Bernoulli **1(1/2)**, 1995, 041–058

On wavelet methods for estimating smooth functions

PETER HALL and PRAKASH PATIL

Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia

Without assuming any prior knowledge of wavelet methods, we develop theory describing their performance as estimators of smooth functions. The linear part of the wavelet estimator is discussed by analogy with classical kernel methods. Concise formulae are developed for its bias, variance and mean square error. These quantities oscillate somewhat erratically on a wavelength that is equivalent to the bandwidth, reflecting the irregular numerical fluctuations that are observed in practice. Nevertheless, the contributions of these oscillations to mean integrated square error tend to dampen one another out, even over very small intervals, with the result that mean integrated square error properties of linear wavelet methods are much closer to those of kernel methods than is perhaps reasonable, given the local behaviour. We illustrate the adaptive qualities of the nonlinear component of a wavelet estimator by describing its performance when the target function is smooth but has high-frequency oscillations. It is shown that the nonlinear component automatically adapts to changing local conditions, to the extent of achieving (except for a logarithmic factor) the same convergence rate as the optimal linear estimator, but without a need to adjust the underlying bandwidth. This makes explicitly clear the way in which the linear part of the estimator takes care of the 'average' characteristics of the unknown curve, while the nonlinear part corrects for more erratic fluctuations, in a manner which is virtually independent of the construction of the linear part.

Keywords: convergence rate; density estimation; differentiability; dilation equation; kernel method; nonparametric curve estimation; orthogonal series; regression; scaling function; smoothness; wavelet

1. Introduction

The classical theory of nonparametric density or regression estimation, dating from the work of Rosenblatt (1956), is founded on assumptions of smoothness. Convergence rates of estimators, indeed their very construction in terms of kernel choice and bandwidth (for a kernel estimator), are often dictated by those assumptions. In these terms, kernel estimators enjoy convergence rates which are the best possible, uniformly over classes of smooth functions – see, for example, Stone (1982; 1983). In view of this superior performance one might perhaps query the need for alternative approaches. However, that view denies the existence of a wide variety of practical problems where the curves in question might not be smooth in the classical sense. Those curves may contain high-frequency oscillations, or discontinuities, and require techniques which should be easily adapted to differing local levels of smoothness.

Wavelet methods, introduced to statistics by Doukhan (1988), Donoho (1995), Donoho and Johnstone (1992; 1994a; 1994b) and Kerkyacharian and Picard (1992; 1993a; 1993b; 1993c), enjoy exceptional potential for adaptive smoothing. They permit two different levels of smoothing: one

1350-4864

globally, in terms of the frequency of the scaling function; and the other locally, via the scale of the wavelet function. These two levels are quite literally non-overlapping, in that the scaling and wavelet functions are orthogonal. The global level, analogous to bandwidth choice for a kernel estimator, might be selected so as to provide an amount of smoothing which is appropriate, in an 'average' sense, for lower-frequency parts of the curve. On the other hand, the local level provides crucial fine-tuning, allowing a wide variety of adjustments and corrections in different places.

When viewed in this way, wavelet methods might be seen not so much as an alternative to the kernel approach but as a way of enhancing that technique. The 'basic' wavelet estimator is none other than a kernel estimator, albeit in generalized form. This view is not totally new – it appears, for example, in the lecture notes of Kerkyacharian and Picard (1993c) and the paper of Tribouley (1993). However, there does not exist an account of wavelet methods from the familiar statistical viewpoint, with concise asymptotic expressions for bias, variance and mean square error, as distinct from simply upper bounds for these quantities, uniformly over function classes. One of our aims in the present paper is to provide such a development, in the familiar context of estimation of smooth functions. Perhaps unexpectedly, we use this smooth-function approach even to explain the local adaptation properties of wavelet estimators. By way of contrast, the virtues of wavelet methods are usually discussed in terms of their ability to cope well with the *failure* of smoothness assumptions.

Thus, we describe both the structure and the performance of wavelet methods in classical terms, the structure being that of kernel estimators (for the linear part of the wavelet methods) plus a degree of enhancement (the nonlinear part). In an effort to bridge the gulf between classical smooth-function approaches to analysing curve estimators (see, for example, Silverman 1986), and the functional-analytic flavour of recent work on wavelet methods, we avoid emphasis of function classes such as Besov spaces, hopefully without violence to the views of those who might have taken other routes.

Section 2 describes the first stage of our project, dealing with the 'linear' part of wavelet methods. There our main results are as follows. Contrary to the case of classical kernel estimators, the terms representing the bias and variance of generalized kernel estimators, such as those based on the wavelet scaling function, oscillate erratically with a wavelength of the same order as the bandwidth, h. Thus the classical bias and variance formulae,

$$\text{bias}\,(x) = \mathrm{E}\,\{\hat{f}(x)\} - f(x) \simeq a_1(x)h^r$$

and

$$\text{var}\,(x) = \text{var}\,\{\hat{f}\,(x)\} \simeq a_2(x)(nh)^{-1},$$

for smooth functions a_1 and a_2, are no longer valid. They must be replaced by

$$\text{bias}\,(x) \simeq a_1(x)a_3(x/h)h^r$$

and

$$\text{var}\,(x) \simeq a_2(x)a_4(x/h)(nh)^{-1},$$

for new, non-degenerate functions a_3 and a_4. The erratic oscillations represented by $a_3(x/h)$ and $a_4(x/h)$ may be readily observed in practice, in a numerical study of wavelet estimators. That they exist at all is not a particularly endearing feature of the wavelet method. An exclusive focus on *integrated* square error obscures this important issue.

Nevertheless, precisely because the oscillations of bias and variance are so vigorous they tend to cancel when averaged over even a small interval. Thus, formulae for mean integrated square error over arbitrary intervals are virtually identical to their counterparts in the case of classical kernel estimators.

Still in Section 2, we address the issue of empirical bandwidth choice. We show that it is practically feasible, although not necessarily desirable, to track very closely the oscillations of local mean square error. Global bandwidth choice is a more straightforward matter.

Section 3 describes the ability of the nonlinear component of wavelet methods to adapt to local features of an unknown curve. We model irregular features by high-frequency oscillations. It is shown that the nonlinear component of a wavelet estimator adjusts to these irregularities as well as would a classical kernel estimator, but without the need to adjust the bandwidth of the linear component, which may be virtually arbitrary. This makes explicitly clear the way in which one may tailor the linear part of the estimator to the 'average' part of the curve, of 'average' smoothness, and rely on the nonlinear part to correct for more erratic features of the curve.

In Section 2 the bandwidth h of the generalized kernel estimator (or, in the wavelet context, the linear portion of the wavelet estimator) is permitted to be a general positive number. However, in some applications we have seen the bandwidth is taken to be an integer power of $\frac{1}{2}$, meaning that it is perhaps within only 30% to 40% of its most desirable value. That fact, and the oscillations of bias and variance noted earlier, explain a large part of the 'roughness' of wavelet estimators of smooth functions.

