The Annals of Statistics
1995, Vol. 23, No. 3, 905–928

FORMULAE FOR MEAN INTEGRATED SQUARED ERROR OF NONLINEAR WAVELET-BASED DENSITY ESTIMATORS

BY PETER HALL AND PRAKASH PATIL

Australian National University

We provide an asymptotic formula for the mean integrated squared error (MISE) of nonlinear wavelet-based density estimators. We show that, unlike the analogous situation for kernel density estimators, this MISE formula is relatively unaffected by assumptions of continuity. In particular, it is available for densities which are smooth in only a piecewise sense. Another difference is that in the wavelet case the classical MISE formula is valid only for sufficiently small values of the bandwidth. For larger bandwidths MISE assumes a very different form and hardly varies at all with changing bandwidth. This remarkable property guarantees a high level of robustness against oversmoothing, not encountered in the context of kernel methods. We also use the MISE formula to describe an asymptotically optimal empirical bandwidth selection rule.

1. Introduction. In this paper we investigate mean integrated squared error (MISE) properties of nonlinear, thresholded, wavelet-type density estimators applied to both continuous and discontinuous curves. We derive an analogue of the classical MISE formula familiar in the context of linear, kernel-type estimators, where MISE admits an expansion with distinct variance and squared bias components. The reader may recognize this in the form

$$\text{MISE} \sim C_1(nh)^{-1} + C_2 h^{2r}, \tag{1.1}$$

where n denotes sample size, h is the bandwidth of the kernel estimator, r is the order of the kernel and C_1 and C_2 are constants depending on both the kernel and the unknown density. The first term derives from variance, the next from squared bias [see, e.g., Rosenblatt (1971)]. In particular, C_2 is proportional to the integral of the square of the rth derivative of the density, and the MISE expansion for kernel estimators generally fails if f does not have r derivatives. We show that an analogue of (1.1) also holds in the case of nonlinear wavelet estimators, at least in the case where the order of h is sufficiently close to that which minimizes the right-hand side of (1.1). However, strikingly for people who are not familiar with the properties of wavelets, (1.1) is also valid for wavelet estimators when the underlying density is only piecewise continuous.

This result provides an explicit illustration of the extraordinary local adaptability of wavelet estimators: they do an excellent job of taking care of

Received July 1993; revised September 1994.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Bandwidth, density estimation, kernel, mean integrated squared error, multiresolution, robustness, wavelet.

discontinuities in the target function, and in consequence they enjoy very good convergence rates even if smoothness conditions are imposed only in a piecewise sense. While this property has been known for some time in the context of upper bounds for general function approximation [e.g., Mallat (1989) provides a relatively recent and sophisticated account of upper bounds to wavelet approximations in Sobolev spaces, which do not demand the existence of derivatives], the present paper gives the first demonstration that discontinuities do not affect even concise asymptotic MISE properties of nonlinear wavelet density estimators, not even to the extent of influencing the constants C_1 and C_2. By way of contrast, even in the commonly considered, elementary case of a second-order kernel estimator (where $r = 2$), formula (1.1) fails for kernel estimators when f is only piecewise continuous.

To be more explicit, if f has two bounded derivatives and $r = 2$, the optimal convergence rate of the right-hand side of (1.1) in the case of kernel estimators is achieved with a bandwidth h of size $n^{-1/5}$. The rate is $n^{-4/5}$. However, if f is only piecewise continuous and the assumption that f is twice differentiable is available only in a piecewise sense, then jump discontinuities may be shown to reduce the convergence rate to only $n^{-3/5}$. See van Eeden (1985) for a more detailed account of kernel estimation in the presence of discontinuities. By way of comparison, the convergence rate of a second-order wavelet estimator is preserved at $n^{-4/5}$, even in the presence of jump discontinuities of f, and the wavelet analogue of (1.1) is not affected.

Our results lead to other important conclusions about wavelet estimators. In particular, we show that wavelet density estimators are relatively robust against oversmoothing, that is, against using too large a bandwidth. The extent of robustness depends on choice of threshold, but for commonly used thresholds the robustness is considerable. In particular, in the case $r = 2$ considered earlier, MISE never rises above the order of $(n^{-1} \log n)^{4/5}$ through using too large a bandwidth. This is scarcely greater than the optimal order of $n^{-4/5}$. It is in stark contrast with the case of kernel estimators, where the rate of convergence of MISE can rise to $n^{-\varepsilon}$, for any $\varepsilon \in (0, \frac{4}{5})$, through oversmoothing. We show that the degree of robustness is greater for smaller thresholds.

One consequence of the result announced in the previous paragraph is that the analogue of (1.1) for wavelet estimators is not available for large bandwidths. That is to say, the traditional MISE formula is valid only for h sufficiently small. This result is unique in the context of nonparametric curve estimators. We quantify it by describing the behaviour of MISE for bandwidths that are relatively large.

We also describe how to modify our results in the more general case of estimation of density derivatives and show that our MISE formulae lead directly to asymptotically optimal bandwidth selection rules.

Wavelet methods have been introduced to statistics by Donoho (1992), Donoho and Johnstone (1992, 1994a, b) and Kerkyacharian and Picard (1992, 1993a–c). These authors have demonstrated the virtues of wavelet methods from the viewpoint of adaptive smoothing, typically in the context of the achievability of very good convergence rates uniformly over exceptionally large function classes. An excellent review of the work of these four authors is given in a recent discussion paper [Donoho, Johnstone, Kerkyacharian and Picard (1995)] wherein another seven papers describing their contributions to wavelet methods are cited. One of these, an unpublished manuscript on the subject of wavelet-based density estimation [Donoho, Johnstone, Kerkyacharian and Picard (1993)] became available to us after the first version of the present paper had been submitted. Like the other fine work of these four authors, it describes upper bounds uniformly over function classes. By way of contrast, our results discuss the performance of wavelet-based density estimators for a fixed density, rather than for a very large number of candidates for the density. By narrowing the focus in this way we are able to provide more detail than has previously been available about the MISE properties of nonlinear wavelet-based density estimators.

The estimators that we employ differ from those suggested by Donoho, Johnstone, Kerkyacharian and Picard [(1995) and other papers], in that they employ an explicit smoothing parameter, denoted by p in our work. As a consequence the truncation parameter, which we denote by q, is also different in our context. An advantage of introducing p is that it avoids logarithmic factors in convergence rates and achieves genuine first-order smoothing, with variance and squared bias balanced against one another. A disadvantage is that, in practice, p has to be chosen empirically.

It is straightforward to derive versions of our results uniformly over a large class of r-times piecewise-differentiable densities, achieving the mean square convergence rate $n^{-2r/(2r+1)}$ simultaneously over all elements of that class.

Our results are described in Section 2. Proofs are collected together in Section 3.

2. Main results.

2.1. *Summary.* We begin in Section 2.2 by describing elements of the basic theory of wavelet methods and introducing nonlinear wavelet-based density estimators. Section 2.3 discusses our main MISE formulae in the context of smooth densities. There we expand on the robustness properties outlined in Section 1, and we also present theory which describes the minimum allowable choice of threshold. The effect of lack of smoothness is described in Section 2.4, in a way which we feel provides valuable intuition into the reason why the fundamental MISE formula (1.1) is unaffected by discontinuities in the case of wavelet estimators. Sections 2.5 and 2.6 address the issues of derivative estimation and empirical bandwidth choice, respectively.

2.2. *Wavelets and wavelet-based density estimators.* Let ϕ, the "father wavelet," be a solution of the dilation equation,

$$\phi(x) = \sum_j c_j \phi(2x - j),$$

where the constants c_j satisfy $\sum j^{2r} c_j^2 < \infty$, $\sum c_j = 2$ and, for a maximal $r \geq 1$,

$$\sum_j (-1)^j j^k c_j = 0, \qquad 0 \leq k \leq r-1. \tag{2.1}$$

We normalize ϕ so that $\int \phi = 1$. Assume too that translates of ϕ are orthonormal, that is,

$$\int \phi(x)\phi(x+j)\, dx = \delta_{0j}, \qquad -\infty < j < \infty, \tag{2.2}$$

where $\delta_{j_1 j_2}$ is the Kronecker delta, and define

$$\psi(x) = \sum_j (-1)^j c_{j+1} \phi(2x+j),$$

the "mother wavelet." We suppose that

(2.3) ϕ and ψ are bounded and compactly supported.

The vast majority of the wavelets used in practice satisfy these conditions [see Daubechies (1992), Chapter 6].

