Statistica Sinica **12**(2002), 1241-1273

ON BLOCK THRESHOLDING IN WAVELET REGRESSION: ADAPTIVITY, BLOCK SIZE, AND THRESHOLD LEVEL

T. Tony Cai

University of Pennsylvania

Abstract: In this article we investigate the asymptotic and numerical properties of a class of block thresholding estimators for wavelet regression. We consider the effect of block size on global and local adaptivity and the ch oice of thresholding constant. The optimal rate of convergence for block thresholding with a given block size is derived for both the global and local estimation. It is shown that there are conflicting requirements on the block size for achieving the global and local adaptivity. We then consider the choice of thresholding constant for a given block size by treating the block thresholding as a hypothesis testing problem. The combined results lead naturally to an optimal choice of block size and thresholding constant. We conclude with a numerical study which compares the finite-sample performance among block thresholding estimators as well as with other wavelet methods.

Key words and phrases: Block thresholding, convergence rate, global adaptivity, local adaptivity, minimax estimation, nonparametric regression, smoothing parameter, wavelets.

1. Introduction

Consider the nonparametric regression model:

$$y_i = f(x_i) + \sigma\, z_i, \tag{1}$$

i = 1, ..., n (= 2^J), $x_i = i/n$, σ is the noise level and z_i's are i.i.d. $N(0,1)$. The function $f(\cdot)$ is an unknown function of interest. We measure the estimation accuracy both globally by the mean integrated squared error

$$R(\hat{f}, f) = E\|\hat{f} - f\|_2^2\,, \tag{2}$$

and locally by the expected loss at a point

$$R(\hat{f}(x_0), f(x_0)) = E(\hat{f}(x_0) - f(x_0))^2. \tag{3}$$

Wavelets are an effective tool for nonparametric regression. Wavelet methods achieve adaptivity through shrinkage of the empirical wavelet coefficients. Standard wavelet shrinkage procedures estimate wavelet coefficients term by term,

on the basis of their individual magnitudes. Other coefficients have no influence on the treatment of particular coefficients. The commonly used VisuShrink of Donoho and Johnstone (1994) is a good example of the term-by-term thresholding procedures. Other term by term shrinkage rules include firm shrinkage (Gao and Bruce (1997)), non-negative garrote shrinkage (Gao (1998)), and Bayesian shrinkage rules based on independent priors on empirical coefficients (see, e.g., Clyde, Parmigiani, and Vidakovic (1998) and Abramovich, Sapatinas, and Silverman (1998)).

Hall, Kerkyacharian and Picard (1998, 1999a) introduced a local block thresholding estimator which thresholds empirical wavelet coefficients in groups rather than individually. The procedure first divides the wavelet coefficients at each resolution level into nonoverlapping blocks and then simultaneously keeps or kills all the coefficients within a block, based on the magnitude of the sum of the squared empirical coefficients with that block. The block size is chosen to be of order $(\log n)^2$ where n is the sample size. They demonstrate that the block thresholding estimator enjoys a number of advantages over the conventional term-by-term thresholding. See also Hall, Kerkyacharian and Picard (1999b) and Härdle, Kerkyacharian, Picard, and Tsybakov (1998). Other block shrinkage estimators have been considered in Cai (1999a), Cai and Silverman (2001) and Efromovich (2000a, b).

Block thresholding is conceptually appealing. It increases estimation precision by utilizing information about neighboring wavelet coefficients and allows the balance between variance and bias to be varied along the curve, resulting in adaptive smoothing. The degree of adaptivity, as we will show however, depends on the choice of block size and threshold level.

In the present paper, we consider the asymptotic and numerical properties of a class of block thresholding estimators for wavelet nonparametric regression with independent and identically distributed Gaussian errors. We have four objectives. The first is to study the effect of block length on both the global and local adaptivity. The second is to derive an appropriate threshold level for any given choice of block size. The third objective is to determine the "optimal" choice of block size and threshold level and investigate the asymptotic properties of the resulting estimator. And finally we wish to study the numerical performance of the block thresholding estimators.

The block size and the threshold level play important roles in the performance of a block thresholding estimator. After Section 2, in which basic notation and the block thresholding method are introduced, we consider in Section 3 the effect of block length on both the global and local adaptivity. The results reveal that there are conflicting demands on the block size for achieving the global and

local adaptivity. To achieve the optimal global adaptivity, the block size must be at least of the order $\log n$. On the other hand, to achieve the optimal local adaptivity, the block size must be no more than order $\log n$. Therefore no block thresholding estimator can achieve simultaneously the optimal global and local adaptivity if the block size is larger or smaller than order $\log n$.

In Section 4, we treat block thresholding as a hypothesis testing problem. We derive an appropriate threshold level for any given block size by imposing the condition that the type I error of the blockwise test for zero signal vanishes asymptotically. This condition also implies that the resulting estimator enjoys a desirable denoising property. With the selected threshold level, the block thresholding estimator is fully specified and attains the optimal rate of convergence for a given choice of block size.

The results obtained in Sections 3 and 4 lead naturally to the consideration in Section 5 of a possible optimal choice of block thresholding estimator. Asymptotic results show that the estimator is indeed optimal in the sense that it achieves simultaneously the exact global and local adaptivity. More specifically, it achieves the exact minimax convergence rate, under the global risk measure (2), over a wide range of function classes of inhomogeneous smoothness. The estimator also optimally adapts to the local smoothness of the underlying function; it achieves the adaptive minimax rate over an interval of local Hölder classes for estimating a function at a point.

We then consider the finite-sample performance of block thresholding estimators in Section 6. The block thresholding estimators are compared among themselves as well as with other wavelet methods. It is shown that the estimator with the "optimal" choice of block size and thresholding constant has superior numerical performance among the block thresholding estimators and in comparison to the other wavelet estimators. Section 7 discusses modifications and extensions of the block thresholding estimators. Proofs are given in Section 8.

2. Block Thresholding Estimators

Let $\{\phi, \psi\}$ be a pair of father and mother wavelets. The functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$. Dilation and translation of ϕ and ψ generates an orthonormal wavelet basis. We use the periodized wavelet bases on [0, 1] in the present paper. See Daubechies (1994) and Cohen, Daubechies, Jawerth, and Vial (1993) for more on wavelet bases on an interval.

A special family of compactly supported wavelets is the so-called Coiflets, constructed by Daubechies (1992), which can have arbitrary number of vanishing moments for both ϕ and ψ. Denote by $W(D)$ the collection of Coiflets $\{\phi, \psi\}$ of order D. So if $\{\phi, \psi\} \in W(D)$, then ϕ and ψ are compactly supported and satisfy

$\int x^i\phi(x)dx = 0$ for $i = 1, \ldots, D-1$; and $\int x^i\psi(x)dx = 0$ for $i = 0, \ldots, D-1$. An orthonormal wavelet basis has an associated orthogonal Discrete Wavelet Transform (DWT). See Daubechies (1992) and Strang (1992) for further details on wavelets and the DWT.

Suppose we observe noisy data $Y = \{y_i\}$ as in (1). Let $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$, and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. Denote the true wavelet coefficients of f by $\xi_{j,k} = \langle f, \phi_{j,k}\rangle$ and $\theta_{j,k} = \langle f, \psi_{j,k}\rangle$. Let $\tilde{Y} = W \cdot n^{-1/2}Y$ be the discrete wavelet transform of $n^{-1/2}Y$. Write

$$\tilde{Y} = (\tilde{\xi}_{j_0 1}, \ldots, \tilde{\xi}_{j_0 2^{j_0}}, \tilde{y}_{j_0 1}, \ldots, \tilde{y}_{j_0 2^{j_0}}, \ldots, \tilde{y}_{J-1,1}, \ldots, \tilde{y}_{J-1,2^{J-1}})'. \tag{4}$$

Here $\tilde{\xi}_{j_0 k}$ are the gross structure terms at the lowest resolution level, and $\tilde{y}_{j,k}$ $(j = 1, \ldots, J-1, k = 1, \ldots, 2^j)$ are empirical wavelet coefficients at level j which represent detail structure at scale 2^j. The $\tilde{y}_{j,k}$ are independent with noise level $n^{-1/2}\sigma$ and can be written as

$$\tilde{y}_{j,k} = \theta'_{j,k} + n^{-1/2}\sigma z_{j,k}, \tag{5}$$

where the $\theta'_{j,k}$ are approximately the true coefficients of f, and the $z_{j,k}$'s are i.i.d. $N(0,1)$.

A term-by-term thresholding procedure estimates the function f by

$$\hat{f}_t(x) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \tilde{y}_{j,k} I(|\tilde{y}_{j,k}| > T)\psi_{j,k}(x).$$

Here, each wavelet coefficient $\theta_{j,k}$ is estimated separately and the estimate $\hat{\theta}_{j,k}$ depends solely on $\tilde{y}_{j,k}$, other coefficients have no influence on $\hat{\theta}_{j,k}$. The threshold $T = (2n^{-1}\log n)^{1/2}\sigma$ is used in Donoho and Johnstone (1994).

A block thresholding estimator thresholds wavelet coefficients in groups instead of individually. Block thresholding aims to increase estimation accuracy by utilizing information about neighboring wavelet coefficients and making simultaneous decisions on all the coefficients within a block. Local block thresholding rules were first introduced by Hall, Kerkyacharian, and Picard (1998, 1999a). The procedure is as follows.

At each resolution level j, the empirical wavelet coefficients $\tilde{y}_{j,k}$ are divided into nonoverlapping blocks of length L. Denote (jb) the indices of the coefficients in the b-th block at level j, i.e., $(jb) = \{(j,k) : (b-1)L + 1 \le k \le bL\}$. Let $S^2_{jb} = \sum_{k\in(jb)} \tilde{y}^2_{j,k}$ denote the sum of squares of the empirical coefficients in the block. A block (jb) is deemed important if S^2_{jb} is larger than a threshold $T = \lambda L n^{-1}\sigma^2$

and then all the coefficients in the block are retained; otherwise the block is considered negligible and all the coefficients in the block are discarded.That is,

$$\hat{\theta}_{j,k} = \tilde{y}_{j,k} \cdot I(S_{jb}^2 > \lambda L n^{-1}\sigma^2), \qquad \text{for } (j,k) \in (jb). \tag{6}$$

The estimator of the whole function is given by

$$\hat{f}(x) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1}\sum_{b}\Big(\sum_{k\in(jb)} \tilde{y}_{j,k}\psi_{j,k}(x)\Big) I(S_{jb}^2 > \lambda L n^{-1}\sigma^2). \tag{7}$$

Block thresholding estimators depend on the choice of the block size L and thresholding constant λ. The term-by-term estimator VisuShrink is a special case of (7) with $L = 1$ and $\lambda = 2\log n$. In the regression case, Hall, Kerkyacharian, and Picard (1999a) suggest to choose $L = (\log n)^2$ and $\lambda \geq 48$. It is shown that, under the global risk measure (2), the block thresholding estimator with the chosen parameters attains the minimax rate of convergence over a range of function classes $\mathcal{H}$ considered in Section 3.1. In the density estimation case, Hall, Kerkyacharian, and Picard (1998) choose $L = C(\log n)^2$ for sufficiently large constant $C > 0$. See also Härdle, Kerkyacharian, Picard, and Tsybakov (1998).

