The Annals of Statistics
1997, Vol. 25, No. 5, 1912–1925

INTERPOLATION METHODS FOR NONLINEAR WAVELET REGRESSION WITH IRREGULARLY SPACED DESIGN

BY PETER HALL AND BERWIN A. TURLACH

Australian National University

We introduce interpolation methods that enable nonlinear wavelet estimators to be employed with stochastic design, or nondyadic regular design, in problems of nonparametric regression. This approach allows relatively rapid computation, involving dyadic approximations to wavelet-after-interpolation techniques. New types of interpolation are described, enabling first-order variance reduction at the expense of second-order increases in bias. The effect of interpolation on threshold choice is addressed, and appropriate thresholds are suggested for error distributions with as few as four finite moments.

1. Introduction. Nonlinear wavelet methods in statistics provide a uniquely adaptive tool, offering unsurpassed levels of utility for estimating a wide range of both regular and irregular functions. See, for example, Donoho, Johnstone, Kerkyacharian and Picard (1995, 1996), Kerkyacharian and Picard (1993) and Donoho and Johnstone (1994, 1998). Despite their adaptivity, however, wavelet methods are typically restricted by assumptions about the type of design that may be employed in problems of nonparametric regression. Usually the design must be not only regularly spaced, but dyadic.

This contrasts with other approaches to curve estimation, where a great deal of attention has been paid to the problem of irregularly spaced design points. One of the earliest nonparametric curve estimating techniques, that of Nadaraya and Watson [cf. Härdle (1990), page 25, and Wand and Jones (1995), page 119], is constructed as a ratio that explicitly cancels out most of the effect of stochastic design. Alternative approaches, for example the convolution method of Gasser and Müller (1979) and local polynomial smoothing [e.g. Hastie and Loader (1993)], conduct the cancellation implicitly.

In the present paper we describe an interpolation-based approach to nonlinear wavelet methods in the case of stochastic, or regular but nondyadic, design. We identify threshold parameters that are sufficient for good performance when using interpolation, and describe mean squared error properties for appropriate thresholds. Particular attention is paid to developing results that hold under minimal conditions on the error distribution, so as to demonstrate wide applicability of the interpolation approach. Two different interpolation rules are considered, and their properties elucidated using both theoretical and numerical arguments. It is shown that, in addition to allow-

Received November 1995; revised September 1996.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62G30.
Key words and phrases. Bias, mean squared error, nonparametric regression, piecewise smooth, stochastic design, threshold, variance.

ing wavelets to be employed with stochastic, or regular but nondyadic, design sequences, interpolation methods admit fast computation. This is achieved by approximating the interpolation-based estimator, defined in the continuum, using a dyadic grid. In the context of kernel methods an early interpolation approach was considered by Clark (1977, 1980). A technique for accommodating stochastic design, modelled on the Nadaraya–Watson kernel method, was considered by Hall and Patil (1996a).

We approach the analysis of performance rather differently from Donoho, Johnstone, Kerkyacharian and Picard in their work cited earlier, since we wish to stress the impact which different interpolation rules have on performance. Critically, choice of interpolation rule affects the size of the variance term through a constant factor, rather than by changing the rate of convergence. Interpolation has no first-order impact on bias. Such properties cannot be described accurately by deriving only upper bounds. The approach taken in the present paper is to produce concise asymptotic formulas, in effect upper and lower bounds that are asymptotically identical. We do this for a single, piecewise-smooth target function, although our results may be shown to hold simultaneously over a large class of such functions. Our focus on functions that are only piecewise smooth, rather than smooth everywhere, still allows us to demonstrate that wavelet methods achieve a degree of adaptivity not enjoyed by traditional techniques such as those based on kernels or splines, which (for example) perform poorly with functions that have jump discontinuities.

We should stress that the most distinctive features of different interpolation rules, which are present only in constant-factor changes to variance, do not emerge at all in a traditional minimax description of "pure thresholded" wavelet estimators. This is because pure thresholded estimators are oversmoothed by an order of magnitude [see, e.g., Hall and Patil (1996b)], with the result that the effect of variance is swamped by that of bias. Hence, the main conclusions reached in the present paper do not emerge from more familiar analyses.

2. Methodology.

2.1. *Model for data.* Let data $\mathcal{Y} \equiv \{(X_m, Y_m),\ 1 \le m \le n\}$ be generated by the model $Y_m = g(X_m) + \xi_m$ for $1 \le m \le n$, where the design sequence $\mathcal{X} \equiv \{X_m,\ 1 \le m \le n\}$ represents the *ordered* values of a random sample from a distribution with density f having support $\mathcal{I} = [0, 1]$, and the ξ_m's are independent but not necessarily identically distributed random variables. In our asymptotic model we should, strictly speaking, denote X_m and ξ_m by X_{nm} and ξ_{nm}, respectively, since the value of the mth among them will vary with increasing sample size.

2.2. *Interpolation rules.* Let $w_m, 1 \le m \le n$, denote weight functions depending on the X_m's but not on the Y_m's, and such that for integers ν_1, ν_2

satisfying $\nu_1 < 0 \le \nu_2$ we have $w_j \equiv 0$ unless $\nu_1 \le j \le \nu_2$. Define the interpolant

$$Y(x) = \sum_m w_m(x) Y_m \quad \text{for } x \in \left(X_{-\nu_1}, X_{n-\nu_2}\right]. \tag{2.1}$$

At the ends of the design interval, that is, outside the region $(X_{-\nu_1}, X_{n-\nu_2}]$ to which the above definition applies, Y may be defined in any of several ways. For definiteness we shall use horizontal extrapolation, meaning that $Y(t) \equiv Y(X_{-\nu_1})$ on $[0, X_{-\nu_1}]$, and $Y(t) \equiv Y(X_{n-\nu_2})$ on $(X_{n-\nu_2}, 1]$.