Conditions (2.1) and (2.2) ensure that, for all integers $i \geq 0$ and $-\infty < j_1, j_2 < \infty$,

$$\int \psi(x+j_1)\psi(2^i x + j_2)\, dx = \delta_{0i}\delta_{j_1 j_2} \quad \text{and} \quad \int \phi(x+j_1)\psi(2^i x + j_2)\, dx \equiv 0,$$

and that $\int y^k \psi(y)\, dy = 0$ for $0 \leq k \leq r-1$. [See Strang (1989).] Therefore, the functions

$$\phi_j(x) = p^{1/2}\phi(px-j), \qquad \psi_{ij}(x) = p_i^{1/2}\psi(p_i x - j),$$

for arbitrary $p > 0$, $-\infty < j < \infty$, $i \geq 0$ and $p_i = p2^i$, are orthonormal: $\int \phi_{j_1}\phi_{j_2} = \delta_{j_1 j_2}$, $\int \psi_{i_1 j_1}\phi_{i_2 j_2} = \delta_{i_1 i_2}\delta_{j_1 j_2}$, $\int \phi_{j_1}\psi_{i j_2} = 0$. Furthermore, an arbitrary square-integrable function f may be expanded in a generalized Fourier series of the form

$$f(x) = \sum_j b_j \phi_j(x) + \sum_{i=0}^{\infty} \sum_j b_{ij}\psi_{ij}(x), \tag{2.4}$$

where $b_j = \int f\phi_j$ and $b_{ij} = \int f\psi_{ij}$. The generalized Fourier series (2.4) converges in L^2. The wavelet analogue of the bandwidth h of a kernel density estimator is p^{-1}.

If f is a probability density and $X_1, \ldots, X_n$ denote independent data values from that distribution, then $\hat{b}_j = n^{-1}\sum_{m=1}^n \phi_j(X_m)$ and $\hat{b}_{ij} = n^{-1}\sum_{m=1}^n \psi_{ij}(X_m)$ are unbiased estimators of b_j and b_{ij}, respectively. A nonlinear wavelet estimator of f has the form

$$\hat{f}(x) = \sum_j \hat{b}_j \phi_j(x) + \sum_{i=0}^{q-1} \sum_j \hat{b}_{ij} I\big(|\hat{b}_{ij}| > \delta\big)\psi_{ij}(x), \tag{2.5}$$

where $\delta > 0$ is a "threshold" and $q \geq 1$ is another smoothing parameter.

(Alternative ways of thresholding will be discussed in Remark 2.5.) The first series in (2.5) represents an unbiased estimator of the first series in (2.4), and converges absolutely under condition (2.3). However, if the second series were not thresholded and truncated in the manner suggested here, or in reasonably similar fashion, then that series would not converge. The more terms are induced in the series, the less is the bias but the greater is the variance. The parameters δ and q adjust this trade-off between bias and variance.

2.3. *Case of smooth f.* We assume initially that f is r times differentiable, and delay until Section 2.4 the case of a piecewise-smooth f. This approach simplifies both our exposition and proof.

Define $\kappa = (r!)^{-1} \int y^r \psi(y)\, dy = (r!)^{-1}(-\frac{1}{2})^{r+1} \Sigma(-1)^j j^r c_j$.

THEOREM 2.1. *Assume the conditions on ϕ and ψ stated in Section 2.2; assume that $f^{(r)}$ exists in a piecewise sense, and is bounded and piecewise continuous on $(-\infty, \infty)$, with finite and well-defined left- and right-hand limits, and monotone on $(-\infty, -u)$ and on (u, ∞) for sufficiently large u; and assume that $p \to \infty$, $q \to \infty$, $p_q \delta^2 \to 0$, $p_q^{2r+1}\delta^2 \to \infty$ and $\delta \ge C(n^{-1} \log n)^{1/2}$, where $C > C_0 \equiv 2\{r(\sup f)/(2r+1)\}^{1/2}$. Then the following hold*:

(i) *if* $p_q^{2r+1}\delta^2 \to \infty$,

$$E\left|\int \left(\hat{f} - f\right)^2 - \left\{n^{-1}p + p^{-2r}\kappa^2(1 - 2^{-2r})^{-1} \int f^{(r)2}\right\}\right| = o\left(n^{-1}p + p^{-2r}\right); \tag{2.6}$$

(ii) *if* $p = O(\delta^{-2/(2r+1)})$,

$$\delta^{4r/(2r+1)} = O\left\{\int E\left(\hat{f} - f\right)^2\right\}. \tag{2.7}$$

Note that the regularity conditions imposed on f imply that any power of $f^{(i)}$ is integrable over $(-\infty, \infty)$ and that $f^{(i)}$ is monotone in the extreme tails, for any $0 \le i \le r$. Observe too that since we are asking only that $f^{(r)}$ exist in a piecewise sense, we do not need f to be continuous.

REMARK 2.1 (Comparison with traditional MISE formulae). Result (2.6) is an unusually strong version of the traditional asymptotic formula for MISE. By taking the expected value on the left-hand side inside the modulus signs we obtain a wavelet version of the traditional MISE formula:

$$\int E\left(\hat{f} - f\right)^2 \sim n^{-1}p + p^{-2r}\kappa^2(1 - 2^{-2r})^{-1} \int f^{(r)2}, \tag{2.8}$$

where "$\sim$" means that the ratio of the left- and right-hand sides converges to 1 as $n \to \infty$. Here, the $n^{-1}p$ term derives from variance, and the p^{-2r} term from squared bias, exactly as in the case of classical formulae for kernel estimators. To obtain the wavelet formula from its counterpart for kernel methods, albeit with different constants multiplying the bias contribution, simply replace bandwidth h by p^{-1}. See, for example, Silverman [(1986), Chapter 3] for a detailed account of the kernel case.

Of course, the right-hand side of (2.8) is asymptotically minimized by taking $p \sim an^{1/(2r+1)}$, where $a = \{2r\kappa^2(1-2^{-2r})^{-1}\int f^{(r)2}\}^{1/(2r+1)}$; and the minimum size of (2.8) is const.$n^{-2r/(2r+1)}$.

REMARK 2.2 (Limitations to the applicability of the traditional MISE formula). Part (i) of Theorem 2.1 makes it clear that the traditional MISE formula (2.8) is valid for all sufficiently large p, but perhaps not for smaller values of p. Part (ii) of the theorem states that if

$$p = O(\delta^{-2/(2r+1)}) = O\{(n/\log n)^{1/(2r+1)}\}$$

then $\int E(\hat{f}-f)^2$ is at least of size $\delta^{4r/(2r+1)} \geq$ const.$(n^{-1}\log n)^{2r/(2r+1)}$. Since the latter quantity is still of larger order than $n^{-2r/(2r+1)}$ then, despite this apparent drawback to the validity of (2.8), the *overall* minimum of MISE does nevertheless occur at the value $p \sim an^{1/(2r+1)}$ introduced in Remark 2.1.

Remark 2.3 will demonstrate that the limitations suggested by part (ii) of Theorem 2.1 are real, not technical artifacts of our method of proof. The main practical implication of this result is the following. The wavelet estimator is relatively robust against oversmoothing, that is, against choosing p too small; and the smaller the threshold, the greater the level of robustness. For example, taking $\delta = n^{-1/2}(\log n)^{(1/2)+\varepsilon}$ for some $\varepsilon > 0$, produces a ceiling of $n^{-2r/(2r+1)}(\log n)^{\{(1/2)+\varepsilon\}4r/(2r+1)}$ on the size of MISE which can be obtained by accidental oversmoothing. This is smaller for smaller ε, that is, smaller δ. Remark 2.4 addresses the issue of the smallest value of δ for which our results are valid.

REMARK 2.3 [Sharpening of part (ii) of Theorem 2.1]. Here we demonstrate that, for a class of compactly supported densities, the order-of-magnitude result in (2.7) may be refined to a concise asymptotic relation.

PROPOSITION 2.1. *Assume the conditions of Theorem* 2.1 *and, in addition, that* supp $f = [c,d]$, *a compact interval. Suppose that* $f^{(r)}$ *restricted to* $[c,d]$ *is bounded away from zero in neighbourhoods of points of discontinuity, and has only a finite number of zeros, in neighbourhoods of which* $f^{(r+1)}$ *exists and is bounded away from zero; and that* $p^{2r+1}\delta^2 \to \ell$, *where* $0 \leq \ell < \infty$. *Then*

$$\int E(\hat{f}-f)^2 \sim C(\ell)\,\delta^{4r/(2r+1)},$$

where

$$C(\ell) = \begin{cases} \ell^{-2r/(2r+1)}\kappa^2 \int f^{(r)2}\left[\sum_{i=0}^{\infty} 2^{-2ri} I\left\{\left(\kappa f^{(r)}\right)^2 \le \ell 2^{(2r+1)i}\right\}\right], & \text{if } \ell > 0, \\ \left(1-2^{-2r}\right)^{-1}|\kappa|^{2/(2r+1)}\int |f^{(r)}|^{2/(2r+1)}, & \text{if } \ell = 0. \end{cases}$$

Proposition 2.1 may be proved under a very wide variety of different regularity conditions on f. The main difficulty in the proof arises from places where $f^{(r)}$ vanishes. In this respect, the conditions imposed above illustrate only one of many alternatives. In particular, versions of the proposition may be proved for infinitely supported densities, where $f^{(r)}$ might be assumed to be regularly varying at $\pm\infty$.

