling we define irregular episodes as varying frequencies of the target function f. The frequencies may be very large, and may be quite different at different points. We demonstrate that, up to a logarithmic factor, wavelet methods manage to adjust optimally to varying frequencies by changing the level of resolution. Our results are quite different from those of other authors, primarily because the target functions in our estimation problems do not come from traditional function classes (e.g. Besov spaces). Instead, the smoothness of the functions is allowed to vary with sample size, reflecting the fact that with larger sample sizes one might hope to be able to capture more erratic episodes in a curve.

In the remainder of this section we summarise our main results and discuss their relationship to more traditional descriptions of curve estimation. Adapting ideas from Hall and Patil (1993a) we take the target function to have the form

$$f(x) = f_0(x) + \sum_{j=1}^{N} \gamma_j\{\omega_j(x - x_j)\}, \tag{1.1}$$

where $x_1, \cdots, x_N$ are distinct fixed points, the functions γ_j are fixed, but the frequencies $\omega_j \geq 1$ may depend on sample size n. The fact that ω_j may vary with n, and may diverge to infinity as n increases, is critical to our analysis. It allows the irregularity of the episode in f at x_j to depend on sample size, and in fact the smoothness of f is governed by the sizes of $\omega_1, \cdots, \omega_N$. (There are no serious technical problems in allowing N and the x_j's to vary with sample size, but doing so would significantly complicate our discussion.) Within the context of fixed x_j's, and assuming that the functions f_0 and γ_j have r bounded derivatives, we shall show that an r'th order wavelet estimator of f achieves the mean square convergence rate $(\omega_j n^{-1} \log n)^{2r/(2r+1)}$ *simultaneously* at each respective point x_j; and that the mean square convergence rate at points between the x_j's is $O\{(n^{-1} \log n)^{2r/(2r+1)}\}$. These properties identify, and indeed are characteristic of, the spatial adaptation features of wavelet shrinkage. The rates are obtainable using estimators with primary resolution of order $\{n(\log n)^{2r}\}^{1/(2r+1)}$ or smaller (including the case where primary resolution is fixed, not depending on n). They should be compared with the rate $O\{(\omega_j n^{-1})^{2r/(2r+1)}\}$, which is optimal in this setting. Thus, wavelet methods achieve optimal local convergence rates within a logarithmic factor.

The logarithmic factor that appears in our rates arises from the form of the threshold, which is generally $\delta = \text{const.}(n^{-1} \log n)^{1/2}$. Only those estimated wavelet coefficients that are larger than δ are included in the empirical wavelet transform. If the logarithmic factor could be removed from δ then the logarithm would also vanish from the convergence rates above. However, removing the $\log n$ from δ would dramatically alter the convergence properties of the infinite series that defines the wavelet estimator – see for example Hall and Patil (1993b), where the issue of smallest possible δ is addressed in detail. Thus, the desired improvement in accuracy is not achievable by simply adjusting the threshold.

As indicated two paragraphs above, a wavelet estimator achieves the convergence rate $(\omega_j n^{-1} \log n)^{2r/(2r+1)}$ at x_j by fixing a global level of primary resolution, p, say, and using thresholding to capture local features, appealing to the multiresolution property of the wavelet transform. The operation of thresholding influences only the bias of the estimator, to first order, and does not appreciably affect the variance. As such, it does not produce statistical smoothing in the usual sense. In particular, it does not achieve the usual trade-off between bias and variance (to first order) that is usually the hallmark of statistical smoothing. For this reason the estimator is often used in an oversmoothed form, its variance contribution to mean squared error, of order $n^{-1}p$, usually being of smaller order than its bias contribution.

This fact provides further insight into why the estimator does not achieve optimality – the estimator oversmooths, with consequent loss of performance. Performance may be improved, and the logarithm removed from the convergence rate, by utilizing appropriate smoothing as well as thresholding. If $p = p_j$ is chosen so that $n^{-1}p$ and $(\omega_j p^{-1})^{2r}$ (representing variance and squared bias, respectively) are of the same size, then the overall convergence rate of mean squared error of the estimator of $f(x_j)$ is $(\omega_j n^{-1})^{2r/(2r+1)}$. As a result, the logarithmic factor has been removed from the convergence rate, which has been correspondingly improved. However, this enhancement has been achieved at the expense of altering the primary threshold at each point x_j where the function f has a significantly different frequency. Such an approach is computationally awkward, and requires an empirical device for choosing the primary resolution adaptively according to location. For many purposes the logarithmic factor is a small price to pay for computational expediency. In contrast, however, we shall show that adaptive local linear smoothing methods compete favourably with wavelet methods, in terms of both accuracy and ease of application. Indeed, we shall demonstrate in Section 4 that adaptive linear estimation methods, such as those investigated by Fan and Gijbels (1993), are able to attain optimal pointwise convergence rates even in the presence of unusually high frequencies in the curve. In this critical sense adaptive local linear smoothing methods outperform globally thresholded wavelet methods.

The case of nonparametric density estimation, where f is a probability density, is the simplest to discuss. Hence we begin by treating that setting, in Section 2. The case of regression is addressed in Section 3, and that of adaptive linear smoothing in Section 4. All proofs are deferred to Section 5.

Work of Hall and Patil (1993a) is closely related to our own in that functions with increasingly high-frequency episodes are treated there. However, those authors address only global convergence rates, and assume that high-frequency episodes in the curve are present over relatively large intervals. In contrast, the oscillations represented by the term $\gamma_j\{\omega_j(x - x_j)\}$ in

(1.1) vanish very quickly, outside an interval of width $O(\omega_j^{-1})$, and so are inherently more difficult to detect. Fan, Hall, Martin and Patil (1993) examine the performance of wavelet methods in the context of more subtle local features of a very smooth, fixed curve, not depending on sample size.

2 Density estimators based on wavelet shrinkage

2.1 Wavelet transforms. We summarize here the basic theory of wavelet transforms. In the next subsection we put it into an empirical framework for estimating density functions. Our main theoretical results are presented in subsection 2.3.

The key ingredients of our analysis are discussed in much more detail by Strang (1989, 1993), Meyer (1990) and Daubechies (1992). We first review some key features of the multiresolution analysis of Meyer (1990). See also Section 5.1 of Daubechies (1992). Suppose there exists a "scale function" or "father wavelet" ϕ such that

1. $V_k \subset V_{k+1}$, where V_k denotes the space spanned by $\{2^{k/2}\phi(2^k x - \ell), \ell \in \mathbb{Z}\}$;

2. the sequence $\{2^{k/2}\phi(2^k x - \ell), \ell \in \mathbb{Z}\}$ is an orthonormal family in $L^2(\mathbb{R})$.

A necessary condition of the above requirements is that ϕ satisfy the so-called scaling equation,

$$\phi(x) = \sum_{\ell} c_\ell \phi(2x - \ell), \tag{2.1}$$

where the constants c_ℓ have the property

$$\sum_{\ell} c_\ell c_{\ell-2m} = 2\delta_{0m}, \tag{2.2}$$

and δ_{ij} is the Kronecker delta. Conditions (2.1) and (2.2) correspond respectively to requirements 1 & 2; see Strang (1989). Then $\cap_{k\in\mathbb{Z}} V_k = \{0\}$, and if, in addition, $\phi \in L^2(\mathbb{R})$ and $\int \phi = 1$, $L^2(\mathbb{R}) = \cup_{k\in\mathbb{Z}} V_k$. The scale of V_k becomes increasingly fine as k increases.

The scaling coefficients $\{c_j\}$ uniquely determine the function ϕ under appropriate regularity conditions. Further, if $\{c_\ell\}$ has bounded support, so does ϕ.

