Statistica Sinica **9**(1999), 85-102

ADAPTATION TO HIGH SPATIAL INHOMOGENEITY USING WAVELET METHODS

Jianqing Fan[*⋆], Peter Hall[⋆†], Michael Martin[⋆] and Prakash Patil[⋆‡]

[*] *University of North Carolina,* [⋆] *Australian National University,* [†] *CSIRO and* [‡] *University of Birmingham*

Abstract: Many of the signals to which wavelet methods are applied, including those encountered in simulation experiments, are essentially smooth but contain a small number of high-frequency episodes such as spikes. In principle it is possible to employ a different amount of smoothing at different spatial locations, but in the context of wavelets this is so awkward to implement that it is not really practicable. Instead, it is attractive to select the primary resolution level (or smoothing parameter) so as to give good performance for smooth parts of the signal. While this is readily accomplished using a cross-validation argument, it is unclear whether it has a deleterious impact on performance at high-frequency episodes. In this paper we show that it does not. We derive upper and lower bounds to pointwise rates of convergence for functions whose "spikiness" increases with sample size. (This allows us to model contexts where wavelet methods have to work hard to recover high-frequency events.) We show that, in order to achieve optimal rates of convergence, it is necessary for the primary resolution level of the empirical wavelet transform to vary with location, sometimes extensively. Nevertheless, the convergence rate penalty incurred through using a non-varying resolution level, chosen to provide good performance for coarse-scale features, equals a factor that is less than the logarithm of sample size.

Key words and phrases: Convergence rate, fine-scale, local adaptivity, resolution, wavelet.

1. Introduction

1.1. Motivation

Wavelet-based curve estimators have particularly attractive adaptivity properties, usually expressed through performance in global metrics uniformly over large function classes. Donoho and Johnstone (1994, 1995, 1996) and Donoho, Johnstone, Kerkyacharian and Picard (1993, 1995) pioneered that type of analysis, demonstrating the extraordinary ability of wavelet transforms to approximate — even in the presence of noise — large sets of functions whose individual complexity defies simple description.

Nevertheless, we contend that some of the attractive properties of wavelet methods are unclear from this approach to the problem. In the present paper

the perspective on the performance of wavelet methods is unconventional from at least two viewpoints. Instead of looking at function classes we consider a single target function, f, whose complexity increases with sample size, n. Also, we examine pointwise convergence rates rather than convergence rates in global metrics.

Let us first motivate the context where f varies with n. By allowing the difficulty of the problem to increase with sample size we challenge wavelet methods in a relatively stringent way. Fixed targets, such as those employed for kernel methods applied in Sobolev spaces, are arguably "too easy" for wavelet methods. An analogous idea appears in the hypothesis testing literature, where one challenges a test procedure by considering its discrimination power at contiguous alternatives that depend on the sample size n. This has led to a considerable literature on power under local alternatives.

Moreover, results on uniform convergence over function classes are effectively studying wavelet methods applied to functions whose complexity diverges with n, since the "worst case" functions, at which minimax bounds are achieved, become rapidly more complicated as n diverges. However, identifying these especially pathological targets, and analysing their properties, is an unrewarding task because they are too abstruse. Instead, we suggest examining the sorts of functions to which wavelet methods might be applied in practice, and also the test functions that are used to assess the performance of wavelet methods in numerical studies, with the aim of allowing them to depend on n so as to make good asymptotic performance a real challenge for wavelets.

Actual targets, as well as numerical models for them, are typically smooth functions with a number of sharp aberrations superimposed. See for example the Blocks, Bumps, HeaviSine and Doppler functions that are used in numerical work of Donoho and Johnstone (1994, Figure 1) and Donoho, Johnstone, Kerkyacharian and Picard (1995, Figure 2). The Bumps function in particular may be written in the form

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

where in the "Bumps" formulation, $f_0 \equiv 0$, $N = 11$, the γ_j's are positive multiples (all between 3.1 and 5.1) of a single, smooth, bell-shaped function, the x_j's are points between 0.10 and 0.81, and the ω_j's represent large positive frequencies, all lying between 33 and 200. As a result of the latter specification, the Bumps function consists of N peaks, or in effect, N sharp spikes. Broadly similar characteristics are also evident in the HeaviSine and Doppler functions, where there are sharply defined peaks and troughs. We shall use the model (1.1) for our targets, keeping the functions γ_j fixed and allowing the ω_j's to diverge

with sample size. In this manner, we effectively model different degrees of "spikiness" of f around locations x_j through different orders of magnitude of $\{\omega_j\}$, and to a lesser extent through the smoothness of functions γ_j. Our challenge to wavelet techniques is to see if they can cope well with different degrees of spikiness, without knowing the sizes of the ω_j's.

Next we motivate our focus on pointwise rates of convergence. In order to give an adequate description of the performance of wavelet estimators for "spike" functions we must determine the estimators' ability to track those functions over a range of arguments, x. Therefore, we wish to determine the accuracy with which wavelet methods estimate f at points x_j where spikes occur, as well as at other points x where f is relatively smooth. It is of particular interest to know how well an estimator that is "tuned" to the background function f_0, for example through selection of the smoothing parameter, performs in estimating the spikes. We cannot really solve this problem by studying performance in a global metric, such as an integral L^p metric, or (for densities) Hellinger distance, since then the error of the estimator at any one of the points x_j is confounded with the accumulation of error at all those points x that are well away from the spikes. So, we contend, an effective examination of the ability of wavelet methods to recover functions such as (1.1) requires pointwise analysis, at the points x_j and at other places.

We refer to (1.1) as a model for "fine-scale phenomena", since the spikes at x_j become increasingly fine as sample size increases. All the quantities in (1.1) are assumed unknown. In particular, N, x_j and the functions γ_j are fixed and unknown, and no attempt is made to fit the model (1.1).