Block thresholding may also be regarded as an automatic model selection procedure, which selects a set of important variables (wavelet coefficients) by omitting insignificant ones and fits to the data a model consisting of only the important variables. The distinctive feature of block thresholding is that it retains or deletes variables group-by-group rather than one-by-one.

Because the choice of block size L and thresholding constant λ largely determines the performance of the resulting estimator, it is important to study in detail the effect of L and λ on the properties of the estimator and derive the optimal L and λ if such values exist. Between the two parameters L and λ, the block size L is more important. It plays a similar role as the bandwidth in the traditional kernel estimation. In the present paper, we consider the block thresholding estimator (7) with general block size $L = (\log n)^s$ with some $s \geq 0$ and denote by $\hat{f}_{s,\lambda}$ an estimator with block size $L = (\log n)^s$ and thresholding constant λ. We begin by investigating the relationship between block size and adaptivity.

3. The Effect of Block Length on Adaptivity

We consider both global and local adaptivity. An estimator that is globally adaptive can automatically adjust to varying level of overall regularity of the target function; and a locally adaptive estimator can optimally adapt to subtle, spatial changes in smoothness along the curve. An estimator that achieves simultaneously the optimal global and local adaptivity permits the trade-off between

variance and bias to be varied along the curve in an optimal way, resulting in spatially adaptive smoothing in classical sense.

3.1. The function classes $\mathcal{H}$

We consider the global adaptivity of block thresholding estimators over a family of large function classes which was used in Hall, Kerkyacharian, and Picard (1999a). The classes contain functions of inhomogeneous smoothness and are different from other traditional smoothness classes. Functions in these classes can be regarded as the superposition of smooth functions with irregular perturbations such as jump discontinuities and high frequency oscillations.

Definition 1. Let $\mathcal{H} = \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, v)$, where $0 \le \alpha_1 < \alpha \le D$, $0 \le \gamma < \frac{1+2\alpha_1}{1+2\alpha}$, and $M_1, M_2, M_3, v \ge 0$, denote the class of functions f such that for any $j \ge j_0 > 0$ there exists a set of integers A_j with $card(A_j) \le M_3 2^{j\gamma}$ for which the following are true:

- For each $k \in A_j$, there exist constants $a_0 = f(2^{-j}k), a_1, \ldots, a_{D-1}$ such that for all $x \in [2^{-j}k, 2^{-j}(k+v)]$, $|f(x) - \sum_{m=0}^{D-1} a_m (x - 2^{-j}k)^m| \le M_1 2^{-j\alpha_1}$;
- For each $k \notin A_j$, there exist constants $a_0 = f(2^{-j}k), a_1, \ldots, a_{D-1}$ such that for all $x \in [2^{-j}k, 2^{-j}(k+v)]$, $|f(x) - \sum_{m=0}^{D-1} a_m (x - 2^{-j}k)^m| \le M_2 2^{-j\alpha}$.

Roughly speaking, the intervals with indices in A_j are "bad" intervals which contain less smooth parts of the function. The number of the "bad" intervals is controlled by M_3 and γ so that the irregular parts do not overwhelm the fundamental structure of the function. The function class $\mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, v)$ contains the Besov class $B^{\alpha}_{\infty\infty}(M_2)$ as a subset for any given α_1, γ, M_1, M_3, D, and v. Loosely speaking, the Besov space $B^{\alpha}_{p,q}$ contains functions having α bounded derivatives in L^p space, the second parameter q gives a finer gradation of smoothness. See Meyer (1992) and Triebel (1983) for definitions and properties of Besov spaces.

A function $f \in \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, v)$ can be regarded as the superposition of a regular smooth function f_s in $B^{\alpha}_{\infty\infty}(M_2)$ and an irregular perturbation τ: $f = f_s + \tau$. The perturbation τ can be, for example, jump discontinuities or high frequency oscillations such as chirp and Doppler of the form: $\tau(x) = \sum_{k=1}^{K} a_k (x - x_k)^{\beta_k} \cos(x - x_k)^{-\gamma_k}$. See Hall, Kerkyacharian, and Picard (1998, 1999a) for further discussions about the function classes $\mathcal{H}$.

3.2. Effect on global adaptivity

An estimator is said to achieve the optimal global adaptivity over some function classes $\mathcal{F}_\alpha$ for a range of smoothness index $\alpha \in \mathcal{A}$ if, under the global

risk measure (2), it attains the minimax rate of convergence simultaneously for all $\alpha \in \mathcal{A}$.

Define the traditional Hölder class $\Lambda^\alpha(M)$ in the usual way:

$$\Lambda^\alpha(M) \;=\; \{f : |f^{(\lfloor\alpha\rfloor)}(x) - f^{(\lfloor\alpha\rfloor)}(y)| \le M\,|x-y|^{\alpha'}\}$$

where $\lfloor\alpha\rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor\alpha\rfloor$. Relative to the function class $\mathcal{H}$, the Hölder class is smaller and contains relatively simple functions.

It is well known that the minimax rate of convergence for global estimation over $\Lambda^\alpha(M)$ and the Besov class $B^\alpha_{\infty\infty}(M)$ is $n^{-2\alpha/(1+2\alpha)}$. Because $\mathcal{H} \equiv \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, v)$ contains $B^\alpha_{\infty\infty}(M_2)$ as a subset, the convergence rate over $\mathcal{H}$ cannot exceed $n^{-2\alpha/(1+2\alpha)}$. The results below shows the significant effect of block length on the global adaptivity.

Theorem 1. *Suppose the wavelets* $\{\phi, \psi\} \in W(D)$ *and* $supp(\phi) = supp(\psi) = (0, N)$. *Let* $\mathcal{H} = \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, v)$.

(i) *If* $0 \le s < 1$, *then for any* $\lambda = \lambda(n)$ *and for all* $0 < \alpha \le D$ *and* $0 < M < \infty$,

$$\varlimsup_{n\to\infty}\; n^{\frac{2\alpha}{1+2\alpha}} \cdot (\log n)^{-\frac{2\alpha(1-s)}{1+2\alpha}} \cdot \sup_{f\in\Lambda^\alpha(M)} E\|\hat{f}_{s,\lambda} - f\|_2^2 > 0. \tag{8}$$

(ii) *On the other hand, if* $s > 1$, *then for any fixed* $\lambda > 1$ *and for all* $0 < \alpha \le D$ *and* $v \ge N$,

$$\varlimsup_{n\to\infty}\; n^{\frac{2\alpha}{1+2\alpha}} \cdot \sup_{f\in\mathcal{H}} E\|\hat{f}_{s,\lambda} - f\|_2^2 < \infty. \tag{9}$$

Theorem 1 shows the striking difference in asymptotic behavior between block thresholding estimators $\hat{f}_{s,\lambda}$ with $s < 1$ and those with $s > 1$. When the block size is small, i.e., $s < 1$, the rate of convergence for $\hat{f}_{s,\lambda}$ over $\Lambda^\alpha(M)$ cannot exceed $(\log^{1-s} n/n)^{2\alpha/(1+2\alpha)}$, and so it is impossible for the estimator $\hat{f}_{s,\lambda}$ to achieve the optimal global adaptivity even over simple function classes $\Lambda^\alpha(M)$. The extra logarithmic factor in (8) is due to the fact that the block size is too small and consequently information on neighboring coefficients within a block is not sufficient to precisely estimate the coefficients. On the other hand, with $s > 1$ and any fixed thresholding constant $\lambda > 1$, a block thresholding estimator $\hat{f}_{s,\lambda}$ is globally adaptive over a wide range of function classes $\mathcal{H}$ of inhomogeneous smoothness.

Theorem 1(i) gives a lower bound for the global risk of block thresholding estimators $\hat{f}_{s,\lambda}$ with $0 \le s < 1$. The lower bound is sharp, i.e., with an appropriately chosen λ, the rate $(\log^{1-s} n/n)^{2\alpha/(1+2\alpha)}$ is attained. So,

$$\inf_{\lambda}\sup_{f\in\Lambda^\alpha(M)} E\|\hat{f}_{s,\lambda} - f\|_2^2 \asymp (\log^{1-s} n/n)^{2\alpha/(1+2\alpha)}. \tag{10}$$

In fact, (10) holds over the larger function class $\mathcal{H}$. The choice of thresholding constant λ is discussed in Section 4.

Remark. A special case is $L = 1$. Theorem shows that the rate of convergence over Hölder classes $\Lambda^\alpha(M)$ cannot exceed $(\log n/n)^{2\alpha/(1+2\alpha)}$ for any term by term thresholding estimator. This rate is attained by the VisuShrink estimator of Donoho and Johnstone (1994).

3.3. Effect on local adaptivity

For functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point and global risk measures such as (2) cannot wholly reflect the performance of an estimator locally. The local risk measure (3) is more appropriate for measuring the spatial adaptivity, where $x_0 \in (0,\ 1)$ is any fixed point of interest.

Define the local Hölder class $\Lambda^\alpha(M, x_0, \delta)$ by

$$\Lambda^\alpha(M, x_0, \delta) \ = \ \{f : |f^{(\lfloor\alpha\rfloor)}(x) - f^{(\lfloor\alpha\rfloor)}(x_0)| \le M\,|x - x_0|^{\alpha'},\ x \in (x_0 - \delta,\ x_0 + \delta)\}$$

where $\lfloor\alpha\rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor\alpha\rfloor$.

There is an interesting and important distinction between global estimation and local estimation. In global estimation, it is possible to achieve complete success of adaptation across a range of function classes in terms of convergence rate, in some case, even at the level of the constant. That is, one can do as well when the degree of smoothness is unknown as one could do if the degree of smoothness is known.

For local estimation, however, one must pay a price for adaptation. When α is known, the local minimax risk over $\Lambda^\alpha(M, x_0, \delta)$ converges at the rate of n^{-r} where $r = 2\alpha/(1 + 2\alpha)$. When α is unknown, as shown by Lepski (1990) and Brown and Low (1996), one has to pay a price for adaptation of at least a logarithmic factor; the best one can do in this case is $(\log n/n)^r$. We call $(\log n/n)^r$ the adaptive minimax rate for local estimation.

We now consider the effect of block size on local adaptivity of $\hat{f}_{s,\lambda}$.

Theorem 2. *Suppose the wavelets $\{\phi, \psi\} \in W(D)$ and $x_0 \in (0, 1)$ is fixed.*

(i) *If $0 \le s < 1$, then there exists $\lambda = \lambda(L)$ such that for all $0 < \alpha \le D$ and $0 < M < \infty$,*

$$\overline{\lim_{n\to\infty}}\ (\frac{n}{\log n})^{\frac{2\alpha}{1+2\alpha}} \cdot \sup_{f\in\Lambda^\alpha(M,x_0,\delta)} E(\hat{f}_{s,\lambda}(x_0) - f(x_0))^2 < \infty. \qquad (11)$$

(ii) *If $s > 1$, then for any fixed thresholding constant $\lambda > 1$ and for all $0 < \alpha \le D$ and $0 < M < \infty$,*

$$\overline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^{\frac{2\alpha}{1+2\alpha}} \cdot (\log n)^{-\frac{2\alpha(s-1)}{1+2\alpha}} \cdot \sup_{f\in\Lambda^\alpha(M,x_0,\delta)} E(\hat{f}_{s,\lambda}(x_0) - f(x_0))^2 > 0. \quad (12)$$

In words, when $s > 1$, no block thresholding estimator can achieve the optimal local adaptivity. The extra logarithmic factor in (12) is due to the fact that the block size is too large and consequently the estimator is not well localized. Intuitively, it is clear that the block length should not be too large in order to well adapt to the local behavior of the underlying function. On the other hand, if $s < 1$, then, with an appropriate choice of λ, the optimal local adaptivity can be achieved. The choice of λ will be discussed in Section 4.