We shall consider two specific rules, respectively, local averaging and local linear interpolation. Assuming $x \in (X_l, X_{l+1}]$, in the first rule we define $w_m(x) = (2\nu)^{-1}$ if $-\nu + 1 \le m - l \le \nu$, and $w_m(x) = 0$ otherwise, and in the second,

$$w_m(x) = \begin{cases} \nu^{-1}(X_{2l-m+1} - x)/(X_{2l-m+1} - X_m), & \text{if } -\nu + 1 \le m - l \le 0, \\ \nu^{-1}(x - X_{2l-m+1})/(X_m - X_{2l-m+1}), & \text{if } 1 \le m - l \le \nu, \\ 0, & \text{otherwise.} \end{cases}$$

Substituting into (2.1), we obtain, respectively, for $x \in (X_l, X_{l+1}]$,

$$Y(x) = (2\nu)^{-1} \sum_{m=-\nu+1}^{\nu} Y_{l+m}, \tag{2.2}$$

$$Y(x) = \nu^{-1} \sum_{m=1}^{\nu} \left(\frac{x - X_{l-m+1}}{X_{l+m} - X_{l-m+1}} Y_{l+m} + \frac{X_{l+m} - x}{X_{l+m} - X_{l-m+1}} Y_{l-m+1} \right). \tag{2.3}$$

2.3. *Empirical wavelet transform for interpolated data.* Write ϕ and ψ for the "father" and "mother" wavelet functions, let $p = p(n)$ be the primary resolution level, define $p_i = 2^i p$ for $i \ge 0$ and let $\phi_j(x) = p^{1/2}\phi(px + j)$ and $\psi_{ij}(x) = p_i^{1/2}\psi(p_i x + j)$ be the functions that form the orthonormal basis of a wavelet expansion. Put $b_j = \int_{\mathscr{I}} g\phi_j$ and $b_{ij} = \int_{\mathscr{I}} g\psi_{ij}$. We assume that ψ is of order r, meaning that $r \ge 1$ is the smallest integer such that $\int x^i \psi(x)\, dx$ is nonzero. Our estimators of b_j and b_{ij} are, respectively, $\hat{b}_j = \int_{\mathscr{I}} Y\phi_j$ and $\hat{b}_{ij} = \int_{\mathscr{I}} Y\psi_{ij}$, and lead to the empirical wavelet transform,

$$\hat{g} = \sum_j \hat{b}_j \phi_j + \sum_{i=0}^{q-1} \sum_j \hat{b}_{ij} I\left(|\hat{b}_{ij}| \ge \delta\right) \psi_{ij}. \tag{2.4}$$

The empirical coefficients $\hat{b}_j$ and $\hat{b}_{ij}$ may be approximated to arbitrary accuracy on a dyadic grid. Indeed, taking $N = 2^k$ for an integer $k \ge 1$, we may define

$$\tilde{b}_j = N^{-1} \sum_{m=1}^{N} Y(m/N)\, \phi_j(m/N) \quad \text{and} \quad \tilde{b}_{ij} = N^{-1} \sum_{m=1}^{N} Y(m/N)\, \psi_{ij}(m/N),$$

which represent series approximations to $\hat{b}_j$ and $\hat{b}_{ij}$, respectively. Both are calculable using Mallat's pyramid algorithm. The resulting analogue of $\hat{g}$ is $\tilde{g}$, obtained by replacing each empirical coefficient $\hat{b}$ in (2.4) by $\tilde{b}$. In taking

this approach to estimation, we are effectively replacing the data set $\mathcal{Y}$ by $\mathcal{Y}' \equiv \{(m/N, Y(m/N)),\ 1 \le m \le N\}$, and applying a relatively standard wavelet estimator to $\mathcal{Y}'$. Provided that $N/n \to \infty$, the first-order asymptotics of $\tilde{g}$ are identical to those of $\hat{g}$.

2.4. *Main theoretical result.* Assume of g that it enjoys r piecewise-continuous derivatives, in the sense that there exist constants $0 = a_1 < a_2 < \cdots < a_k = 1$ such that g has r continuous derivatives on each interval (a_l, a_{l+1}) for $1 \le l \le k-1$, with left- and right-hand limits at a_l and a_{l+1}, respectively; of f that it is piecewise-continuous in this sense, possibly with a different k and different a_i's, and is bounded away from zero on $\mathcal{I} = [0,1]$; of ψ and ϕ that they are compactly supported and Hölder continuous. Then for some $r \ge 1$ and $\kappa \ne 0$, and all integers $i \in [0, r]$ and $j \in (-\infty, \infty)$,

$$\int \psi^2 = 1, \qquad \int x^i \psi(x)\, dx = \kappa (r!)^{-1} \delta_{ir},$$

$$\int \phi = 1, \qquad \int \phi(x)\, \phi(x + j)\, dx = \delta_{0j},$$

where δ_{jk} is the Kronecker delta; assume of the tuning parameters p and q in the definition of $\hat{g}$, that for some $u > 0$ and $\varepsilon > 0$,

$$p^{-1} = o\left\{\left(n^{-1} \log n\right)^{1/(2r+1)}\right\}, \qquad p_q^{-1} = o\left(n^{-2r/(2r+1)}\right), \tag{2.5}$$
$$p_q = O\left(n^{\min(u + 1/(2r+1),\, 1) - \varepsilon}\right);$$

and of errors $\xi_m = \xi_{nm}$ in the model of section 2.1 that they may be written as $\xi_{nm} = \sigma(X_{nm}) \xi'_{nm}$, where σ is a piecewise-continuous function on $\mathcal{I}$, $\xi'_{1m}, \ldots, \xi'_{nm}$ are stochastically independent of one another and of $X_1, \ldots, X_n$, and each ξ'_{nm} has, for $1 \le m \le n < \infty$, the distribution of ξ', with $E(\xi') = 0$, $E(\xi'^2) = 1$, $E|\xi'|^{2(1+u)} < \infty$, and u as in (2.5).

We refer to the conditions in the previous paragraphs collectively as (C). Condition (2.5) and the assumption $E|\xi'|^{2(1+u)} < \infty$ are satisfied if we take p equal to a constant multiple of $n^{1/(2r+1)}$, q equal to the integer part of $(1 - \varepsilon) \log_2 n$ for some $0 < \varepsilon < 2r/(2r+1)$, and $u = 2r/(2r+1)$; and if $E|\xi'|^4 < \infty$. It will follow from Theorem 2.1 that this size of p is optimal.