REMARK 2.4 (Lower bound to the constant C). In both theory and practice and parameters q and δ play interactive roles. Large values of q demand larger values of δ. Specifically, we claim that if $\delta = C(n^{-1}\log n)^{1/2}$, then, for each $\xi > 0$,

$$p_q \delta^2 n^{-(1/2)(1+\xi)C^2/(\sup f)} = O\left\{\int\left(E\hat{f} - f\right)^2\right\} + o\left(n^{-2r/(2r+1)}\right). \tag{2.9}$$

Therefore, if MISE is to be reducible to $O(n^{-2r/(2r+1)})$ by suitable choice of p, and if $p_q = \text{const.}n^{\beta}$, for some $\beta > 0$, then it is essential that $\beta - 1 - (1/2)(1+\xi)C^2(\sup f)^{-1} < -2r(2r+1)^{-1}$, for some $\xi > 0$, and hence that $\beta \ge 1/(2r+1)$ and

$$C \ge \left[2\left\{\beta - (2r+1)^{-1}\right\}(\sup f)\right]^{1/2}. \tag{2.10}$$

Note that, for larger values of β, the lower bound to C is larger.

In Theorem 2.1 we demanded that $p_q\delta^2 \to 0$, which [if $\delta = C(n^{-1}\log n)^{1/2}$] amounts to insisting only that $\beta < 1$. In this case the lower bound at (2.10) implies

$$C \ge 2\{r(\sup f)/(2r+1)\}^{1/2}.$$

Except for the possibility of equality here, this condition is identical to the restriction on C imposed in Theorem 2.1. Thus, that restriction is seen to be very close to the "best possible."

When $\delta = C(n^{-1}\log n)^{1/2}$, the assumption that $p_q\delta^2 \to 0$ is also close to providing a maximal upper bound on q for the validity of Theorem 2.1. To appreciate why, note that since ψ is compactly supported then the expected number of nonzero terms in the series defining $\hat{b}_{ij}$ is of size np_i^{-1}. Demanding that $p_q\delta^2 \to 0$ is therefore equivalent to asking that this expected number be of larger order than $\log n$, which is a very small order of magnitude.

We conclude this remark with an explicit statement of result (2.9).

PROPOSITION 2.2. *Assume the conditions on the wavelets and the density* f *imposed in Theorem* 2.1. *Suppose too that* $\sup f$ *occurs at a unique point* *which is a continuity point of f and that* $p, q \to \infty$ *and* $p_q \delta^2 \to 0$, *and take* $\delta = C(n^{-1} \log n)^{1/2}$, *where* $C > 0$. *Then* (2.9) *holds for each* $\xi > 0$.

REMARK 2.5 (Alternative forms of thresholding). The thresholding implicit in the definition of $\hat{f}$ at (2.5) is often called hard, in that the term $\hat{b}_{ij}\psi_{ij}$ is either included or completely excluded—there are no half measures. An alternative, more general approach is to replace the term $\hat{b}_{ij} I(|\hat{b}_{ij}| > \delta)\psi_{ij}$, appearing in (2.5), by $\hat{b}_{ij} w(\hat{b}_{ij}/\delta)\psi_{ij}$, where the weight function w satisfies $w(u) = 0$ for $0 < u < c_1$, $w(u) \in [0, 1]$ for $c_1 \le u \le c_2$, and $w(u) = 1$ if $u > c_2$, with $0 < c_1 < c_2 < \infty$ being constants. If the function w is continuous, then the estimator $\hat{f}$ is said to be based on "soft" thresholding. Theorem 2.1 is valid in this context, the only change being that, depending on w, the constant C may have to take a different range of values. The theorem certainly holds if the inequality $C \ge C_0$ is replaced by the requirement that C be sufficiently large.

2.4. *Effects of discontinuities.* In the previous subsection we described the performance of wavelet methods for densities with r derivatives. Clearly that smoothness assumption does have a bearing on our results, not least because the function $f^{(r)}$ appears in formulae for MISE. See, for example, (2.8). Nevertheless, the failure of the smoothness condition at a finite number of points does not affect Theorem 2.1, as our next result shows.

THEOREM 2.2. *Assume all the conditions of Theorem* 2.1, *except that we add the assumption that* $p_q^{2r+1} n^{-2r} \to \infty$, *and impose the condition of r-times differentiability of f only in a piecewise sense; that is, we ask that there exist points* $x_0 = -\infty < x_1 < \cdots < x_N < \infty = x_{N+1}$ *such that the first r derivatives of f exist and are bounded and continuous on* (x_i, x_{i+1}) *for* $0 \le i \le N$, *with left- and right-hand limits; and that* $f^{(r)}$ *is monotone on* $(-\infty, -u)$ *and on* (u, ∞) *for* $u > 0$ *sufficiently large. In particular, f itself may be only piecewise continuous. Then the result of Theorem* 2.1 *holds.*

REMARK 2.6 (Comparison with kernel estimators). This result is strikingly different from its nearest analogues for a variety of other estimator types, such as kernel estimators. There the presence of discontinuities can dramatically increase the order of magnitude of MISE. To illustrate this point, let $\tilde{f}$ be a kernel estimator of f based on an rth-order kernel [see, e.g., Rosenblatt (1971) and Silverman (1986), page 66]. Let h denote the bandwidth of the estimator. Suppose that, at some point x_0, f has just s derivatives and that $f^{(s+1)}$ exists on the left- and right-hand sides of x_0, with different limits at x_0. We assume that $0 \le s \le r - 1$. The usual Taylor expansion argument may be used to show that in a neighbourhood of x_0, with width approximately h, $\tilde{f}$ has bias of size h^{s+1}. The corresponding contribution to MISE is of size

$$(\text{width of neighbourhood}) \times (\text{size of bias})^2 = h(h^{s+1})^2 = h^{2s+3}.$$

The contribution from variance is of size $h(nh)^{-1} = n^{-1}$. Now, if bandwidth is chosen so that the estimator performs well across "most" of the real line, where $f^{(r)}$ is well defined, then h will be of size $n^{-1/(2r+1)}$ [Rosenblatt (1971)]. Therefore the contribution to MISE from jump discontinuities will be of size $n^{-(2s+3)/(2r+1)} + n^{-1}$. In particular, if $s = 0$, meaning that f' has a jump discontinuity at some point, then MISE will be of size at least $n^{-3/(2r+1)}$. Taking into account the contribution from intervals where $f^{(r)}$ is well defined, the actual MISE is of size $\max(n^{-2r/(2r+1)}, n^{-3/(2r+1)})$, which equals $n^{-3/(2r+1)}$ when $r \geq 2$. This convergence rate is inferior to the rate $n^{-2r/(2r+1)}$ provided by wavelet estimators (see Theorem 2.2). The reader is referred to van Eeden (1985) for detailed discussion of convergence rates of kernel estimators when there are discontinuities in f or its derivatives.

This analysis extends only as far as convergence rates and does not address the actual trace of estimators produced by wavelet or kernel methods. There again the wavelet estimator produces more satisfactory results, in that it more satisfyingly captures the actual jump discontinuities.

REMARK 2.7. Remarks 2.1–2.4 all have analogues applicable to Theorem 2.2.

REMARK 2.8 (Intuition behind Theorem 2.2). While Theorem 2.2 is not actually a corollary of Theorem 2.1, its derivation may be accomplished without great difficulty by inspecting portions of the proof of the latter result. Let us briefly outline the argument. It provides at once an explanation of the virtues of smoothing locally in the spatial domain, and a proof of Theorem 2.2.

Observe that, by the orthogonality properties of ϕ and ψ,

$$\int (\hat{f} - f)^2 = I_q(\mathbb{Z}, \mathbb{Z}, \dots),$$

where $\mathbb{Z}$ denotes the set of all integers and

$$I_q(\mathscr{S}, \mathscr{S}_0, \dots) = \sum_{j \in \mathscr{S}} \left(\hat{b}_j - b_j\right)^2 + \sum_{i=0}^{q-1} \sum_{j \in \mathscr{S}_i} \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(|\hat{b}_{ij}| > \delta\right)$$
$$+ \sum_{i=0}^{q-1} \sum_{j \in \mathscr{S}_i} b_{ij}^2 I\left(|\hat{b}_{ij}| \leq \delta\right) + \sum_{i=q}^{\infty} \sum_{j \in \mathscr{S}_i} b_{ij}^2.$$

Let $\mathscr{X}$ denote the finite set of points where $f^{(s)}$ has a point of discontinuity for some $0 \leq s \leq r$. If $\operatorname{supp} \psi \subseteq (-v, v)$, then, unless

$$j \in \mathscr{K}_i = \{k : k \in (p_i x - v, p_i x + v) \text{ for some } x \in \mathscr{X}\},$$

both b_{ij} and $\hat{b}_{ij}$ are constructed entirely from an integral over or an average of data values from an interval where $f^{(r)}$ exists and is bounded. Likewise, if $\operatorname{supp} \phi \subseteq (-v, v)$, then, unless

$$j \in \mathscr{K} = \{k : k \in (px - v, px + v) \text{ for some } x \in \mathscr{X}\},$$

b_j and $\hat{b}_j$ are also constructed solely from such regions. We may write

$$\int (\hat{f}-f)^2 = I_q(\mathscr{K},\mathscr{K}_1,\dots) + I_q(\tilde{\mathscr{K}},\tilde{\mathscr{K}}_1,\dots),$$

where $\tilde{\mathscr{S}}$ denotes the complement of $\mathscr{S}$ in $\mathbb{Z}$. The methods in the proof of Theorem 2.1 may be employed to prove that $I_q(\tilde{\mathscr{K}},\tilde{\mathscr{K}}_1,\dots)$ has precisely the asymptotic properties claimed for $\int(\hat{f}-f)^2$ in Theorem 2.1. Furthermore, noting that both $\mathscr{K}$ and $\mathscr{K}_i$ have no more than $(2v+1)(\#\mathscr{Z})$ elements, for each i, and that $q = O(\log n)$ and $p_q^{-1} = o(n^{-2r/(2r+1)})$, we may show that $E\{I_q(\mathscr{K},\mathscr{K}_1,\dots)\} = o(n^{-1}p + p^{-2r})$ in the context of part (i) of Theorem 2.1, or $= o(\delta^{4r/(2r+1)})$ in the context of part (ii). Combining these results we obtain Theorem 2.2. [The condition $p_q^{-1} = o(n^{-2r/(2r+1)})$ is used to establish the negligibility of the series $\sum_{i\geq q}\sum_j b_{ij}^2$; note that, at discontinuities, b_{ij}^2 is of size p_i^{-1}.]