It is revealing to put Theorems 1 and together. Block size affects the global and local adaptivity in the opposite direction. To attain optimal rate of convergence in global estimation, the block size L needs to be large so the information contained in a block is sufficient for accurate decision making. On the other hand, to attain adaptive rate of convergence in local estimation, the block size L needs to be small so the estimator is well localized. Theorems 1 and combined show that it is impossible to simultaneously achieve both by a block thresholding estimator with $L = (\log n)^s$ and $s \neq 1$.

These results leave the choice of $L = \log n$ as the only possible optimal compromise. We will consider this case in Section 5 and show that $L = \log n$ is indeed the optimal choice in the sense that with $L = \log n$ and an appropriate λ derived in Section 4, the resulting block thresholding estimator achieves simultaneously the optimal global and local adaptivity.

4. The Choice of Thresholding Constant

The aim of block thresholding is to achieve better adaptivity while retaining the smoothing and denoising properties. In particular, we wish to choose the threshold so that the estimator removes pure noise completely, with probability tending to 1. In this section, we treat block thresholding as a hypothesis testing problem and select the thresholding constant so that the resulting estimator achieves these objectives.

Suppose one observes

$$x_i = \theta_i + z_i, \qquad i = 1, \ldots, n,$$

with $z_i \overset{i.i.d.}{\sim} N(0, 1)$. The mean $\theta = (\theta_i)$ is the object of interest. Assume one has reasons to think, although not certain, that the mean θ is zero. Then it is natural first to test

$$H_0 : \theta_1 = \cdots = \theta_n = 0. \quad (13)$$

Term-by-term thresholding can be viewed as a Bonferroni type test which tests the global hypothesis (13) coordinate-wise. In contrast, block thresholding tests

the global hypothesis (13) in groups. Divide the mean vector into block of size L and test the hypothesis $H_0^{(b)}: \theta_{bL-L+1} = \cdots = \theta_{bL} = 0$. On each block (b) for $b = 1, \ldots, n/L$, and estimate $\theta_{(b)}$ by 0 when the hypothesis $H_0^{(b)}$ is not rejected and by $x_{(b)}$ otherwise. Multivariate normal decision theory shows that, for each block, a uniformly most powerful test exists; the best rejection region is of the form $\sum x_i^2 > T$, where T is a constant (see, e.g., Lehmann and Casella (1998, p. 351). Hence the estimator becomes $\hat{\theta}_j = x_j \cdot I(\sum_{i\in(b)} x_i^2 > T)$ for $j \in (b)$, which is exactly a block thresholding estimator.

Rewriting the threshold T as $T = \lambda \cdot L$, it is easy to see that the probability of type I error of the blockwise test under the null hypothesis is

$$p_L(\lambda) = 1 - (1 - P(\chi_L^2 > \lambda \cdot L))^{n/L}, \tag{14}$$

where χ_L^2 denotes a chi-squared distribution with L degrees of freedom. We impose the condition that, under the null, the blockwise test asymptotically makes the correct decision with certainty, i.e., $p_L(\lambda) \to 0$, as $n \to \infty$. In the context of signal detection, this means that if the observations are pure noise without any signal, as the sampling frequency increases, one can tell eventually with certainty that there is no signal. Equivalently, the corresponding estimator removes pure noise completely, with probability tending to 1. The condition $p_L(\lambda) \to 0$ as $n \to \infty$ provides a criterion for the selection of the thresholding constant λ.

Theorem 3. *Let $L = (\log n)^s$ and let $p_L(\lambda)$, as given in (14), be the probability of type I error of the blockwise test. Denote $T_s = 2(\log n)^{1-s}$ for $0 < s < 1$ and $\delta_s = 2(\log n)^{-(s-1)/2}$ for $s > 1$. Let*

(i) $\lambda_s = 2\log n$, *when* $s = 0$; (15)

(ii) $\lambda_s = T_s + \log T_s + 1$, *when* $0 < s \le 1/2$; (16)

(iii) $\lambda_s = T_s + \log T_s + 1 + (\log T_s + 1)/T_s$, *when* $1/2 < s < 1$; (17)

(iv) $\lambda_s = 4.5052$ *the root of* $\lambda - \log\lambda - 3 = 0$, *when* $s = 1$; (18)

(v) $\lambda_s = 1 + \delta_s + \delta_s^2/3 + \delta_s^3/36$, *when* $1 < s < 2$. (19)

(vi) $\lambda_s = 1 + \delta_s + \delta_s^2/3$, *when* $2 \le s < 3$. (20)

(vii) $\lambda_s = 1 + \delta_s$, *when* $s \ge 3$. (21)

Then, for $\lambda \ge \lambda_s$, $p_L(\lambda) \to 0$ as $n \to \infty$. Moreover, the bounds given above are sharp.

For example, in the case of $0 < s \le 1/2$, if $\lambda \le T_s + \log T_s + c$ with a constant $c < 1$, then $p_L(\lambda) \to 1$. In particular, if $\lim_{n\to\infty} \lambda/\lambda_s < 1$, then $p_L(\lambda) \to 1$.

Remark. In the special case of $L = 1$, the bound $\lambda_0 = 2\log n$ given in Theorem 3 is equivalent to the bound $\sqrt{2\log n}$ in the Gaussian case, which motivates the choice of the threshold for VisuShrink (see Donoho and Johnstone (1994)). In the case of $s = 1$ (or $L = \log n$), λ_1 is an absolute constant satisfying $\lambda - \log\lambda - 3 = 0$.

Guided by Theorem 3, for a given block length $L = (\log n)^s$, we choose the thresholding constant λ_s as in (15)−(21) Furthermore, for $L = (\log n)^s$ with $s \neq 1$, the lower bounds of convergence rates as specified in Theorem 1 and Theorem 2 are attained by the block thresholding estimators $\hat{f}_{s,\lambda_s}$. For example, with $L = (\log n)^s$ and $0 < s \leq 1/2$, and λ_s as given in (16), $\hat{f}_{s,\lambda_s}$ satisfies that, for all $0 < \alpha \leq D$ and $0 < M < \infty$,

$$\sup_{f \in \mathcal{H}} E\|\hat{f}_{s,\lambda_s} - f\|_2^2 \leq C(\frac{\log^{1-s} n}{n})^{\frac{2\alpha}{1+2\alpha}},$$

$$\sup_{f \in \Lambda^\alpha(M,x_0,\delta)} E(\hat{f}_{s,\lambda_s}(x_0) - f(x_0))^2 \leq C(\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}.$$

Both rates are optimal for the given choice of block length.

5. $L = \log n$: the Optimal Choice

The results in Section 3 show that for both global and local adaptivity, $s = 1$ is the dividing line. A natural question is what happens in this critical case of $s = 1$, or equivalently $L = \log n$? As derived in Section 4, the corresponding thresholding constant λ in this case is $\lambda_1 = 4.5052$. This block thresholding estimator is denoted by $\hat{f}_{1,\lambda_1}$.

Theorem 4. (i) *Under the conditions of Theorem* 1, $\hat{f}_{1,\lambda_1}$ *is globally adaptive over* $\mathcal{H} = \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, v)$ *for all* $0 < \alpha \leq D$ *and for all* $v \geq N$, *i.e.,*

$$\sup_{f \in \mathcal{H}} E\|\hat{f}_{1,\lambda_1} - f\|_2^2 \leq Cn^{-2\alpha/(1+2\alpha)}. \tag{22}$$

(ii) *Under the conditions of Theorem ,* $\hat{f}_{1,\lambda_1}$ *is locally adaptive for all* $0 < \alpha \leq D$, $\delta > 0$ *and* $0 < M < \infty$, *i.e., for any fixed* $x_0 \in (0,1)$,

$$\sup_{f \in \Lambda^\alpha(M,x_0,\delta)} E(\hat{f}_{1,\lambda_1}(x_0) - f(x_0))^2 \leq C \cdot (\log n/n)^{2\alpha/(1+2\alpha)}. \tag{23}$$

Thus, the estimator $\hat{f}_{1,\lambda_1}$, without knowing the a priori degree or amount of smoothness of the underlying function, attains the adaptive minimax rate simultaneously for global and local estimation. It is easy to generalize the global result over other function classes such as the regular Besov classes. The estimator $\hat{f}_{1,\lambda_1}$ achieves simultaneously the global and local adaptivity, which is impossible to achieve for any block thresholding estimator $\hat{f}_{s,\lambda}$ with $s \neq 1$. In this sense, $\hat{f}_{1,\lambda_1}$ is asymptotically optimal within this class of block thresholding estimators.

Remark. (*Use of Coiflets*): If the following local Lipschitz condition is imposed on $\mathcal{H}$ when functions in $\mathcal{H}$ are relatively smooth, then there is no need for using Coiflets. Indeed simulation shows no particular advantages of using Coiflets in the finite sample case.

(i) If $\alpha > 1 \geq \alpha_1$, then for $k \notin A_j$, $|f(x) - f(2^{-j}k)| \leq M_4 2^{-j}$, for $x \in [2^{-j}k, 2^{-j}(k + v)]$.
(ii) If $\alpha > \alpha_1 > 1$, then $|f(x) - f(2^{-j}k)| \leq M_4 2^{-j}$, for $x \in [2^{-j}k, 2^{-j}(k + v)]$.

6. Numerical Comparisons

We have so far focused on the comparisons of asymptotic properties. In this section, we carry out a simulation study to compare the finite sample performance among the block thresholding estimators as well as with the conventional wavelet methods.

The block thresholding estimator $\hat{f}_{s,\lambda_s}$ can be easily implemented in three steps for estimating f at the sample points, at a computational cost of $O(n)$, as follows.

1. Transform the noisy data via the discrete wavelet transform.
2. At each resolution level, the empirical coefficients are grouped into nonoverlapping blocks of length $L_s = (\log n)^s$. If the sum of the squared empirical coefficients in a block is above the threshold $T = \lambda_s L_s \sigma^2$, then all the coefficients in the block are retained, otherwise all the coefficients in the block are discarded.
3. Obtain the estimate of function f at the sample points by the inverse discrete wavelet transform of the denoised wavelet coefficients.

For numerical comparisons we consider the average mean squared errors (AMSE) of the estimators at the sample points,

$$\text{AMSE} = \frac{1}{N}\sum_{\ell=1}^{N}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_\ell(x_i) - f(x_i))^2\right),$$

where $\hat{f}_\ell$ is the estimate of f in ℓ-th replication and N is the total number of replications.

Eight test functions representing different level of spatial variability are used. The test functions are normalized so that all the functions have the same signal standard deviation of 10. Doppler, HeaviSine, Bumps and Blocks are from Donoho and Johnstone (1994), Blip and Wave are from Marron, Adak, Johnstone, Neumann, and Patil (1998), and Spikes and Corner are from Cai (1999a). Plots of the test functions are given in the appendix. Sample sizes from 512 to 8192 and signal-to-noise ratios (SNR) from 3 to 7 are considered. Different combinations of wavelets and SNRs yield basically the same results. For reasons of space, we only report in detail the results for one particular case, using Daubechies' compactly supported wavelet *Symmlet* 8 and SNR equal to 5. See Cai (1999b) for additional results.