We focus most of our attention on density estimators, since this is the simplest case from the viewpoint of exposition. However, the case of nonparametric regression is identical in all important respects, and is addressed in its own right in subsections of Sections 2 and 3.

2. Generalized kernel estimators and linear wavelet estimators

2.1. SUMMARY

Section 2.2 defines generalized kernel density estimators, of which linear wavelet density estimators are a special case. The latter are introduced as an example in Section 2.3. Several of the properties that are usually associated with wavelets do not play any role in this development, and so are delayed to Section 3. In particular, the only clear advantage of orthogonality, in the context of Section 2, is that it confers greater ease of computation. Strang (1989) gives an example of non-orthogonal wavelets.

Section 2.4 discusses analogues of our results in the context of nonparametric regression, and Section 2.5 describes empirical methods for local bandwidth choice. Global bandwidth choice is discussed in an extended version of this paper, Hall and Patil (1993).

2.2. DEFINITION AND BASIC PROPERTIES

Let $K(\cdot, \cdot)$ denote a symmetric function satisfying

$$|K(x,y)| \leq C_1(1+|x-y|)^{-(1+C_2)}, \qquad \forall x, y \in \mathbb{R}, \tag{2.1}$$

$$\int K(x,y)\mathrm{d}y = 1, \qquad \forall x \in \mathbb{R}, \tag{2.2}$$

where C_1, C_2 are positive constants. Let $X_1, \ldots, X_n$ denote independent and identically distributed random variables from a distribution with density f. A generalized kernel estimator of f, with kernel K and bandwidth h, might be defined by

$$\hat{f}(x) = (nh)^{-1} \sum_{j=1}^{n} K(h^{-1}x, h^{-1}X_j).$$

If f is bounded, and continuous at x, then provided that $h \to 0$ and $nh \to \infty$,

$$\mathrm{E}\{\hat{f}(x)\} - f(x) = o(1), \tag{2.3}$$

$$\mathrm{var}\{\hat{f}(x)\} = (nh)^{-1} f(x) \kappa(h^{-1}x) + o\{(nh)^{-1}\}, \tag{2.4}$$

where $\kappa(x) = \int K(x, x+y)^2 \mathrm{d}y$. Therefore the generalized kernel estimator is mean-square consistent under the usual, minimal conditions on f and h. If f is uniformly continuous then (2.3) and (2.4) hold uniformly in x.

As in the case of traditional kernel estimators, i.e. those where $K(x, y) \equiv K_1(x - y)$ for a univariate function K_1, interest centres on conditions under which bias is of order h^r for some integer $r \geq 1$, rather than simply of size $o(1)$ as indicated by (2.3). When f is bounded and has r uniformly continuous derivatives, and (2.1) holds for a constant $C_2 > r$, it may be shown by Taylor expansion that $E\hat{f} - f = O(h^r)$ for all such fs if and only if

$$\lambda_j(x) = 0, \qquad \forall x \in \mathbb{R}, 1 \leq j \leq r - 1, \tag{2.5}$$

where $\lambda_j(x) = \int y^j K(x, x+y) \mathrm{d}y$. In this circumstance,

$$\mathrm{E}\{\hat{f}(x)\} - f(x) = h^r (r!)^{-1} f^{(r)}(x) \lambda_r(h^{-1}x) + o(h^r). \tag{2.6}$$

In the case of traditional kernel estimators the functions κ and λ_r are constant, and so the asymptotic behaviour of variance and bias, evidenced by (2.4) and (2.6), is exceptionally simple. However, in other circumstances both κ and λ_r can be non-degenerate, and the fluctuations of variance and bias can be quite erratic. To appreciate this point, let us assume that K is piecewise continuous, and periodic in the sense that for some $p > 0$,

$$K(x + mp, y + mp) = K(x, y), \qquad \forall x, y \in \mathbb{R}, m \in \mathbb{Z}. \tag{2.7}$$

(The wavelet kernels introduced in the next subsection are periodic with period $p = 1$.) Then for almost all $x \in \mathbb{R}$, in the sense of Lebesgue measure, and for $\alpha = \kappa$ or λ_r,

$$\limsup_{h \to 0} \alpha(h^{-1}x) = \sup_{-\infty < y < \infty} \alpha(y), \qquad \liminf_{h \to 0} \alpha(h^{-1}x) = \inf_{-\infty < y < \infty} \alpha(y). \tag{2.8}$$

It follows that for almost all x, $\kappa(h^{-1}x)$ and $\lambda_r(h^{-1}x)$ oscillate continually between the smallest and largest values that the functions κ and λ_r, respectively, can take.

Precisely because these oscillations are so rapid and frequent, their average over any small interval is very stable. For example, we may deduce from (2.4) and (2.6) that if K is periodic of period p (see

(2.7)), and if $f^{(r)}$ is uniformly continuous over an open set containing the compact interval $\mathscr{I}$, then

$$\int_{\mathscr{I}} E(\hat{f}-f)^2 = (nh)^{-1}\left(\int_{\mathscr{I}} f\right)\left(p^{-1}\int_{-p/2}^{p/2}\kappa\right) + h^{2r}(r!)^{-2}\left(\int_{\mathscr{I}} f^{(r)^2}\right)\left(p^{-1}\int_{-p/2}^{p/2}\lambda_r^2\right) + o\{(nh)^{-1}+h^{2r}\}. \tag{2.9}$$

The oscillations have now gone.

For estimation over an interval $\mathscr{I}$ the asymptotically optimal bandwidth may be computed using formula (2.9). It is $h = \{A_1 a_1 (2rA_2 a_2 n)^{-1}\}^{1/(2r+1)}$, where $A_1 = \int_{\mathscr{I}} f$, $A_2 = (r!)^{-2}\int_{\mathscr{I}} f^{(r)^2}$ denote unknown but estimable quantities, and $a_1 = p^{-1}\int_{-p/2}^{p/2}\kappa$, $a_2 = p^{-1}\int_{-p/2}^{p/2}\lambda_r^2$ are known constants. In terms of its dependence on a_1 and a_2 the asymptotic minimum of $\int_{\mathscr{I}} E(\hat{f}-f)^2$ is an increasing function of $a_1^{2r}a_2$.

2.3. EXAMPLES

To construct a kernel of 'wavelet type', let ϕ be a univariate function with the properties

$$|\phi(x)| \le C_3(1+|x|)^{-(1+C_2)}, \qquad \forall x \in \mathbb{R}, \tag{2.10}$$

$$\sum_{\ell}\phi(x+\ell) = 1, \qquad \forall x \in \mathbb{R}. \tag{2.11}$$

Put

$$K(x,y) = \sum_{\ell}\phi(x+\ell)\phi(y+\ell). \tag{2.12}$$

Conditions (2.10) and (2.11) imply (2.1) and (2.2), respectively, and (2.5) is equivalent to

$$\sum_{\ell}(x+\ell)^j\phi(x+\ell) \equiv \text{const. and } \iint (x-y)^j\phi(x)\phi(y)\mathrm{d}x\,\mathrm{d}y = 0, \qquad 1 \le j \le r-1. \tag{2.13}$$

The constant on the right-hand side of the first identity in (2.13) must equal $\int y^j\phi(y)\mathrm{d}y$; to see why, integrate both sides over $x \in (-\frac{1}{2}, \frac{1}{2})$. (Applying the same argument to (2.11), we deduce that $\int\phi = 1$.) The second part of (2.13) holds trivially if $r = 2$, and is equivalent to

$$\int y^2\phi(y)\mathrm{d}y = \left\{\int y\phi(y)\mathrm{d}y\right\}^2$$

in the cases $r = 3, 4$.