The most commonly-used wavelet functions are those having bounded support with $r-1$ vanishing moments, for some $r \geq 1$:

$$\int x^j \phi(x)dx = \delta_{0,j} \text{ for } j = 0, \cdots, r-1 \text{ and } \int |x^r \phi(x)|dx < \infty. \tag{2.3}$$

See Daubechies (1992) for constructions of this family. Translated to the scaling parameters c_ℓ, conditions (2.3) entail $\sum c_\ell = 2$, $\sum_\ell (-1)^\ell \ell^j c_\ell = 0$ $(0 < j \leq r-1)$, $\sum \ell^{2r} c_\ell^2 < \infty$; see Strang (1989).

Under these assumptions there exists a function ψ (the "mother" wavelet) given by

$$\psi(x) = \sum_\ell (-1)^\ell c_{1-\ell} \phi(2x - \ell),$$

such that

1. $\{2^{k/2}\psi(2^k x - \ell), \ell \in \mathbb{Z}\}$ is an orthonormal basis of W_k, where W_k is the space such that $V_{k+1} = V_k \oplus W_k$;
2. $\{2^{k/2}\psi(2^k x - \ell), \ell \in \mathbb{Z}, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$;
3. the zero'th and first $r-1$ moments of ψ vanish:

$$\int x^j \psi(x) dx = 0 \text{ for } j = 0, \cdots, r-1 \text{ and } \int |x^r \psi(x)| dx < \infty. \tag{2.4}$$

In practice, ϕ and ψ are typically compactly supported, and so we impose that condition here. The sequence $\{\phi(x-\ell),\ 2^{k/2}\psi(2^k x - \ell),\ \ell \in \mathbb{Z}, k \geq 0\}$ is a complete orthonormal basis of $L^2(\mathbb{R})$.

Let $p > 0$ denote the level of prmary resolution, and define $p_k = p2^k$. Put

$$\phi_\ell(x) = p^{1/2}\phi(px - \ell) \quad \text{and} \quad \psi_{k\ell}(x) = p_k^{1/2}\psi(p_k x - \ell)$$

for an integer $\ell \in \mathbb{Z}$. Then, as noted in the previous paragraph, the bases $\{\phi_\ell(x), \psi_{k\ell}(x), \ell \in \mathbb{Z}, k \in \mathbb{Z}_+\}$ are completely orthonormal for $L^2(\mathbb{R})$: for any $f \in L_2(\mathbb{R})$,

$$f(x) = \sum_\ell b_\ell \phi_\ell(x) + \sum_{k=0}^{\infty} \sum_\ell b_{k\ell}\psi_{k\ell}(x), \tag{2.5}$$

with wavelet coefficients

$$b_\ell = \int f(x)\phi_\ell(x) dx, \qquad b_{k\ell} = \int f(x)\psi_{k\ell}(x) dx. \tag{2.6}$$

2.2 Empirical wavelet transforms for density estimation. The orthonormal bases discussed above can easily be applied to statistical function estimation. The most convenient setting is perhaps density estimation. Let $X_1, \cdots, X_n$ be a random sample from a distribution with density f. Formulae (2.6) suggests unbiased estimates of the wavelet coefficients:

$$\hat{b}_\ell = \frac{1}{n}\sum_{i=1}^{n}\phi_\ell(X_i), \qquad \hat{b}_{k\ell} = \frac{1}{n}\sum_{i=1}^{n}\psi_{k\ell}(X_i). \tag{2.7}$$

For high resolution (i.e. large p_k), the estimate $\hat{b}_{k\ell}$ will basically be noise, since $\psi_{k\ell}$ is supported only in a small neighborhood around ℓ/p_k and hence very few data points are used to calculate $\hat{b}_{k\ell}$. (Indeed, if ψ is compactly supported then $\psi_{k\ell}$ vanishes outside an interval of width $O(p_k^{-1})$.) Following Donoho and Johnstone (1992a), we select useful estimated coefficients $\hat{b}_{k\ell}$ by "thresholding". Considerations of this nature suggest the estimator

$$\hat{f}(x) = \sum_\ell \hat{b}_\ell\phi_\ell(x) + \sum_{k=0}^{\infty}\sum_\ell \hat{b}_{k\ell}I\{|\hat{b}_{k\ell}| \geq \delta\}\psi_{k\ell}(x); \tag{2.8}$$

compare (2.5). Asymptotic theory developed by Donoho, Johnstone, Kerkyacharian and Picard (1993), and Hall and Patil (1993a), suggests taking $\delta = c(n^{-1}\log n)^{1/2}$, where $c > 0$ is a constant. Following Donoho and Johnstone (1992a), the above estimator corresponds to "hard thresholding". "Soft thresholding" involves replacing $\hat{b}_{k\ell}1_{\{|\hat{b}_{k\ell}|\geq\delta\}}$ in (2.8) by $\text{sgn}(\hat{b}_{k\ell})(|\hat{b}_{k\ell}|-\delta)_+$, leading to the estimator

$$\hat{f}_S(x) = \sum_\ell \hat{b}_\ell\phi_\ell(x) + \sum_{k=0}^{\infty}\sum_\ell \text{sgn}(\hat{b}_{k\ell})(|\hat{b}_{k\ell}| - \delta)_+\psi_{k\ell}(x). \tag{2.9}$$

The intuition behind either type of thresholding is based on the 'signal-to-noise' ratio. When this ratio is larger than a certain threshold, the (k, ℓ)'th term is included in the sum; otherwise, the (k, ℓ)'th term is omitted from the sum.

2.3 Asymptotic theory for wavelet density estimators. We begin by addressing the case of densities of the form (1.1). To ensure that f is a density for all sufficiently large choices of the ω_j's we ask that f_0 be a fixed, r-times differentiable density bounded away from zero on an interval $\mathcal{I} = (-B, B)$; that the points $x_1 < ... < x_N$ all lie within $\mathcal{I}$; that the support of γ_j is contained within an interval $\mathcal{I}_j = (-B_j, B_j)$; that γ_j have r bounded derivatives on $\mathcal{I}_j$ with $\int \gamma_j = 0$; and that $\inf_{(-B,B)} f_0 > -\min_j \inf_{\mathcal{I}_j} \gamma_j$. Then there exist constants B_0 and $\omega^* \geq 1$, depending on N and $x_1, ..., x_N$, such that for all $\omega_1, ..., \omega_N \geq \omega^*$, f is bounded above by B_0 on $\mathcal{I}$ and is a proper density function.

To picture the f's that the model (1.1) generates, consider, for example, the class of Normal mixtures. Densities in that class range from very smooth curves to very rough curves – see for example Marron and Wand (1992). The "Doppler" example treated by Donoho and Johnstone (1992a) is also of this type.

Our first result treats the mean squared error of density estimators under model (1.1). We assume throughout that ϕ and ψ are bounded and compactly supported, satisfy (2.3) and (2.4), and are such that the functions $\phi_\ell, \psi_{k\ell}, -\infty < \ell < \infty, k \geq 0$ form a complete orthonormal family.

Theorem 2.1. *Take $\delta = c(||f||_\infty n^{-1} \log n)^{1/2}$, where the constant $c \geq \sqrt{6}$. Let $0 < \varepsilon < 1$, and $\eta_1 \leq \eta_2$ denote positive numbers converging to zero as $n \to \infty$ and such that $\eta_1^{-1}(n^{-1+\varepsilon} \log n)^{2r/(2r+1)}$ is bounded. Let ω_0 and C be fixed positive numbers, and assume that $\max_{0 \leq j \leq N} \omega_j = O(n^\varepsilon)$. Let x_0 be any real number not in the set $(x_1, \ldots, x_N)$. Then for $0 \leq j \leq N$,*

$$E\{\hat{f}(x_j) - f(x_j)\}^2 = O\{n^{-1}p + (\omega_j n^{-1} \log n)^{2r/(2r+1)}\} \tag{2.10}$$

uniformly in values of p and q satisfying $p \geq C$ and $\eta_1 \leq n^{-1}p2^q \log n \leq \eta_2$.