1.2. Practical estimation of a "Bumps" type function

We begin by noting that if a wavelet estimator is constructed so as to effect a trade-off between bias and variance, then the primary resolution level, *not the threshold*, is the principal smoothing parameter (Hall and Patil (1995)). It is essentially the inverse of bandwidth, in the sense that the optimal primary resolution level when using r'th order wavelet methods, for functions with r bounded derivatives, is of size $n^{1/(2r+1)}$, whereas the optimal bandwidth for a kernel estimator in the same setting would be of size $n^{-1/(2r+1)}$. Alternative approaches, which have received considerable attention and take the primary resolution level to be 1, result in oversmoothing by only a logarithmic order of magnitude, with the result that bias dominates error about the mean. See Hall and Patil (1996). (Choice of r is generally determined by a relatively small number of options available to the experimenter, and is typically 4 or 6.)

The "optimal" primary resolution level (in the sense of mean squared error), for smooth parts of a function f, is amenable to estimation using standard statistical methods, such as cross-validation or plug-in. (Smooth parts of f are

easily identified by preliminary analysis, for example using a pilot estimator for which the primary resolution level is 1.) Theoretical verification of the asymptotic validity of cross-validation in this setting follows standard arguments. We show that a primary resolution level that is optimal for smooth parts of f is also optimal for estimating a sharp spike, except for a factor that is only logarithmically large. Hence, in using throughout a primary resolution level that is chosen for smooth parts of f, we incur only a small penalty at the really rough parts. In Section 2.6 we show that this remains true even if the primary resolution level is chosen empirically from smooth parts of the curve as suggested above.

1.3. Main results

In technical detail, with reference to the model at (1.1), the results above may be described as follows. If the functions f_0 and γ_j are r-times differentiable then an r'th order wavelet estimator of $f(x_j)$ may be constructed so that it enjoys a mean square convergence rate of $\rho_j \equiv (\omega_j/n)^{2r/(2r+1)}$ at the specific point x_j. This rate is only achieved when a primary resolution level, p, of size $(n\omega_j^{2r})^{1/(2r+1)}$ is employed to construct the estimator in the vicinity of x_j, and then the rate is optimal in a minimax sense. Now, the optimal primary resolution level in a smooth part of the curve is $p = n^{1/(2r+1)}$. We shall show if we use this level at x_j then the mean square convergence rate is

$$\min\left\{\left(\omega_j^{2r+1}\Big/ n\right)^{2r/(2r+1)}, \left(\omega_j n^{-1}\log n\right)^{2r/(2r+1)}\right\} \le \rho_j(\log n)^{2r/(2r+1)}, \text{ for all } j, \tag{1.2}$$

which differs from the optimal rate only by a factor smaller than $\log n$.

In addition to showing that the quantity at (1.2) is an upper bound to the convergence rate of a wavelet estimator $\hat{f}$ (see Theorem 2.1), we prove that in many circumstances it is also a lower bound (Theorem 2.2). This demonstrates that the results discussed above describe actual properties of $\hat{f}$, not just artifacts of the method of proof. The results apply to estimators of both density and regression functions. A detailed account in the setting of regression is beyond the scope of this short note, however, because treatment of general error distributions demands relatively sophisticated development of large-deviation bounds. Instead we give a rigorous account of the bounds in the context of nonparametric density estimation, and state an analogue of the upper bound (Theorem 2.3) for regression under simplified conditions on the error distribution — specifically, that it be either Normal or essentially bounded. The analogue of the lower bound may be derived similarly.

1.4. Background

In connection with our results for density estimation we mention related work of Kerkyacharian and Picard (1993), on the performance of wavelet density estimators in global metrics uniformly over function classes. Research on

wavelet methods which predates that of Donoho, Johnstone, Kerkyacharian and Picard, although not addressing the issue of thresholding, includes the contribution of Doukhan (1988). Convergence rates in a minimax setting, for general nonparametric estimators, have been discussed by, for example, Stone (1980, 1982), Brown and Low (1991) and Donoho and Johnstone (1998). The results of Low and co-authors also provide explicit lower bounds to the constant multipliers of rates. Fan and Gijbels (1995) have developed locally adaptive methods for selecting the smoothing parameter of kernel estimators, and have compared the performance of their techniques with that of wavelets.

2. 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 subsections 2.3–2.5.

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 the "scale function" or "father wavelet" ϕ has the properties

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 for these properties 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 and 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 bounded and compactly supported, with $r - 1$ vanishing moments for some $r \geq 1$:

$$\int x^j \phi(x) dx = \delta_{0j} \qquad \text{for} \quad j = 0, \ldots, r-1\,. \tag{2.3}$$

See Daubechies (1992) for constructions of this family. We note, however, that (2.3) is not essential to our results — the more important moment condition is (2.4) below. Under such assumptions there exists a function ψ (the "mother" wavelet) given by

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

and which has the properties

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:

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

In practice, ϕ and ψ are typically compactly supported, and 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 primary 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 be applied easily to statistical function estimation. In the case of density estimation, let $X_1, \ldots, X_n$ be a random sample from a distribution with density f. Formulae (2.6) suggest unbiased estimates of the wavelet coefficients:

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

For high resolution (i.e. large p_k), the estimate $\hat{b}_{k\ell}$ will be basically noise, since $\psi_{k\ell}$ is supported only in a small neighbourhood 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 (1994), we select useful estimated coefficients $\hat{b}_{k\ell}$ by "thresholding". Considerations of this nature suggest the estimator

$$\hat{f}(x) = \sum_{\ell} \hat{b}_{\ell}\phi_{\ell}(x) + \sum_{k=0}^{q-1} \sum_{\ell} \hat{b}_{k\ell} I(|\hat{b}_{k\ell}| \geq \delta)\, \psi_{k\ell}(x); \tag{2.8}$$

compare (2.5). In (2.8), q denotes a truncation parameter which may be chosen within a reasonably wide range; see Theorem 2.1 for precise conditions. Asymptotic theory developed by Donoho, Johnstone, Kerkyacharian and Picard (1993, 1995), and extended by Hall and Patil (1995), suggests taking $\delta = c(n^{-1}\log n)^{1/2}$, where $c > 0$ is a constant. Following Donoho and Johnstone (1994), the above estimator corresponds to "hard thresholding". An alternative approach, "soft thresholding", involves replacing $I(|\hat{b}_{k\ell}| \geq \delta)$ in (2.8) by $\operatorname{sgn}(\hat{b}_{k\ell})(|\hat{b}_{k\ell}| - \delta)_+$, leading to the alternative estimator

$$\hat{f}(x) = \sum_{\ell} \hat{b}_{\ell}\phi_{\ell}(x) + \sum_{k=0}^{q-1} \sum_{\ell} \operatorname{sgn}(\hat{b}_{k\ell})(|\hat{b}_{k\ell}| - \delta)_+ \psi_{k\ell}(x). \tag{2.9}$$

The intuition behind either type of thresholding is based on a "signal-to-noise" ratio argument. 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. The amount of smoothing is adjusted through p, to first order.

2.3. Asymptotic theory for wavelet density estimators

We begin by addressing the case of densities of the form (1.1). It suffices to take $N = 1$. We write γ for γ_1 but retain the notation ω_1, since we shall shortly introduce another quantity ω_0. To ensure that f is a density for all sufficiently large choices of ω_1, assume f_0 is a fixed, r-times differentiable density bounded away from zero on an interval $\mathcal{I} = (-B, B)$; that $x_1 \in \mathcal{I}$; that the support of γ is contained within an interval $\mathcal{I} = (-B_1, B_1)$; that γ has r bounded derivatives on $\mathcal{I}_1$ with $\int \gamma = 0$; that $\inf_{(-B,B)} f_0 > -\inf_{I_1} \gamma$. Then there exists $B_0 > 0$, depending on x_1, such that for all $\omega_1 \geq 1$, f is bounded above zero by B_0 on $\mathcal{I}$ and is a proper density function.

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. Let $\hat{f}$ be given by either (2.8) or (2.9).

Theorem 2.1. *Take $\delta = c(\|f\|_\infty n^{-1}\log n)^{1/2}$, with $c \geq \sqrt{6}$. Let $0 \leq \epsilon < 1$, and let $\eta_1 \leq \eta_2$ be positive numbers converging to zero as $n \to \infty$ such that*

$\eta_1^{-1}(n^{-1+\epsilon}\log\ n)^{2r/(2r+1)}$ *is bounded. Let* ω_0 *and* C *be fixed positive numbers, and assume that* $\omega_1 = O(n^\epsilon)$. *Let* x_0 *be any real number not equal to* x_1. *Then for* $j = 0, 1$,

$$E\{\hat f(x_j) - f(x_j)\}^2 = O\Big[(p/n) + \min\Big\{(\omega_j/p)^{2r}, (\omega_j n^{-1}\log n)^{2r/(2r+1)}\Big\}\Big] \quad (2.10)$$

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

In Section 2.4 we demonstrate that the convergence rate described by (2.10) is generally the best possible. Assuming this for the present, the following remarks describe the main implications of (2.10).

Remark 2.1. *Optimal rate of convergence at* x_j. The right-hand side of (2.10) is minimized by taking p to be of size $(n\omega_j^{2r})^{1/(2r+1)}$, yielding

$$\inf_p E\Big\{\hat f(x_j) - f(x_j)\Big\}^2 = O\Big\{(\omega_j/n)^{2r/(2r+1)}\Big\} \quad (2.11)$$

for $j = 0, 1$. In particular, if ω_j is bounded (for example, if $j = 0$) then the optimal convergence rate is $O(n^{-2r/(2r+1)})$, and this rate is achieved by taking p to be of size $n^{1/(2r+1)}$. On the other hand, if ω_j diverges with n, the optimal p is an order of magnitude larger than $n^{1/(2r+1)}$.

Remark 2.2. *The extent to which local adaptivity accommodates different scales,* ω_j. At all points except x_1 it is (asymptotically) optimal to select p to be of size $n^{1/(2r+1)}$. Thus, if we were to employ the same p for all x's, as is generally mandated by considerations of computational efficiency, we would take p to be a constant multiple of $n^{1/(2r+1)}$. Result (2.10) implies that for this selection, and for $j = 0$ or 1,

$$E\Big\{\hat f(x_j) - f(x_j)\Big\}^2 = O\Big[n^{-2r/(2r+1)}\ \min\Big\{\omega_j^{2r}, (\omega_j\ \log\ n)^{2r/(2r+1)}\Big\}\Big]. \quad (2.12)$$

Further, using a fixed primary resolution level $p = C$, which is independent of the degree of smoothness r, we have

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

In other words, even if we use a primary resolution level and a threshold that are independent of the degree of smoothness r, the thresholded wavelet estimator pays at most a price of logarithmic order in terms of convergence rates.

Therefore, using this fixed p throughout, the convergence rate is never worse than $(\omega_j\, n^{-1}\ \log\ n)^{2r/(2r+1)}$, which differs from the optimal rate at (2.11) only by a logarithmic factor. Thus the inherent local adaptivity of $\hat f$ overcomes some of the problems arising from using a global rather than a local choice of the smoothing parameter p, even when fine-scale aberrations are present.