6.1. Comparisons among block thresholding estimators

We consider the block thresholding estimators $\hat{f}_{s,\lambda_s}$ for five different block sizes: $s = 0, 0.5, 1, 1.5$, and 2. There are 2^j empirical wavelet coefficients at a given resolution level j. It is often more convenient to choose the block size to be a dyadic integer and evenly divide the coefficients at each resolution level into nonoverlapping blocks. In the simulation, for a given choice of s, the block size is chosen to be the largest dyadic integer smaller than or equal to $(\log n)^s$, i.e., $L = 2^{[\log_2(\log n)^s]}$. Throughout, the lowest resolution level $j_0 = \text{ceiling}(\log_2 \log n) + 1$ was used. For certain values of n and s, the number of coefficients at the lowest level is smaller than the block size. In that case, thresholding stops at the level where the number of coefficients equals the block size so that there is at least one block at a level. Table 1 reports the AMSEs (rounded to two significant digits) over 500 replications for the block thresholding estimators $\hat{f}_{s,\lambda_s}$ with $s = 0, 0.5, 1, 1.5$, and 2. A graphical presentation is given in Figure 1.

Table 1. Mean squared errors from 500 replications (SNR=5)

n	$s=1$	$s=0$	$s=0.5$	$s=1.5$	$s=2$	$s=1$	$s=0$	$s=0.5$	$s=1.5$	$s=2$
	Doppler					*HeaviSine*				
512	0.98	1.55	1.33	0.89	1.29	0.56	0.56	0.52	0.63	0.60
1024	0.63	0.91	0.75	0.52	0.74	0.36	0.37	0.32	0.45	0.44
2048	0.32	0.53	0.42	0.28	0.39	0.19	0.19	0.16	0.29	0.31
4096	0.17	0.31	0.22	0.17	0.31	0.13	0.12	0.09	0.17	0.19
8192	0.08	0.18	0.12	0.09	0.16	0.07	0.07	0.06	0.10	0.13
	Bumps					*Blocks*				
512	1.81	2.43	1.96	2.04	3.03	1.98	2.28	2.03	2.15	3.16
1024	1.23	1.66	1.34	1.98	2.40	1.18	1.40	1.14	1.83	2.06
2048	0.74	1.08	0.79	1.14	1.42	0.77	0.95	0.78	1.26	1.55
4096	0.55	0.63	0.47	0.67	0.97	0.67	0.61	0.47	0.81	1.23
8192	0.28	0.35	0.25	0.36	0.58	0.41	0.39	0.28	0.51	0.84
	Spikes					*Blip*				
512	0.82	1.09	0.76	0.97	1.34	0.43	0.45	0.39	0.52	1.26
1024	0.50	0.70	0.49	0.65	0.80	0.26	0.26	0.27	0.48	0.77
2048	0.30	0.38	0.29	0.34	0.47	0.17	0.18	0.17	0.26	0.43
4096	0.16	0.21	0.17	0.17	0.26	0.09	0.11	0.09	0.14	0.39
8192	0.08	0.11	0.09	0.09	0.13	0.05	0.06	0.05	0.09	0.24
	Corner					*Wave*				
512	0.36	0.39	0.35	0.40	0.42	0.56	1.34	1.00	0.56	0.52
1024	0.21	0.23	0.20	0.24	0.27	0.31	0.53	0.34	0.28	0.29
2048	0.12	0.13	0.12	0.13	0.15	0.17	0.17	0.18	0.16	0.17
4096	0.06	0.07	0.07	0.07	0.08	0.09	0.10	0.10	0.09	0.09
8192	0.04	0.04	0.04	0.05	0.05	0.06	0.07	0.07	0.07	0.06

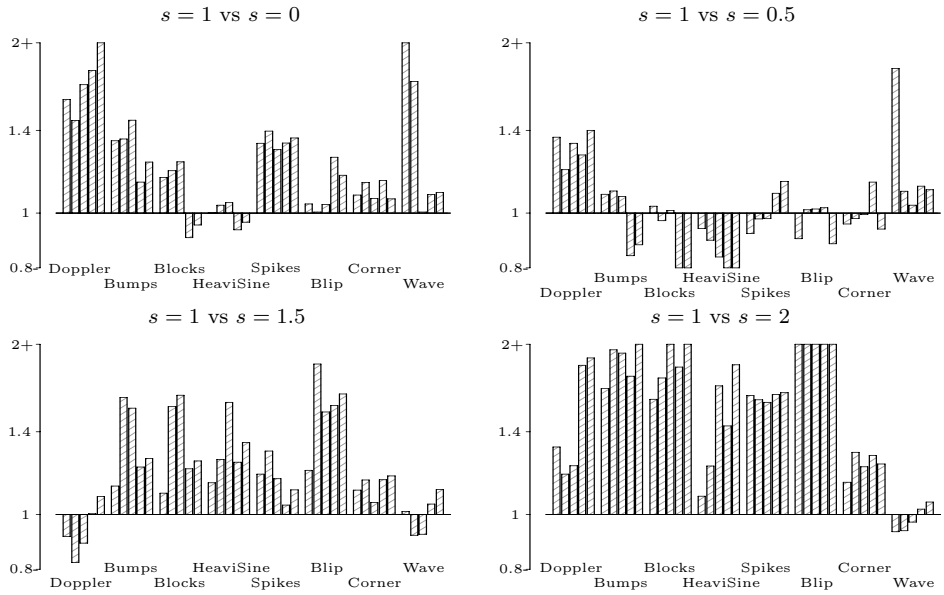


Figure 1. The vertical bars represent the ratios of the AMSEs of $\hat{f}_{s,\lambda_s}$ with $s = 0, 0.5, 1.5$, and 2 to the corresponding AMSE with $s = 1$. The higher the bar the better the relative performance of $\hat{f}_{1,\lambda_1}$. For each signal the bars are ordered from left to right by the sample sizes ($n = 512$ to 8192).

The estimator $\hat{f}_{1,\lambda_1}$ has smaller AMSE than VisuShrink $\hat{f}_{0,\lambda_0}$ in all but four cases, among the total of 40 combinations of signals and sample sizes. The improvement is more significant for functions with significant spatial variability such as Doppler, Bumps, and Spikes. The estimator $\hat{f}_{1,\lambda_1}$ outperforms the other two block thresholding estimators with larger block sizes as well. Among the 40 cases, the AMSE of $\hat{f}_{1,\lambda_1}$ is lower than those of $\hat{f}_{1.5,\lambda_{1.5}}$ and $\hat{f}_{2,\lambda_2}$ in 35 and 37 cases, respectively. The differences between the AMSEs of $\hat{f}_{1,\lambda_1}$ and $\hat{f}_{2,\lambda_2}$ are highly significant. In terms of AMSE, the only competitor to $\hat{f}_{1,\lambda_1}$ among the estimators under consideration is $\hat{f}_{0.5,\lambda_{0.5}}$. These two estimators are comparable. Overall, among the five block thresholding estimators, $\hat{f}_{1,\lambda_1}$ has the best numerical performance.

The numerical results agree to a certain extent with the asymptotic properties of estimators. For example, for $0 \le s \le 1$, the global performance of the estimator $\hat{f}_{s,\lambda_s}$ improves as s increases from 0 to 0.5 and to 1. Both the asymptotic and finite sample results show that the estimator $\hat{f}_{1,\lambda_1}$ has the best performance among the class of estimators under consideration. However, there is also some noticeable discrepancy between the asymptotic and finite sample results. For instance, although it is shown that asymptotically the estimators $\hat{f}_{s,\lambda_s}$ attain the optimal rate of convergence for any $s \ge 1$ under the global risk

measure, both $\hat{f}_{1.5,\lambda_{1.5}}$ and $\hat{f}_{2,\lambda_2}$ do not perform well in the simulation study. It is possible that it requires very large sample sizes for the asymptotics to take effect. In addition the asymptotic results only concern the rate of the convergence; the constant factor in the asymptotic risk is not considered. The discrepancy shows the need of examining both the asymptotic and finite sample performance of estimators.

6.2. Comparisons with other wavelet methods

The estimator $\hat{f}_{1,\lambda_1}$ stands out among the block thresholding estimators, both in terms of asymptotic adaptivity and numerical performance. We now compare $\hat{f}_{1,\lambda_1}$ with four other wavelet methods, RiskShrink, SureShrink, Translation-Invariant (TI) de-noising, and BlockJS.

RiskShrink (Donoho and Johnstone (1994)) is a term-by-term thresholding estimator with the threshold chosen to achieve certain minimaxity for a given sample size n. SureShrink thresholds the empirical coefficients by minimizing the Stein's unbiased risk estimate at each resolution level. We use the hybrid method proposed in Donoho and Johnstone (1995) in the simulations. RiskShrink and SureShrink usually have better mean squared error performance than VisuShrink, but the reconstructions often contain visually unpleasant spurious fine-structure. TI de-noising (Coifman and Donoho (1995)) averages over VisuShrink estimates based on all the shifts of the original data. BlockJS (Cai (1999a)) is a block shrinkage procedure using the James-Stein rule. This estimator is shown to have numerical advantages over several conventional estimators. For further details see the original papers.

The AMSEs over 500 replications is reported in Table 2 with a graphical presentation given in Figure 2. In Table 2, "BJS" stands for BlockJS. The estimator $\hat{f}_{1,\lambda_1}$ outperforms the other methods. It yields better results than RiskShrink in 36 out of the 40 cases; and beats TI de-noising in 35 out of 40 cases. The differences are especially notable when the underlying function is of significant spatial variability. In terms of AMSE, the competitors among the four methods are SureShrink and BlockJS. The estimator $\hat{f}_{1,\lambda_1}$ has similar asymptotic properties as BlockJS. But $\hat{f}_{1,\lambda_1}$ has almost uniformly better numerical performance; it yields smaller AMSE than BlockJS in 38 out of 40 cases. Apart from being better than SureShrink in more than 75% of cases in mean squared error, $\hat{f}_{1,\lambda_1}$ yields noticeably better results visually. See Section 6.3 for a qualitative comparison and see Cai (1999b) for more simulation results. Cai and Silverman (2001) propose two block shrinkage estimators, NeighBlock and NeighCoeff, which are shown to perform well against well-known conventional wavelet estimators. Our numerical study shows that the estimator $\hat{f}_{1,\lambda_1}$ outperforms NeighBlock slightly and is comparable to NeighCoeff. For reasons of space, we omit the detailed numerical comparisons here.