The following heuristic argument in terms of the traditional smoothing notion of balancing squared bias against variance clarifies the origins of, and the roles played by, the various terms on the right-hand side of (2.10). The analogue of bandwidth for wavelet-based estimators is the quantity p^{-1}, where p denotes the level of primary resolution. For kernel estimators with bandwidth h the variance contribution to mean squared error is of size $(nh)^{-1} = n^{-1}p$, which is the first term on the right-hand side of (2.10). The squared bias contribution for kernel estimations is of size $(h^r f^{(r)})^2$, which at the point x_j is of size $\zeta = (p^{-r}\omega_j^r)^2$. If p is of smaller order than $(n/\log n)^{1/(2r+1)}$ then the bias term on the right-hand side of (2.10), $(n^{-1} \log n)^{2r/(2r+1)}$, is actually of smaller order than ζ. The ability of wavelet methods to achieve this result is a consequence of the multiresolution property of the wavelet transform. However, despite multiresolution the squared bias contribution for wavelet estimators is still of larger order than that from the variance, and the density estimator is still oversmoothed. The optimal amount of smoothing is achieved when p is chosen so that the variance and the "nonmultiresolution component" of squared bias are of the same size – that is, when $n^{-1}p$ and

$(p^{-r}\omega_j^r)^2$ are of the same size. It is simplest to discuss this context when ω is fixed, which is the case we shall address next.

Our next theorem addresses the case of densities whose r'th derivatives are bounded. Let $\mathcal{F} = \mathcal{F}(r, B)$ denote the class of r-times differentiable densities f on the real line, such that both $||f||_\infty$ and $||f^{(r)}||_\infty \le B$.

Theorem 2.2. *Assume conditions (2.1) - (2.4) on the wavelet functions ψ and ϕ. Take $\delta = \delta_f = c(||f||_\infty n^{-1} \log n)^{1/2}$, where the constant $c \ge \sqrt{6}$. Let $\eta_1 \le \eta_2$ denote a positive number converging to zero as $n \to \infty$ and such that $\eta_1^{-1}(n^{-1}\log n)^{2r/(2r+1)}$ is bounded. Then*

$$\sup_{-\infty<x<\infty; f\in\mathcal{F}} E\{\hat{f}(x) - f(x)\}^2 = O\{n^{-1}p + (n^{-1}\log n)^{2r/(2r+1)}\} \tag{2.11}$$

uniformly in p, q satisfying $p \ge C$ and $\eta_1 \le p2^q n^{-1} \log n \le \eta_2$, for arbitrary fixed $C > 0$.

The proof of Theorem 2.2 is similar to that of Theorem 2.1 and is omitted.

An immediate consequence of this result is that if $\mathcal{G} = \mathcal{G}(r, B_1, B_2)$ denotes the class of densities f for which both $||f||_\infty$ and $||f^{(r)}||_\infty \le B_1$, and whose support is contained within $(-B_2, B_2)$, then

$$\sup_{f\in\mathcal{G}} \int E(\hat{f} - f)^2 = O\{n^{-1}p + (n^{-1}\log n)^{2r/(2r+1)}\} \tag{2.12}$$

uniformly in p, q as specified in Theorem 2.2. Generalizations to densities with unbounded support are also possible.

It is clear from (2.11) and (2.12) that by choosing p to be of order $\{n(\log n)^{2r}\}^{1/(2r+1)}$, or smaller (for example, fixed p is adequate), we ensure a mean square convergence rate of $O\{(n^{-1}\log n)^{2r/(2r+1)}\}$. This exceeds the optimal convergence rate, $n^{-2r/(2r+1)}$, by only a logarithmic factor. (See Stone (1980, 1982) for an account of optimality in this setting.) Hall and Patil (1993b) show that the rate of convergence in (2.12) may be improved to the optimal one by taking p to be of size $n^{1/(2r+1)}$. Uniformity is not established there, but it may be readily derived. It is also demonstrated by Hall and Patil (1993b) that for fixed f, and for p of size $\{n(\log n)^{2r}\}^{1/(2r+1)}$ or smaller, the bias contribution to the mean integrated squared error of $\hat{f}$ dominates the variance contribution, and

$$\int E(\hat{f} - f)^2 \sim C(f)(n^{-1}\log n)^{2r/(2r+1)}, \tag{2.13}$$

where the constant $C(f)$ does not depend on n, p or q. This last result demonstrates that the upper bound evinced by (2.12) (and also by (2.11), in an average sense) is best possible – compare the right-hand sides of (2.12) and (2.13). Likewise, it is straightforward to establish a lower bound to result (2.10), in which the bound is of size $(n^{-1}\omega_j \log n)^{1/(2r+1)}$.

Theorems 2.1 and 2.2 apply without change to the soft thresholded estimator $\tilde{f}$.

The analogues of (2.11) and (2.12) for r'th order kernel density estimators, with bandwidth h, are of course

$$\sup_{-\infty<x<\infty; f\in\mathcal{F}} E\{\hat{f}(x) - f(x)\}^2 = O\{(nh)^{-1} + h^{2r}\} \tag{2.14}$$

$$\sup_{g\in\mathcal{G}} \int E(\hat{f} - f)^2 = O\{(nh)^{-1} + h^{2r}\}. \tag{2.15}$$

See for example Silverman (1986, Chapter 3). Choosing h to be of size $n^{-1/(2r+1)}$ we obtain optimal convergence rates.

Next we establish a similar result for derivative estimation. Assume that the wavelet functions ψ and ψ have $\nu \geq 1$ bounded derivatives. In view of (2.8) and (2.9), density derivative estimation can be defined by

$$\begin{aligned} \hat{f}^{(\nu)}(x) &= p^{\nu} \sum_{\ell} \hat{b}_{\ell} \phi_{\ell}^{(\nu)}(x) + \sum_{k=0}^{q-1} \sum_{\ell} \hat{b}_{k\ell} I\{|\hat{b}_{k\ell}| \geq \delta\} p_k^{\nu} \psi_{k\ell}^{(\nu)}(x), && (2.16) \\ \hat{f}_S^{(\nu)}(x) &= p^{\nu} \sum_{\ell} \hat{b}_{\ell} \phi_{\ell}^{(\nu)}(x) + \sum_{k=0}^{q-1} \sum_{\ell} \operatorname{sgn}(\hat{b}_{k\ell}) \{|\hat{b}_{k\ell}| - \delta\}_{+} p_k^{\nu} \psi_{k\ell}^{(\nu)}(x). && (2.17) \end{aligned}$$

Another method of estimating derivative function is given by Hall and Patil (1993b). Suppose the wavelet basis is chosen such that

$$f^{(\nu)}(x) = p^{\nu} \sum_{\ell} b_{\ell} \phi_{\ell}^{(\nu)}(x) + \sum_{k=0}^{\infty} \sum_{\ell} b_{k\ell} p_k^{\nu} \psi_{k\ell}^{(\nu)}(x),$$

uniformly in $x \in [-B, B]$. The following Theorem shows that these derivative estimators also possess spatial adaptation.

Theorem 2.3. *Let $\delta = c(||f||_{\infty} n^{-1} \log n)^{1/2}$, where the constant $c \geq \sqrt{6 + 4\nu}$. Under conditions of Theorem 2.1,*

$$E\{\hat{f}^{(\nu)}(x_j) - f^{(\nu)}(x_j)\}^2 = O\{n^{-1} p^{2\nu+1} + \omega_j^{(2\nu+1)2r/(2r+1)} (n^{-1} \log n)^{2(r-\nu)/(2r+1)}\}$$

uniformly in values of p and q satisfying $p \geq C$ and $\eta_1 \leq n^{-1} p 2^q \log n \leq \eta_2$.