Remark 2.3. *Uniform convergence rates.* The uniform convergence rate x is never worse than the worst rate described by (2.10). Indeed, under the conditions of Theorem 2.1 it may be proved that

$$\sup_{-\infty<x<\infty} E\{\hat{f}(x) - f(x)\}^2 = O\Big[(p/n) + \min\Big\{(\omega_1/p)^{2r}, (\omega_1\, n^{-1}\, \log\, n)^{2r/(2r+1)}\Big\}\Big]\,. \tag{2.13}$$

The case of fixed f is of classical interest, and there we may strengthen (2.13) to hold uniformly over f as well as x, as follows. Given $B > 0$, 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 $\|f^{(r)}\|_\infty$ do not exceed B. Then, under the conditions and with the parameter configurations of Theorem 2.1,

$$\sup_{-\infty<x<\infty; f\in\mathcal{F}} E\{\hat{f}(x) - f(x)\}^2 = O\Big[(p/n) + \min\Big\{p^{-2r}, (n^{-1}\, \log\, n)^{2r/(2r+1)}\Big\}\Big]\,.$$

Remark 2.4. *Convergence rates in other metrics.* It is straightforward to generalize our results to rates in any L^t metric, for each $1 \le t < \infty$. In particular, if we write $\zeta_j = (p/n)^{1/2} + \min\{(\omega_j/p)^r, (\omega_j\, n^{-1}\, \log\, n)^{r/(2r+1)}\}$, then for $j = 0, 1$, Theorem 2.1 continues to hold if (2.10) is replaced by

$$E\Big|\hat{f}(x_j) - f(x_j)\Big|^t = O\Big(\zeta_j^t\Big)\,, \quad t \ge 1\,. \tag{2.14}$$

Only minor modifications to the proof are required. Indeed, our calculation of the bias contribution to the left-hand side remains unchanged. Calculation of the error-about-the-mean contribution uses Rosenthal's inequality and manipulations that are standard in the theory of sums of independent random variables.

Remark 2.5. *Adaptation to various degrees of smoothness.* In the classical formulation, the density f is assumed to be a fixed target function with a bounded rth derivative. This corresponds to our case with a bounded ω_1. Taking $p = C$ and using (2.10), the convergence rate is $O\{(n^{-1}\log\, n)^{r/(2r+1)}\}$. In other words, even without full knowledge of the degree of smoothness r, the wavelet estimator achieves the optimal rate of convergence within a logarithmic factor.

2.4. Lower bounds to convergence rates

We state a converse to Theorem 2.1, showing that the convergence rate described there is generally the best possible. For convenience, we make the following assumptions, noting that they can be weakened by using similar arguments:

$$p = 2^m, \text{ where } m \text{ is an integer}\,; \tag{2.15}$$

$$x_0 \text{ and } x_1 \text{ have finite dyadic expansions}\,; \tag{2.16}$$

$$\omega_1 = \omega_1(n) \to \infty \text{ as } n \to \infty\,; \tag{2.17}$$

$$f_0^{(r)} \text{ and } \gamma^{(r)} \text{ are continuous, and } f_0^{(r)}(x_1) \neq 0 \neq \gamma^{(r)}(0); \tag{2.18}$$

$$\int x^r \, \psi(x) \, dx \neq 0 \neq \sum_\ell \psi(\ell) \, . \tag{2.19}$$

These assumptions are mild, and might reasonably be expected to hold in practice: the primary resolution levels that are used in practice usually satisfy (2.15); the values of x_0 and x_1 that satisfy (2.16) are dense in the real line; in the context of (2.17), if ω_1 is bounded there is no loss of generality in taking it to be fixed, and treating x_1 as though it were x_0; condition (2.18) ensures that the contribution to bias from the r'th derivative of f is non-negligible; and the first part of (2.19) asks that, while the wavelet is of r'th order, it is not of $(r+1)$'th order.

Theorem 2.2. *Assume* (2.15)–(2.19) *as well as the conditions of Theorem* 2.1. *Then, for* $j = 0$ *or* 1,

$$(p/n) + \min\Big\{(\omega_j/p)^{2r}, (\omega_j \, n^{-1} \, \log \, n)^{2r/(2r+1)}\Big\} = O\Big[E\Big\{\hat f(x_j) - f(x_j)\Big\}^2\Big] \tag{2.20}$$

uniformly in values $p = 2^m \geq C$ *and* $\eta_1 \leq n^{-1} \, p 2^q \, \log \, n \leq \eta_2$.

Combining Theorems 2.1 and 2.2 we see that, under the conditions of Theorem 2.2, for $j = 0$ or 1, the ratio

$$\frac{E\{\hat f(x_j) - f(x_j)\}^2}{(p/n) + \min\{(\omega_j/p)^{2r}, (\omega_j \, n^{-1} \, \log \, n)^{2r/(2r+1)}\}}$$

is bounded away from zero and infinity.

Theorem 2.2 is readily extended to L^t metrics, for arbitrary $1 \leq t < \infty$. Indeed, if we define ζ_j as in Remark 2.4 then Theorem 2.2 continues to hold if (2.20) is changed to

$$\zeta_j^t = O\Big\{E\Big|\hat f(x_j) - f(x_j)\Big|^t\Big\}, \quad t \geq 1 \, ;$$

compare (2.14).

2.5. The case of nonparametric regression

The results described in Sections 2.3 and 2.4 have direct analogues in the context of nonparametric regression. We content ourselves with stating a regression version of Theorem 2.1, from which our conclusions follow as before.

Suppose data $Y_1, \ldots, Y_n$ are generated by the model $Y_i = f(i/n) + \epsilon_i$, in which the ϵ_i's are independent and identically distributed with zero mean and variance σ^2, and f is given by (1.1) with $N = 1$, f_0 has r bounded derivatives on $[0, 1]$, γ has compact support and r bounded derivatives, and $0 < x_1 < 1$. The

wavelet expansion of f, and the coefficients b_ℓ and $b_{k\ell}$, are given by (2.5) and (2.6). Estimators of b_ℓ and $b_{k\ell}$ are

$$\hat{b}_\ell = n^{-1} \sum_{i=1}^{n} \phi_\ell(i/n) Y_i \,, \qquad \hat{b}_{k\ell} = n^{-1} \sum_{i=1}^{n} \psi_{k\ell}(i/n) Y_i \tag{2.21}$$

(compare (2.7)), and wavelet estimators of f are defined by (2.8) and (2.9). We assume the same conditions on ϕ and ψ as in the paragraph preceding Theorem 2.1, except that we further ask that both functions be Hölder continuous. For simplicity suppose that the ϵ_i's are either Normally distributed or bounded. Let $x_0 \in (0, 1)$.