We shall assume that the interpolation rule is given by either (2.2) or (2.3), although any of several other approaches could be employed. In the case of the rule at (2.2) put $d_\nu \equiv 1 + (2\nu)^{-1}$, and for the rule at (2.3), let

$$d_\nu \equiv (2\nu)^{-2} E\left\{ \sum_{l=-\nu}^{-1} Z_l \left(2 \sum_{r=2l+1}^{l-1} Z_r + Z_l\right) \left(\sum_{r=2l+1}^{-1} Z_r\right)^{-1} \right.$$
$$\left. + \sum_{l=0}^{\nu-1} Z_l \left(2 \sum_{r=l+1}^{2l} Z_r + Z_l\right) \left(\sum_{r=0}^{2l} Z_r\right)^{-1} \right\}^2,$$

where $\{Z_r, -\infty < r < \infty\}$ are independent exponentially distributed random variables. For this definition, $d_1 = 3/2$ and $d_\nu = 1 + O(\nu^{-1})$ as $\nu \to \infty$.

Construct $\hat{g}$ using tuning parameters p, q satisfying (2.5), and employing the threshold $\delta = (Dn^{-1} \log n)^{1/2}$, where the constant D satisfies

$$D > 2u\, d_\nu \sup(\sigma^2/f) \tag{2.6}$$

and u is as in (2.5). Define $D_1 \equiv d_\nu \int \sigma^2 f^{-1}$ and $D_2 \equiv \kappa^2(1 - 2^{-2\nu})^{-1} \int (g^{(r)})^2$.

THEOREM 2.1. *Under conditions* (C),

$$\int E(\hat{g} - g)^2 = D_1 n^{-1} p + D_2 p^{-2r} + o(n^{-1}p + p^{-2r}). \tag{2.7}$$

2.5. *Discussion.*

REMARK 2.1 (Variance and bias). The first and second terms on the right-hand side of (2.7) represent, respectively, the variance and squared bias contributions to mean integrated squared error (MISE). If we replace p in (2.7) by h^{-1}, where h is a bandwidth, then (2.7) is transparently an analogue of the classical formula for mean squared error of a kernel or local polynomial estimator, albeit with different values of the constants D_1 and D_2. However, such formulas for kernel methods fail in the context of g's that are only piecewise smooth.

Of course, the fact that we can decompose MISE so neatly into variance and squared-bias components is a consequence of our decision to adapt the order of the wavelet to that of the target function in places where the latter is smooth. This has enabled us to give a detailed analysis of the effect that different interpolation rules have on performance. More traditional minimax analysis would not have permitted us to reach such concise conclusions.

REMARK 2.2 (Effects of ν). Larger values of ν produce slightly better mean square performance, since the value of d_ν, and hence D_1, decreases as ν increases. However, first-order properties of mean squared error do not capture all the qualitative features of the estimator, and in particular do not indicate the detrimental second-order effect that using too large a value of ν can have on bias. If interpolation is over a wide range, then a wavelet estimator applied to interpolated data will recover a slope rather than a jump. Our numerical work in Section 3 will address this point.

The variance inflation represented by the fact that $d_\nu > 1$ for finite ν is related to that discussed by Chu and Marron (1991) in the case of convolution-based kernel methods. If $\nu = \nu(n) \to \infty$ so slowly that $\nu = O(n^\varepsilon)$ for all $\varepsilon > 0$, then Theorem 2.1 remains true without any changes to the regularity conditions. We may then replace d_ν by $d_\infty = 1$ in (2.6) for the threshold constant.

REMARK 2.3 (Choice of p). We see from (2.7) that the optimal value of p, in the sense of minimizing the right-hand side, is asymptotic to $(2rnD_2/D_1)^{1/(2r+1)}$. The first part of condition (2.5) asks that p be of larger

order than $\theta_n = (n/\log n)^{1/(2r+1)}$, which is only marginally less than the order of the optimal p. However, it may be proved that (2.7) fails if p is of size θ_n or smaller; and that, under the conditions of Theorem 2.1 excluding the first part of (2.5), and assuming that $p \to \infty$,

$$\left(n^{-1} \log n\right)^{2r/(2r+1)} = O\left\{\int_{\mathscr{I}} E(\hat{g} - g)^2\right\}.$$

REMARK 2.4 (Locally varying thresholds). Theorem 2.1 remains valid, under identical conditions, if we use a varying threshold in which the indicator $I(|\hat{b}_{ij}| \geq \delta)$ in (2.4) is replaced by $I(|\hat{b}_{ij}| \geq \delta_{ij})$, where $\delta_{ij} = (D_{ij} n^{-1} \log n)^{1/2}$ and, in analogy to (2.6),

$$D_{ij} > 2u(1 + \varepsilon)\, d_\nu \sigma^2(-j/p_i)/f(-j/p_i)$$

for some $\varepsilon > 0$. (This definition is appropriate if σ^2 and f are continuous, but should be modified at jump discontinuities if those functions are discontinuous.)

REMARK 2.5 (Empirical choice of threshold). Theorem 2.1 is readily extended to the case of an empirical chosen threshold, as follows. For simplicity, assume that σ is constant on $\mathscr{I}$. Let $\hat{\gamma}$, $\hat{\sigma}$ denote estimators of $\gamma \equiv \inf f$, σ, respectively, let $A > 2ud_\nu$ be a constant, put $\hat{D} = A\hat{\sigma}^2/\hat{\gamma}$, and let $\tilde{g}_1$ denote the version of $\hat{g}$ in which the threshold is the random variable $\hat{\delta} = (\hat{D} n^{-1} \log n)^{1/2}$. Assume we know a constant B such that $|g| \leq B$, and define $\tilde{g} = \tilde{g}_1$ if $\int \tilde{g}_1^2 \leq B^2$, and $\tilde{g} = 0$ otherwise. Put $\nu = 2r/(2r + 1)$.