Table 2. Average Mean Squared Error From 500 Replications (SNR=5).

n	$\hat{f}_{1,\lambda_1}$	Risk	Sure	TI	BJS	$\hat{f}_{1,\lambda_1}$	Risk	Sure	TI	BJS
	Doppler					*HeaviSine*				
512	0.98	1.72	1.59	2.64	1.21	0.56	0.47	0.50	0.52	0.55
1024	0.63	1.17	0.89	1.66	0.74	0.36	0.33	0.34	0.39	0.39
2048	0.32	0.77	0.54	1.02	0.40	0.19	0.23	0.23	0.28	0.23
4096	0.17	0.45	0.34	0.56	0.21	0.13	0.14	0.13	0.16	0.17
8192	0.08	0.29	0.18	0.34	0.11	0.07	0.09	0.07	0.10	0.09
	Bumps					*Blocks*				
512	1.81	4.99	2.23	7.53	2.37	1.98	3.05	2.61	5.35	2.64
1024	1.23	3.16	1.69	4.50	1.66	1.18	2.08	1.59	3.61	1.67
2048	0.74	2.04	1.12	2.70	0.90	0.77	1.48	1.04	2.40	1.02
4096	0.55	1.18	0.57	1.47	0.72	0.67	0.94	0.71	1.39	0.92
8192	0.28	0.74	0.34	0.86	0.38	0.41	0.64	0.44	0.89	0.56
	Spikes					*Blip*				
512	0.82	1.45	1.05	2.11	1.08	0.43	0.56	0.63	0.75	0.53
1024	0.50	0.94	0.56	1.25	0.55	0.26	0.40	0.42	0.51	0.32
2048	0.30	0.61	0.33	0.72	0.32	0.17	0.27	0.24	0.32	0.18
4096	0.16	0.34	0.15	0.30	0.22	0.09	0.16	0.15	0.19	0.13
8192	0.08	0.21	0.08	0.17	0.10	0.05	0.10	0.09	0.11	0.07
	Corner					*Wave*				
512	0.36	0.35	0.29	0.30	0.38	0.56	1.77	2.95	2.62	0.90
1024	0.21	0.21	0.17	0.20	0.24	0.31	1.04	3.20	1.56	0.38
2048	0.12	0.13	0.09	0.12	0.13	0.17	0.62	3.38	0.90	0.16
4096	0.06	0.06	0.05	0.06	0.07	0.09	0.25	0.09	0.11	0.09
8192	0.04	0.04	0.03	0.03	0.04	0.06	0.16	0.06	0.06	0.06

6.3. A qualitative example

The sunspots data are well-known and have been analyzed, for example, by Anderson (1971), Brockwell and Davis (1991) and recently by Efromovich (1999). We consider 1024 consecutive monthly means of daily numbers of sunspots from January, 1749 to March, 1834. (The data is available in the standard Splus package.) See Figure 3. Using the model in Section 3.1, we can envision the true underlying function f as the superposition of two components: a smooth part f_s and a high frequency oscillation part τ. In this example, the smooth part f_s can be thought of as the well-known periodic, seasonal component (with a period of about 11 years).

Three wavelet estimators, $\hat{f}_{1,\lambda_1}$, VisuShrink, and SureShrink, are applied to the data. Figure 4 displays the reconstructions, and their residuals, of $\hat{f}_{1,\lambda_1}$, SureShrink and VisuShrink. The $\hat{f}_{1,\lambda_1}$ reconstruction shows remarkable spatial adaptivity The reconstruction is smooth near the valleys and the sixth peak where the volatility is low; at the same time, it captures the high frequency oscillation

part very well near the other peaks where the volatility is high. The estimator $\hat{f}_{1,\lambda_1}$ permits the balance between variance and bias to be varied along the curve. The reconstruction confirms the theoretical results derived in Section 5.

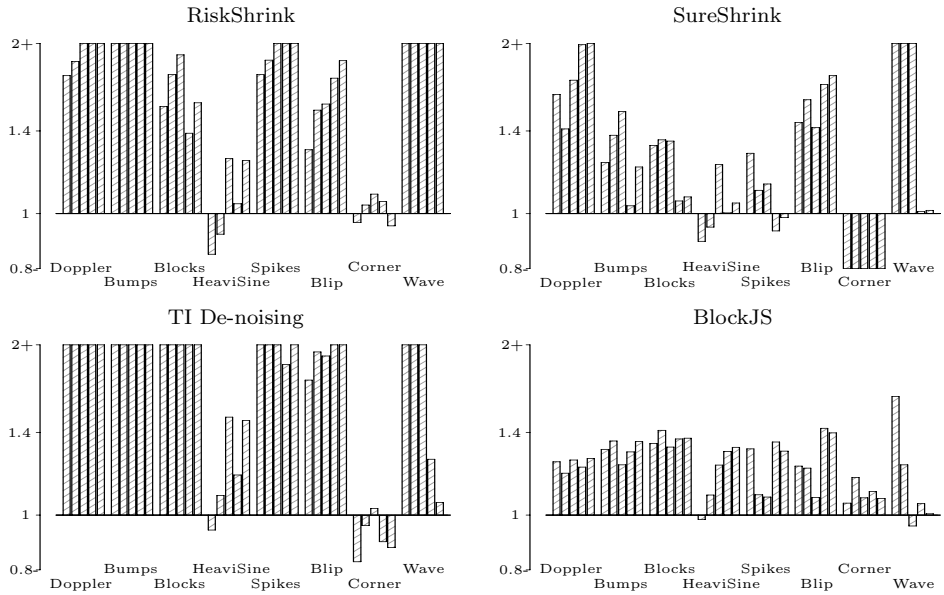


Figure 2. The vertical bars represent the ratios of the AMSEs of the estimators to the corresponding AMSE of $\hat{f}_{1,\lambda_1}$. The higher the bar the better the relative performance of $\hat{f}_{1,\lambda_1}$. For each signal the bars are ordered from left to right by the sample sizes ($n = 512$ to 8192).

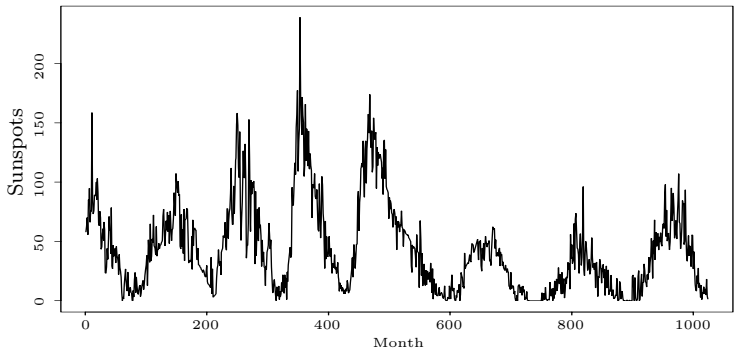


Figure 3. Monthly number of sunspots from January, 1749 to March, 1834.

In comparison, VisuShrink over-smoothes the data; it captures the smooth seasonal component well but misses almost all the fine details. It does not show the local oscillations around the peaks. SureShrink performs better than Vis-

uShrink. But SureShrink smoothes out some oscillations around the peaks, noticeably near the fourth and the seventh peaks, but still retains a fair amount of noise near the valleys. The reconstructions of VisuShrink and SureShrink fail to show the significant difference in volatilities between the peaks and valleys. The reconstructions of RiskShrink and TI de-noising, not shown here for the reason of space, are very similar to that of VisuShrink.

A look at the residual plots is also revealing. The residuals of both VisuShrink and SureShrink have a clear pattern−they cluster around the peaks; in comparison the residuals of $\hat{f}_{1,\lambda_1}$ are much more uniform. SureShrink keeps many wavelet coefficients at the high resolution levels around the areas in which the underlying function is smooth. In fact, an examination of the wavelet coefficients shows that SureShrink uses 345 coefficients while $\hat{f}_{1,\lambda_1}$ keeps 63 blocks of size 4 with a total of 252 coefficients.

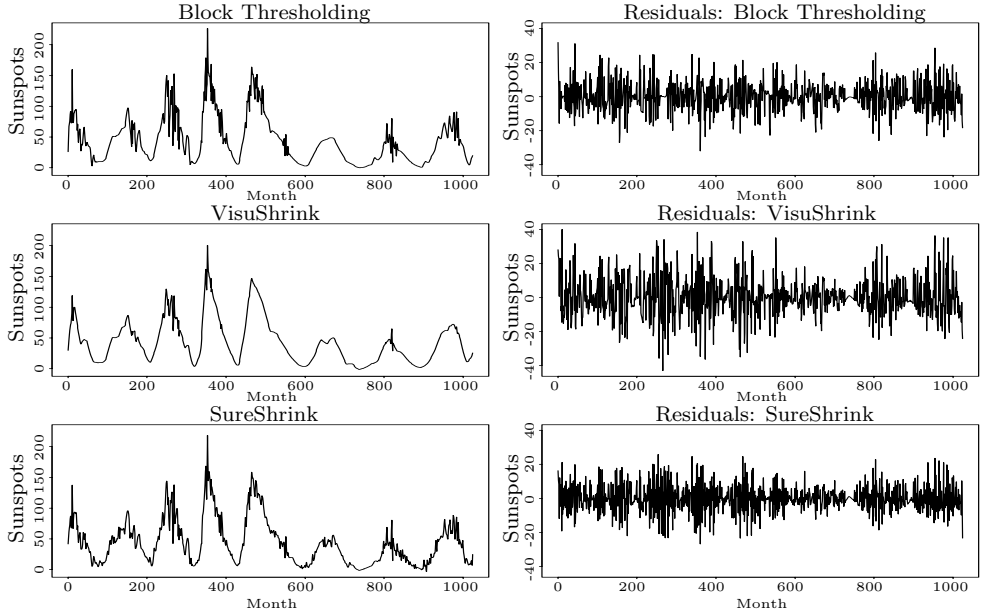


Figure 4. Comparison of reconstructions and residuals. The block thresholding estimator used here is $\hat{f}_{1,\lambda_1}$.

7. Discussions

7.1. Modifications and extensions

Modifications and extensions of the block thresholding estimators discussed in the earlier sections are possible. The modified estimators in many cases have better numerical performance than the original version. We briefly discuss two such modifications below.

Cai and Silverman (2001) introduce a technique for enhancing numerical performance of wavelet estimators by incorporating information on neighboring coefficients. The technique can be readily used on the block thresholding estimators discussed in this paper. For example, we can use the construction of the NeighBlock estimator in Cai and Silverman (2001) to obtain a new version of $\hat{f}_{1,\lambda_1}$. The procedure can be summarized in four steps.

1. Transform the data into the wavelet domain via the discrete wavelet transform.

2. At each resolution level j, group the empirical wavelet coefficients into disjoint blocks b_i^j of length $L_c = [(\log n)/2]$. Extend each block b_i^j by an amount $L_h = \max(1, [L_c/2])$ in each direction to form overlapping larger blocks B_i^j of length $L_1 = L_c + 2L_h$.

3. If the sum of the squared empirical coefficients in the larger block B_i^j is above the threshold $T = \lambda_1 L_1 \sigma^2$, then all the coefficients in the smaller block b_i^j are retained; otherwise all the coefficients in b_i^j are discarded.

4. Obtain the estimate of the function via the inverse discrete wavelet transform of the denoised wavelet coefficients.

We can envision B_i^j as a sliding window which moves L_c positions each time and, for each given window, only the half of the coefficients in the center of the window are estimated. Let us denote the resulting estimator by $\hat{f}^*_{1,\lambda_1}$. This estimator often has numerical advantages over the original version. See Figure 5.