If K is given by (2.12), and the function ϕ satisfies (2.10) and (2.11), then K is periodic with period $p = 1$, in the sense defined by (2.7). Therefore, the quantities $\kappa(h^{-1}x)$ and $\lambda_r(h^{-1}x)$ exhibit the erratic, oscillatory behaviour described by (2.8); and averaging over an interval $\mathscr{I}$ removes these oscillations, as illustrated by (2.9). Explicitly, κ and λ_r may be defined by

$$\kappa(x) = \sum_{\ell_1}\sum_{\ell_2} a_{\ell_1-\ell_2}\phi(x+\ell_1)\phi(x+\ell_2), \tag{2.14}$$

$$\lambda_r(x) = (-1)^r\left\{\sum_{\ell}(x+\ell)^r\phi(x+\ell)\right\} + b_r + (-1)^{r+1}d_r, \tag{2.15}$$

where $a_\ell = \int \phi(x)\phi(x+\ell)\mathrm{d}x$, $b_r = \int\int (u-v)^r \phi(u)\phi(v)\mathrm{d}u\,\mathrm{d}v$, $d_r = \int y^r \phi(y)\mathrm{d}y$. The integrals κ and λ_r^2 appearing in (2.9) are given by

$$\int_{-1/2}^{1/2} \kappa = \sum_\ell a_\ell^2,$$

$$\int_{-1/2}^{1/2} \lambda_r^2 = b_r^2 - d_r^2 + \int_{-\infty}^{\infty} x^r \phi(x) \left\{ \sum_\ell (x+\ell)^r \phi(x+\ell) \right\} \mathrm{d}x.$$

One class of functions ϕ satisfying the key conditions (2.10), (2.11) and (2.13) is that of 'scale functions' or 'father wavelets', determined by the orthonormality relation

$$\int \phi(x)\phi(x+\ell)\mathrm{d}x = \delta_{0\ell}, \qquad -\infty < \ell < \infty, \tag{2.16}$$

and the dilation equation,

$$\phi(x) = \sum_\ell c_\ell \phi(2x-\ell), \tag{2.17}$$

where $\delta_{0\ell}$ is the Kronecker delta and the scaling parameters c_ℓ satisfy $\sum \ell^{2r} c_\ell^2 < \infty$, $\sum c_\ell = 2$ and

$$\sum_\ell (-1)^\ell \ell^j c_\ell = 0, \qquad 0 \le j \le r-1. \tag{2.18}$$

If (2.18) holds with $r = 1$ then a non-degenerate solution ϕ of (2.17) exists, and satisfies (2.11) if it is normalized so that $\int \phi = 1$. If (2.18) holds for some $r \ge 2$ then ϕ (determined by (2.16) and (2.17)) also satisfies (2.13), for the same value of r. The remaining condition, (2.10), is trivially satisfied if ϕ is compactly supported, which is typically the case in practice. Cases where ϕ does not have compact support include the Meyer wavelet, but that also obeys (2.10) if the appended function ν used in its construction is sufficiently smooth (Daubechies 1992, pp. 117ff.). The simplest scaling function is $\phi(x) = 1$ for $0 < x \le 1$ and 0 otherwise. It satisfies (2.17) with $c_0 = c_1 = 1$ and $c_\ell = 0$ otherwise, and gives rise to the Haar wavelet sequence. The resulting density estimator $\hat{f}$ is the usual histogram estimator with bin-width h. Formula (2.18) holds with $r = 1$.

In view of (2.16) the function κ in (2.14) simplifies to $\kappa(x) = \sum \phi(x+\ell)^2$, and $\int_{-1/2}^{1/2} \kappa = 1$.

2.4. GENERALIZATION TO NONPARAMETRIC REGRESSION

Extension of these ideas to the context of nonparametric regression is straightforward, and the results are direct analogues of those in the density estimation case. That is, local behaviour of both bias and variance can be very erratic, but this very property ensures that the average of local behaviour over even a small interval is particularly stable. For example, let us consider nonparametric regression in which the design points x_i are conditioned upon, and where the data (x_i, Y_i), $1 \le i \le n$, are generated by the model $Y_i = g(x_i) + \epsilon_i$, with the errors ϵ_i being independent with zero mean and variances $\sigma(x_i)^2$, and g is to be estimated. A generalized kernel estimator of g is given by

$$\hat{g}(x) = \left\{ \sum_{i=1}^n K(h^{-1}x, h^{-1}x_i) Y_i \right\} \left\{ \sum_{i=1}^n K(h^{-1}x, h^{-1}x_i) \right\}^{-1}.$$

Assuming that g, the design density f and the variance function σ^2 are sufficiently smooth, and that the generalized kernel satisfies the usual conditions (2.1), (2.2) and (2.5), the latter in the case $r = 2$ for the sake of simplicity, we may deduce that

$$\operatorname{var}\{\hat{g}(x)\} = (nh)^{-1} b_1(x) \kappa(h^{-1}x) + o\{(nh)^{-1}\},$$

$$\mathrm{E}\{\hat{g}(x)\} - g(x) = \tfrac{1}{2} h^2 b_2(x) \lambda_2(h^{-1}x) + o(h^2),$$

where $b_1 = \sigma^2 f^{-1}$, $b_2 = g'' + 2g'f'f^{-1}$ and κ, λ_2 are as defined in Section 2.2. If the kernel K is periodic (see (2.7)) and if f, f^{-1}, g'' and σ^2 are bounded and uniformly continuous over an interval $\mathscr{I}$, then the wild oscillations of $\kappa(h^{-1}x)$ and $\lambda_2(h^{-1}x)$ may be damped by averaging variance and bias over $\mathscr{I}$. For example, the average mean square error of $\hat{g}$ is given by

$$\int_{\mathscr{I}} \mathrm{E}(\hat{g} - g)^2 = (nh)^{-1} \left(\int_{\mathscr{I}} b_1 \right) \left(p^{-1} \int_{-p/2}^{p/2} \kappa \right) + \frac{1}{4} h^4 \left(\int_{\mathscr{I}} b_2^2 \right) \left(p^{-1} \int_{-p/2}^{p/2} \lambda_r^2 \right) + o\{(nh)^{-1} + h^4\}, \tag{2.20}$$

in direct analogy with (2.9).