THEOREM 2.2. *If conditions* (C) *hold, with* σ *constant, and if for all* $\varepsilon > 0$ *and some* $C > 0$,

$$\begin{aligned} P\left(C^{-1} \leq \hat{\gamma} \leq \gamma + \varepsilon\right) &= 1 - o(n^{-\nu}), \\ P(\sigma - \varepsilon \leq \hat{\sigma} \leq C) &= 1 - o(n^{-\nu}), \end{aligned} \tag{2.8}$$

then (2.7) *holds with* $\tilde{g}$ *replacing* $\hat{g}$.

For example, let U equal the maximal spacing between adjacent X_i's, and let V equal the variance of the Y_i's (an overestimate of σ^2). Put $\hat{\gamma} = (\log n)/(nU)$ and $\hat{\sigma}^2 = V$. Then (2.8) holds if in addition to (C) we assume that f is continuous on $\mathscr{I}$; see Section 5. The resulting threshold has a particularly simple formula: $\hat{\delta} = (AUV)^{1/2}$.

Likewise, one may develop empirical, locally varying thresholds, and prove that they have properties similar to those of their nonempirical counterparts discussed in Remark 2.4.

REMARK 2.6 (Interpolating nonrandom designs). If the design variables X_i may be written as $X_i = F^{-1}(i/n)$, where F has a piecewise-continuous derivative f on $\mathscr{I}$, and f is bounded above zero on $\mathscr{I}$ and integrates to 1 on $\mathscr{I}$, then Theorem 2.1 holds if we replace d_ν by 1 throughout. In the case of

regularly-spaced design, $f \equiv 1$. There is now no advantage in using higher values of ν, with the accompanying second-order exacerbation of bias that they entail. Methods for this case have also been studied by Cai (1996).

This point bears restating, for emphasis: the cases of stochastic design and deterministic-but-irregular design are distinctly different. They have identical first-order bias properties, but the former produces estimators with greater variance. These points are analysed in more detail in a longer manuscript [Hall and Turlach (1995)], obtainable from the authors, on which this paper is based.

3. Numerical properties. We performed an extensive simulation study using the software package `wavethresh` [Nason and Silverman (1994)]. Three different targets were employed: the polynomial-with-jump function from Nason and Silverman (1994), and the functions "HeaviSine" and "Blocks" of Donoho and Johnstone (1994). Following Donoho and Johnstone's lead we chose the noise level so that the signal-to-noise ratio on a standard deviation scale was seven. The design points X_i were drawn from a Uniform or a truncated Normal distribution. We used sample sizes $n = 25$, 50, 100, 250, 500 and 1000, and varied ν between 1 and 9 in steps of 2. We employed the threshold $\hat{\delta} = (AUV)^{1/2}$ suggested in Remark 2.5, with $A = 3$. Computation was performed on a dyadic grid, as suggested in Section 2.3, with $N = 2^k$ and $k = \max(8, \lceil 1.2 \log_2 n \rceil)$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x, and $\lfloor x \rfloor$ the largest integer less than or equal to x. We used $p = 2^i$, $q = 1, \ldots, k - i$ $(i = 1, \ldots, \lfloor k/2 \rfloor - 1)$ and Daubechies' "extremal phase" wavelet of order six [Nason and Silverman (1994)].

For each setup we simulated 1000 realizations. In each case we calculated the wavelet decomposition, performed the thresholding and computed the estimator. We calculated integrated squared error, using the trapezoidal rule, and estimated mean integrated squared error (MISE) by averaging over these 1000 values.

We implemented both interpolation rules, (2.2) and (2.3). For small sample sizes, we obtained slightly better mean integrated squared performance by using (2.3), but usually, from $n = 250$ onward, performance of (2.2) and (2.3) was equivalent. In the following summary of our results we shall concentrate on (2.2) and X_i uniformly distributed. Results in other cases are available from the authors upon request.

Figure 1 shows, for different choices of ν and $n = 500$, five typical realizations of the wavelet estimator for the polynomial-with-jump target. The values used for p and q are those which minimized observed MISE. The positive (decreasing) influence on variance for larger values of ν is clear. At the same time the negative (increasing) influence on (finite sample) bias is apparent; smaller jumps are "smoothed away," while larger jumps turn into slopes. Similar features were observed in other cases.

For $n \geq 250$ the tendency of bias to increase with ν had an effect on MISE only in the case of the "Blocks" target. For the other two targets, the decrease in variance was sufficient to compensate for the increase in bias.