2.5. *Estimators of derivatives.* If f, ϕ and ψ have $s \geq 0$ derivatives, then the generalized Fourier coefficients of $f^{(s)}$ are given by

$$b_j^{(s)} = \int f^{(s)}\phi_j = (-1)^s p^{s+(1/2)} \int f(x)\phi^{(s)}(px-j)\,dx,$$

$$b_{ij}^{(s)} = \int f^{(s)}\psi_{ij} = (-1)^s p_i^{s+(1/2)} \int f(x)\psi^{(s)}(p_i x-j)\,dx.$$

Therefore, unbiased estimators of these quantities are provided by

$$\hat{b}_j^{(s)} = (-1)^s p^{s+(1/2)} n^{-1} \sum_{m=1}^{n} \phi^{(s)}(pX_m - j),$$

$$\hat{b}_{ij}^{(s)} = (-1)^s p_i^{s+(1/2)} \sum_{m=1}^{n} \psi^{(s)}(p_i X_m - j).$$

The corresponding estimator of $f^{(s)}$ is

$$\hat{f}^{(s)} = \sum_j \hat{b}_j^{(s)}\phi_j + \sum_{i=0}^{q-1}\sum_j \hat{b}_{ij}^{(s)} I\left(|\hat{b}_{ij}^{(s)}| > p_i^s\delta\right)\psi_{ij}.$$

Versions of Theorems 2.1 and 2.2, and Propositions 2.1 and 2.2, may be established for $\hat{f}^{(s)}$. In particular, if ϕ and ψ satisfy the conditions of Theorem 2.1 and have s bounded derivatives; if f has $r+s$ bounded continuous derivatives in a piecewise sense, with well-defined left- and right-hand limits; if $f^{(r+s)}$ is monotone on $(-\infty,-u)$ and on (u,∞) for sufficiently large u; and if $p\to\infty$, $q\to\infty$, $p_q\delta^2\to 0$, $p^{2r+2s+1}\delta^2\to\infty$ and $\delta \geq C(n^{-1}\log n)^{1/2}$, where $C \geq 2\{r(\sup f)(\int\psi^{(s)2})/(2r+1)\}^{1/2}$; then

$$E\left|\int(\hat{f}^{(s)}-f^{(s)})^2 - \left\{n^{-1}p^{2s+1}\int\phi^{(s)^2} + p^{-2r}\kappa^2(1-2^{-2r})^{-1}\int f^{(r+s)^2}\right\}\right|$$
$$= o(n^{-1}p^{2s+1}+p^{-2r}).$$

This result represents an analogue of part (i) of Theorem 2.2.

2.6. *Empirical smoothing parameter choice.* The value of p may be selected empirically using a plug-in rule, as follows. Let $\hat{J}$ denote an estimator of $J = \int f^{(r)2}$, weakly consistent in the sense that $\hat{J} \to J$ in probability. Let a be as defined in Remark 2.1, and define $\hat{a} = \{2r\kappa^2(1 - 2^{-2r})^{-1}\hat{J}\}^{1/(2r+1)}$, being no more than the definition of a with $\hat{J}$ substituted for J. Write p_0 and $\hat{p}_0$ for $an^{1/(2r+1)}$ and $\hat{a}n^{1/(2r+1)}$, respectively, let $\hat{f}_p$ denote the estimator previously defined by $\hat{f}$ [see (2.5)], and set $I(p) = \int(\hat{f}_p - f)^2$. We claim that, assuming the conditions of Theorem 2.2 and that ϕ and ψ are both Hölder continuous,

$$I(\hat{p}_0) = I(p_0) + o_p(n^{-2r/(2r+1)}). \tag{2.11}$$

It then follows from (2.8), which is a consequence of Theorem 2.2 in precisely the same way that it was of Theorem 2.1, that

$$I(\hat{p}_0) = b_n^{-2r/(2r+1)} + o_p(n^{-2r/(2r+1)}),$$

where the constant $b > 0$ is defined by

$$\inf_{p>0}\left\{n^{-1}p + p^{-2r}\kappa^2(1 - 2^{-2r})^{-1}\int f^{(r)2}\right\} = bn^{-2r/(2r+1)}.$$

Therefore the empirically chosen bandwidth $\hat{p}_0$ asymptotically achieves the mean integrated squared error associated with p_0.

Since $\hat{J}$ is weakly consistent for J then if $\varepsilon_n \downarrow 0$ sufficiently slowly, $P(|\hat{J} - J| > \varepsilon_n) \to 0$. Therefore, to establish (2.11) it suffices to prove that, for each sequence ε_n converging to zero,

$$\sup_{p:|pp_0^{-1}-1|\le\varepsilon_n} |I(p) - I(p_0)| = o_p(n^{-2r/(2r+1)}). \tag{2.12}$$

To do this, let $A > 0$ be a very large positive constant, let $\mathscr{P}$ denote the set of points kn^{-A} for integer $k \ge 1$ and, given $p > 0$, let $\pi(p)$ be that value in $\mathscr{P}$ which minimizes $|p - \pi(p)|$, with any tie broken in an arbitrary manner. Then (2.12) will follow if we show that, for A sufficiently large,

$$\sup_{p:|pp_0^{-1}-1|\le\varepsilon_n} |I(p) - I\{\pi(p)\}| = o_p(n^{-2r/(2r+1)}), \tag{2.13}$$

$$\sup_{p\in\mathscr{P}:|pp_0^{-1}-1|\le\varepsilon_n} |I(p) - E\{I(p)\} = o_p(n^{-2r/(2r+1)}), \tag{2.14}$$

$$\sup_{p\in\mathscr{P}:|pp_0^{-1}-1|\le\varepsilon_n} \left|E\{I(p)\} - \left\{n^{-1}p + p^{-2r}\kappa^2(1 - 2^{-2r})^{-1}\int f^{(r)2}\right\}\right| = o_p(n^{-2r/(2r+1)}). \tag{2.15}$$

Result (2.13) may be proved using the Hölder continuity of ϕ and ψ, and approximating the estimator very closely using a soft thresholding rule. [In the context of deriving (2.13) it is actually simpler, but not essential, to treat soft thresholding.] Result (2.14) may be derived by showing, using large deviation arguments, that, for each η, $\lambda > 0$,

$$\sup_{p \in \mathscr{P}: |pp_0^{-1} - 1| \le \varepsilon_n} P\big[|I(p) - E\{I(p)\}| > \eta n^{-2r/(2r+1)}\big] = O(n^{-\lambda}).$$

Finally, result (2.15) follows from Theorem 2.2.

Cross-validation also produces an asymptotically optimal value of p. A more detailed account of theory for smoothing parameter choice is given by Hall and Patil (1993).

3. Proofs of Theorem 2.1 and Propositions 2.1 and 2.2.

PROOF OF THEOREM 2.1. The proof is broken into five parts. Symbols $C_1, C_2, \dots$ denote positive constants.

Step 1. *Bound for*

$$s_1 = \sum_{i=0}^{q-1} \sum_j E\Big\{\big(\hat{b}_{ij} - b_{ij}\big)^2 I\big(|\hat{b}_{ij}| > \delta\big)\Big\}.$$

Let α and β denote positive numbers satisfying $\alpha + \beta = 1$, and set

$$s_{11} = \sum_{i=0}^{q-1} \sum_j E\Big\{\big(\hat{b}_{ij} - b_{ij}\big)^2\Big\} I\big(|b_{ij}| > \alpha\delta\big),$$

$$s_{12} = \sum_{i=0}^{q-1} \sum_j E\Big\{\big(\hat{b}_{ij} - b_{ij}\big)^2 I\big(|\hat{b}_{ij} - b_{ij}| > \beta\delta\big)\Big\}.$$

Since $I(|\hat{b}_{ij}| > \delta) \le I(|b_{ij}| > \alpha\delta) + I(|\hat{b}_{ij} - b_{ij}| > \beta\delta)$, then

$$s_1 \le s_{11} + s_{12}. \tag{3.1}$$

We shall bound s_{11} and s_{12}, in turn.