When the design points x_i are equally spaced, and the argument x equals one of the design points, the denominator in (2.19) does not depend on x, provided that one is not close to the boundary of the set of design points. More generally, in the context of equally spaced design but for general x one would typically replace (2.19) by

$$\hat{g}(x) = (nh)^{-1} \sum_{i=1}^{n} K(h^{-1}x, h^{-1}x_i) Y_i,$$

in direct analogy with the kernel density estimator $\hat{f}$ defined in Section 2.2.

2.5. SHOULD WE TRY TO TRACK LOCAL OSCILLATIONS, AND IF SO, HOW?

In order to formulate this question clearly we return to our expressions for variance and bias, (2.4) and (2.6). Together they produce an expression for the mean square error of f at a fixed point x:

$$D(h) = \mathrm{E}\{\hat{f}(x) - f(x)\}^2 = D_1(h) + o\{(nh)^{-1} + h^{2r}\},$$

where

$$D_1(h) = (nh)^{-1} f(x) \kappa(h^{-1}x) + h^{2r} (r!)^{-2} f^{(r)}(x)^2 \lambda_r(h^{-1}x)^2.$$

In the context of classical kernel density estimation, the bandwidth $h_1 = h_1(x)$ which minimizes D_1, and asymptotically minimizes D, is equal to a constant multiple of $n^{-1/(2r+1)}$. However, this result does not hold if the kernel K is periodic and the functions κ and λ_r are non-degenerate, both of which statements are typically true for a wavelet kernel. There, for almost all real numbers x (in the sense of Lebesgue measure), the value of h_1 oscillates continually between two different multiples of $n^{-1/(2r+1)}$ as $n \to \infty$. Similarly, the minimum value of D_1, and the asymptotic minimum value of D, oscillate between two different multiples of $n^{-2r/(2r+1)}$.

Should we try to track these oscillations by producing an empirical bandwidth $\hat{h}_1 = \hat{h}_1(x)$ which closely emulates the properties of $h_1(x)$? We argue against such a procedure on at least practical

grounds. However, in sheer theoretical terms the reader may find that this approach is attractive. So we shall briefly describe an empirical algorithm which produces a version of h_1.

An empirical rule for minimizing D_1 may be constructed by substituting estimates $\tilde{f}(x)$ and $\tilde{f}^{(r)}(x)$ for $f(x)$ and $f^{(r)}(x)$, respectively, in the definition of D_1. (These may be taken to be kernel estimators.) The resulting function,

$$\hat{D}_1(h) = (nh)^{-1}\tilde{f}(x)\kappa(h^{-1}x) + h^{2r}(r!)^{-2}\tilde{f}^{(r)}(x)^2\lambda_r(h^{-1}x)^2,$$

may be minimized with respect to h, producing an estimator $\hat{h}_1$ of h_1. However, in order to perform in the manner expected this procedure must use estimators $\tilde{f}$ and $\tilde{f}^{(r)}$ of particularly high quality. To appreciate why, note that since h_1 is of size $n^{-1/(2r+1)}$ then the value of $\hat{h}_1^{-1}x$ will be close to $h_1^{-1}x$, in the sense that $\hat{h}_1^{-1}x - h_1^{-1}x \to 0$ with probability 1, if and only if $\hat{h}_1/h_1 = 1 + o(n^{-1/(2r+1)})$ with probability 1. In order to ensure that this is the case we should select our estimators of f and $f^{(r)}$ such that

$$\tilde{f}(x) - f(x) = o(n^{-1/(2r+1)}) \text{ and } \tilde{f}^{(r)}(x) - f^{(r)}(x) = o(n^{-1/(2r+1)}) \tag{2.21}$$

with probability 1. Under this condition it may be proved that the quantity $\hat{h}_1$ which minimizes $\hat{D}_1$ satisfies both $\hat{h}_1/h_1 = 1 + o(n^{-1/(2r+1)})$ and that $D_1(\hat{h}_1)/D_1(h_1) = 1 + o(1)$ with probability 1. On the other hand, these conditions will fail if the quality of the estimators $\tilde{f}$ and $\tilde{f}^{(r)}$ is so poor that condition (2.21) is violated.

Such an approach can be highly computer-intensive, which is one argument against it. Note that the values of $\kappa(h^{-1}x)$ and $\lambda_r(h^{-1}x)$, as functions of x, have period ph, where p denotes the period of the kernel K. Therefore $h_1(x)$ changes substantially within an interval whose width is of order $n^{-1/(2r+1)}$. This means that $\hat{h}_1(x)$ has to be varied very frequently as x changes, in order to track $h_1(x)$. Local bandwidth rules for traditional kernel estimators may be constructed on a much more *ad hoc* basis, with typically only a small number (e.g. two to four) bench-mark values computed at different xs, and the bandwidths at intervening points computed by simple interpolation.

From other points of view, too, it is disconcerting to have a bandwidth selection rule which is so sensitive to location. For example, if one wished to estimate a density at a given point x then it would be common practice to translate the data by the amount $-x$, so that the point of interest became the origin. But at the origin the functions $\kappa(h^{-1}x)$ and $\lambda_r(h^{-1}x)$ do not oscillate at all, and so the approach to bandwidth choice would be very different.

The extreme sensitivity of the bandwidth selector to the accuracy of the pilot estimators $\tilde{f}$ and $\tilde{f}^{(r)}$ is another drawback to tracking local oscillations. It can be awkward to ensure good accuracy at such a preliminary stage of the algorithm.

There are obvious and immediate analogues of all these points in the context of nonparametric regression, based on the parallels drawn in Section 2.4.

3. Nonlinear wavelet estimators

3.1. SUMMARY

Section 3.2 introduces the wavelet function and develops its basic properties. Wavelet density estimators are described in Section 3.3, and their properties discussed in Section 3.4. A formal limit

theorem describing the extraordinary adaptability of wavelet methods is stated in Section 3.5. Finally, Section 3.6 outlines versions of our results in the context of nonparametric regression. Related work, particularly on bounds for large deviations of estimators uniformly over Besov spaces, may be found in Donoho *et al.* (1993; 1995).

3.2. THE WAVELET FUNCTION

The wavelet function ψ may be obtained from the scaling function ϕ by taking differences in the dilation equation (2.17):

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

The functions ϕ and ψ are sometimes termed 'father' and 'mother' wavelets, respectively. Examples are given by Daubechies (1992, Chapters 5–7). If the scaling and orthonormality relations (2.16)–(2.18) hold, then for $0 \le j \le r - 1$ and $-\infty < \ell < \infty$,

$$\int x^j \psi(x) \mathrm{d}x = 0, \qquad \int \psi(x) \phi(x + \ell) \mathrm{d}x = 0,$$

$$\int \psi(x) \psi(x + \ell) \mathrm{d}x = \delta_{0\ell}, \qquad \int \psi(x) \psi(2x + \ell) \mathrm{d}x = 0.$$

Let $k \ge 0$ be an integer, let $p > 0$, and put $p_k = 2^k p$, and define

$$\phi_\ell(x) = p^{1/2} \phi(px + \ell), \qquad \psi_{k\ell}(x) = p_k^{1/2} \psi(p_k x + \ell)$$

for $-\infty < \ell < \infty$. The results noted earlier imply that the functions ϕ_ℓ and $\psi_{k\ell}$ are orthonormal:

$$\int \phi_{\ell_1} \phi_{\ell_2} = \delta_{\ell_1 \ell_2}, \qquad \int \psi_{k_1 \ell_1} \psi_{k_2 \ell_2} = \delta_{k_1 k_2} \delta_{\ell_1 \ell_2}, \qquad \int \phi_{\ell_1} \psi_{k\ell_2} = 0.$$

Orthogonality of the functions $\psi_{k\ell}$ is the chief element of so-called 'multiresolution analysis' – see Daubechies (1992, pp. 13ff.).