We omit the proof of Theorem 2.3.

3 Wavelet shrinkage in nonparametric regression

3.1 Canonical nonparametric regression. As Donoho and Johnstone (1992a,b) have shown, wavelet shrinkage can be readily applied to regression problems with regularly spaced design. To appreciate how, suppose we observe data

$$Y_i = m(X_i) + \varepsilon_i, \tag{3.1}$$

where $X_i = i/n$ and the ε_i's are independent and identically distributed with zero mean and variance σ^2, and the mean regression function is as given in (1.1), with f replaced by m. Then the wavelet expansion of m and its coefficients, b_ℓ's and $b_{k\ell}$'s, are given in (2.5) and (2.6). Estimators of b_ℓ and $b_{k\ell}$ are

$$\hat{b}_\ell = \frac{1}{n}\sum_{i=1}^{n} \phi_\ell(X_i)Y_i, \qquad \hat{b}_{k\ell} = \frac{1}{n}\sum_{i=1}^{n} \psi_{k\ell}(X_i)Y_i, \tag{3.2}$$

and the wavelet shrinkage estimators of m are defined by

$$\begin{aligned} \hat{m}(x) &= \sum_\ell \hat{b}_\ell\phi_\ell(x) + \sum_{k=0}^{\infty}\sum_\ell \hat{b}_{k\ell}1_{\{|\hat{b}_{k\ell}|\ge\delta\}}\psi_{k\ell}(x), \\ \hat{m}_S(x) &= \sum_\ell \hat{b}_\ell\phi_\ell(x) + \sum_{k=0}^{\infty}\sum_\ell \operatorname{sgn}(\hat{b}_{k\ell})(|\hat{b}_{k\ell}| - \delta)_+\psi_{k\ell}(x). \end{aligned}$$

(If n is a power of 2, then Mallat's pyramid algorithm can effect the above computation very rapidly.)

Next we state the analogue of Theorem 2.1 for the above regression setting. The conditions on ϕ and ψ are as before. In addition, we assume that ϕ and ψ are Hölder continuous. For simplicity we assume that ε_i's are either Normally distributed or bounded.

Theorem 3.1. *Take $\delta = c\sigma(n^{-1}\log n)^{1/2}$, where $c > 0$ is sufficiently large. Let $\varepsilon, \eta_1, \eta_2, \omega_0, C$ and x_0 be as Theorem 2.1. If $\max_{0\le j\le N}\omega_j = O(n^\varepsilon)$, then for $0 \le j \le N$,*

$$E_f\{\hat{m}(x_j) - m(x_j)\}^2 = O\{n^{-1}p + (\omega_j n^{-1}\log n)^{2r/(2r+1)}\}, \tag{3.3}$$

uniformly in values of p and q satisfying $p \ge C$ and $\eta_1 \le n^{-1}p2^q\log n \le \eta_2$.

An heuristic argument illustrating from where the various terms on the right-hand side of (2.10) originate has an analogue in (3.3), which we shall not pursue here.

Our results on derivative estimation are readily extended to the present regression setting. We omit details.

3.2. Options in highly inhomogeneous designs. Wavelet shrinkage is more difficult to apply to nonuniform designs. To appreciate this problem, let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a random sample from a bivariate distribution with mean and variance respectively given by

$$m(x) = E(Y|X = x), \qquad \mathrm{var}(Y|X = x) = \sigma^2(x).$$

One approach to estimating m is to use the wavelet expansion to a certain resolution q

$$m(x) \approx \sum_{\ell} b_\ell \phi_\ell(x) + \sum_{k=0}^{q-1} \sum_{\ell} b_{k\ell} \psi_{k\ell}(x),$$

then apply the least squares method

$$\min \sum_{i=1}^{n} \left\{ Y_i - \sum_{\ell} b_\ell \phi_\ell(X_i) - \sum_{k=0}^{q-1} \sum_{\ell} b_{k\ell} \psi_{k\ell}(X_i) \right\}^2$$

to find wavelet coefficients, and finally use "soft" or "hard" thresholding to construct an appropriate estimator. We remark that estimators (3.2) in the uniform design case can be thought of having been obtained in this way. Given the local nature of wavelets, heteroscedasticity is not a major concern in the above least-squares problem. However, the property of orthogonality got lost — the vectors $\{\phi_\ell(X_i)\}_{i=1}^n$ and $\{\psi_{k\ell}(X_i)\}_{i=1}^n$ and are no longer nearly orthogonal unless design is uniform, and thus wavelets do not help much in computation. Besides, difficulties in choosing the terms in ℓ arise (in particular, those ℓ for which only a portion of data are in the support of $\psi_{k\ell}$), again because of the non-orthogonality.

A second option is to use

$$m(x) = \left\{ \int y f(x, y) dy \right\} / f(x) \equiv m_1(x)/f(x)$$

and to use wavelet shrinkage separately to estimate $m_1(x)$ and $f(x)$. The estimate of $f(x)$ was studied in Section 2.3. To see how to estimate $m_1(x)$, note that wavelet coefficients for the function $m_1(x)$ are

$$b_\ell = \int \int \phi_\ell(x) y f(x, y) dx dy \quad b_{k\ell} = \int \int \psi_{k\ell}(x) y f(x, y) dx dy,$$

and may be estimated as in (3.2). Now apply "soft" or "hard" shrinkage to these coefficients, to obtain an estimate of $\hat{m}_1(x)$. Spatial adaptation results can similarly be obtained, but two separate estimates of $m_1(x)$ and $f(x)$ makes the idea less attractive. In particular, the local neighborhoods used by $\hat{m}_1(x)$ and $\hat{m}(x)$ can be very different. The extension to estimating higher-order derivatives $m^{(\nu)}(x)$ are not very convenient. Note too that $\text{var}(\hat{b}_{k\ell})$ is no longer nearly constant unless design is uniform and errors are homoscedastic.

A third option involves binning. Partition the x-axis into to N equispaced bins. Let $x_j, c_j, \bar{Y}_j$ be respectively the bin center, bin counts and the average of Y_i's in that center. Now apply wavelet shrinkage to the data $(x_j, \bar{Y}_j)$ as in the canonical regression setup. Owing to the different number of observations in different bins, $\bar{Y}_j$ exhibits heteroscedasticity even if the original data was homoscedastic. This makes wavelet shrinkage more difficult.

While wavelet shrinkage can in principle be used with the above options, we must admit the limitations of this application when the design is nonuniform, in particular, not least primary interest focuses on derivative estimation. In this case, local polynomial regression can easily be used. With variable smoothing, it readily adapts to spatial inhomogeneity. See Fan and Gijbels (1993) for a data-driven choice of these bandwidths.

4 Locally adaptive kernel methods

In this section we show how locally adaptive kernel estimators achieve optimal convergence rates in the context of relatively extreme episodes in a curve, such as those described by model (1.1). In particular, the convergence rates do not contain an extraneous logarithmic factor, which as we noted in Section 2 is an inherent consequence of the thresholding method by means of which wavelet estimators achieve their local adaptability. Our main results are formulated for a general procedure for bandwidth choice, and are later shown to be applicable to a simple "plug in" rule.

As in Section 2, the unknown density f is to be estimated from data $\{X_1, \cdots, X_n\}$. In a slight abuse of notation we redefine

$$\hat{f}(x) = \hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

where h denotes bandwidth and K is of bounded variation, compactly supported and of r'th order – the latter constraint is equivalent to asking that

$$\int x^j K(x)dx = \delta_{0,j}, \quad j = 0, \cdots, r-1; \quad \int x^r K(x)dx = \kappa/r!,$$

where κ is nonzero. We shall choose the bandwidth h to depend on location, x, and shall do that in a data-dependent way. To reflect these dependences we shall write $\hat{h}(x)$ for h.