Theorem 2.3. *Take $\delta = c\sigma(n^{-1} \log n)^{1/2}$, where $c > 0$ is sufficiently large. Let $\epsilon, \eta_1, \eta_2, \omega_0, C$ and x_0 be as in Theorem* 2.1. *If $\max(\omega_0, \omega_1) = O(n^\epsilon)$ then* (2.10) *holds uniformly in values of p and q satisfying $p \ge C$ and $\eta_1 \le n^{-1} p 2^q \log n \le \eta_2$.*

2.6. Empirical choice of primary resolution level

In Section 1.2 we argued that the primary resolution level might be selected empirically by using a method such as cross-validation or plug-in adapted to smooth parts of the curve. Such an approach will generally produce a random version of p, $\hat{p}$ say, with the property that $\hat{p}/p_0 \to 1$ in probability, where p_0 denotes a deterministic threshold satisfying the conditions imposed on p in Theorems 2.1 and 2.3. It would usually be the case that $p_0/n^{1/(2r+1)}$ converges to a positive constant, although this is not essential for developing empirical versions of (2.10). See Remark 2.2 for discussion of deterministic resolution levels of size $n^{1/(2r+1)}$.

Let $\bar{f}$ denote the estimator $\hat{f} = \hat{f}_p$ in which the resolution level p is replaced by its empirical form $\hat{p}$. There are several approaches to proving that $\bar{f}$ is asymptotically as good as $\hat{f}_{p_0}$. One argument borrows ideas from Krieger and Pickands (1981) and, by first establishing an invariance principle for the stochastic process $\hat{f}_p$ indexed by $p \in [p_0(1-\delta), p_0(1+\delta)]$ for some $\delta > 0$, shows that the following result holds:

$$|\bar{f}(x_j) - f(x_j)| = O_p\biggl(\Bigl[(p_0/n) + \min\Bigl\{ (\omega_j/p_0)^{2r}, (\omega_j n^{-1} \log n)^{2r/(2r+1)} \Bigr\} \Bigr]^{1/2} \biggr) .$$

This is essentially (2.10) for $\bar{f}$ rather than $\hat{f}_{p_0}$, but without the additional strength conferred by the expectation at (2.10). It may be derived for both $j = 0$ and $j = 1$. However, in the most important case $j = 1$ we retain the full force of (2.10) while using the empirical resolution level $\hat{p}$, as we now show.

Assume that $\hat{p}$ satisfies

$$P(C_1 \le \hat{p}/p_0 \le C_2) = 1 - O(n^{-1}) \tag{2.22}$$

for constants $0 < C_1 < C_2 < \infty$. In the setting of density estimation, where $\hat{p}$ is chosen by cross-validation or a plug-in rule over a region $\mathcal{R} = (-\infty, x_1 - \eta) \cup (x_1 + \eta, \infty)$, with $\eta > 0$, (2.22) actually holds in the stronger form $P(|\hat{p}\, p_0^{-1} - 1| > \delta) = O(n^{-\lambda})$ for all $\delta, \lambda > 0$. For nonparametric regression, the stronger form is also valid for cross-validation or plug-in forms of $\hat{p}$ provided we suppose that all moments of the error distribution are finite. In addition to (2.22), assume the conditions imposed on f and on the wavelet basis in Theorems 2.1 or 2.3, and that $\hat{p}$ satisfies the conditions imposed on p there. That is, with C, η_1, η_2, q as in those results, both $\hat{p} \ge C$ and $\eta_1 \le n^{-1}\hat{p}2^q \log n \le \eta_2$ with probability 1. (For this it is adequate that deterministic thresholds $p_1 < p_2$ satisfy these conditions, and $P(p_1 \le \hat{p} \le p_2) = 1$.) Then, we claim that the following version of (2.10) holds:

$$E\{\bar{f}(x_1) - f(x_1)\}^2 = O\Big[(p_0/n) + \min\Big\{(\omega_1/p_0)^{2r}, (\omega_1 n^{-1} \log n)^{2r/(2r+1)}\Big\}\Big]. \tag{2.23}$$

To appreciate why (2.23) holds, note that since the wavelets are assumed to be compactly supported, and since $\hat{p}$ is constructed using data X_i (in the density estimation context) or (X_i, Y_i) (for nonparametric regression, with $X_i = i/n$) for which $X_i \in \mathcal{R}$, (and so is distant at least η from x_0), then there exists $n_0 \ge 1$ such that $\hat{f}_p(x_1)$ is stochastically independent of $\hat{p}$, for all $n \ge n_0$ and all values of p satisfying $p \ge C$ and $\eta_1 \le n^{-1}p2^q \log n \le \eta_2$. For such values of n we may evaluate $E\{\bar{f}(x_1) - f(x_1)\}^2$ by first conditioning on $\hat{p}$ and then taking expectation in the distribution of $\hat{p}$. Arguing in this way we find that

$$\begin{aligned} E\{\bar{f}(x_1) - f(x_1)\}^2 \le \{1 - P(C_1 \le \hat{p}/p_0 \le C_2)\} \sup_p{}^{(1)}\, E\{\hat{f}_p(x_1) - f_p(x_1)\}^2 \\ + \sup_p{}^{(2)}\, E\{\hat{f}_p(x_1) - f_p(x_1)\}^2, \end{aligned} \tag{2.24}$$

where $\sup_p{}^{(1)}$ [respectively, $\sup_p{}^{(2)}$] denotes the supremum over p satisfying $p \ge C$ and

$$\eta_1 \le n^{-1}p2^q \log n \le \eta_2$$

[respectively, $C_1 \le \hat{p}/p_0 \le C_2$]. The desired result (2.23) is immediate from (2.22), (2.24) and Theorem 2.1 (in the case of density estimation) or Theorem 2.3 (for nonparametric regression).