3.3. WAVELET EXPANSIONS AND WAVELET DENSITY ESTIMATORS

If f is a square-integrable function, and ϕ satisfies (2.16)–(2.18), then, for each $p > 0$, f may be expanded as a generalized Fourier series in the orthonormal functions ϕ_ℓ and $\psi_{k\ell}$:

$$f(x) = \sum_{\ell} b_\ell \phi_\ell(x) + \sum_{k=0}^{\infty} \sum_{\ell} b_{k\ell} \psi_{k\ell}(x), \tag{3.1}$$

where $b_\ell = \int f \phi_\ell$, $b_{k\ell} = \int f \psi_{k\ell}$ – see Strang (1989). When f is a probability density, and $X_1, \ldots, X_n$ are independent data values drawn from the corresponding distribution, unbiased estimators of b_ℓ and $b_{k\ell}$ are given by

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

One approach to estimating f is simply to ignore the (mother) wavelet terms in (3.4), focusing attention on that part of the expansion which derives from the scaling function ϕ. This argument produces the estimator

$$\hat{f}(x) = \sum_{\ell} \hat{b}_{\ell} \phi_{\ell}(x),$$

which is identical to the generalized kernel estimator defined in Section 2, with kernel K given by (2.12) and bandwidth $h = p^{-1}$. The performance of this estimator may often be enhanced, even for smooth densities f, by incorporating the (mother) wavelet terms from (3.1). Following Kerkyacharian and Picard (personal communication) we do this by 'thresholding'. This method endeavours to exclude from (3.1) those terms where accurate estimation of $b_{k\ell}$ is not possible, owing to the size of the error about the mean, $\hat{b}_{k\ell} - b_{k\ell}$. Considerations of this nature suggest the estimator

$$\tilde{f}(x) = \sum_{\ell} \hat{b}_{\ell} \phi_{\ell}(x) + \sum_{k=0}^{q-1} \sum_{\ell} \hat{b}_{k\ell} w(\hat{b}_{k\ell}/\delta) \psi_{k\ell}(x), \tag{3.2}$$

where $q = 1, 2, \ldots$ and $\delta > 0$ are adjustable constants, and the 'weight' or 'threshold' function w satisfies

$$w(u) \begin{cases} = 0 & \text{if } 0 < u < c_1 \\ \in [0,1] & \text{if } c_1 \le u \le c_2 \\ = 1 & \text{if } u > c_2 \end{cases}$$

for constants $0 < c_1 < c_2 < \infty$. Examples of w include

$$w(u) = \begin{cases} 0 & \text{if } 0 < u < 1 \\ 1 & \text{if } u > 1, \end{cases} \tag{3.3}$$

which corresponds to 'hard thresholding', and

$$w(u) = \begin{cases} 0 & \text{if } 0 < u < c_1 \\ (u - c_1)/c_2 & \text{if } c_1 \le u \le c_1 + c_2 \\ 1 & \text{if } u > c_2, \end{cases}$$

corresponding to 'soft thresholding'. We suggest taking $\delta = C(n^{-1} \log n)^{1/2}$. The n^{-1} term here is motivated by the fact that each $\hat{b}_{k\ell}$ has variance approximately n^{-1}, the log n term arises from the 'large deviation' nature of the problem; and the constant C may be any sufficiently large number. The integer q should be chosen large, but not so large that the series defining $\hat{b}_{k\ell}$ contains so few non-zero terms that accurate estimation of $b_{k\ell}$ is not feasible.

3.4. ADVANTAGES OF WAVELETS AND THRESHOLDING

The analysis in Section 2 shows that the estimator $\hat{f}$ has several important properties that are virtually identical to those of a classical kernel estimator based on an rth-order kernel. In particular, it has variance of size $(nh)^{-1} f$ and bias of size $h^r f^{(r)}$, where $h = p^{-1}$. To appreciate the limitations of

any estimator with these features, particularly when applied to smooth densities, consider estimating an f which, in the vicinity of x_j for $1 \le j \le \nu$, has rather severe oscillations with frequency ω_j. For example, near x_j, f might have the form $c_j + \cos(\omega_j x)$, where $c_j \ge 1$ and $\omega_j \ge 1$ is large. Now, $f^{(r)}(x) \simeq \omega_j^r$ when x is near x_j, and so the variance and squared bias of $\hat{f}$ are, respectively, of size $(nh)^{-1}$ and $(\omega_j h)^{2r}$ there. Equating these two quantities we see that $h \simeq (n\omega_j^{2r})^{-1/(2r+1)}$ is the optimal size for the bandwidth. It produces an estimator whose mean square error in the vicinity of x_j is of size

$$(\omega_j/n)^{2r/(2r+1)}. \tag{3.4}$$

If the frequencies ω_j are quite different from one another then, to achieve the optimal convergence rate described by (3.4), we must choose very different bandwidths $h = h_j$ in neighbourhoods of the different points x_j. Thus, the 'optimal' verion of the density estimator $\hat{f}$ may have to be based on many different smoothing parameters, each of which has to be selected empirically. These different versions of $\hat{f}$ need to be spliced together to produce a final estimator of f. Such an approach demands a considerable amount of effort, both in formulating an algorithm for defining such an estimator and in actually calculating it for real data.

The virtue of wavelet methods and thresholding is that, provided the parameters q and δ are chosen appropriately, the estimator $\tilde{f}$ achieves this end automatically, with relatively little effort. By including the second series in (3.2) – that is, by appending appropriate nonlinear wavelet terms to the estimator $\hat{f}$, which was formerly linear and based solely on the scaling function ϕ – we produce an estimator $\tilde{f}$ which automatically adapts itself to the varying frequency of f. The thresholding approach selects the required extra terms in a manner which is virtually optimal, in that it attains the convergence rate described by (3.4) except for the inclusion of a logarithmic factor. In the next subsection we shall state a result which demonstrates this fact. The wavelet estimator $\tilde{f}$ achieves this level of performance in a 'smooth' way, in the sense that, provided ψ, or equivalently ϕ, is smooth, the estimator $\tilde{f}$ is continuous. That is, the splicing referred to in the previous paragraph is achieved automatically. This follows on inspection of the formula for $\tilde{f}$.