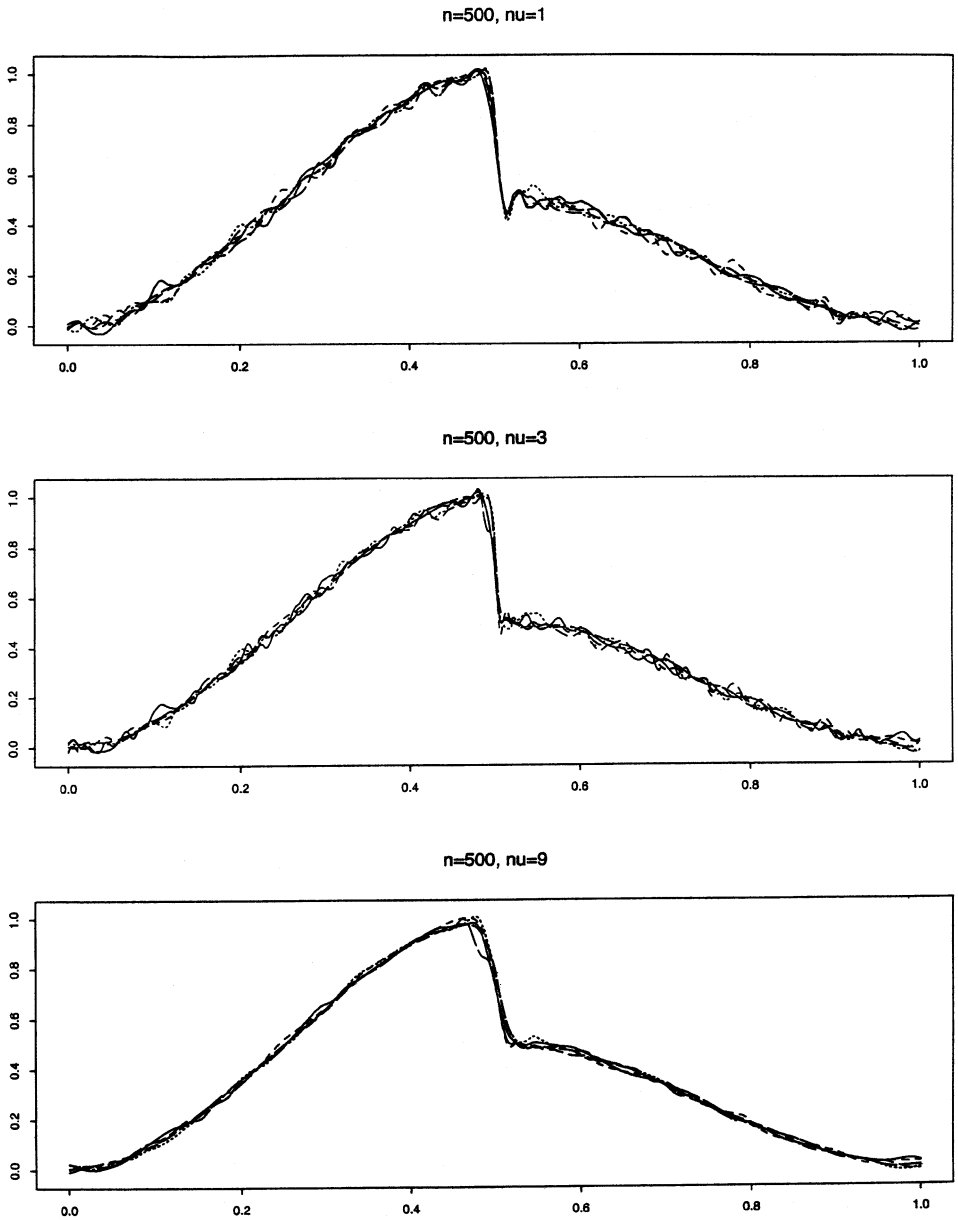

FIG. 1. *Five typical realizations of the wavelet estimator, defined at* (2.4), *for the polynomial-with-jump target. Each realization is based on* $n = 500$ *observations, interpolation rule* (2.2) *and* $\nu = 1, 3$ *or* 9.

4. Outline proof of Theorem 2.1. A fuller account of the argument is available in Hall and Turlach (1995), obtainable from the authors.

4.1. *Preliminaries*. For the sake of brevity we shall give the proof only in the case where f is uniformly continuous on $\mathscr{S}$ and the function σ^2 is a constant. An additional argument would allow us to overcome the inconvenience of jump discontinuities in f and a varying σ^2. Let $C, C_1, C_2, \ldots$ denote generic positive constants.

In view of the orthogonality properties of ϕ and ψ,

$$\int (\hat{g} - g)^2 = A_1 + A_2 + A_3 + A_4,$$

where

$$A_1 \equiv \sum_j \bigl(\hat{b}_j - b_j\bigr)^2, \qquad A_2 \equiv \sum_{i=0}^{q-1} \sum_j \bigl(\hat{b}_{ij} - b_{ij}\bigr)^2 I\bigl(|\hat{b}_{ij}| > \delta\bigr),$$

$$A_3 \equiv \sum_{i=0}^{q-1} \sum_j b_{ij}^2 I\bigl(|\hat{b}_{ij}| \le \delta\bigr), \qquad A_4 \equiv \sum_{i=q}^{\infty} \sum_j b_{ij}^2.$$

4.2. *Bounds for moderate deviations*. Let $v_1, \ldots, v_n$ denote weights, which we shall take here to be nonrandom, and suppose that for some $0 \le \varepsilon_1 < 1/20$ they satisfy

$$|v_n| \le C_1 n^{\varepsilon_1}, \qquad n^{-1} \sum_{m=1}^{n} v_m^2 \ge C_2 > 0. \tag{4.1}$$

Let $\xi_1, \ldots, \xi_n$ be independent and identically distributed random variables satisfying

$$E(\xi_1) = 0, \qquad E\bigl(\xi_1^2\bigr) = \sigma^2 > 0, \qquad E|\xi_1|^{C_3+2} \le C_4 \tag{4.2}$$

for some $C_3 > 4\varepsilon_1/(1 - 2\varepsilon_1)$. Define $S_n \equiv n^{-1/2} \sum_m v_m \xi_m$ and $\tau^2 \equiv n^{-1}\sigma^2 \sum_m v_m^2$.

LEMMA 4.1. *Assume* (4.1) *and* (4.2). *Then for each* $\varepsilon_2 > 0$ *there exist* $C_5, C_6 > 0$, *depending only on* $\varepsilon_1, \varepsilon_2$ *and* $C_1, \ldots, C_4$, *such that*

$$E\bigl\{S_n^2 I(S_n \ge z)\bigr\} \le C_5\bigl[\exp\bigl\{-(1 - \varepsilon_2) z^2/(2\tau^2)\bigr\} + n^{2\varepsilon_1 + \varepsilon_2 - C_3(1/2 - \varepsilon_1)}\bigr]$$

uniformly in $0 \le z \le n^{C_6}$ *for all* n.

PROOF. In this proof the constants $C_7, \ldots, C_{13}$ depend only on σ^2, ε_1 and $C_1, \ldots, C_4$. The argument is via two variants of Bernstein's and Bennett's inequalities, both of which are available in Hoeffding (1963). To state these results, let $Z_1, \ldots, Z_n$ denote independent random variables with zero means

and satisfying $|Z_m| \le b < \infty$ for each $1 \le m \le n$. Put $\tau'^2 \equiv n^{-1}\sum_m E(Z_m^2)$ and $T_n \equiv n^{-1/2}\sum_m Z_m$, and for $z > 0$ define $\eta = \eta(z) = bz/(n^{1/2}\tau'^2)$. Then

$$P(T_n > z) \le \exp\bigl(-\tfrac{1}{2}b^{-2}z^2\bigr) \quad \text{for all } z \ge 0, \tag{4.3}$$

$$P(T_n > z) \le \exp\bigl[-(n^{1/2}z/b)\{(1+\eta^{-1})\log(1+\eta) - 1\}\bigr] \quad \text{for all } 0 \le z \le b. \tag{4.4}$$