A numerical study of the relative performance of different empirical rules for selecting the primary resolution level is beyond the scope of this paper. However, the numerical benefits of reducing bias (for example, by using a resolution level that diverges with sample size rather than remains constant) are clear from pre-existing work. See Hall et al. (1997) for a numerical study in this context. There it is shown that bias reduction improves the sharpness of the response of an estimator to irregularities in the curve, and reduces the impact of spurious Gibbs

phenomenon "wiggles". The latter are associated with excessive bias, and, while not directly apparent in first-order theoretical studies such as our own, can reduce the performance of an estimator. On the negative side, as Hall et al. show, redressing the bias-variance trade-off more in favour of variance tends to increase the likelihood of mistaking the signal for spurious effects due to noise.

3. Proofs

Proof of Theorem 2.1. For the sake of definiteness, assume that $\hat{f}$ is given by (2.8). Put $\xi = u(\|f\|_\infty n^{-1} \log n)^{1/2}$, with $0 < u < c$. Write $\hat{f}(x) = \hat{f}_1(x) + \Delta(x)$ where

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

$$\Delta(x) = \sum_{k=0}^{q-1} \sum_\ell \hat{b}_{k\ell} \Big\{ I(|\hat{b}_{k\ell}| > \xi) - I(|b_{k\ell}| > \xi) \Big\} \psi_{k\ell}(x)\,.$$

We 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 inequalities. Observe that by Taylor expansion, there exists a function x' of x such that, interpreting γ as zero if $j = 0$,

$$\begin{aligned} |b_{k\ell}| &= \frac{1}{r!} \Big| p_k^{-\{r+(1/2)\}} \int \psi(x) x^r f_0^{(r)}\{(x' + \ell)/p_k\} dx \\ &\qquad + p_k^{-\{r+(1/2)\}} \omega_j^r \int \psi(x) x^r \gamma^{(r)} \left\{ \omega_j(\ell p_k^{-1} - x_j) + \omega_j x' p_k^{-1} \right\} dx \Big| \\ &\le \max\left(\|f_0^{(r)}\|_\infty, \|\gamma^{(r)}\|_\infty \right) (r!)^{-1} \int |x^r \psi(x)|\, dx \\ &\qquad \times p_k^{-\{r+(1/2)\}} \left[1 + \omega_j^r I\left\{ |\ell p_k^{-1} - x_j| \le 2A(p_k^{-1} + \omega_j^{-1}) \right\} \right]. \end{aligned} \tag{3.1}$$

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

$$E\{\psi_{k\ell}(X)^2\} = \int \psi(x)^2 f\{p_k^{-1}(x + \ell)\}\, dx \le \|f\|_\infty\,. \tag{3.2}$$

By (3.2) and Bernstein's or Bennett's inequality (see for example, Pollard (1984), pp. 192-3), for each $y, \zeta > 0$ and for all sufficiently large n,

$$\begin{aligned} &\max_{0 \le k \le q-1; \ell} P\Big\{ \Big| \hat{b}_{k\ell} - b_{k\ell} \Big| > y(n^{-1} \log n)^{1/2} \Big\} \\ &\le 2 \exp\{-\tfrac{1}{2}(1 - \zeta) \|f\|_\infty^{-1} y^2 \log n\} \le 2n^{-(1-\zeta)y^2/(2\|f\|_\infty)}\,. \end{aligned} \tag{3.3}$$

(Here we used the fact that $p 2^q n^{-1} \log n \to 0$.)

Next, we describe the convergence rate of $E\hat{f}_1(x_j)$ to $f(x_j)$ for $j = 0$ or 1. If ψ vanishes outside $[-A, A]$, then $\psi_{k\ell}(x)$ vanishes unless $|x - \ell p_k^{-1}| \le A p_k^{-1}$, and there are at most $2A + 1$ values of ℓ with this property for any given x. Furthermore by (3.1) there exist constants $C_1, C_2 > 0$ such that if $\psi_{k\ell}(x_j) \neq 0$ (which confers dependence of b_{kl} on j) and $p_k \ge C_1$, then $|b_{k\ell}| \le C_2 p_k^{-\{r+(1/2)\}} \omega_j^r$, for $j = 0$ or 1, while if $\psi_{k\ell}(x_j) \neq 0$ and $p_k < C_1$,

$$|b_{k\ell}| \le p_k^{-1/2} \|f\|_\infty \int |\psi| \le C_2 p_k^{-\{r+(1/2)\}} \omega_j^r \,.$$

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,

$$\begin{aligned} |E\hat{f}_1(x_j) - f(x_j)| &= \Big| \sum_{k=0}^{q-1} \sum_\ell b_{k\ell} I(|b_{k\ell}| \le \xi) \psi_{k\ell}(x_j) + \sum_{k=q}^{\infty} \sum_\ell b_{k\ell} \psi_{k\ell}(x_j) \Big| \\ &\le (2A+1) \|\psi\|_\infty \sum_{k=0}^{q-1} p_k^{1/2} \min\Big(C_2 p_k^{-\{r+(1/2)\}} \omega_j^r, \xi \Big) \\ &\quad + (2A+1) C_2 \|\psi\|_\infty \omega_j^r \sum_{k=q}^{\infty} p_k^{-r} . \end{aligned} \tag{3.4}$$