3.5. LIMIT THEOREM FOR $\tilde{f}$

We could state our result in the case of a density f which had any (finite) number of different oscillation frequencies ω_j, possibly all varying with n, as postulated in the previous subsection. The overall convergence rate of $\tilde{f}$ would then be $n^{-1}p + (\omega n^{-1}\log n)^{2r/(2r+1)}$, in mean square error terms, where $\omega = \max\{\omega_j\}$. It is, however, notationally simpler to treat the case of a single frequency, and allow the reader mentally to make the simple extension to arbitrary ν. To simplify the notation further we shall take the density f to be defined in the unit interval, although more complex settings are easily treated using identical arguments.

Let a_1, a_2, a_3 denote positive constants with $a_3 \le a_1^{-1}$. We next define a class $\mathscr{F}_\omega = \mathscr{F}_\omega(a_1, a_2, a_3)$ of densities on (0,1), of which all members have 'frequency' ω. Let $\gamma : (-\infty, \infty) \to (-\infty, \infty)$ denote an r-times-differentiable periodic function with period 1 satisfying

$$\int_0^1 \gamma = 0, \qquad |\gamma| \le a_1, |\gamma^{(r)}| \le a_2.$$

Then $\mathscr{F}_\omega$ is the class of densities f that may be written in the form

$$f(x) = 1 + a_3\gamma(\omega x), \qquad 0 < x < 1,$$

for such a γ.

Theorem 3.1 Assume conditions (2.16)–(2.18), that the wavelet function ψ is bounded and compactly supported, and that ψ^2 is Hölder continuous on its support. Suppose that $\omega = \omega(n) = o(n/\log n)$ as $n \to \infty$. Take $\delta = C(n^{-1}\log n)^{1/2}$. If $C > 0$ is sufficiently large, and if $p = p(n) \to \infty$ and $q = q(n) \to \infty$ in such a manner that

$$2^q n^{-1} p \log n \to 0 \text{ and } 2^{-q} n p^{-2+(1/2r)} \log n = O(1), \tag{3.5}$$

then

$$\sup_{f \in \mathscr{F}_\omega} \int \mathrm{E}_f(\hat{f} - f)^2 = \{1 + o(1)\} n^{-1} p + O\{(\omega n^{-1}\log n)^{2r/(2r+1)}\} \tag{3.6}$$

as $n \to \infty$.

A proof of Theorem 3.1 is given in the Appendix.

Remark 3.1
Our proof shows that for any $r \geq 1$, and in the case of hard thresholding, it suffices to take $C \geq 12^{1/2}$.

Remark 3.2
The first part of (3.5) is equivalent to asking that the estimators $\hat{b}_{k\ell}$ be based on a larger number than $O(\log n)$ sample values. This is a very weak assumption – one can hardly hope to obtain an accurate estimate of $b_{k\ell}$ from only $O(\log n)$ data values. To appreciate the validity of the claim in the first sentence of this remark, note that since ψ is compactly supported the expected number of non-zero terms in the series $\Sigma_i \psi(p_k X_i + \ell) = n\hat{b}_{k\ell}$ is no greater than const. np_k^{-1}, which is of larger order than $\log n$ for each $k \leq q$ if and only if the first part of (3.5) holds.

Remark 3.3
Complementing the first part of (3.5), which asks that q not be too large, the second part demands that q not be too small. To appreciate that it, too, is a weak condition, let us take $p_q \simeq$ const. $n^c(\log n)^d$ for constants $c > 0$ and d. Then the second part of (3.5) requires that $n^{1-c}(\log n)^{1-d} = O(p^{(2r-1)/2r})$. In practice, p would typically be asymptotic to a constant multiple of $n^{1/(2r+1)}$ – see Remark 3.5 below – and then the second part of (3.5) would be equivalent to asking that

$$c > \{(2r)^2 + 1\}/\{2r(2r+1)\}, \qquad \text{or } c = \{(2r)^2 + 1\}/\{2r(2r+1)\} \text{ and } d > 1.$$

On the other hand, the first part of (3.5) is equivalent to

$$c < 1, \qquad \text{or } c = 1 \text{ and } d < -1.$$

Thus, both parts of (3.5) are satisfied if $p \simeq$ const. $n^{1/(2r+1)}$ and $p_q \simeq$ const. n^c, with $\{(2r)^2 + 1\}/\{2r(2r+1)\} < c < 1$.

Remark 3.4
Our assumptions on ϕ and ψ are satisfied by the wavelet functions that are typically used in practice – see Daubechies (1992, Chapter 6). In particular, they are satisfied in the case of the Haar sequence.

Remark 3.5
In practical density estimation one would choose the effective bandwidth, $h = p^{-1}$, of the 'kernel estimator' part of $\tilde{f}$ so that it was appropriate in places where f was smooth, without high-frequency oscillations. This would, of course, require $p \simeq \text{const.}\, n^{1/(2r+1)}$; see, for example, Section 2.5. In this instance the first term on the right-hand side of (3.6) is dominated by the last, and so (3.6) is equivalent to

$$\sup_{f \in \mathscr{F}_\omega} \int \mathrm{E}_f(\tilde{f} - f)^2 = O\{(\omega n^{-1} \log n)^{2r/(2r+1)}\}.$$

Remark 3.6
Since the first version of this manuscript was prepared, related work of Donoho *et al.* (1995) (see, for example, their Theorem 6) has addressed related problems in the context of large deviations of density estimators uniformly over Besov spaces. See also Donoho *et al.* (1993).

3.6. NONPARAMETRIC REGRESSION

We treat the case where $n = 2^k$ for a positive integer k, and the x_is are equally spaced on the interval (0,1).

Let a_1, a_2 denote positive constants; let $\gamma : (-\infty, \infty) \to (-\infty, \infty)$ be a periodic function with period 1, satisfying $\int \gamma = 1$, $|\gamma| \le a_1$, $|\gamma^{(r)}| \le a_2$; and let $\mathscr{G}_\omega = \mathscr{G}_\omega(a_1, a_2)$ denote the class of functions g on (0,1) that may be expressed in the form $g(x) = \gamma_1(x) + \gamma_2(\omega x)$, with γ_1 and γ_2 both having the properties ascribed to γ. Let $k \ge 0$ be an integer, let $p > 0$ and put $p_k = 2^k p$. Define ϕ_ℓ and $\psi_{k\ell}$ as in Section 3.2. The generalized Fourier expansion of g is

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

where $b_\ell = \int g\phi$, $b_{k\ell} = \int g\psi_{k\ell}$. Suppose we observe $Y_i = g(x_i) + \epsilon_i$, $1 \le i \le n$, where $x_i = i/n$ and the ϵ_is are independent and identically distributed with zero mean. Our estimators of b_ℓ and $b_{k\ell}$ are

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

A nonlinear wavelet estimator of g is

$$\tilde{g}(x) = \sum_\ell \hat{b}_\ell \phi_\ell(x) + \sum_{k=0}^{q-1} \sum_\ell \hat{b}_{k\ell} w(\hat{b}_{k\ell}/\delta) \psi_{k\ell}(x),$$

where q, δ and w are as in (3.2).