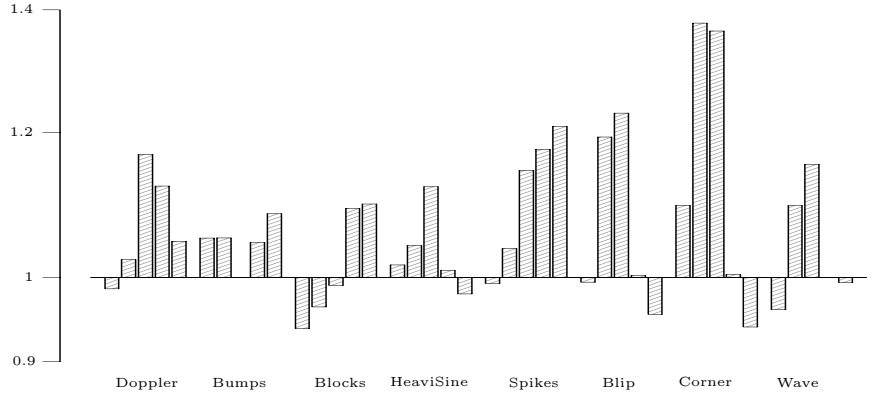


Figure 5. The bars represent the ratios of the AMSE of $\hat{f}_{1,\lambda_1}$ to the AMSE of $\hat{f}^*_{1,\lambda_1}$. For each signal the bars are ordered from left to right by the sample sizes ($n = 512$ to 8192).

The block thresholding estimators can also be modified by averaging over different block centers. For each given $0 \le i \le L-1$, partition the indices at each resolution level j into blocks $\{(j,k) : (b-1)L+i+1 \le k \le bL+i\}$. In the original estimator, we take $i=0$. Let $\hat{f}^{(i)}_{s,\lambda_s}$ be the version of $\hat{f}_{s,\lambda_s}$ for a given i. Define the modified estimator $\hat{f}^{**}_{s,\lambda_s} = \sum_{i=0}^{L-1} \hat{f}^{(i)}_{s,\lambda_s}/L$. This technique was also used in Hall, Penev, Kerkyacharian and Picard (1997).

Other possible modifications include incorporating block thresholding with translation-invariant denoising (Coifman and Donoho (1995)). These modified estimators often have better numerical performance, at the cost of higher computational complexity. For reasons of space, we leave the detailed numerical study to future work.

7.2. Concluding remarks

We study the effect of block size on global and local adaptivity and derive the optimal rate of convergence for block thresholding with a given choice of block size for both the global and local estimation. The results lead naturally to a possible optimal choice of block thresholding estimator. Asymptotic and numerical results show that the estimator $\hat{f}_{1,\lambda_1}$ with block size $L = \log n$ and thresholding constant $\lambda_1 = 4.5052$ indeed enjoys excellent performance both among the class of block thresholding estimators and in comparisons to other wavelet estimators.

Block thresholding is a way to pool information on neighboring coefficients for simultaneous decision-making. It is shown in Cai (2000b) that information pooling is a necessity rather than an option for achieving optimal adaptivity. For instance, no separable rules can achieve the optimal rate of convergence adaptively for global estimation. Block thresholding provides an easy and convenient tool for information pooling.

In the present paper, our main concern is with the nonparametric regression estimation of a function observed at regular intervals with independent homoscedastic Gaussian noise. Our results rely on these assumptions. Detailed study under more general structures on both design and errors is an interesting topic for future work.

Besides nonparametric regression, block thresholding techniques can be applied to other statistical problems such as linear inverse problems (see Cai (2000a) and Cavalier and Tsybakov (2000)). For example, block thresholding can be used to improve the asymptotic results obtained in Abramovich and Silverman (1998) for linear inverse problems.

8. Proofs

8.1. Preparatory results

We prove the main results in the order of Theorems 4, 1, 2, and 3. A key result used in the proofs is Proposition 1 which is proved at the end. Besides Proposition 1, we also need a number of preparatory results given below.

Proposition 1. *Suppose that* $x_i \overset{ind.}{\sim} N(\theta_i,\ \sigma_*^2),\ i = 1, \ldots, L$. *Let* $\hat{\theta}_i = x_i\ I(S^2 > \lambda L\sigma_*^2)$, *where* $S^2 = \sum_{i=1}^{L} x_i^2$ *and* $\lambda \geq 4$. *Then*

$$E\|\hat{\theta} - \theta\|_2^2 \leq (2\lambda + 2)(\|\theta\|_2^2 \wedge L\sigma_*^2) + 2\lambda L(\lambda^{-1}e^{\lambda-1})^{-L/2}\sigma_*^2. \tag{24}$$

In particular, if $\lambda = 4.5052$, *the root of* $\lambda - \log\lambda - 3 = 0$, $L = \log n$ *and* $\sigma_*^2 = n^{-1}\sigma^2$, *then*

$$E\|\hat{\theta} - \theta\|_2^2 \leq (2\lambda + 2)(\|\theta\|_2^2 \wedge L\sigma^2) + 2\lambda\sigma^2 n^{-2}\log n. \tag{25}$$

The second term on the right hand side of (25) is negligible. Thus the risk inequality shows that the estimator achieves, within a constant factor, the optimal balance between the variance and the squared bias over the blocks.

Lemma 1. (i) *Let* $f \in \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, v)$. *Assume the wavelets* $\{\phi, \psi\} \in W(D)$ *with* $supp(\phi) = supp(\psi) \subseteq [0, v]$. *Let* $n = 2^J$. *Then*

$$\begin{aligned}
|\xi_{J,k} - n^{-\frac{1}{2}} f(k/n)| &\leq M_1\|\phi\|_1 n^{-(1/2+\alpha_1)} \quad \textit{for all } k \in A_J;\\
|\xi_{J,k} - n^{-\frac{1}{2}} f(k/n)| &\leq M_2\|\phi\|_1 n^{-(1/2+\alpha)} \quad \textit{for all } k \notin A_J;\\
|\theta_{j,k}| &\leq M_1\|\psi\|_1 2^{-j(1/2+\alpha_1)} \textit{ for all } k \in A_j;\\
|\theta_{j,k}| &\leq M_2\|\psi\|_1 2^{-j(1/2+\alpha)} \quad \textit{for all } k \notin A_j.
\end{aligned}$$

(ii) *For all functions* $f \in \Lambda^\alpha(M)$, *the wavelet coefficients of* f *satisfy* $|\theta_{j,k}| \leq C' \cdot 2^{-j(1/2+\alpha)}$ *where the constant* C' *depends on the wavelets,* α *and* M *only.*

Lemma 1 (i) is a direct consequence of the vanishing moments conditions on the wavelets $\{\phi,\ \psi\}$, see Hall, Kerkyacharian and Picard (1999a). Lemma 1 (ii) bounds the wavelet coefficients of a function based on the smoothness, see, e.g., Daubechies (1992).

Lemma 2. *If* $\|u\|_{\ell_2}^2 \leq \gamma^2 t$ *with* $0 < \gamma < 1$, *then*
(i) $\{x :\ \|x + u\|_{\ell_2}^2 \leq t\} \supseteq \{x :\ \|x\|_{\ell_2}^2 \leq (1-\gamma)^2 t\}$;
(ii) $\{x :\ \|x + u\|_{\ell_2}^2 \geq t\} \subseteq \{x :\ \|x\|_{\ell_2}^2 \geq (1-\gamma)^2 t\}$.

Lemma 3. *Let* Y *and* X_i *be random variables, then*

$$\text{(i)}\quad E(\textstyle\sum X_i)^2 \leq (\sum (EX_i^2)^{1/2})^2; \tag{26}$$

$$\text{(ii)}\quad (E(Y + \textstyle\sum X_i)^2)^{1/2} \geq (EY^2)^{1/2} - \sum (EX_i^2)^{1/2}. \tag{27}$$

Lemma 4. *Let $Y_L \sim \chi^2_L$ and $\lambda > 1$. Then*
(i) $\frac{2}{5}\lambda^{-1}L^{-1/2}(\lambda^{-1}e^{\lambda-1})^{-L/2} \le P(Y_L > \lambda L)$
$\le \pi^{-1/2}(\lambda-1)^{-1}L^{-1/2}(\lambda^{-1}\,e^{\lambda-1})^{-L/2}$; (28)
(ii) $EY_L I(Y_L \ge \lambda L) \le \lambda L\,(\lambda^{-1}\,e^{\lambda-1})^{-L/2}$. (29)

Proof. Denote by $f_m(y)$ the pdf of a χ^2_m variable. Integration by parts yields $P(Y_m > x) = 2f_m(x) + P(Y_{m-2} > x)$ and by recursion, $P(Y_L > \lambda L) \le 2\sum_{k=0}^{[(L-1)/2]} f_{L-2k}(\lambda L)$. It is easy to see that, for $m \le L$, $f_m(\lambda L) = \frac{m}{\lambda L} f_{m+2}(\lambda L) \le \lambda^{-1} f_{m+2}(\lambda L)$. Then

$$P(Y_L > \lambda L) \le 2 \sum_{k=0}^{[(L-1)/2]} \lambda^{-k} f_L(\lambda L) \le \frac{2\lambda}{\lambda - 1} \cdot \frac{1}{2^{L/2}\Gamma(L/2)} (\lambda L)^{L/2-1} e^{-\lambda L/2}.$$

Now Stirling's formula in the form $\Gamma(x+1) = \sqrt{2\pi}\; x^{x+1/2}\; e^{-x+\theta/(12x)}$, with $0 < \theta < 1$, yields $P(Y_L > \lambda L) \le \pi^{-1/2}(\lambda-1)^{-1}L^{-1/2}(\lambda^{-1}e^{\lambda-1})^{-L/2}$. On the other hand,

$$P(Y_L \ge \lambda L) = \frac{1}{2^{L/2}\,\Gamma(L/2)} \int_{\lambda L}^{\infty} x^{L/2-1} e^{-\frac{x}{2}}\, dx \ge \frac{(\lambda L)^{L/2-1}\; 2e^{-\lambda L/2}}{2^{L/2}\,\Gamma(L/2)}.$$

Again, it follows from Stirling's formula, after some simple algebra, that $P(Y_L \ge \lambda L) \ge \frac{2}{5}\,\lambda^{-1}\,L^{-1/2}\,(\lambda^{-1}\,e^{\lambda-1})^{-L/2}$. The proof of (29) is straightforward.