Set $f_{ij} = \sup_{y \in \operatorname{supp} \psi} f\{(y + j)/p_i\}$. Since ψ is compactly supported and f is bounded and monotone in the extreme tails, then

$$\sup_{n; i \ge 0} p_i^{-1} \sum_j f_{ij} < \infty. \tag{3.2}$$

Now,

$$\begin{aligned} nE\big(\hat{b}_{ij} - b_{ij}\big)^2 &\le p_i E\big\{\psi(p_i X - j)^2\big\} \\ &= \int \psi(y)^2 f\bigg\{\frac{y + j}{p_i}\bigg\}\, dy \le f_{ij}, \end{aligned} \tag{3.3}$$

and, by Taylor expansion,

$$
\begin{aligned}
p_i^{1/2}|b_{ij}| = \bigg| \int \psi(y) \bigg[\sum_{\ell=0}^{r-1} (\ell!)^{-1} \bigg(\frac{y}{p_i} \bigg)^{\ell} f^{(\ell)}\bigg(\frac{j}{p_i} \bigg) \\
+ \{(r-1)!\}^{-1} \bigg(\frac{y}{p_i} \bigg)^{r} \int_0^1 (1-t)^{r-1} f^{(r)}\bigg\{ \frac{j+ty}{p_i} \bigg\} \, dt \bigg] \, dy \bigg| \\
\le C_1 p_i^{-r};
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
s_{11} &\le n^{-1} \sum_{i=0}^{q-1} \sum_j f_{ij} I\big(C_2 p_i^{-(r+(1/2))} > \delta\big) \\
&\le n^{-1} \bigg(\sup_{0 \le i \le q-1} p_i^{-1} \sum_j f_{ij} \bigg) \sum_{i=0}^{q-1} p_i I \bigg\{ p_i \le \bigg(\frac{C_2}{\delta} \bigg)^{2/(2r+1)} \bigg\} \\
&= O\big(n^{-1}\delta^{-2/(2r+1)}\big) = o\big(n^{-2r/(2r+1)}\big),
\end{aligned} \tag{3.5}
$$

since $n^{1/2}\delta \to \infty$. The second-to-last identity uses (3.2).

Let a and b denote positive numbers satisfying $a^{-1} + b^{-1} = 1$. By Rosenthal's inequality [Hall and Heyde (1980), page 23],

$$
\begin{aligned}
E|\hat{b}_{ij} - b_{ij}|^{2a} &\le C_3(a) \Big[\Big\{ E\big(\hat{b}_{ij} - b_{ij}\big)^2 \Big\}^a + n^{1-2a} E|p_i^{1/2}\psi(p_i X - j) - b_{ij}|^{2a} \Big] \\
&\le C_4(a, \psi, f)\big(n^{-a} + n^{1-2a} p_i^{a-1}\big).
\end{aligned}
$$

The summands of $n\hat{b}_{ij}$ are bounded by $C_5 p_q^{1/2}$, uniformly in i and j, and by (3.3) the sum of their variances does not exceed nf_{ij}. By hypothesis, $p_q \delta^2 \to 0$, and so in view of Bennett's or Bernstein's inequality [Pollard (1984), pages 192–193] we have, for any $0 < \varepsilon < 1$ and all sufficiently large n,

$$
P\big(|\hat{b}_{ij} - b_{ij}| > \beta\delta\big) \le 2 \exp\big\{ -\tfrac{1}{2}(1-\varepsilon)\beta^2 f_{ij}^{-1} n \delta^2 \big\}
$$

uniformly in $0 \le i \le q-1$ and j. Hence, by Hölder's inequality,

$$
\begin{aligned}
s_{12} &\le \sum_{i=0}^{q-1} \sum_j \big(E|\hat{b}_{ij} - b_{ij}|^{2a} \big)^{1/a} P\big(|\hat{b}_{ij} - b_{ij}| > \beta\delta\big)^{1/b} \\
&= O\bigg[\sum_{i=0}^{q-1} \big(n^{-1} + n^{(1/a)-2} p_i^{1-(1/a)}\big) \sum_j \exp\big\{ -\tfrac{1}{2}(1-\varepsilon)\beta^2 b^{-1} f_{ij}^{-1} n \delta^2 \big\} \bigg].
\end{aligned} \tag{3.6}
$$

Recall that, by assumption, $\delta \ge C(n^{-1} \log n)^{1/2}$, where $C > 2\{r(\sup f)/(2r + 1)\}^{1/2}$. Choose $\varepsilon \in (0, \frac{1}{3})$ so small, $\beta < 1$ so close to 1, and $a > 1$ so large (or, equivalently, $b > 1$ so close to 1) that $\frac{1}{2}(1 - 3\varepsilon)(\beta C)^2 b^{-1}(\sup f)^{-1} \ge d \equiv$

$2r(2r+1)^{-1}+a^{-1}$. Then $\frac{1}{2}(1-3\varepsilon)\beta^2 b^{-1}(\sup f)^{-1} n\delta^2 \ge d\log n$, and so, uniformly in $0 \le i \le q-1$,

$$\begin{aligned}\sum_j \exp\left\{-\frac{1}{2}(1-\varepsilon)\beta^2 b^{-1} f_{ij}^{-1} n\delta^2\right\} &= O\left[\sum_j f_{ij}(n\delta^2)^{-1}\exp\left\{-\tfrac{1}{2}(1-2\varepsilon)\beta^2 b^{-1} f_{ij}^{-1} n\delta^2\right\}\right] \\ &= O\left[\sum_j f_{ij}(n\delta^2)^{-1}\exp\left\{-\tfrac{1}{2}(1-3\varepsilon)\beta^2 b^{-1}(\sup f)^{-1} n\delta^2\right\}\right] \\ &= O\left\{\sum_j f_{ij}(n\delta^2)^{-1} n^{-d}\right\} = O\left(p_i n^{-d-1}\delta^{-2}\right),\end{aligned}$$

the last identity following from (3.2). Therefore, by (3.6), and since $p_q\delta^2 \to 0$ and $n^{1/2}\delta \to \infty$ imply $n^{-1}p_q \to 0$,

$$\begin{aligned}s_{12} &= O\left\{\sum_{i=0}^{q-1}\left(n^{-1}p_i + n^{(1/a)-2}p_i^2\right)n^{-d-1}\delta^{-2}\right\} \\ &= O\left\{\left(p_q + n^{(1/a)-1}p_q^2\right)n^{-d-2}\delta^{-2}\right\} \\ &= o\left(n^{(1/a)-d-1}\delta^{-2}\right) = o\left(n^{(1/a)-d}\right) = o\left(n^{-2r/(2r+1)}\right).\end{aligned} \tag{3.7}$$

Combining (3.1), (3.5) and (3.7), we deduce that

$$s_1 = o\left(n^{-2r/(2r+1)}\right). \tag{3.8}$$

Step 2. Bounds involving

$$s_2 = \sum_{i=0}^{q-1}\sum_j b_{ij}^2 I\left(|\hat{b}_{ij}| \le \delta\right).$$

Let $\varepsilon > 0$, and define

$$s_{21} = \sum_{i=0}^{q-1}\sum_j b_{ij}^2 I\left\{|b_{ij}| \le (1+\varepsilon)\delta\right\},$$

$$s_{22} = \sum_{i=0}^{q-1}\sum_j b_{ij}^2 I\left\{|b_{ij}| \le (1-\varepsilon)\delta\right\},$$

$$s_{23} = \sum_{i=0}^{q-1}\sum_{j\in\mathscr{J}} b_{ij}^2 I\left(|b_{ij}| \le \tfrac{1}{2}\delta\right),$$

$$\Delta_1 = \sum_{i=0}^{q-1}\sum_j b_{ij}^2 I\left(|\hat{b}_{ij} - b_{ij}| > \varepsilon\delta\right),$$

$$\Delta_2 = \sum_{i=0}^{q-1}\sum_{j\in\mathscr{J}} b_{ij}^2 I\left(|\hat{b}_{ij} - b_{ij}| > |b_{ij}|\right),$$

for any set $\mathscr{J}$ of indices. Since

$$I\left(|\hat{b}_{ij}| \le \delta\right) \le I\left\{|b_{ij}| \le (1+\varepsilon)\delta\right\} + I\left(|\hat{b}_{ij} - b_{ij}| > \varepsilon\delta\right)$$

and

$$I\left(|b_{ij}| \le (1-\varepsilon)\delta\right) \le I\left(|\hat{b}_{ij}| \le \delta\right) + I\left(|\hat{b}_{ij} - b_{ij}| > \varepsilon\delta\right),$$

then

$$s_{22} - \Delta_1 \le s_2 \le s_{21} + \Delta_1; \tag{3.9}$$

and since

$$I\left(|b_{ij}| \le \tfrac{1}{2}\delta\right) \le I\left(|\hat{b}_{ij}| \le \delta\right) + I\left(|\hat{b}_{ij} - b_{ij}| > |b_{ij}|\right),$$