Suppose initially that the unknown density f is fixed, not depending on sample size n. Then the optimal bandwidth $h^* = h^*(x)$, in the sense of minimising mean square error at the point x, is asymptotic to

$$h_0 = h_0(x) = A\left\{f(x)/f^{(r)}(x)^2\right\}^{1/(2r+1)} n^{-1/(2r+1)}, \tag{4.1}$$

where the constant $A = \{(\int K^2)/(2r\kappa^2)\}^{1/(2r+1)}$ depends only on K, not on the unknown density. Here, asymptotic optimality means that both h_0/h^* and the ratio of the respective mean squared errors converge to 1. See for example Rossenblatt (1971). It may be shown that a closely related formula remains valid for densities of the type (1.1), which contain erratic episodes described by high frequencies ω_j. For example, if f is given by (1.1) and γ_j has $r+1$ bounded derivatives then the asymptotically optimal bandwidth for estimating $f(x_j)$ is still $h_0(x_j)$, provided that $\omega_j h^*(x_j) \to 0$, or equivalently that $\omega_j/n \to 0$.

Let $\hat{h}(x)$ denote an empirical approximation to $h_0(x)$ (and so to $h^*(x)$), and suppose that the approximation is sufficiently accurate to ensure that for a sequence of constants $b_n > 0$, and fixed constants $C_1, C_2 > 0$,

$$P\{b_n \le \hat{h}(x_j) \le C_1\} = 1, \tag{4.2}$$

and

$$P\{|\hat{h}(x)h_0(x)^{-1} - 1| > C_2/\log n\} = O\left(b_n^2 n^{-2r/(2r+1)}\right). \tag{4.3}$$

The next theorem shows that provided the ω_j's do not increase too rapidly, the locally adaptive kernel estimator $\hat{f}_{\hat{h}}$ achieves the optimal convergence rates discussed in Section 2 – without the $\log n$ factors that are needed for thresholded wavelet estimators. As in Theorem 2.1 we take x_0 to be any number not equal to one of $x_1, \cdots, x_N$, and ω_0 to be any fixed frequency.

Theorem 4.1. *Let f be given by (1.1), where the functions f_0 and γ_j satisfy the conditions asserted in subsection 2.3, and* $\max \omega_j = o\{n(\log n)^{-3(2r+1)/(2r)}\}$. *Assume that K satisfies the conditions stated above, and that (4.2) and (4.3) hold. Then for $0 \le j \le N$,*

$$E\left\{\hat{f}_{\hat{h}(x_j)}(x_j) - f(x_j)\right\}^2 = O\left\{(\omega_j n^{-1})^{2r/(2r+1)}\right\} \tag{4.4}$$

as $n \to \infty$.

Next we offer a concrete construction of $\hat{h}(x)$, satisfying the conditions of the theorem. First we define estimates of f and $f^{(r)}$, and then we plug them into formula (4.1) to obtain an empirical version of h_0.

Let $\bar{f}(x)$, our estimator of $f(x)$ for the purposes of bandwidth construction, denote a quantity with the property that for constants $0 < C_1 < C_2 < \infty$,

$$P\{C_1 \le \bar{f}(x) \le C_2\} = O(n^{-1})$$

for $x = x_0, \cdots, x_N$. A great many estimators – for example, kernel estimators – have this elementary property. For our purposes we may even take $\bar{f}$ to be a fixed constant. (We are primarily interested in addressing first-order properties of the adaptive estimator in tracking high-order frequency episodes of f, and for this purpose the estimators of f and $f^{(r)}$ need only be of the right order of magnitude with sufficiently high probability.) Let L denote a compactly supported function satisfying

$$\int u^j L(u)du = j!\delta_{j,r}, \quad 0 \le j \le r.$$

In effect, L is a kernel suitable for nonparametric estimation of the r'th derivative of a function. Put $h_1 = C_3 n^{-1/(2r+1)} \log n$, for arbitrary $C_3 > 0$, and

$$\check{f}^{(r)}(x) = (nh_1^{r+1})^{-1} \sum_{i=1}^{n} L\{(x - X_i)/h_1\},$$

our estimator of $f^{(r)}$. Let $\check{h}$ denote the version of h_0 that is obtained on replacing the pair $(f, f^{(r)})$ in (4.1) by $(\bar{f}, \check{f}^{(r)})$ and let $\hat{h} = \min[C_1, \max\{\check{h}, n^{-1}\}]$.

Our final theorem confirms that, provided the frequencies ω_j do not increase too rapidly, the convergence rates claimed at (4.4) hold with this particular version of $\hat{h}$.

Theorem 4.2. *Assume the conditions of Theorem 4.1, and in addition that $f_0^{(r)}$ and each $\gamma_j^{(r)}$ are Hölder continuous and that* $\max \omega_j = o\{n^{1/(2r+1)}(\log n)^{-2}\}$. *Then (4.4) holds for* $0 \le j \le N$.

5 Proofs

Proof of Theorem 2.1. Put $\xi = u(||f||_\infty n^{-1}\log n)^{\frac{1}{2}}$, with $0 < u < c$. Write $\hat{f}(x) = \hat{f}_1(x) + \Delta(x)$ where

$$\begin{aligned} \hat{f}_1(x) &= \sum_\ell \hat{b}_\ell \phi_\ell(x) + \sum_{k=0}^{q-1}\sum_\ell \hat{b}_{k\ell} I(|b_{k\ell}| > \xi)\psi_{k\ell}(x), \\ \Delta(x) &= \sum_{k=0}^{q-1}\sum_\ell \hat{b}_{k\ell}\Big\{I(|\hat{b}_{k\ell}| > \xi) - I(|b_{k\ell}| > \xi)\Big\}\psi_{k\ell}(x). \end{aligned}$$

We shall prove that $\hat{f}_1$ and Δ converge to f and 0, respectively, at the rate described in Theorem 2.1.

We preface the proof with three useful inequalities. Observe that by Taylor expansion, there exists a function x' of x such that

$$\begin{aligned} |b_{k\ell}| &= \frac{1}{r!}\left|p_k^{-1/2}\int \psi(x) f\{(x+\ell)/p_k\}\,dx\right| \\ &= \frac{1}{r!}\Bigg|p_k^{-(r+\frac{1}{2})}\int \psi(x)x^r f_0^{(r)}\{(x'+\ell)/p_k\}dx \\ &\qquad + p_k^{-(r+\frac{1}{2})}\sum_j \omega_j^r \int \psi(x)x^r\gamma_j^{(r)}\{\omega_j(\ell p_k^{-1} - x_j) + \omega_j x' p_k^{-1}\}\,dx\Bigg| \\ &\le \max\Big\{||f_0^{(r)}||_\infty, ||\gamma_1^{(r)}||_\infty, \ldots, ||\gamma_N^{(r)}||_\infty\Big\}(r!)^{-1}\int |x^r\psi(x)|\,dx \\ &\qquad \times p_k^{-(r+\frac{1}{2})}\left[1 + \sum_{j=1}^N \omega_j^r I\{|\ell p_k^{-1} - x_j| \le 2A(p_k^{-1} + \omega_j^{-1})\}\right]. \qquad (5.1) \end{aligned}$$

Here we used the fact that ψ has bounded support $[-A, A]$. Note that

$$E\{\psi_{k\ell}(X)^2\} = \int \psi(x)^2 f\{p_k^{-1}(x+\ell)\}\,dx \le ||f||_\infty. \qquad (5.2)$$

By (5.2) and Bernstein's or Bennett's inequality (for example, Pollard 1984, pp. 192-3), for each $y, \varepsilon > 0$ and for all sufficiently large n,

$$\max_{0\le k\le q-1;\ell} P\{|\hat{b}_{k\ell} - b_{k\ell}| > y(n^{-1}\log n)^{1/2}\}$$

$$
\begin{aligned}
&= \max_{0\le k\le q-1;\ell} P\left[\left|\sum_{i=1}^{n}\{\psi_{k\ell}(X_i) - E\psi_{k\ell}(X)\}\right| > y(n\log n)^{1/2}\right] \\
&\le 2\exp\{-\tfrac{1}{2}(1-\varepsilon)||f||_\infty^{-1} y^2 \log n\} \le 2n^{-(1-\varepsilon)y^2/(2||f||_\infty)}. \qquad (5.3)
\end{aligned}
$$

(Here we have used the fact that $p2^q n^{-1}\log n \to 0$.)