Defining $\xi_j \equiv \min\{(\omega_j/p)^r, (\omega_j n^{-1} \log n)^{r/(2r+1)}\}$ we find that

$$\sum_{k=0}^{q-1} p_k^{1/2} \min\Big(C_2 p_k^{-\{r+(1/2)\}} \omega_j^r, \xi \Big) = O(\xi_j) \,,$$

$$\sum_{k=q}^{\infty} p_k^{-r} = O(p^{-r} 2^{-qr}) = O(\omega_j^{-r} \xi_j) \,, \tag{3.5}$$

where the latter holds if $p 2^q > n\eta_1 / \log n$ where $\eta_1^{-1} (n^{-1+\epsilon} \log n)^{2r/(2r+1)}$ is bounded. Therefore,

$$|E\hat{f}_1(x_j) - f(x_j)| = O(\xi_j) \,. \tag{3.6}$$

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

$$\begin{aligned} \mathrm{Var}\{\hat{f}_1(x_j)\} &= n^{-1} \mathrm{Var} \Big\{ \sum_\ell \phi_\ell(X) \phi_\ell(x_j) + \sum_{k=0}^{q-1} \sum_\ell \psi_{k\ell}(X) \psi_{k\ell}(x_j) I(|b_{k\ell}| > \xi) \Big\} \\ &\le n^{-1} \|f\|_\infty \Big\{ \sum_\ell |\phi_\ell(x_j)| + \sum_{k=0}^{q-1} \sum_\ell |\psi_{k\ell}(x_j)| I(|b_{k\ell}| > \xi) \Big\}^2 . \end{aligned} \tag{3.7}$$

The inequality in (3.7) follows from the Cauchy-Schwartz inequality and the fact that $E\{\phi_\ell(X)^2\}$, $E\{\psi_{k\ell}(X)^2\} \le \|f\|_\infty$. Arguments similar to those used to

derive (3.6) may now be employed to prove that

$$\mathrm{Var}\{\hat f_1(x_j)\} = O\{(p/n) + \xi_j^2\}\,. \tag{3.8}$$

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

$$\Big|I(|\hat b_{k\ell}| > \delta) - I(|b_{k\ell}| > \xi)\Big| \le I(|\hat b_{k\ell} - b_{k\ell}| > \delta - \xi) + I(|\hat b_{k\ell}| \le \delta, |b_{k\ell}| > \xi),$$
$$\Big|\hat b_{k\ell}\psi_{k\ell}(x)\Big| \le p_k(\|\psi\|_\infty)^2 I(|x - \ell p_k^{-1}| \le A p_k^{-1}),$$

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

$$\begin{aligned}|\Delta(x_j)| \le (\|\psi\|_\infty)^2 \sum_{k=0}^{q-1} p_k \sum_{\ell\in I_{jk}} I(|\hat b_{k\ell} - b_{k\ell}| > \delta - \xi)\\ +\delta \sum_{k=0}^{q-1}\sum_\ell |\psi_{k\ell}(x_j)| I(|b_{k\ell}| > \xi)\,.\end{aligned} \tag{3.9}$$

Since I_{jk} contains at most $2A+1$ elements then by (3.3), for all $\zeta > 0$,

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

where ϵ' may be rendered arbitrarily small by choosing u and ζ sufficiently small. The arguments leading to (3.6) show that

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

Combining the estimates from and below (3.9), we deduce that

$$E\{\Delta(x_j)^2\} = o\{(p/n) + \xi_j^2\}\,. \tag{3.10}$$

The theorem follows from (3.6), (3.8) and (3.10).

Proof of Theorem 2.2. Let $\hat f_1$ and Δ be as in the proof of Theorem 2.1. Let $\mathcal{I}$ denote the set of integer pairs (k,ℓ) such that $\psi_{k\ell}(x_j) \ne 0$ (which confers

dependence of b_{kl} on j), and note that

$$b_{k\ell} = \{(r-1)!\}^{-1}\, p_k^{-\{r+(1/2)\}} \Big[\int_0^1 (1-t)^{r-1}\, dt \int x^r \psi(x)\, f_0^{(r)}\{(tx+\ell)/p_x\}\, dx \\ + \omega_j^r \int_0^1 (1-t)^{r-1}\, dt \int x^r \psi(x)\, \gamma^{(r)}\{\omega_j(tx+\ell-p_k x_j)/p_k\}\, dx \Big]\,. \quad (3.11)$$

When $j = 0$, define $\omega_j = 1$ and $a = (r!)^{-1}\, \kappa f^{(r)}(x_0)$, where $\kappa = \int x^r \psi(x)\, dx$; and when $j = 1$, put $a = (r!)^{-1}\, \kappa \gamma^{(r)}(0)$. In both cases, (3.11) implies that

$$b_{k\ell} = p_k^{-\{r+(1/2)\}}\, \omega_j^r \{a + o(1)\} \quad (3.12)$$

uniformly in $(k, \ell) \in \mathcal{I}_j$. (The second term on the right-hand side of (3.11) may be dropped when $j = 0$.)