Fix $\delta \in ((1/10) - \varepsilon_1, (1/2) - \varepsilon_1)$, and put $\xi_m' \equiv \xi_m I(|\xi_m| \le n^\delta)$, $u_1 = u_1(n) \equiv E(\xi_1')$ and $Z_m \equiv v_m(\xi_m' - u_1)$. In this notation, the conditions imposed on Z_m in the previous paragraph hold with $b \equiv 2C_1 n^{\delta+\varepsilon_1}$, and so we may prove from (4.3) that for all $n \ge C_7$, say,

$$P(T_n > z) \le \exp(-C_8 z) \quad \text{for all } z > n^{2(\delta+\varepsilon_1)}, \tag{4.5}$$

$$P(T_n > z) \le \exp\bigl(-C_9 n^{\min(1/2-\delta-\varepsilon_1, C_3\delta)}\bigr) \quad \text{for all } n^{\min(1/2-\delta-\varepsilon_1, C_3\delta)/2} < z \le n^{2(\delta+\varepsilon_1)}; \tag{4.6}$$

and from (4.4) that for $n \ge C_7$,

$$P(T_n > z) \le \exp\bigl\{-\tfrac{1}{2}z^2\tau^{-2}\bigl(1 - C_{10}n^{\max(\delta+\varepsilon_1-1/2, -C_3\delta)}\bigr)\bigr\} \quad \text{for all } 0 \le z \le n^{\min(1/2-\delta-\varepsilon_1, C_3\delta)/2}. \tag{4.7}$$

Put $s \equiv n^{-1}\sum v_m$. Then for all $z > 0$,

$$\begin{aligned} E\{S_n^2 I(S_n \ge z)\} &\le E\bigl\{\bigl(T_n + n^{1/2}su_1\bigr)^2 I\bigl(T_n + n^{1/2}su_1 \ge z\bigr)\bigr\} \\ &\quad + (\sigma^{-2}\tau^2 + \tau^2 n^{1-2\delta})E\{\xi_1^2 I(|\xi_1| > n^\delta)\} \\ &\le 2E\{T_n^2 I(T_n \ge z - n^{1/2}su_1)\} + C_{11}n^{1-\delta(C_3+2)}, \end{aligned} \tag{4.8}$$

$$E\{T_n^2 I(T_n \ge z)\} = z^2 P(T_n \ge z) + 2\int_z^\infty yP(T_n > y)\,dy. \tag{4.9}$$

We may use (4.5)–(4.7) and (4.9), with $\delta < (1/2) - \varepsilon_1$ chosen sufficiently close to $(1/2) - \varepsilon_1$, to prove that for a constant C_{14} depending on σ^2, ε_1, ε_2 and $C_1, \dots, C_4$,

$$E\{T_n^2 I(T_n > z - n^{1/2}u_1)\} \le C_{14}\exp\{-\tfrac{1}{2}(1-\varepsilon_2)z^2/(2\tau^2)\}, \tag{4.10}$$

provided $0 \le z \le n^{\min(1/2-\delta-\varepsilon_1, C_3\delta)/2}$. Choosing C_5 sufficiently large to remove the condition $n \ge C_7$, we may deduce the lemma from (4.8) and (4.10).

4.3. *Approximations to empirical wavelet coefficients.* Observe that $\hat{b}_{ij} = b_{ij} + B_{ij} + \hat{\xi}_{ij}$, where $B_{ij} \equiv \int_{\mathscr{F}} \Delta\psi_{ij}$, $\hat{\xi}_{ij} \equiv n^{-1/2}S_{ij}$,

$$S_{ij} \equiv (p_i/n)^{1/2}\sum_m v_{ij;m}\xi_m, \qquad \tau_{ij}^2 \equiv p_i n^{-1}\sum_{m=\nu+1}^{n-\nu} v_{ij;m}^2,$$

$\Delta \equiv E(Y\,|\,\mathscr{X}) - g$ and $\nu_{ij;\,m} \equiv (n/p_i^{1/2}) \int w_m \psi_{ij}$. Properties of spacings of order statistics may be used to prove that

$$E\left(|B_{ij}|^k\right) = \begin{cases} O\left\{\left(p_i^{1/2}/n\right)^k n^\eta\right\}, & \text{uniformly in } j \in \mathscr{J}_i(\varepsilon), \\ O(n^{\eta-k}), & \text{uniformly in } j \notin \mathscr{J}_i(\varepsilon). \end{cases} \tag{4.11}$$

Lemma 4.1 may be applied to show that if $E|\xi_1|^{2(1+t)+\eta} < \infty$ for some $t, \eta > 0$, then for each $\varepsilon > 0$, and for each of the interpolation rules at (2.2) and (2.3),

$$\sup_{i,\,j} E\left(S_{ij}^2 I\left[|S_{ij}| > \left\{2t(1+\varepsilon)\, d_\nu \sigma^2 (\sup f^{-1}) \log n\right\}^{1/2}\right]\right) = O(n^{-t}). \tag{4.12}$$

[For each pair (i, j) the conditions of the lemma are readily checked, noting that the values of v_m, τ^2 and n there are replaced by $v_{ij;\,m}$, τ_{ij}^2 and a constant multiple of n/p_i, respectively. The latter is an upper bound to the number of nonzero terms in, for example, the series defining τ_{ij}^2.]