Next, we describe the convergence rate of $E\hat{f}_1(x_j)$ to $f(x_j)$. If ψ vanishes outside $[-A, A]$, then $\psi_{k\ell}(x)$ vanishes unless $|x - \ell p_k^{-1}| \le Ap_k^{-1}$, and there are at most $2A+1$ values of ℓ with this property for any given x. Furthermore, by (5.1) there exist constants $C_1, C_2 > 0$ such that if $\psi_{k\ell}(x_j) \ne 0$ and $p_k \ge C_1$ then $|b_{k\ell}| \le C_2 p_k^{-(r+\frac{1}{2})}\omega_j^r$, while if $\psi_{k\ell}(x_j) \ne 0$ and $p_k < C_1$,

$$
\begin{aligned}
|b_{k\ell}| &\le p_k^{-1/2}||f||_\infty \int|\psi| \le ||f||_\infty\left(\int|\psi|\right)p_k^{-1/2}(C_1/p_k)^r\omega_j^r/\omega^{*r} \\
&\le C_2 p_k^{-(r+\frac{1}{2})}\omega_j^r.
\end{aligned}
$$

Also, $|\psi_{k\ell}(x_j)| \le ||\psi_{k\ell}||_\infty \le p_k^{1/2}||\psi||_\infty$, and $|b_{k\ell}|I(|b_{k\ell}| \le \xi) \le \min(|b_{k\ell}|, \xi)$. Therefore,

$$
\begin{aligned}
|E\hat{f}_1(x_j) - f(x_j)| &= \left|\sum_{k=0}^{q-1}\sum_{\ell} b_{k\ell}I(|b_{k\ell}| \le \xi)\psi_{k\ell}(x_j) + \sum_{k=q}^{\infty}\sum_{\ell} b_{k\ell}\psi_{k\ell}(x_j)\right| \\
&\le (2A+1)||\psi||_\infty \sum_{k=0}^{q-1} p_k^{1/2}\min(C_2 p_k^{-(r+\frac{1}{2})}\omega_j^r, \xi) \\
&\qquad + (2A+1)C_2||\psi||_\infty\omega_j^r \sum_{k=q}^{\infty} p_k^{-r}.
\end{aligned}
$$

Now,

$$
\begin{aligned}
\sum_{k=0}^{q-1} p_k^{1/2}\min\left(C_2 p_k^{-(r+\frac{1}{2})}\omega_j^r, \xi\right) &\le \min(C_2, \xi)\sum_{k=0}^{q-1} p_k^{1/2}\min(p_k^{-(r+\frac{1}{2})}\omega_j^r, 1) \\
&= O\{(\omega_j n^{-1}\log n)^{r/(2r+1)}\},
\end{aligned}
$$

and

$$
\sum_{k=q}^{\infty} p_k^{-r} = O(p^{-r}2^{-qr}) = O\{\omega_j^{-r}(\omega_j n^{-1}\log n)^{r/(2r+1)}\},
$$

provided that $p2^q > n\eta_1/\log n$ where $\eta_1^{-1}(n^{-1+\varepsilon}\log n)^{2r/(2r+1)}$ is bounded. Therefore,

$$
|E\hat{f}_1(x_j) - f(x_j)| = O\{(\omega_j n^{-1}\log n)^{r/(2r+1)}\}. \qquad (5.4)
$$

In the next step of the proof, we examine the variance of $\hat{f}_1(x_j)$:

$$\begin{aligned}
\operatorname{Var}\{\hat{f}_1(x_j)\} &= n^{-1}\operatorname{var}\left\{\sum_\ell \phi_\ell(X)\phi_\ell(x_j) + \sum_{k=0}^{q-1}\sum_\ell \psi_{k\ell}(X)\psi_{k\ell}(x_j)I(|b_{k\ell}| > \xi)\right\} \\
&\leq n^{-1}E\left\{\sum_\ell \phi_\ell(X)\phi_\ell(x_j) + \sum_{k=0}^{q-1}\sum_\ell \psi_{k\ell}(X)\psi_{k\ell}(x_j)I(|b_{k\ell}| > \xi)\right\}^2 \\
&\leq n^{-1}||f||_\infty\left\{\sum_\ell |\phi_\ell(x_j)| + \sum_{k=0}^{q-1}\sum_\ell |\psi_{k\ell}(x_j)|I(|b_{k\ell}| > \xi)\right\}^2 ,
\end{aligned}$$

the last inequality following from the Cauchy-Schwartz inequality and the fact that $E\{\phi_{k\ell}(X)^2\}$, $E\{\psi_{k\ell}(X)^2\} \leq ||f||_\infty$. Arguments similar to those used to derive (5.4) may now be employed to prove that

$$\operatorname{Var}\{\hat{f}_1(x_j)\} = O\{n^{-1}p + (\omega_j n^{-1})^{2r/(2r+1)}(\log n)^{-1/(2r+1)}\}. \tag{5.5}$$

Finally, we show that Δ converges to zero at the desired rate. Since

$$\begin{aligned}
\left|I(|\hat{b}_{k\ell}| > \delta) - I(|b_{k\ell}| > \xi)\right| &\leq I(|\hat{b}_{k\ell} - b_{k\ell}| > \delta - \xi) + I(|\hat{b}_{k\ell}| \leq \delta, |b_{k\ell}| > \xi), \\
\left|\hat{b}_{k\ell}\psi_{k\ell}(x)\right| &\leq p_k(||\psi||_\infty)^2 I(|x - \ell p_k^{-1}| \leq Ap_k^{-1})
\end{aligned}$$

then, with I_{jk} denoting the class of integers ℓ such that $|x_j - \ell p_k^{-1}| \leq Ap_k^{-1}$, we have

$$\begin{aligned}
|\Delta(x_j)| \leq\ & (||\psi||_\infty)^2 \sum_{k=0}^{q-1} p_k \sum_{\ell \in I_{jk}} I(|\hat{b}_{k\ell} - b_{k\ell}| > \delta - \xi) \\
&+ \delta \sum_{k=0}^{q-1}\sum_\ell |\psi_{k\ell}(x_j)|I(|b_{k\ell}| > \xi).
\end{aligned} \tag{5.6}$$