then

$$s_{23} - \Delta_2 \le s_2. \tag{3.10}$$

Set $g_{ij} = f^{(r)}(j/p_i)$. In view of the identity at (3.4),

$$b_{ij} = \kappa p_i^{-(r+(1/2))}(g_{ij} + \eta_{ij}),$$

where

$$\sup_{0 \le i \le q-1; j} |\eta_{ij}| \to 0.$$

We shall assume that

$$p^{2r+1}\delta^2 \to \ell, \tag{3.11}$$

where $0 \le \ell \le \infty$. (The case where such convergence is only along a subsequence may be treated similarly.) Suppose first that $\ell < \infty$. Let $C_6, C_7 > 0$ be such that the set $\mathscr{J}'$, of integers j with $|g_{ij}| > 2|\kappa|^{-1}C_6$, has at least $2C_7 p_i$ elements for all $i \ge 0$ and all sufficiently large n. Then a certain subset $\mathscr{J}$ of $\mathscr{J}'$ consists entirely of integers j such that $|b_{ij}| > C_6 p_i^{-(r+(1/2))}$, and has between $C_7 p_i$ and $2C_7 p_i$ elements for all $i \ge 0$ and all large n. We shall use this $\mathscr{J}$ in the definition of s_{23} and Δ_2. Note too that, for some $C_8 > 0$ and all i and j, $|b_{ij}| \le \frac{1}{2}C_8^{1/2} p_i^{-(r+(1/2))}$. Therefore,

$$\begin{aligned} s_{23} &\ge \sum_{i=0}^{q-1} \sum_{j \in \mathscr{J}} C_6^2 p_i^{-(2r+1)} I\left(C_8 p_i^{-(2r+1)} \le \delta^2\right) \\ &\ge C_6^2 C_7 \sum_{i=0}^{q-1} p_i^{-2r} I\left\{p_i \ge \left(C_8/\delta^2\right)^{1/(2r+1)}\right\} \\ &\ge C_9 \delta^{4r/(2r+1)}, \end{aligned} \tag{3.12}$$

the latter inequality requiring our hypothesis that $p = O(\delta^{-2/(2r+1)})$. Furthermore, by Bernstein's or Bennett's inequality,

$$
\begin{aligned}
E(\Delta_2) &= \sum_{i=0}^{q-1} \sum_{j\in\mathscr{J}} b_{ij}^2 P\left(|\hat{b}_{ij} - b_{ij}| > |b_{ij}|\right) \\
&\le 2 \sum_{i=0}^{q-1} \sum_{j\in\mathscr{J}} b_{ij}^2 \exp\left\{-C_{10} n \min\left(b_{ij}^2, |b_{ij}| p_i^{-1/2}\right)\right\} \\
&= O\left[\sum_{i=0}^{q-1} p_i^{-2r} \exp\{-C_{11} n \min(p_i^{-(2r+1)}, p_i^{-(r+1)})\}\right] \\
&= O\left\{p^{-2r} \sum_{i=0}^{\infty} 2^{-2ri} \exp\left(-C_{11} n p^{-(r+1)} 2^{-(r+1)i}\right)\right\} \\
&= O\left\{p^{-2r} \int_1^{\infty} x^{-2r-1} \exp\left(-C_{12} n p^{-(r+1)} x^{-(r+1)}\right) dx\right\} \\
&= O\left\{n^{-2r/(r+1)} \int_0^{\infty} x^{(r-1)/(r+1)} \exp(-C_{12} x)\, dx\right\} \\
&= O(n^{-2r/(r+1)}).
\end{aligned} \tag{3.13}
$$

Combining (3.10), (3.12) and (3.13) and noting that by hypothesis $\delta \ge C(n^{-1} \log n)^{1/2}$, we deduce that, for sufficiently large n,

$$E(s_2) \ge s_{23} - E(\Delta_2) \ge \tfrac{1}{2} C_9 \delta^{4r/(2r+1)}. \tag{3.14}$$

Suppose next that, in (3.11), $\ell = \infty$. Then, using (3.4),

$$\sup_j |b_{ij}| \le C_1 p_i^{-(r+(1/2))} \le C_1 p^{-(r+(1/2))} \ll \delta,$$

whence it follows that, for all sufficiently large n,

$$s_{21} = s_{22} = \sum_{i=0}^{q-1} \sum_j b_{ij}^2 = t(\infty),$$

where

$$t(u) = \sum_{i=0}^{q-1} \sum_{|j|\le up_i} \kappa^2 p_i^{-(2r+1)} (g_{ij} + \eta_{ij})^2.$$

For any finite $u > 0$,

$$
\begin{aligned}
t(u) &= \sum_{i=0}^{q-1} \sum_{|j|\le up_i} \kappa^2 p_i^{-(2r+1)} g_{ij}^2 + o\left(\sum_{i=0}^{q-1} \sum_{|j|\le up_i} p_i^{-(2r+1)}\right) \\
&= \left(\sum_{i=0}^{q-1} \kappa^2 p_i^{-2r}\right) \int_{-u}^{u} f^{(r)^2} + o\left(\sum_{i=0}^{\infty} p_i^{-2r}\right) \\
&= \kappa^2 (1 - 2^{-2r})^{-1} p^{-2r} \int_{-u}^{u} f^{(r)^2} + o(p^{-2r}),
\end{aligned}
$$

and since $f^{(r)}$ is monotone in the extreme tails then, by a similar argument, for sufficiently large $u > 0$,

$$t(\infty) - t(u) = \kappa^2(1 - 2^{-2r})^{-1} p^{-2r}\left(\int_{-\infty}^{-u} + \int_u^{\infty}\right) f^{(r)2} + o(p^{-2r}).$$

Therefore,

$$s_{21} = s_{22} \sim \kappa^2(1 - 2^{-2r})^{-1} p^{-2r} \int f^{(r)2}. \tag{3.15}$$

By Bernstein's or Bennett's inequality and for all sufficiently large n,

$$\begin{aligned} E(\Delta_1) &= \sum_{i=0}^{q-1} \sum_j b_{ij}^2 P\left(|\hat{b}_{ij} - b_{ij}| > \varepsilon\delta\right) \\ &\le 2 \sum_{i=0}^{q-1} \sum_j b_{ij}^2 \exp\left(-\tfrac{1}{3}\varepsilon^2 f_{ij}^{-1} n\delta^2\right) \\ &= o\left(\sum_{i=0}^{q-1} \sum_j b_{ij}^2\right) = o(s_{21}), \end{aligned} \tag{3.16}$$

the second-to-last identity following since $n\delta^2 \to \infty$. By (3.9), (3.15) and (3.16),

$$E\left| s_2 - \kappa^2(1 - 2^{-2r})^{-1} p^{-2r} \int f^{(r)2} \right| = o(p^{-2r}), \tag{3.17}$$

as $n \to \infty$.

Step 3. *Bound for*

$$s_3 = \sum_{i=q}^{\infty} \sum_j b_{ij}^2.$$

Observe that, for large n,

$$\begin{aligned} s_3 &= \sum_{i=q}^{\infty} \sum_j \kappa^2 p_i^{-(2r+1)} (g_{ij} + \eta_{ij})^2 \\ &\le 2\kappa^2 \sum_{i=q}^{\infty} p_i^{-(2r+1)} \sum_j g_{ij}^2 \\ &= O\left(\sum_{i=q}^{\infty} p_i^{-2r}\right) \\ &= O\left(p_q^{-2r}\right) = o\{\min(p^{-2r}, \delta^{4r/(2r+1)})\}, \end{aligned} \tag{3.18}$$

using the fact that $p_q^{2r+1}\delta^2 \to \infty$ and $q \to \infty$.

Step 4. *Bound for*

$$s_4 = E\left| \sum_j \left(\hat{b}_j - b_j\right)^2 - n^{-1}p \right|.$$

Observe that

$$nE\left(\hat{b}_j - b_j\right)^2 = \int \phi(y)^2 f\{(y+j)/p\}\,dy - b_j^2,$$

so that, since $\int \phi^2 = 1$ and the extreme tails of f are monotone,

$$\sum_j E\left(\hat{b}_j - b_j\right)^2 = n^{-1}p + o(n^{-1}p). \tag{3.19}$$

Set $Y_{jk} = p^{1/2}\phi(pX_k - j) - b_j$. Then

$$n^2 \sum_j \left(\hat{b}_j - b_j\right)^2 = \sum_{k=1}^{n} \sum_j Y_{jk}^2 + \sum\sum_{\substack{1 \le k_1, k_2 \le n \\ k_1 \ne k_2}} \sum_j Y_{jk_1} Y_{jk_2}. \tag{3.20}$$

Now,

$$n^{-1}E\left\{\sum_{k=1}^{n} \sum_j \left(Y_{jk}^2 - EY_{jk}^2\right)\right\}^2 = E\left\{\sum_j \left(Y_{j1}^2 - EY_{j1}^2\right)\right\}^2$$

and

$$\begin{aligned}\tfrac{1}{2}\sum_j Y_{j1}^2 &\le p\sum_j \phi(pX_1 - j)^2 + \sum b_j^2 \\ &\le p(\sup \phi^2)(\operatorname{supp} \phi + 2) + \int f^2.\end{aligned}$$