8.2. Proof of Theorem 4

We first consider global estimation. Denote $L_* = \log n$ and $\lambda_* = 4.5052$. For simplicity, in all the proofs we assume that the sample size n is divisible by the block size L. Let $\tilde{Y}$ be the discrete wavelet transform of $\{n^{-1/2}Y\}$ and be written as in (4). One may write

$$\tilde{y}_{j,k} = \theta_{j,k} + a_{j,k} + n^{-1/2}\sigma z_{j,k} \tag{30}$$

where $\theta_{j,k}$ is the true wavelet coefficients of f, $a_{j,k}$ is some approximation error which is considered "small" by the results of Lemma 1 (i), and $z_{j,k}$'s are i.i.d. $N(0,1)$. Denote $\tilde{f}(x) = \sum_{i=1}^n n^{-1/2} y_i \phi_{Ji}(x)$. The function $\tilde{f}(x)$ can be written as

$$\begin{aligned}\tilde{f}(x) &= \sum_{i=1}^{n} [\xi_{Ji} + (n^{-1/2}f(x_i) - \xi_{Ji}) + n^{-1/2}\sigma z_i]\phi_{Ji}(x) \\ &= \sum_{k=1}^{2^{j_0}} [\xi_{j_0k} + \tilde{a}_{j_0k} + n^{-1/2}\sigma\tilde{z}_{j_0k}]\phi_{j_0k}(x) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} [\theta_{j,k} + a_{j,k} + n^{-1/2}\sigma z_{j,k}]\psi_{j,k}(x).\end{aligned}$$

Here, ξ_{j_0k} and $\theta_{j,k}$ are the orthogonal transform of $\{\xi_{Ji}\}$ via W, likewise $\tilde{a}_{j_0k}$ and $a_{j,k}$ the transform of $\{n^{-1/2}f(x_i) - \xi_{Ji}\}$, and $\tilde{z}_{j_0k}$ and $z_{j,k}$ the transform of

$\{z_i\}$. Thus $\tilde{z}_{j_0k}$ and $z_{j,k}$ are i.i.d. $N(0,1)$. Let $\tilde{\xi}_{j_0k} = \xi_{j_0k} + \tilde{a}_{j_0k} + n^{-1/2}\sigma\tilde{z}_{j_0k}$ and $\tilde{y}_{j,k} = \theta_{j,k} + a_{j,k} + n^{-1/2}\sigma z_{j,k}$. Lemma 1 (i) and the orthogonality of the discrete wavelet transform yield that

$$\sum_{k=1}^{2^{j_0}} \tilde{a}_{j_0k}^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} a_{j,k}^2 = \sum_{i=1}^{n}(n^{-1/2}f(x_i) - \xi_{Ji})^2 = o(n^{-2\alpha/(1+2\alpha)}). \tag{31}$$

See Hall, Kerkyacharian and Picard (1999a, p.43) for more details on the derivation of 31. Let $\hat{\xi}_{j_0k} = \tilde{\xi}_{j_0k}$ and $\hat{\theta}_{j,k} = \tilde{y}_{j,k}I(S_{jb}^2 > \lambda_* L_* n^{-1}\sigma^2)$, for $(j,k) \in (jb)$. By the isometry of the function norm and the sequence norm, the risk of $\hat{f}_{1,\lambda_1}$ can be written as

$$E\|\hat{f}_{1,\lambda_1} - f\|_2^2 = \sum_k E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 + \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2 + \sum_{j=J}^{\infty}\sum_k \theta_{j,k}^2. \tag{32}$$

Lemma 1 (i) and (31) yield that

$$\sum_k E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 + \sum_{j=J}^{\infty}\sum_k \theta_{j,k}^2 = o(n^{-2\alpha/(1+2\alpha)}). \tag{33}$$

Denote by C a generic constant that varies from place to place and let

$G_j = \{$blocks at level j contain at least one coefficient with indices in $A_j\}$;
$G_j' = \{$blocks at level j contain no coefficients with indices in $A_j\}$.

The term $S \equiv \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2$ can be bounded by using Proposition 1 and 31.

$$\begin{aligned} S &\le (2\lambda_* + 2)\sum_{j=j_0}^{J-1}\sum_k (\theta_{j,k} + a_{j,k})^2 \wedge L_* n^{-1}\sigma^2 + \lambda_* L_* n^{-1}\sigma^2 \\ &\le C\sum_{j=j_0}^{J-1}\sum_k \theta_{j,k}^2 \wedge L_* n^{-1} + o(n^{-2\alpha/(1+2\alpha)}). \end{aligned}$$

Denote $S_1 = \sum_{j=j_0}^{J-1}\sum_{(jb)\in G_j}\sum_{(j,k)\in(jb)} \theta_{j,k}^2 \wedge L_* n^{-1}$; $S_2 = \sum_{j=j_0}^{J-1}\sum_{(jb)\in G_j'}$ $\sum_{(j,k)\in(jb)} \theta_{j,k}^2 \wedge L_* n^{-1}$. Note that $card(G_j) \le M_3 2^{j\gamma}$ and let J_1 and J_2 be two integers satisfying $2^{J_1} \asymp n^{1/(1+2\alpha_1)}$ and $2^{J_2} \asymp n^{1/(1+2\alpha)}$ respectively. Also note that by the assumptions $\delta \equiv \frac{1+2\alpha_1}{1+2\alpha} - \gamma > 0$, so

$$\begin{aligned} S_1 &\le \sum_{j=J_0}^{J_1-1}\sum_{(jb)\in G_j} L_* n^{-1} + \sum_{j=J_1}^{J-1}\sum_{(jb)\in G_j}\sum_{(j,k)\in(jb)} \theta_{j,k}^2 \\ &\le L_* n^{-1}2^{J_1\gamma} + CL_* 2^{-J_1(1+2\alpha_1-\gamma)} = o(n^{-2\alpha/(1+2\alpha)}), \end{aligned} \tag{34}$$

$$S_2 \le \sum_{j=J_0}^{J_2-1} \sum_{(jb)\in G_j'} L_* n^{-1} + \sum_{j=J_2}^{J-1} \sum_{(jb)\in G_j'} \sum_{(j,k)\in(jb)} \theta_{j,k}^2 \le C n^{-2\alpha/(1+2\alpha)}. \tag{35}$$

Now (22) follows from (33), (34) and (35).

Now consider local estimation. For brevity, we prove the result for Hölder class $\Lambda^\alpha(M)$. It follows from Lemma 3 (i) that

$$\begin{aligned} E(\hat{f}_{1,\lambda_1}(x_0) - f(x_0))^2 \le \Bigg\{ & \sum_{k=1}^{2^{j_0}} (E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2)^{1/2} |\phi_{j_0k}(x_0)| \\ & + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} (E(\hat{\theta}_{j,k} - \theta_{j,k})^2)^{1/2} |\psi_{j,k}(x_0)| \\ & + \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{j,k}||\psi_{j,k}(x_0)| \Bigg\}^2 \equiv (Q_1 + Q_2 + Q_3)^2. \end{aligned}$$

Let us consider the three terms separately. First note that at each resolution level j, there are at most N basis functions $\psi_{j,k}$ such that $\psi_{j,k}(x_0) \neq 0$, where N is the length of the support of ψ. Denote $K(j, x_0) = \{k : \psi_{j,k}(x_0) \neq 0\}$. Then $|K(j, x_0)| \le N$. Therefore,

$$Q_1 = \sum_{k=1}^{2^{j_0}} (E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2)^{1/2} |\phi_{j_0k}(x_0)| \le 2^{j_0/2} \|\phi\|_\infty N n^{-1/2} \sigma = o(n^{-\alpha/(1+2\alpha)}). \tag{36}$$

For the third term, it follows from Lemma 1 (ii) that

$$Q_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{j,k}||\psi_{j,k}(x_0)| \le \sum_{j=J}^{\infty} N \|\psi\|_\infty 2^{j/2} C 2^{-j(1/2+\alpha)} \le C n^{-\alpha}. \tag{37}$$

Consider the second term Q_2. Note that for function $f \in \Lambda^\alpha(M)$, the approximation error $a_{j,k}$ satisfies $|a_{j,k}| \le C n^{-\alpha} 2^{-j/2}$. By applying Lemma 1 (ii) and Proposition 1, we have

$$\begin{aligned} Q_2 &\le \sum_{j=j_0}^{J-1} \sum_{k\in K(j,x_0)} 2^{j/2} \|\psi\|_\infty (E(\hat{\theta}_{j,k} - \theta_{j,k})^2)^{1/2} \\ &\le C \sum_{j=j_0}^{J-1} 2^{j/2} [(2^{-j(1+2\alpha)} + 2^{-j} n^{-2\alpha}) \wedge L_* n^{-1} \sigma^2 + L_* n^{-2} \sigma^2]^{1/2} \\ &= C (\log n / n)^{\alpha/(1+2\alpha)}. \end{aligned} \tag{38}$$

Combining (36), (37) and (38), we have $E(\hat{f}_{1,\lambda_1}(x_0) - f(x_0))^2 \le C(\log n/n)^{2\alpha/(1+2\alpha)}$.

8.3. Proof of theorem 1

We prove only part (i) in detail. The proof of part (ii) is similar to that of Theorem 4 (i). Denote $w = 2\alpha/(1+2\alpha)$ and $s = 1-\gamma$ with $0 < \gamma \le 1$. The proof is divided into two cases.

Case 1. For all $n > 0$, the threshold $\lambda_n > w(\log n)^\gamma$. Let J_1 be an integer such that $2^{J_1} \asymp n^{1/(1+2\alpha)}(\log n)^{-\gamma/(1+2\alpha)}$. Let $f_n(t) = \sum_k \theta_{J_1k}\psi_{J_1k}(t)$ where $\theta_{J_1k} = c_0(\log n)^{\gamma/2}n^{-1/2} \asymp 2^{-J_1(1/2+\alpha)}$ with $c_0 > 0$. Then $f_n \in \Lambda^\alpha(M)$ when the constant c_0 is chosen small enough. We again use the decomposition (30). Since $f_n \in \Lambda^\alpha(M)$, Lemma 1 (i) and the fact that the wavelets are compactly supported yield that $|d(x)| \equiv |\sum_{k=1}^n [\xi_{J,k} - n^{-1/2}f(k/n)]\phi_{J,k}(x)| \le Cn^{-\alpha}$. Hence the approximation error satisfies $|a_{j,k}| \equiv |\int d(x)\psi_{j,k}(x)dx| \le Cn^{-\alpha}2^{-j/2}$. For a given block (J_1b) at level J_1,

$$\sum_{(j,k)\in(J_1b)} E(\hat\theta_{jk}-\theta_{jk})^2 \ge \sum_{(j,k)\in(J_1b)} (\frac{1}{2}E[\hat\theta_{jk}-(\theta_{jk}+a_{j,k})]^2 - a_{jk}^2)$$

$$= \sum_{(j,k)\in(J_1b)} [\frac{1}{2}E(\tilde y_{jk}-\theta_{jk})^2 I(S_{J_1b}^2 > \lambda_n L) + \frac{1}{2}(\theta_{j,k}+a_{j,k})^2 P(S_{J_1b}^2 \le \lambda_n L) - a_{jk}^2]$$

$$\ge \frac{1}{4}\sum_{(j,k)\in(J_1b)} \theta_{j,k}^2 P(S_{J_1b}^2 \le \lambda_n L) - 2\sum_{(j,k)\in(J_1b)} a_{jk}^2 \tag{39}$$

To get a lower bound for $P(S_{J_1b}^2 \le \lambda_n L)$, we apply Lemma 2 to $S_{J_1b}^2$. Since $|a_{j,k}| \le Cn^{-\alpha}2^{-j/2}$, $\sum_{(J_1b)} a_{j,k}^2 \le Cn^{-1-4\alpha^2/(1+2\alpha)}(\log n)^{(1-2\alpha\gamma)/(1+2\alpha)}$. Hence there exists $N > 0$ such that for $n > N$, $\sum_{(J_1b)} a_{j,k}^2 \le \frac{1}{16}\lambda_n L n^{-1}\sigma^2$. Now $\sum_{(J_1b)} \theta_{j,k}^2 = c_0^2 n^{-1}\log n$, so for small $c_0 > 0$, $\sum_{(J_1b)} \theta_{j,k}^2 \le \frac{1}{16}\lambda_n L n^{-1}\sigma^2$. Choosing the constant $c_0 > 0$ small enough, we have, for $n > N$,

$$\sum_{(J_1b)} (\theta_{j,k}+a_{j,k})^2 \le 2\sum_{(J_1b)} \theta_{j,k}^2 + 2\sum_{(J_1b)} a_{j,k}^2 \le \frac{1}{4}\lambda_n L n^{-1}\sigma^2.$$