Since I_{jk} contains at most $2A + 1$ elements then by (5.3), for all $\varepsilon > 0$,

$$\begin{aligned}
& E\left\{\sum_{k=0}^{q-1} p_k \sum_{\ell \in I_{jk}} I(|\hat{b}_{k\ell} - b_{k\ell}| > \delta - \xi)\right\}^2 \\
&\quad \leq \sum_{k_1=0}^{q-1}\sum_{k_2=0}^{q-1} p_{k_1}p_{k_2} \sum_{\ell_1 \in I_{jk_1}} \sum_{\ell_2 \in I_{jk_2}} \left\{\prod_{i=1}^{2} P(|\hat{b}_{k_i\ell} - b_{k_i\ell}| > \delta - \xi)\right\}^{1/2} \\
&\quad = O\left\{p^2 2^{2q} q \max_{0 \leq k \leq q; \ell} P(|\hat{b}_{k\ell} - b_{k\ell}| > \delta - \xi)\right\} \\
&\quad = O(p^2 2^{2q} q n^{-(1-\varepsilon)(c-u)^2/2}) \\
&\quad = O(n^{-1+\varepsilon'}),
\end{aligned}$$

where ε' may rendered arbitrarily small by choosing u and ε sufficiently small. The arguments leading to (5.4) may be employed to prove that

$$\delta \sum_{k=0}^{q-1} \sum_{\ell} |\psi_{k\ell}(x_j)| I(|b_{k\ell}| > \xi) = O\{(n^{-1}\log n)^{\frac{1}{2}}(\omega_j n^{-1}\log n)^{r/(2r+1)}\}.$$

Combining the estimates from (5.6) down we deduce that

$$E\{\Delta(x_j)^2\} = o\{n^{-1}p + (\omega_j n^{-1}\log n)^{2r/(2r+1)}\}. \tag{5.7}$$

The theorem follows from (5.4),(5.5) and (5.7). □

Proof of Theorem 4.1. We first establish a lemma. Let $B > 0$ denote an arbitrary constant, let x be any real number, and let $\mathcal{F}$ be any class of densities f such that $||f||_\infty \le B$ uniformly in all $f \in \mathcal{F}$, and

$$\lim_{\varepsilon \to 0} \sup_{f \in \mathcal{F}, |x-y| < \epsilon} |f(x) - f(y)| = 0.$$

Define $\mu(h) = E\hat{f}_h(x)$.

Lemma 5.1. *Let K satisfy the conditions imposed in Theorem 4.1, and let α_n, β_n denote positive numbers converging to zero and such that $n\beta_n \to \infty$. Then*

$$\begin{aligned}\sup_{f\in\mathcal{F}} \left| E \sup_{|h\beta_n^{-1}-1| \le \alpha_n} \{\hat{f}_h(x) - \mu(h)\}^2 - (n\beta_n)^{-1} f(x) \int K^2 \right| \\ = o\left\{(n\beta_n)^{-1} + (\log n)^3 (n\beta_n)^{-2} + \alpha_n |\log \beta_n| (n\beta_n)^{-1}\right\},\end{aligned}$$

as $n \to \infty$.

Proof of Lemma 5.1. Write

$$\hat{f}_h(x) - \mu(h) = -h^{-1} \int \{\hat{F}_n(x + hz) - F(x + hz)\} dK(z),$$

where $\hat{F}$ is the empirical distribution function. By the "Hungarian Embedding" of Komlós, Major and Tusnády (1975), there exist a Brownian bridge B_n and positive universal constants c_1, c_2 and λ such that, with $D_n = ||n\{\hat{F}_n - F\} - n^{1/2} B_n(F)||_\infty$,

$$P(D_n > c_1 \log n + t) \le c_2 \exp(-\lambda t).$$

It follows that for any $k > 0$, $ED_n^k \le d_k \log^k n$, for universal constants d_k. Thus, by changing probability spaces if necessary,

$$\hat{f}_h(x) - \mu(h) = -(n^{1/2}h)^{-1} \int B_n\{F(x + hz)\} dK(z) + o_2\{(n\beta_n)^{-1} \log n\} \tag{5.8}$$

uniformly in h such that $|h\beta_n^{-1} - 1| \leq \alpha_n$, where $o_2(\eta_n)$ denotes a random variable whose second moment is of order $o(\eta_n^2)$. Put

$$S_n(h) = -(n^{1/2}h)^{-1} \int B_n\{F(x + hz)\} dK(z).$$

By the covariance structure of the Brownian bridge,

$$ES_n^2(\beta_n) = (n\beta_n^2)^{-1} \int\int [\min\{F(x+\beta_n z_1), F(x+\beta_n z_2)\} - F(x+\beta_n z_1)F(x+\beta_n z_2)\}]\, dK(z_1)dK(z_2).$$

By Taylor expansion,

$$ES_n^2(\beta_n) = (n\beta_n)^{-1} f(x) \int K^2(u)du + o\{(n\beta_n)^{-1}\}. \tag{5.9}$$

Write

$$\begin{aligned} S_n(h) - S_n(\beta_n) &= -(n^{1/2}h)^{-1} \int \Big[B_n\{F(x+hz)\} - B_n\{F(x+\beta_n z)\}\Big] dK(z) \\ &\quad + \Big\{(n^{1/2}\beta_n)^{-1} - (n^{1/2}h)^{-1}\Big\} \int B_n\{F(x+\beta_n z)\} dK(z) \\ &\equiv I_1(h) + I_2(h). \end{aligned} \tag{5.10}$$

Let $\mathcal{H} = \{h : |h/\beta_n - 1| \leq \alpha_n\}$. By Theorems 1 and 2 of Garsia (1970), we deduce that for Brownian Bridge B_n,

$$\sup_{|s-t|\leq u} |B_n(s) - B_n(t)| \leq D_n u^{1/2} + 32u^{1/2}\{\log(1/u)\}^\delta,$$

for $0 < u < 1$, where $\delta > 1/2$ and D_n is a random variable such that $ED_n^2 \leq 384 \log 2$. Using this modulus of continuity and Taylor expansion we may show that

$$\sup_{h\in\mathcal{H}} |I_1(h)| \leq c_3(n^{1/2}\beta_n)^{-1} \left[D_n(\alpha_n\beta_n)^{1/2} + (\alpha_n\beta_n)^{1/2}\{\log(\alpha_n\beta_n)^{-1}\}^\delta\right],$$

for some finite constant c_3, depending only on $||f||_\infty$ and K. Thus,

$$E\{\sup_{h\in\mathcal{H}} I_1^2(h)\} \leq c_4\alpha_n(\log\beta_n^{-1})^\delta (n\beta_n)^{-1}, \tag{5.11}$$

where c_4 depends only on $||f||_\infty$ and K.

Now, $I_2(h)$ can easily be bounded from (5.9):

$$E\sup_{h\in\mathcal{H}} I_2^2(h) \leq \alpha_n^2 ES_n^2(\beta_n) = O\{\alpha_n^2(n\beta_n)^{-1}\}. \tag{5.12}$$

Combining (5.8) – (5.12), we obtain Lemma 5.1. □

We now return to prove Theorem 4.1. Let $\mathcal{H} = \{h : |hh_0(x_j)^{-1} - 1| \leq C_2/\log n\}$ and $\mu(h) = E\hat{f}_h(x_j)$. Then, by (4.2) and (4.3),

$$E\hat{h}^{2r}(x_j) \leq \{1 + C_2(\log n)^{-1}\}^{2r} h_0^{2r}(x_j) + C_1^{2r} b_n^2 n^{-2r/(2r+1)} = O\{(\omega_j n^{-1})^{2r/(2r+1)}\}.$$

Consequently, by Taylor expansion,

$$E\Big[\mu\{\hat{h}(x_j)\} - f(x_j)\Big]^2 = O\{\omega_j^{2r} E\hat{h}^{2r}(x_j)\} = O\{(\omega_j n^{-1})^{2r/(2r+1)}\}.$$

This together with Lemma 5.1 and (4.3) gives,

$$\begin{aligned} E\Big\{\hat{f}_{\hat{h}(x_j)}(x_j) - f(x_j)\Big\}^2 &\leq 2E\Big[\mu\{\hat{h}\}(x_j) - f(x_j)\Big]^2 \\ &\quad + 2E\Big[\hat{f}_{\hat{h}(x_j)}(x_j) - \mu\{\hat{h}(x_j)\}\Big]^2 \Big[I\{\hat{h}(x_j) \in \mathcal{H}\} + I\{\hat{h}(x_j) \notin \mathcal{H}\}\Big] \\ &= O\{(\omega_j n^{-1})^{2r/(2r+1)}\}. \end{aligned}$$