Therefore,

$$\operatorname{var}\left(n^{-2}\sum_{k=1}^{n} \sum_j Y_{jk}^2\right) = O(n^{-3}p^2) = o(n^{-2}p^2). \tag{3.21}$$

Furthermore, if $\operatorname{supp}\phi \subseteq (-v, v)$, then

$$\begin{aligned}
&E\left(\sum\sum_{\substack{1 \le k_1, k_2 \le n \\ k_1 \ne k_2}} \sum_j Y_{jk_1} Y_{jk_2}\right)^2 \\
&\quad = \sum_{j_1} \sum_{j_2} \sum\sum_{\substack{1 \le k_{11}, k_{12} \le n \\ k_{11} \ne k_{12}}} \sum\sum_{\substack{1 \le k_{21}, k_{22} \le n \\ k_{21} \ne k_{22}}} E\left(Y_{j_1 k_{11}} Y_{j_2 k_{12}} Y_{j_2 k_{21}} Y_{j_2 k_{22}}\right) \\
&\quad = O\left\{\sum_{j_1} \sum_{j_2} \sum\sum_{\substack{1 \le k_1, k_2 \le n \\ k_1 \ne k_2}} E\left(Y_{j_1, k_1} Y_{j_1, k_2} Y_{j_2 k_1} Y_{j_2 k_2}\right)\right\} \\
&\quad = O\Big(n^2 \sum_{j_1} \sum_{j_2} \Big| E\Big[\{p^{1/2}\phi(pX_1 - j_1) - b_{j1}\}\{p^{1/2}\phi(pX_2 - j_1) - b_{j1}\} \\
&\qquad\qquad \times \{p^{1/2}\phi(pX_1 - j_2) - b_{j2}\} \\
&\qquad\qquad \times \{p^{1/2}\phi(pX_2 - j_2) - b_{j2}\}\Big]\Big|\Big)
\end{aligned} \tag{3.22}$$

$$= O\left[n^2 \sum_{j_1} \sum_{j_2:|j_1-j_2|\le 2v} \{pE|\phi(pX_1 - j_1)\phi(pX_2 - j_2)| + |b_{j1}b_{j2}|\}^2\right]$$

$$= O\left\{n^2\left(\sum_j \left[\int_{|y|\le v} f\{(y+j)/p\}\, dy\right]^2 + \sum_j b_j^2\right)\right\}$$

$$= O(n^2 p).$$

Combining (3.20)–(3.22), we deduce that

$$\operatorname{var}\left\{\sum_j \left(\hat{b}_j - b_j\right)^2\right\} = o(n^{-2}p^2),$$

which together with (3.19) implies that

$$s_4 = o(n^{-1}p). \tag{3.23}$$

Step 5. *Conclusion.* Observe that

$$\begin{aligned}\int (\hat{f} - f)^2 = \sum_j \left(\hat{b}_j - b_j\right)^2 + \sum_{i=0}^{q-1} \sum_j \left(\hat{b}_{ij} - b_{ij}\right)^2 I\left(|\hat{b}_{ij}| > \delta\right) \\ + \sum_{i=0}^{q-1} \sum_j b_{ij}^2 I\left(|\hat{b}_{ij}| \le \delta\right) + \sum_{i=q}^{\infty} \sum_j b_{ij}^2.\end{aligned} \tag{3.24}$$

Combining this formula with (3.8), (3.17), (3.18) and (3.23), we deduce that, provided $p^{2r+1}\delta^2 \to \infty$,

$$E\left|\int (\hat{f} - f)^2 - \left\{n^{-1}p + \kappa^2(1 - 2^{-2r})^{-1} p^{-2r} \int f^{(r)2}\right\}\right| = o(n^{-1}p + p^{-2r}).$$

If $p^{2r+1}\delta^2$ is bounded, then, by the same sequence of results but with (3.14) replacing (3.17), we obtain

$$\int E(\hat{f} - f)^2 \ge n^{-1}p + \tfrac{1}{2}C_9 \delta^{4r/(2r+1)} + o(n^{-2r/(2r+1)} + \delta^{4r/(2r+1)}).$$

When $p^{2r+1}\delta^2$ is bounded,

$$p = O(\delta^{-2/(2r+1)}) = O\left\{(n/\log n)^{1/(2r+1)}\right\},$$

so that $n^{-1}p = o(n^{-2r/(2r+1)})$ and

$$\int E(\hat{f} - f)^2 \ge \tfrac{1}{2}C_9 \delta^{4r/(2r+1)} + o(n^{-2r/(2r+1)} + \delta^{4r/(2r+1)}).$$

It follows that

$$\inf_{p:\, p^{2r+1}\delta^2 \le C_{13}} \int E(\hat{f} - f)^2 \ge C_{14}\, \delta^{4r/(2r+1)},$$

for all $C_{13} > 0$. □

PROOF OF PROPOSITION 2.1. We refer to parts of the previous proof for estimates of various quantities. In particular, by (3.8), (3.16), (3.23) and (3.24),

$$\int E(\hat{f}-f)^2 = n^{-1}p + s_5 + o(n^{-1}p + \delta^{4r/(2r+1)}),$$

where

$$s_5 = \sum_{i=0}^{q-1} \sum_j b_{ij}^2 P(|\hat{b}_{ij}| \le \delta).$$

[Note that $n^{-2r/(2r+1)} = o(\delta^{4r/(2r+1)})$.] Since $p = o(n^{1/(2r+1)})$, then

$$\int E(\hat{f}-f)^2 = s_5 + o(\delta^{4r/(2r+1)}). \tag{3.25}$$

Let $\varepsilon \in (0,1)$ and set

$$s_{51} = \sum_{i=0}^{q-1} \sum_j b_{ij}^2 I\{|b_{ij}| \le (1+\varepsilon)^{-1}\delta\},$$

$$s_{52} = \sum_{i=0}^{q-1} \sum_j b_{ij}^2 I\{|b_{ij}| \le (1-\varepsilon)^{-1}\delta\},$$

$$\Delta_3 = \sum \sum b_{ij}^2 I(|\hat{b}_{ij} - b_{ij}| > \varepsilon |b_{ij}|).$$

Since

$$I\{|b_{ij}| \le (1+\varepsilon)^{-1}\delta\} \le I(|\hat{b}_{ij}| \le \delta) + I(|\hat{b}_{ij} - b_{ij}| > \varepsilon|b_{ij}|)$$

and

$$I(|\hat{b}_{ij}| \le \delta) \le I\{|b_{ij}| \le (1-\varepsilon)^{-1}\delta\} + I(|\hat{b}_{ij} - b_{ij}| > \varepsilon|b_{ij}|),$$

then

$$s_{51} - E(\Delta_3) \le s_5 \le s_{52} + E(\Delta_3). \tag{3.26}$$

Arguments similar to those leading to (3.13) may be used to prove that

$$E(\Delta_3) = o(\delta^{4r/(2r+1)}). \tag{3.27}$$

Set $\alpha = (1 \pm \varepsilon)^{-1}$ and let $\beta = |\kappa|\alpha^{-1}$. Tedious but straightforward calculations may be used to prove that, under the conditions of Proposition 2.1,

$$\begin{aligned}
\sum_{i=0}^{q-1} \sum_j b_{ij}^2 I(|b_{ij}| \le \alpha\delta) &\sim \sum_{i=0}^{q-1} \sum_j \{\kappa f^{(r)}(j/p_i) p_i^{-(r+(1/2))}\}^2 \\
&\qquad \times I\{|\kappa f^{(r)}(j/p_i) p_i^{-(r+(1/2))}| \le \alpha\delta\} \\
&\sim \kappa^2 \sum_{i=0}^{q-1} p_i^{-2r} \int f^{(r)}(x)^2 I\{\beta\delta^{-1}|f^{(r)}(x)| \le p_i^{(r+(1/2))}\}\, dx \\
&= s_{53},
\end{aligned}$$

say. If $p^{2r+1}\delta^2 \to \ell > 0$, then

$$s_{53} = \kappa^2 \int f^{(r)^2} \left[\sum_{i=0}^{q-1} p_i^{-2r} I\left\{ \left(\beta\delta^{-1} |f^{(r)}| \right)^{2/(2r+1)} \le p_i \right\} \right]$$

$$\sim \kappa^2 p^{-2r} \int f^{(r)^2} \left[\sum_{i=0}^{\infty} 2^{-2ri} I\left\{ \left(\beta f^{(r)} \right)^2 \le \ell 2^{(2r+1)i} \right\} \right],$$

while if $p^{2r+1}\delta^2 \to 0$,

$$s_{53} \sim \kappa^2 \int f^{(r)^2} \left(1 - 2^{-2r}\right)^{-1} \left(\beta\delta^{-1} |f^{(r)}| \right)^{-4r/(2r+1)}$$

$$= \kappa^2 \left(1 - 2^{-2r}\right)^{-1} (\delta/\beta)^{4r/(2r+1)} \int |f^{(r)}|^{2/(2r+1)}.$$