4.4. *Calculation of* $E(A_1)$. As in step 4.3, put $\Delta \equiv E(Y\,|\,\mathscr{X}) - g$. Define $v_{j;\,m} \equiv (n/p^{1/2}) \int w_m \phi_j$,

$$B_j \equiv \int_{\mathscr{I}} \Delta \phi_j, \qquad \hat{\xi}_j \equiv n^{-1/2} S_j, \qquad S_j \equiv (p_i/n)^{1/2} \sum_m v_{j;\,m}\, \xi_m.$$

Then $\hat{b}_j = b_j + B_j + \hat{\xi}_j$. Let $[-c, c]$ be a compact interval containing the support of ψ, and let $\mathscr{J}(\varepsilon)$ denote the set of indices j such that, for some x that is a point of discontinuity of g, $px + j \in (-c - pn^{\varepsilon-1}, c + pn^{\varepsilon-1})$. The analogue of (4.11) for B_j is, for all $\varepsilon, \eta > 0$,

$$E\left(|B_j|^k\right) = \begin{cases} O\left\{\left(p^{1/2}/n\right)^k n^\eta\right\}, & \text{uniformly in } j \in \mathscr{J}(\varepsilon), \\ O(n^{\eta-k}), & \text{uniformly in } j \notin \mathscr{J}(\varepsilon) \end{cases}$$

Hence, for all $\epsilon, \eta > 0$,

$$E\left(\hat{b}_j - b_j\right)^2 = \begin{cases} O\left\{E\left(\hat{\xi}_j^2\right) + pn^{\eta-2}\right\}, & \text{uniformly in } j \in \mathscr{J}(\varepsilon), \\ E\left(\hat{\xi}_j^2\right) + O\left\{n^{\eta-2} + E\left(\hat{\xi}_j^2\right)\right\}, & \text{uniformly in } j \notin \mathscr{J}(\varepsilon). \end{cases} \tag{4.13}$$

It may be proved that

$$\sup_{(1)} \left| E\left(\hat{\xi}_j^2\right) - n^{-1}\sigma^2\, d_\nu f(-j/p)^{-1} \right| = o(n^{-1}),$$

$$\limsup_{n\to\infty} \sup_{(2)} E\left(\hat{\xi}_j^2\right) \le n^{-1}\sigma^2 d_\nu \sup f^{-1},$$

where $\sup_{(1)}$ is taken over j such that $\mathscr{J}_j \equiv (-(c+j)/p, (c-j)/p) \subseteq \mathscr{I}$, and $\sup_{(2)}$ is taken over all j such that $\mathscr{J}_j \cap \mathscr{I}$ is nonempty. Combining the results from (4.13) down we conclude that for all $\eta > 0$,

$$E(A_1) = \sum_j E\left(\hat{b}_j - b_j\right)^2 \sim n^{-1} p \sigma^2 d_\nu \int f^{-1}. \tag{4.14}$$

4.5. *Bound for* $E(A_2)$. Let $\mathscr{K}_{i1}$ denote the set of indices j that are contained in an interval $(p_i x - 2c, p_i x + 2c)$ for at least one of the discontinuity points x of at least one of the functions $g^{(0)}, \dots, g^{(r)}$, and let $\mathscr{K}_{i2}$ be the set of all other j's. Write $A_2 = A_{21} + A_{22}$, where

$$A_{2k} \equiv \sum_{i=0}^{q-1} \sum_{j \in \mathscr{K}_{ik}} \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(|\hat{b}_{ij}| > \delta\right).$$

It may be proved by routine calculations from the formula $\hat{b}_{ij} = b_{ij} + B_{ij} + \hat{\xi}_{ij}$ (see step 4.3 for definitions of the terms) that for all $\eta > 0$,

$$E(A_{21}) = O\left\{q \sup_{0 \le i \le q-1,\, j \in \mathscr{K}_{1i}} E\left(\hat{b}_{ij} - b_{ij}\right)^2\right\} = O(qn^{\eta - 1});$$

and by applying (4.15) to bound $E[\hat{\xi}_{ij}^2 I\{|\hat{\xi}_{ij}| > (1 - \varepsilon)\delta\}]$ and $P\{|\hat{\xi}_{ij}| > (1 - \varepsilon)\delta\}$, and assuming that the threshold satisfies

$$\delta > \left\{2t(1 + \varepsilon')\, d_\nu \sigma^2 (\sup f^{-1}) n^{-1} \log n\right\}^{1/2} \tag{4.15}$$

for some $\varepsilon' > 0$, and that $E|\xi_1|^{2(1+t)+\varepsilon''} < \infty$ for some $\varepsilon'' > 0$; that

$$E(A_{22}) = O\left(\sum_{i=0}^{q-1} p_i n^{-(t+1)}\right) = O\left(p_q n^{-(t+1)}\right).$$

Combining these bounds we deduce that

$$E(A_2) = O\left(qn^{-1} + n^{\eta - 1} + p_q n^{-(t+1)}\right). \tag{4.16}$$

4.6. *Calculation of* $E(A_3)$. Write $E(A_3) = E(A_{31}) + E(A_{32})$, where

$$A_{3k} \equiv \sum_{i=0}^{q-1} \sum_{j \in \mathscr{K}_{ik}} b_{ij}^2 I\left(|\hat{b}_{ij}| \le \delta\right).$$

Since $b_{ij}^2 \le 2\{(\hat{b}_{ij} - b_{ij})^2 + \hat{b}_{ij}^2\}$, and since the number of elements of $\mathscr{K}_{i1}$ is uniformly bounded, then we have for all $\eta > 0$,

$$E(A_{31}) = O\left[\sum_{i=0}^{q-1} \left\{\sup_{j \in \mathscr{K}_{i1}} E\left(\hat{b}_{ij} - b_{ij}\right)^2 + \delta^2\right\}\right] = O(qn^{\eta - 1}).$$

Now, $b_{ij} = \kappa p_i^{-(2r+1)/2} g^{(r)}(-j/p_i) + o(p_i^{-(2r+1)/2})$, since $g^{(r)}$ is piecewise continuous. Therefore, $E(A_{32}) = \kappa^2 (1 - 2^{-2r})^{-1} p^{-2r} \int (g^{(r)})^2 + o(p^{-2r})$. Combining these results we deduce that

$$E(A_3) = \kappa^2 (1 - 2^{-2r})^{-1} p^{-2r} \int \left(g^{(r)}\right)^2 + o\left(p^{-2r}\right) + O(qn^{\eta - 1}). \tag{4.17}$$