Theorem 3.2 Assume the conditions of Theorem 3.1, and that the ϵ_is are either normally distributed or have a compactly supported distribution, with variance σ^2. If C, in the definition $\delta = C(n^{-1}\log n)^{1/2}$, is sufficiently large then

$$\sup_{g\in\mathscr{G}_\omega} \int \mathrm{E}(\tilde{g}-g)^2 = \{1+o(1)\}\sigma^2 n^{-1}p + O\{(\omega n^{-1}\log n)^{2r/(2r+1)}\}$$

as $n \to \infty$.

This result is a close analogue of Theorem 3.1, and has a similar proof. Its implications are those of the earlier result – that nonlinear wavelet methods adapt readily to relatively high-frequency, local irregularities in the curve, and thereby compensate for the excessive smoothing that is sometimes inherent in a global choice of the bandwidth $h = p^{-1}$.

Each of Remarks 3.1–3.5 has an analogue in the present setting, which we shall not pursue here.

Acknowledgements

We are particularly grateful to Gérard Kerkyacharian and Dominique Picard for helpful conversations, and to a referee for constructive criticism.

Appendix: Proof of Theorem 3.1

We shall prove only the version of the theorem for the case of hard thresholding, where the function w is given by (3.3). Throughout we shall take $\delta = 12^{1/2}(n^{-1}\log n)^{1/2}$; it will be clear that all our conclusions remain valid if $12^{1/2}$ is replaced by any larger positive number.

Let $C_1, C_2, \ldots$ denote generic positive constants depending only on a_1, a_2, a_3 and ψ, and in one instance on $y > 0$. For $j = 1, 2, \ldots$, let $\{\epsilon_{jn}, n \ge 1\}$ denote a sequence of positive numbers converging to zero. Write $\max'$ for the maximum of a quantity over $0 \le k \le q$, $0 \le \ell \le p_k - 1$ and $f \in \mathscr{F}_\omega$. If $|\psi| \le C_1$ then $|\psi_{k\ell}(x) - b_{k\ell}| \le 2p_k^{1/2}C_1$, and also

$$\begin{aligned} |b_{k\ell}| &= \left| p_k^{-1/2} \int \psi(x) f\{p_k^{-1}(x-\ell)\}\mathrm{d}x \right| \\ &\le C_1(1+a_3)p_k^{-1/2}, \\ \max' \mathrm{E}\{\psi_{k\ell}(X)^2\} &= \max' \int \psi(x)^2 f\{p_k^{-1}(x-\ell)\}\mathrm{d}x \\ &\le 1 + a_1 a_3 \le 2. \end{aligned}$$

(Hölder continuity of ψ^2 is used to obtain the inequality here.) Hence for all $n \ge C_2$,

$$\max' \mathrm{var}\{\psi_{k\ell}(X)\} \le 2.$$

Using Bernstein's or Bennett's inequality (see, for example, Pollard 1984, pp. 192–193) and the bounds established above we see that for each $y > 0$ and all $n \geq C_3(y)$,

$$\begin{aligned}\max{}' P_f\{|\hat{b}_{k\ell} - b_{k\ell}| > (n^{-1}\log n)^{1/2} y\} \\ &= \max{}' P_f\left[\left|\sum_{i=1}^{n}\{\psi_{k\ell}(X_i) - b_{k\ell}\}\right| > (n\log n)^{1/2} y\right] \\ &\leq 2\exp\{-\tfrac{1}{2}(1-\epsilon_{3n})y^2\log n\} = 2n^{-(1-\epsilon_{3n})y^2/4}. \qquad \text{(A.1)}\end{aligned}$$

(To obtain this bound from Bernstein's or Bennett's inequality we need the first relation in (3.5).)

Observe next that

$$\begin{aligned}\int(\hat{f}-f)^2 = \sum_{\ell=0}^{p-1}(\hat{b}_\ell - b_\ell)^2 + \sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1}(\hat{b}_{k\ell} - b_{k\ell})^2 I(|\hat{b}_{k\ell}| > \delta) \\ + \sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1} b_{k\ell}^2 I(|\hat{b}_{k\ell}| \leq \delta) + \sum_{k=q}^{\infty}\sum_{\ell=0}^{p_k-1} b_{k\ell}^2, \qquad \text{(A.2)}\end{aligned}$$

$$\mathrm{E}\left\{\sum_{\ell=0}^{p-1}(\hat{b}_\ell - b_\ell)^2\right\} = n^{-1}p + \epsilon_{4n}n^{-1}p. \qquad \text{(A.3)}$$

Recall that $\delta = 12^{1/2}(n^{-1}\log n)^{1/2}$, let $0 < u < 2^{3/2}$, and put $\xi = u(n^{-1}\log n)^{1/2}$ and $\eta = (12^{1/2} - u)(n^{-1}\log n)^{1/2}$. Since $\mathrm{E}(\hat{b}_{k\ell} - b_{k\ell})^2 \leq C_4 n^{-1}$ and

$$I(|\hat{b}_{k\ell}| > \delta) \geq I(|b_{k\ell}| > \xi) + I(|\hat{b}_{k\ell} - b_{k\ell}| > \eta)$$

then, for any $a, b > 1$ with $a^{-1} + b^{-1} = 1$,

$$\begin{aligned}s_1 &\equiv \mathrm{E}\left\{\sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1}(\hat{b}_{k\ell} - b_{k\ell})^2 I(|\hat{b}_{k\ell}| > \delta)\right\} \\ &= O\left\{n^{-1}\sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1} I(|b_{k\ell}| > \xi) + \sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1}(E|\hat{b}_{k\ell} - b_{k\ell}|^{2a})^{1/a} P(|\hat{b}_{k\ell} - b_{k\ell}| > \eta)^{1/b}\right\}. \qquad \text{(A.4)}\end{aligned}$$

(Here and below the 'big oh' notation is valid uniformly in $f \in \mathscr{F}_\omega$.) Now, by (A.1),

$$\begin{aligned}\sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1}(\mathrm{E}|\hat{b}_{k\ell} - b_{k\ell}|^{2a})^{1/a} P(|\hat{b}_{k\ell} - b_{k\ell}| > \eta)^{1/b} \\ \leq \sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1}(C_4 p_k^{1-(1/a)})(2n^{-(1/2)(1-\epsilon_{3n})(6^{1/2}-u)^2})^{1/b}\end{aligned}$$

$$= O\left(\sum_{k=0}^{q-1} p_k^{2-(1/a)} n^{-(1/2b)(1-\epsilon_{3n})(6^{1/2}-u)^2}\right)$$

$$= O(p_q^{2-(1/a)} n^{-(1/2b)(1-\epsilon_{3n})(6^{1/2}-u)^2})$$

$$= O\{(n/\log n)^{2-(1/a)} n^{-(1/2b)(1-\epsilon_{3n})(6^{1/2}-u)^2}\} = O(n^{-1+c}), \tag{A.5}$$

for any given $c > 0$, if $u > 0$ is sufficiently small and a is sufficiently large.