□

Proof of Theorem 4.2. We need only prove that $\hat{h}$ satisfies (4.3). Since $h_1\omega_j \to 0$, then by Taylor expansion,

$$|E\check{f}^{(r)}(x_j) - f^{(r)}(x_j)| = O(\omega_j^{r+1} h_1).$$

Since $f^{(r)}(x_j) > C_3\omega_j^r$ for a positive constant C_3, independent of x_j, then

$$|E\check{f}^{(r)}(x_j) f^{(r)}(x_j)^{-1} - 1| = O(\omega_j h_1) = o(1/\log n). \tag{5.13}$$

It can be shown that

$$\sup_x \operatorname{var}\Big\{n^{-1} {h_1}^{-2r-1} L\{(X_i - x)/h_1\}\Big\} \leq \|f\|_\infty (\int L^2)(n^2 {h_1}^{2r+1})^{-1}.$$

Using this and Bennett's inequality (e.g. Pollard 1984, pp. 192-3), for any $\varepsilon > 0$,

$$\sup_x P\{|\check{f}^{(r)}(x) - E\check{f}^{(r)}(x)| > \xi_n\} \leq 2n^{-(1-\varepsilon)u^2/2}, \tag{5.14}$$

where

$$\xi_n = u\Big\{\|f\|_\infty (\int L^2) n^{-1} {h_1}^{-2r-1} \log n\Big\}^{1/2} = o(\log^{-1} n). \tag{5.15}$$

We are now in a position to prove that $\hat{h}$ satisfies (4.3) with $b_n = n^{-1}$. Note that by the definition of $\hat{h}$,

$$\begin{aligned} & P\left\{\left|\hat{h}(x_j)h(x_j)^{-1} - 1\right| > C_2 \log^{-1} n\right\} \\ \leq \;& P\{\check{h}(x_j) < n^{-1}\} + P\{\check{h}(x_j) > C_1\} + P\left\{\left|\check{h}(x_j)h(x_j)^{-1} - 1\right| > C_2 \log^{-1} n\right\}. \end{aligned} \quad (5.16)$$

The first term on the right-hard side is bounded by

$$P\{|\check{f}^{(r)}(x_j)| > C_4 n^r\} \leq P\{|\check{f}^{(r)}(x_j) - E\check{f}^{(r)}(x_j)| > \xi_n\} \leq 2n^{-(1-\varepsilon)u^2/2},$$

since

$$|E\check{f}^{(r)}(x_j)| = O(\omega_j^{r+1} h_1) = O(n^{1/2}).$$

The second term on the right-hand side of (5.16) is bounded by

$$P\Big\{|\check{f}^{(r)}(x_j)| < C_5 n^{-1/2}\Big\} \leq P\Big\{|\check{f}^{(r)}(x_j) - E\check{f}^{(r)}(x_j)| \geq |E\check{f}^{(r)}(x_j)| - C_5 n^{-1/2}\Big\} \leq 2n^{-(1-\varepsilon)u^2/2},$$

since

$$|E\check{f}^{(r)}(x_j)| \geq C_6 \geq \xi_n + C_5 n^{-1/2}. \quad (5.17)$$

The last term in (5.16) can be bounded by

$$\begin{aligned} & P\left\{\left|\left|f^{(r)}(x_j)\check{f}^{(r)}(x_j)^{-1}\right|^{-2/(2r+1)} - 1\right| > C_2 \log^{-1} n\right\} \\ \leq \;& P\{|\check{f}^{(r)}(x_j) - E\check{f}^{(r)}(x_j)| > \xi_n f^{(r)}(x_j)\} + P(E_n) \\ \leq \;& 2n^{-(1-\varepsilon)u^2/2} + P(E_n), \end{aligned}$$

where C_6 is given by (5.17) and

$$E_n = \left\{\left|\check{f}^{(r)}(x_j) - E\check{f}^{(r)}(x_j)\right| \leq \xi_n f_0^{(r)}(x_j), \quad \left|\left|f^{(r)}(x_j)\check{f}^{(r)}(x_j)^{-1}\right|^{-2/(2r+1)} - 1\right| > C_2 \log^{-1} n\right\}.$$

When n is large, the set E_n is empty since by (5.13) and (5.15), on E_n

$$|\check{f}^{(r)}(x_j)f^{(r)}(x_j)^{-1} - 1| = o(\log^{-1} n).$$

Therefore, we conclude that

$$P\left\{\left|\hat{h}(x_j)h(x_j)^{-1} - 1\right| > C_2 \log^{-1} n\right\} = O(n^{-3}),$$

by taking $u = \sqrt{6/(1-\varepsilon)}\max(1, 1/C_6)$. This completes the proof. □

Acknowledgements. The work of Fan was partially supported by an Australian Research Council grant, NSF grant DMS-9203135 and an NSF postdoctoral fellowship; that of Martin was partially supported by NSF grant INT-8913333 under the NSF Cooperative Program; and the work of both Fan and Martin was conducted during visits to the Centre for Mathematics and its Applications at the ANU.

REFERENCES

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Philadelphia: SIAM

Donoho, D.L. (1992). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Technical report 403*, Department of Statistics, Stanford University.

Donoho, D.L. and Johnstone, I.M. (1992a). Ideal spatial adaptation by wavelet shrinkage. *Technical report 400*, Department of Statistics, Stanford University.

Donoho, D.L. and Johnstone, I.M. (1992b). Minimax estimation via wavelet shrinkage. *Technical report 402*, Department of Statistics, Stanford University.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1993). Density estimation by wavelet thresholding. *Manuscript.*

Fan, J. and Gijbels, I. (1993). Data-driven bandwidth selection in local polynomial regression: variable bandwidth selection and spatial adaptation. *Manuscript*

Fan, J., Hall, P., Martin, M. and Patil, P. (1993). On the local smoothing of nonparametric curve estimators. *Research report SMS- -93*, Australian National University.

Garsia, A. M. (1970). Continuity properties of Gaussian processes with multidimensional time parameter. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **2**, 369-374. University of California Press.

Hall, P. and Patil, P. (1993a). On wavelet methods for estimating smooth functions. *Research report SMS-37-93*, Australian National University.

Hall, P. and Patil, P. (1993b). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Research report SMS-49-93*, Australian National University.

Kerkyacharian, G. and Picard, D. (1993a). Density estimation by kernel and wavelet methods, optimality in Besov Spaces. *Manuscript.*

Kerkyacharian, G. and Picard, D. (1993b). Linear wavelet methods and other periodic kernel methods. *Manuscript.*

Komlós, J., P. Major and Tusnády, G. (1975). An approximation of partial sums of independent RV's, and the sample DF. I. *Z. Wahrscheinlichkeitsheorie verw. Gebiete* **32**, 111-131.

Kooperberg, C. and Stone, C. J. (1990). A study of logspline density estimation. *manuscript.*

Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet regresentation. *IEEE Trans. Pattern Anal. and Machine Intell.* **11** 674-693.

Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.*, **20**, 712 - 736.

Meyer, Y. (1990). *Ondelettes.* Hermann, Paris.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Rossenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.* **42**, 1815-1842.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348-1360.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040-1053.

Strang, G. (1989). Wavelets and dilation equations: a brief introduction. *SIAM Review* **31** 614-627.

Strang, G. (1993). Wavelet transforms versus fourier transforms. *Bulletin Amer. Math. Soc.* **28**, 288–305.