The proposition follows on combining these two results with (3.25)–(3.27) and taking ε arbitrarily small. □

PROOF OF PROPOSITION 2.2. Let $0 < \varepsilon < 1$, let s_1 be as in the proof of Theorem 2.1 and define

$$s_{61} = \sum_{i=0}^{q-1} \sum_j E\left\{ \left(\hat{b}_{ij} - b_{ij} \right)^2 \right\} I\left(|b_{ij}| > \varepsilon\delta \right),$$

$$s_{62} = \sum_{i=0}^{q-1} \sum_j E\left[\left(\hat{b}_{ij} - b_{ij} \right)^2 I\left\{ |\hat{b}_{ij} - b_{ij}| > (1+\varepsilon)\delta \right\} \right].$$

Since

$$I\left\{ |\hat{b}_{ij} - b_{ij}| > (1+\varepsilon)\delta \right\} \le I\left(|\hat{b}_{ij}| > \delta \right) + I\left(|b_{ij}| > \varepsilon\delta \right),$$

then

$$s_{62} \le s_1 + s_{61}. \tag{3.28}$$

The argument leading to (3.5) may be used to show that

$$s_{61} = o\left(n^{-2r/(2r+1)}\right). \tag{3.29}$$

More simply,

$$s_{62} \ge \delta^2 \sum_{i=0}^{q-1} \sum_j P\left\{ |\hat{b}_{ij} - b_{ij}| > (1+\varepsilon)\delta \right\}. \tag{3.30}$$

We shall provide a lower bound to the probability on the right-hand side.

Let μ denote the model of f, and choose $\eta > 0$ so that f is continuous and bounded away from zero on the interval $(\mu - 2\eta, \mu + 2\eta)$. Set $\mathcal{J}_i = \{j: |\mu - jp_i^{-1}| \le \eta\}$, which set has at least $C_{15} p_i$ elements for some $C_{15} > 0$. By selecting η sufficiently small we may ensure that $\operatorname{var}(\hat{b}_{ij}) \ge n^{-1}(1+\varepsilon)^{-2} \sup f$ for all $j \in \mathcal{J}_i$ and $0 \le i \le q-1$. If $\hat{b}_{ij} - b_{ij}$ were normally distributed, then it would follow that, with N denoting a standard normal random variable,

$$\begin{aligned}P\left\{|\hat{b}_{ij} - b_{ij}| > (1+\varepsilon)\delta\right\} &\geq P\left\{|N| > (1+\varepsilon)^2(\sup f)^{-1/2}n^{1/2}\delta\right\}\\ &\sim 2(1+\varepsilon)^{-2}(\sup f)^{1/2}(n^{1/2}\delta)^{-1}\\ &\quad\times \exp\left\{-\tfrac{1}{2}(1+\varepsilon)^4(\sup f)^{-1}n\delta^2\right\}\\ &\geq C_{16}(\log n)^{-1/2}\exp\left\{-\tfrac{1}{2}(1+\varepsilon)^4(\sup f)^{-1}C^2\log n\right\}.\end{aligned}$$

A variant of this result may be established rigorously using methods of large deviation theory:

$$\begin{aligned}&P\left\{|\hat{b}_{ij} - b_{ij}| > (1+\varepsilon)\delta\right\}\\ &\qquad\geq C_{17}(\log n)^{-1/2}\exp\left\{-\tfrac{1}{2}(1+\varepsilon)^6(\sup f)^{-1}C^2\log n\right\} - C_{18}n^{-1},\end{aligned}\tag{3.31}$$

uniformly in $j \in \mathscr{J}_i$ and $0 \leq i \leq q-1$. Combining (3.24) and (3.28)–(3.31) we see that, with $\beta = 1+\varepsilon$,

$$\begin{aligned}\int E(\hat{f}-f)^2 &\geq C_{15}C_{17}(\log n)^{-1/2}\delta^2\sum_{i=0}^{q-1} p_i\exp\left\{-\tfrac{1}{2}\beta^6(\sup f)^{-1}C^2\log n\right\}\\ &\quad - o(n^{-2r/(2r+1)})\\ &\geq C_{19}(\log n)^{-1/2}\delta^2 p_q n^{-(1/2)\beta^6C^2/(\sup f)} - o(n^{-2r/(2r+1)}),\end{aligned}$$

as required.

A difficulty in deriving (3.31) using classical arguments is that the summands in the series represented by $n(\hat{b}_{ij} - b_{ij})$ do not have bounded moment-generating function. Essentially, the reason for this is that the series contains a great many terms which are identically zero, and it has been normalized to allow for this fact. To circumvent this difficulty, we condition on a quantity which is virtually equal to the number of nonzero terms. In more detail, suppose supp $\psi \subseteq (-v, v)$ and let Y_{ij} have the distribution of X conditional on $X \in ((j-v)/p_i, (j+v)/p_i)$. Write π_{ij} for probability that X lies in the latter interval and define

$$N_{ij} = \sum_{k=1}^{n} I\{X_k \in ((j-v)/p_i, (j+v)/p_i)\},$$

for $j \in \mathscr{J}_i$ and $0 \leq i \leq q-1$. Write Y_{kij}, $k \geq 1$, for independent random variables, independent also of N_{ij}, with the common distribution of Y_{ij}. Set $\nu_{ij} = E\{\psi(p_iY_{ij} - j)\}$,

$$S_{ij} = \sum_{k=1}^{N_{ij}} \psi(p_iY_{kij} - j), \qquad T_{ij} = \sum_{k=1}^{N_{ij}} \left\{\psi(p_iY_k - j) - \nu_{ij}\right\}.$$

Then S_{ij} has the same distributions as $np_i^{-1/2}\hat{b}_{ij}$, and so

$$\begin{aligned}P\big(|\hat{b}_{ij}-b_{ij}|>\beta\delta\big) &= P\big(|S_{ij}-np_i^{-1/2}\pi_{ij}^{-1}b_{ij}|>\beta np_i^{-1/2}\delta\big)\\ &= P\big\{|T_{ij}+(N_{ij}-EN_{ij})\nu_{ij}|>\beta np_i^{-1/2}\delta\big\}\\ &\geq P\big(|T_{ij}|>\beta^2 np_i^{-1/2}\delta\big)-P\big(|N_{ij}-EN_{ij}|>\varepsilon\beta np_i^{-1/2}\delta/|\nu_{ij}|\big)\\ &\geq P\big(|T_{ij}|/N_{ij}^{1/2}>\beta^{5/2}np_i^{-1/2}\delta\big)\\ &\quad -P\big(|N_{ij}-EN_{ij}|>\varepsilon\beta np_i^{-1/2}\delta/|\nu_{ij}|\big)-P(\beta N_{ij}\leq EN_{ij}).\end{aligned}$$

Result (3.31) follows from this inequality, from application of Bennett's or Bernstein's inequality to large deviations of $N_{ij}-EN_{ij}$ and from classicial arguments of large deviation theory (conducted initially conditional on N_{ij}). For the latter, see, for example, Linnik (1961a, b, 1962), Rubin and Sethuraman (1965) and Petrov [(1975), Chapter VIII]. □

Acknowledgments. The comments of a referee and an Associate Editor have been particularly helpful.

REFERENCES

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

DONOHO, D. L. (1992). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. Technical Report 403, Dept. Statistics, Stanford Univ.

DONOHO, D. L. and JOHNSTONE, I. M. (1992). Minimax estimation via wavelet shrinkage. Technical Report 402, Dept. Statistics, Stanford Univ.

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 426, Dept. Statistics, Stanford Univ.

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.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

HALL, P. and PATIL, P. (1993). Empirical bandwidth choice for nonlinear wavelet estimators. In preparation.

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. *Statist. Probab. Lett.* **18** 327–336.

KERKYACHARIAN, G. and PICARD, D. (1993b). Linear wavelet methods and other periodic kernel methods. Unpublished manuscript.

KERKYACHARIAN, G. and PICARD, D. (1993c). *Introduction aux Ondelettes et Estimation de Densité 1: Introduction aux Ondelettes et a l'Analyse Multiresolution.* Lecture Notes.

LINNIK, YU. V. (1961a). Limit theorems for sums of independent random variables, taking account of large deviations, I. *Theory Probab. Appl* **6** 131–148.

LINNIK, YU. V. (1961b). Limit theorems for sums of independent random variables, taking account of large deviations, II. *Theory Probab. Appl.* **6** 343–360.

LINNIK, YU. V. (1962). Limit theorems for sums of independent random variables, taking account of large deviations, III. *Theory Probab. Appl.* **7** 115–129.

MALLAT, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc.* **315** 69–87.

PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, Berlin.

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

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

RUBIN, H. and SETHURAMAN, J. (1965). Probabilities of moderate deviations. *Sankhyā Ser. A* **27** 325–346.

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

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

VAN EEDEN, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Ann. Inst. Statist. Math.* **37** 461–472.

STATISTICAL SCIENCE PROGRAM
CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
JOHN DEDMAN BUILDING
CANBERRA, ACT 0200
AUSTRALIA