It is straightforward to show that

$$\max_{0 \le \ell \le p_k - 1,\, f \in \mathscr{F}_\omega} |b_{k\ell}| \le C_5 p_k^{-1/2} (\omega/p_k)^r. \tag{A.6}$$

Therefore

$$\begin{aligned} n^{-1} \sum_{k=0}^{q-1} \sum_{\ell=0}^{p_k-1} I(|b_{k\ell}| > \xi) &\le n^{-1} \sum_{k=0}^{q-1} p_k I\{C_5^2 p_k^{-1} (\omega/p_k)^{2r} > \xi^2\} \\ &= n^{-1} \sum_{k=0}^{q-1} 2^k p I\{2^k p \le (C_5^2 \xi^{-2} \omega^{2r})^{1/(2r+1)}\} \\ &= O\{n^{-1} (\xi^{-2} \omega^{2r})^{1/(2r+1)}\} \\ &= O[\{\omega n^{-1} (\log n)^{-1/2r}\}^{2r/(2r+1)}] \\ &= O\{(\omega n^{-1})^{2r/(2r+1)}\}. \end{aligned} \tag{A.7}$$

Combining (A.4), (A.5) and (A.7), and taking c sufficiently small, we deduce that

$$s_1 = o\{(\omega n^{-1})^{2r/(2r+1)}\}. \tag{A.8}$$

Next, redefine $\xi = (12^{1/2} + 2)(n^{-1} \log n)^{1/2}$ and $\eta = 2(n^{-1} \log n)^{1/2}$. Since

$$I(|\hat{b}_{k\ell}| \le \delta) \le I(|b_{k\ell}| \le \xi) + I(|\hat{b}_{k\ell} - b_{k\ell}| > \eta),$$

then

$$\begin{aligned} s_2 &\equiv E\left\{\sum_{k=0}^{q-1} \sum_{\ell=0}^{p_k-1} b_{k\ell}^2 I(|\hat{b}_{k\ell}| \le \delta)\right\} \\ &\le \sum_{k=0}^{q-1} \sum_{\ell=0}^{p_k-1} b_{k\ell}^2 I(|b_{k\ell}| \le \xi) + \sum_{k=0}^{q-1} \sum_{\ell=0}^{p_k-1} b_{k\ell}^2 P(|\hat{b}_{k\ell} - b_{k\ell}| > \eta). \end{aligned} \tag{A.9}$$

By (A.1), for any $c > 0$,

$$\begin{aligned} \sum_{k=0}^{q-1} \sum_{\ell=0}^{p_k-1} b_{k\ell}^2 P(|\hat{b}_{k\ell} - b_{k\ell}| > \eta) &= O\left(\sum_{k=0}^{q-1} \sum_{\ell=0}^{p_k-1} p_k^{-1} n^{-(1-\epsilon_{3n})}\right) \\ &= O(q n^{-(1-\epsilon_{3n})}) = O(n^{-1+c}). \end{aligned} \tag{A.10}$$

By (A.6),

$$\begin{aligned}\sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1} b_{k\ell}^2 I(|b_{k\ell}| \le \xi) &\le \sum_{k=0}^{q-1}\sum_{\ell=0}^{p_k-1} \min(b_{k\ell}^2, \xi^2) \\ &= O\left[\sum_{k=0}^{q-1} p_k \min\{p_k^{-1}(\omega/p_k)^{2r}, \xi^2\}\right] \\ &= O\left[pn^{-1}(\log n)\sum_{k=0}^{q-1} 2^k \min\{(n/p)(\log n)^{-1}(\omega/p)^{2r}2^{-(2r+1)k}, 1\}\right] \\ &= O[pn^{-1}(\log n)\{(n/p)(\log n)^{-1}(\omega/p)^{2r}\}^{1/(2r+1)}] \\ &= O\{(\omega n^{-1}\log n)^{2r/(2r+1)}\}. \end{aligned} \tag{A.11}$$

Combining (A.9)–(A.11) we see that

$$s_2 = O\{(\omega n^{-1}\log n)^{2r/(2r+1)}\}. \tag{A.12}$$

By (A.6),

$$\begin{aligned}\sum_{k=0}^{\infty}\sum_{\ell=0}^{p_k-1} b_{k\ell}^2 &= O\left\{\sum_{k=q}^{\infty} p_k p_k^{-1}(\omega/p_k)^{2r}\right\} = O\left\{(\omega/p)^{2r}\sum_{k=q}^{\infty} 2^{-2rk}\right\} \\ &= O\{(\omega/p)^{2r} p p_q^{-2r}\} = O\{(\omega n^{-1})^{2r} p^{-(2r-1)}(n/p_q)^{2r}\} \\ &= O\{(\omega n^{-1}\log n)^{2r}\}, \end{aligned} \tag{A.13}$$

using the second relation in (3.5). Combining (A.2), (A.3), (A.8), (A.12) and (A.13) we deduce that

$$\sup_{f\in\mathcal{F}_\omega}\int E(\hat{f}-f)^2 = n^{-1}p + o(n^{-1}p) + O\{(\omega n^{-1}\log n)^{2r/(2r+1)}\},$$

as had to be shown.

References

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

Donoho, D.L. (1995) Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Herm. Anal.*, to appear.

Donoho, D.L. and Johnstone, I.M. (1992) Minimax estimation via wavelet shrinkage. Technical Report No. 402, Department of Statistics, Stanford University.

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

Donoho, D.L. and Johnstone, I.M. (1994b) Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields*, **99**, 277–303.

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.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, **57**, 301–370.

Doukhan, P. (1988) Formes de Toeplitz associées à une analyse multi-échelle. *C.R. Acad. Sci. Paris*, **306**, 663–668.

Hall, P. and Patil, P. (1993) On wavelet methods for estimating smooth functions. Research Report No. CMA-SR12-93, Centre for Mathematics and its Applications, Australian National University.

Kerkyacharian, G. and Picard, D. (1992) Density estimation in Besov spaces. *Statist. Probab. Lett.*, **13**, 15–24.

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.

Kerkyacharian, G. and Picard, D. (1993c) Introduction aux Ondelettes et Estimation de Densité, 1: Introduction aux Ondelettes et à l'Analyse Multiresolution. Lecture notes.

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

Rosenblatt, M. (1956) Remarks on some nonparametric estimators of a density function. *Ann. Math. Statist.* **27**, 832–837.

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

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

Stone, C.J. (1983) Optimal uniform rate of convergence for nonparametric estimate of a density function and its derivatives. In M.H. Rizvi, J.S. Rustagi and D. Siegmund (eds), *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, pp. 393–406. New York: Academic Press.

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

Tribouley, K. (1993) Practical estimation of multivariate density using wavelet methods. Manuscript.

Received November 1993 and revised December 1994