4.7. *Bound for* $E(A_4)$. Divide the series into two portions, $E(A_4) = E(A_{41}) + E(A_{42})$, where

$$A_{4k} \equiv \sum_{i=q}^{\infty} \sum_{j \in \mathscr{K}_{ik}} b_{ij}^2.$$

Since $|b_{ij}| = O(p_i^{-1})$ uniformly in $j \in \mathscr{I}_{i1}$, and the number of such j's is uniformly bounded, then $A_{41} = O(\Sigma_{i \geq q} p_i^{-1}) = O(p_q^{-1})$. Furthermore, $|b_{ij}| = O(p_i^{-(2r+1)/2})$ uniformly in $j \in \mathscr{I}_{i2}$, and the number of such j's for which b_{ij} does not vanish equals $O(p_i)$. Hence, $A_{42} = O(\Sigma_{i \geq q} p_i^{-2r}) = O(p_q^{-2r})$. Combining these bounds we deduce that

$$A_4 = O\big(p_q^{-1}\big). \tag{4.18}$$

4.8. *Conclusion*. Combining (4.14) and (4.16)–(4.18), we deduce that for all $\eta > 0$,

$$\begin{aligned} \int E(\hat{g} - g)^2 &= D_1 n^{-1} p + D_2 p^{-2r} + o\big(n^{-1}p + p^{-2r}\big) \\ &\quad + O\big(p_q n^{-(t+1)} + q n^{\eta - 1} + p_q^{-1}\big). \end{aligned} \tag{4.19}$$

By taking $t = u - \zeta$, where $u > 0$ is as in condition (2.5) and $\zeta > 0$ is sufficiently small, we see from conditions (C) imposed in Theorem 2.1 that $E|\xi_1|^{2(1+t)+\eta} < \infty$ for some $\eta > 0$ (which condition is needed in steps 4.3 and 4.4 of the proof) and that (4.15) holds for some $\varepsilon' > 0$. Furthermore, for such a t it follows from (2.5) that the $O(\cdots)$ remainder term on the right-hand side of (4.19) equals $o(n^{-2r/(2r+1)})$, and so may be incorporated into the $o(n^{-1}p + p^{-2r})$ term. Result (2.7) is immediate.

5. Outline proof of Theorem 2.2. Let $\delta_D = (Dn^{-1} \log n)^{1/2}$ be the threshold used in Theorem 2.1; let $D^{(1)}, D^{(2)}$ be constants satisfying $2nd_\nu \sup(\sigma^2/f) < D^{(1)} < D^{(2)} < \infty$; and recall the expansion $\int(\hat{g} - g)^2 = A_1 + A_2(\delta) + A_3(\delta) + A_4$ derived in Section 4.1. Using the fact that $A_2(\delta)$, $A_3(\delta)$ are monotone in δ, and A_1, A_4 do not depend on δ, we may modify the proofs of (4.14) and (4.16)–(4.18) to show that for any constant $K > 0$, and any random variable $\tilde{D}$ taking values in $[D^{(1)}, D^{(2)}]$, we have $E\{A_1 I(A_1 \leq K)\} \sim E(A_1) \sim D_1 n^{-1}$, $E\{A_2(\delta_{\tilde{D}})\} + E(A_4) = o(n^{-1}p + p^{-2r})$ and $E[A_3(\delta_{\tilde{D}}) I\{A_3(\delta_{\tilde{D}}) \leq K\}] \sim E(A_3(\delta_{\tilde{D}})\} \sim D_2 p^{-2r}$. In view of (2.8) we may choose $D^{(1)}, D^{(2)}$ so that $P(\hat{D} \in [D^{(1)}, D^{(2)}]) = 1 - o(n^{-\nu})$. The theorem follows from these results, on defining $\tilde{D} = \hat{D}$ if $\hat{D} \in [D^{(1)}, D^{(2)}]$, and $\tilde{D} = D^{(1)}$ (say) otherwise.

To establish the validity of the first part of (2.8) for the case $\hat{\gamma} = (\log n)/(nU)$, it suffices to prove that for independent exponential random variables $Z_1, Z_2, \ldots$, all $\varepsilon > 0$ and some $C > 1$,

$$P\left\{1 - \varepsilon \leq \Big(\max_{1 \leq i \leq m} Z_i\Big) \Big/ \log m \leq C\right\} = 1 - O(m^{-1})$$

as $m \to \infty$. In fact, the left-hand side equals $1 - O(n^{1-C})$.

Acknowledgments. We are grateful to Iain Johnstone, Gerard Kerkyacharian, Steve Marron, Guy Nason and Dominique Picard for helpful discussion at various stages of this project. The `wavethresh` package used for the numerical work was obtained from StatLib (an electronic archive of statistical algorithms and data).

REFERENCES

CAI, T. (1996). Nonparametric function estimation via wavelets. Ph.D. dissertation, Cornell Univ.

CHU, C.-K. and MARRON, J. S. (1991). Choosing a kernel regression estimator. *Statist. Sci.* **6** 404–436.

CLARK, R. M. (1977). Nonparametric estimation of a smooth regression function. *J. Roy. Statist. Soc. Ser. B* **39** 107–113.

CLARK, R. M. (1980). Calibration, cross-validation and carbon-14. II. *J. Roy. Statist. Soc. Ser. A* **143** 177–194.

DONOHO, D. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. Unpublished manuscript.

DONOHO, D., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539.

DONOHO, D., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

GASSER, TH. and MÜLLER, H.-J. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, Heidelberg.

HALL, P. and PATIL, P. (1996a). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods. *J. Roy. Statist. Soc. Ser. B* **58** 361–377.

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

HALL, P. and TURLACH, B. A. (1995). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. Unpublished manuscript.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HASTIE, T. and LOADER, C. (1993). Local regression: automatic kernel carpentry. *Statist. Sci.* **8** 120–143.

HOEFFDING, W. (1963). On sequences of sums of independent random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by kernel and wavelet methods, optimality in Besov Spaces. *Statist. Probab. Lett.* **18** 327–336.

NASON, G. P. and SILVERMAN, B. W. (1994). The discrete wavelet transform. *J. Comput. Graph. Statist.* **3** 163–191.

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

CENTRE FOR MATHEMATICS AND ITS APPLICATION
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA
E-MAIL: halpstat@durras.anu.edu.au
berwin.turlach@anu.edu.au