Define

$$\alpha_{jk} = p_k^{-1/2} \sum_\ell \psi_{k\ell}(x_j) = \sum_\ell \psi(p 2^k x_j - \ell)\,.$$

Since $x_j = m_1/2^{m_2}$ for (fixed) integers m_1 and m_2, and since $p = 2^m$, then for all sufficiently large p, $\alpha_{jk} = s \equiv \sum \psi(\ell)$ uniformly in $k \ge 0$. Therefore,

$$\Big| \sum_{k=0}^{q-1} \sum_\ell b_{k\ell}\, I(|b_{k\ell}| \le \xi)\, \psi_{k\ell}(x_j) \Big| \ge \tfrac{1}{2}\, |as|\, \omega_j^r \sum_{k=0}^{q-1} p_k^{-r}\, I\Big(2 p_k^{-\{r+(1/2)\}}\, \omega_j^r |a| \le \xi\Big)$$

for all sufficiently large p. (The case of smaller p is easily treated separately.) It may be proved that the right-hand side is bounded below by a constant multiple of ξ_j. Much as in the proof of Theorem 2.1,

$$\Big| \sum_{k=q}^{\infty} \sum_\ell b_{k\ell}\, \psi_{k\ell}(x_j) \Big| = O\Big(\omega_j^r \sum_{k=q}^{\infty} p_k^{-r}\Big) = o(\xi_j)\,. \quad (3.13)$$

(Compare (3.5). That result gives only an upper bound of $O(\xi_j)$, but since we may assume that $\max(\omega_0, \omega_1) = o(n^\epsilon)$, we obtain the bound at (3.13).) Combining these results, and noting (3.4), we see that

$$|E\hat{f}_1(x_j) - f(x_j)| \ge C_3\, \xi_j\,, \quad (3.14)$$

where $C_3, C_4, , \ldots$ are positive constants.

Next we treat the variance of $\hat{f}_1(x_j)$. Observe that, by (3.12),

$$\sum_{k=0}^{q-1} \sum_\ell |\psi_{k\ell}(x_j)|\, I(|b_{k\ell}| > \xi) \le C_4 \sum_{k=0}^{q-1} p_k^{1/2}\, I\Big\{2^k \le C_5\, p^{-1} \Big(\omega_j^{2r}\, n / \log\, n\Big)^{1/(2r+1)}\Big\} \\ \le C_6\, n^{1/2} \Big\{n^{-1}\, \omega_j (\log\, n)^{-1/(2r)}\Big\}^{r/(2r+1)}\,.$$

Also, $E\{\psi_{k\ell}(X)^2\} \le \|f\|_\infty$ and

$$E\Big\{\sum_\ell \phi_\ell(X)\,\phi_\ell(x_j)\Big\}^2 = p\int \Big\{\sum_\ell \phi(px_j + v - \ell)\,\phi(px_j - \ell)\Big\}^2 f(x_j + p^{-1}v)\,dv$$
$$\ge C_7\,p\int \Big\{\sum_\ell \phi(v+\ell)\,\phi(\ell)\Big\}^2 dv \ge C_8\,p\,.$$

(Here we have again used the fact that $x_j = m_1/2^{m_2}$ and $p = 2^m$.) Hence, with

$$T \equiv \sum_\ell \phi_\ell(X)\,\phi_\ell(x_j) + \sum_{k=0}^{q-1}\sum_\ell \psi_{k\ell}(X)\,\psi_{k\ell}(x_j)\,I(|b_{k\ell}| > \xi),$$

we have

$$n^{-1}\,E(T^2) \ge C_8(p/n) + O\Big[(p/n)^{1/2}\Big\{n^{-1}\,\omega_j(\log\,n)^{-1/(2r)}\Big\}^{r/(2r+1)}$$
$$+\Big\{n^{-1}\,\omega_j(\log\,n)^{-1/(2r)}\Big\}^{2r/(2r+1)}\Big]$$
$$= C_8(p/n) + o(\beta_j)\,,$$

where $\beta_j = (p/n) + (\omega_j/n)^{2r/(2r+1)}$. It may be proved that $n^{-1}(ET)^2 = o(\beta_j)$, and so (noting (3.7)),

$$\mathrm{Var}\,\{\hat f_1(x_j)\} = n^{-1}\{E(T^2) - (ET)^2\} \ge C_8(p/n) + o(\beta_j)\,. \tag{3.15}$$

Combining (3.14) and (3.15) we find

$$E\{\hat f_1(x_j) - f(x_j)\}^2 \ge C_9\{(p/n) + \xi_j^2\} + o\{(p/n) + \xi_j^2\}\,.$$

The theorem follows from this result and (3.10).

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; the work of both Fan and Martin was conducted during visits to the Centre for Mathematics and its Applications at the ANU. The comments of two referees and an associate editor have been very helpful in improving presentation.

References

Brown, L. D. and Low, M. G. (1991). Information inequality bounds on the minimax risk (with application to nonparametric regression). *Ann. Statist.* **19**, 329-337.

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

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Donoho, D. L. and Johnstone, I. M. (1995). Neoclassical minimax problems, thresholding and adaptive function estimation. *Bernoulli* **2**, 39-62.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1993). Density estimation by wavelet thresholding. Technical Report No. 426, Department of Statistics, Stanford University, Stanford.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion.) *J. Roy. Statist. Soc. Ser. B* **57**, 301-369.

Doukhan, P. (1988). Formes de Toeplitz associées à une analyse multi-échele. *Comptes Rendus Acad. Sci. Paris* **306**, 663-668.

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

Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23**, 905-928.

Hall, P. and Patil, P. (1996). Effect of threshold rules on performance of wavelet-based curve estimators. *Statist. Sinica* **6**, 331-345.

Hall, P., Penev, S., Kerkyacharian, G. and Picard, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statist. Comput.* **7**, 115-124.

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

Krieger, A. M. and Pickands, J. III (1981). Weak convergence and efficient density estimation at a point. *Ann. Statist.* **9**, 1066-1078.

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

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

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

Stone, C. J. (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.

Department of Statistics, University of North Carolina at Chapel Hill, U.S.A.

E-mail: jfan@stat.unc.edu

Centre for Mathematics and its Applications, Australian National University, Canberra, Australia.

CSIRO Division of Mathematics and Statistics, Sydney, Australia.

E-mail: halpstat@pretty.anu.edu.au

Department of Statistics and Econometrics, Australian National University, Canberra, Australia.

E-mail: martin@beatbox.anu.edu.au

School of Mathematics and Statistics, University of Birmingham, UK.

E-mail: p.n.patil@bham.ac.uk

(Received October 1996; accepted December 1997)