Then it follows from Lemma 2 that

$$\{S_{J_1b}^2 \le \lambda L n^{-1}\sigma^2\} = \Big\{\sum_{(J_1b)} (\theta_{j,k}+a_{j,k}+n^{-1/2}\sigma z_{j,k})^2 \le \lambda_n L n^{-1}\sigma^2\Big\} \supseteq \Big\{\sum_{(J_1b)} z_{j,k}^2 \le \frac{1}{4}\lambda_n L\Big\}.$$

For large n, $\lambda_n/4 \ge (w/4)(\log n)^\gamma \ge 2$. Hence

$$P(S_{J_1b}^2 \le \lambda_n L n^{-1}\sigma^2) \ge P(\sum_{(J'b)} z_{j,k}^2 \le 2L) \ge 1/2. \tag{40}$$

Combining (39) and (40), we have, for large n,

$$E\|\hat f_n - f_n\|^2 \ge \sum_k E(\hat\theta_{J_1k}-\theta_{J_1k})^2 \ge \frac{1}{8}\sum_k \theta_{J_1k}^2 - 2\sum_k a_{J_1k}^2$$
$$= (c_0^2/8)(n/\log^\gamma n)^{-2\alpha/(1+2\alpha)}(1+o(1)).$$

So, in this case, $\overline{\lim}_{n\to\infty} n^{\frac{2\alpha}{1+2\alpha}} \cdot (\log n)^{-\frac{2\alpha\gamma}{1+2\alpha}} \cdot \sup_{f\in\Lambda^\alpha(M)} E\|\hat{f}_n - f\|_2^2 \geq c_0^2/8 > 0$.

Case 2. There exists a subsequence (n_m) such that the threshold $\lambda_{n_m} \leq w(\log n_m)^\gamma$.

Without loss of generality, assume that $\lambda_n \leq w(\log n)^\gamma$ for all n. Consider $f_n \equiv 0$. Then all $\theta_{j,k} = 0$ and all $a_{j,k} = 0$ and for each block (jb), $\sum_{(j,k)\in(jb)} E(\hat{\theta}_{j,k} - \theta_{j,k})^2 = n^{-1}\sigma^2 EYI(Y > \lambda_n L)$, where $Y \sim \chi_L^2$. Hence

$$E\|\hat{f}_n - f_n\|_2^2 \geq \sum_{j=j_0}^{J-1} \sum_b \sum_{(j,k)\in(jb)} E(\hat{\theta}_{j,k} - \theta)^2 = (n - 2^{j_0})n^{-1}\sigma^2 EYI(Y > \lambda_n L).$$

Let $\lambda_n' = \max(\lambda_1, 1)$, then Lemma 4 yields

$$EYI(Y > \lambda_n L) \geq \lambda_n' LP(Y > \lambda_n' L) \geq \frac{2}{5} L^{1/2} (\lambda_n' e)^{L/2} n^{-r/2}.$$

Hence in this case $\overline{\lim}_{n\to\infty} n^{\frac{2\alpha}{1+2\alpha}} \cdot (\log n)^{-\frac{2\alpha\gamma}{1+2\alpha}} \cdot \sup_{f\in\Lambda^\alpha(M)} E\|\hat{f}_n - f\|_2^2 = \infty$.

8.4. Proof of Theorem 2

We give the proof of part (ii) in detail. With the thresholding constant λ_s chosen as in Section 4, the proof of part (i) is similar to that of Theorem 4 (ii). Let J' be an integer satisfying $2^{J'} \asymp (n/L)^{1/(1+2\alpha)}$ and let k' be an integer such that $|\psi(2^{J'}x_0 - k')| \geq c_0 > 0$. Let $f_n^*(x) = \theta_{J'k'}^* \psi_{J'k'}(x)$ where $\theta_{J'k'}^* = c_1(n^{-1}L)^{1/2} \asymp 2^{-J'(1/2+\alpha)}$. The function f_n^* has only one "large" wavelet coefficient and all other coefficients are zero. It is easy to show that $f_n \in \Lambda^\alpha(M)$ if the constant $c_1 > 0$ is sufficiently small. Noting that $\xi_{j_0k} = \langle f_n^*, \phi_{j,k}\rangle = 0$ for all k and $\theta_{j,k} = \langle f_n^*, \psi_{j,k}\rangle = 0$ for all $(j,k) \neq (J',k')$, we have

$$\begin{aligned}\mathcal{S} &\equiv \{\sup_{f\in\Lambda^\alpha(M)} E_f(\hat{f}_n(x_0) - f(x_0))^2\}^{1/2} \geq (E_{f_n^*}(\hat{f}_n(x_0) - f_n^*(x_0))^2)^{1/2} \\ &= \left\{E[(\hat{\theta}_{J'k'} - \theta_{J'k'})\psi_{J'k'}(x_0) + \sum_{k=1}^{2^{j_0}} \hat{\xi}_{j_0k}\phi_{j_0k}(x_0) + \sum_{(j,k)\in\mathcal{J}} \hat{\theta}_{j,k}\psi_{j,k}(x_0)]^2\right\}^{1/2} \quad (41)\end{aligned}$$

where $\mathcal{J} = \{(j,k) : j_0 \leq j \leq J-1, 1 \leq k \leq 2^j \text{ and } (j,k) \neq (J',k')\}$. Applying Lemma 3 (ii) to the RHS of (41), we have

$$\begin{aligned}\mathcal{S} &\geq (E(\hat{\theta}_{J'k'} - \theta_{J'k'})^2)^{1/2}|\psi_{J'k'}(x_0)| - \sum_{k=1}^{2^{j_0}} (E\hat{\xi}_{j_0k}^2)^{1/2}|\phi_{j_0k}(x_0)| - \sum_{(j,k)\in\mathcal{J}} (E\hat{\theta}_{j,k}^2)^{1/2}|\psi_{j,k}(x_0)| \\ &\equiv T_1 - T_2 - T_3. \quad (42)\end{aligned}$$

We show that the first term T_1 is dominating and that T_2 and T_3 are "small". We first derive a lower bound for T_1. Denote by $(J'b)$ the block containing (J', k'), then

$$E(\hat{\theta}_{J'k'}-\theta_{J'k'})^2=E(\tilde{y}_{J'k'}-\theta_{J'k'})^2 I(S^2_{J'b}>\lambda L n^{-1}\sigma^2)+\theta^2_{J'k'}P(S^2_{J'b}\le\lambda L n^{-1}\sigma^2)$$
$$\ge \theta^2_{J'k'}P(S^2_{J'b}\le \lambda L n^{-1}\sigma^2). \tag{43}$$

As in the proof of Theorem 1, we apply Lemma 2 to $S^2_{J_1b}$ to get a lower bound for $P(S^2_{J_1b}\le\lambda_n L)$. Since $|a_{j,k}|\le Cn^{-\alpha}2^{-j/2}$, $\sum_{(J'b)}a^2_{j,k}\le Cn^{-1-4\alpha^2/(1+2\alpha)}L^{(2+2\alpha)/(1+2\alpha)}$. Hence there exists a constant $N_*>0$ such that for $n>N_*$

$$\sum_{(J'b)} a^2_{j,k}\le\frac{1}{4}(1-\lambda^{-1/2})^2\lambda L n^{-1}\sigma^2. \tag{44}$$

By choosing $c_1\le\frac{\sigma}{2}(\lambda^{1/2}-1)$, we have for $n>N_*$,

$$\sum_{(J'b)}(\theta_{j,k}+a_{j,k})^2\le 2\theta^2_{J'k'}+2\sum_{(J'b)}a^2_{j,k}\le(1-\lambda^{-1/2})^2\lambda L n^{-1}\sigma^2.$$

It follows from Lemma 2 that

$$\{S^2_{J'b}\le\lambda L n^{-1}\sigma^2\}=\{\sum_{(J'b)}(\theta_{j,k}+a_{j,k}+n^{-1/2}\sigma z_{j,k})^2\le\lambda L n^{-1}\sigma^2\}\supseteq\{\sum_{(J'b)}z^2_{j,k}\le L\}.$$

So, $P(S^2_{J'b}\le\lambda L n^{-1}\sigma^2)\ge P(\sum_{(J'b)}z^2_{j,k}\le L)\ge 1/2$. Now (43) yields $E(\hat{\theta}_{J'k'}-\theta_{J'k'})^2\ge\frac{1}{2}\theta^2_{J'k'}=\frac{1}{2}c_1^2 n^{-1}L$. Therefore

$$T_1=(E(\hat{\theta}_{J'k'}-\theta_{J'k'})^2)^{1/2}2^{J'/2}|\psi(2^{J'}x_0-k')|\ge\frac{1}{\sqrt{2}}c_0c_1n^{-\alpha/(1+2\alpha)}L^{\alpha/(1+2\alpha)}. \tag{45}$$

For T_2, as in the proof of Theorem 4 (ii), we have

$$T_2=\sum_{k=1}^{2^{j_0}}(E\hat{\xi}^2_{j_0k})^{1/2}|\phi_{j_0k}(x_0)|=o(n^{-\alpha/(1+2\alpha)}). \tag{46}$$

Now consider the term T_3. Let J_1 be an integer satisfying $2^{j_1}\asymp\max(1, n^{(1-\alpha^2)/(1+2\alpha)})$. Denote $\mathcal{J}_1=\{(j,k)\in\mathcal{J}\text{ and }j\le j_1\}$, and $\mathcal{J}_2=\{(j,k)\in\mathcal{J}\text{and}j>j_1\}$. First consider $(j,k)\in\mathcal{J}_1$. It is easy to see that

$$E\hat{\theta}^2_{j,k}=E\tilde{y}^2_{jk}I(S^2_{jb}>\lambda L n^{-1}\sigma^2)\le E\tilde{y}^2_{j,k}$$
$$=a^2_{j,k}+n^{-1}\sigma^2\le Cn^{-2\alpha}2^{-j}+n^{-1}\sigma^2. \tag{47}$$
$$T_{31}\equiv\sum_{(j,k)\in\mathcal{J}_1}(E\hat{\theta}^2_{j,k})^{1/2}|\psi_{j,k}(x_0)|$$
$$\le Cn^{-\alpha}\log n+Cn^{-1/2}2^{j_1/2}\log n=o(n^{-\alpha/(1+2\alpha)}). \tag{48}$$

Now consider $(j,k) \in \mathcal{J}_2$. In this case, similar to (44), for large n, we have

$$\sum_{(jb)} a_{j,k}^2 \le (1 - (\frac{1+\lambda}{2\lambda})^{1/2})^2 \lambda L n^{-1} \sigma^2.$$

It then follows from Lemma 2 that

$$\{S_{jb}^2 \ge \lambda L n^{-1}\sigma^2\} = \{\sum_{(jb)} (a_{j,k} + n^{-1/2}\sigma z_{j,k})^2 \ge \lambda L n^{-1}\sigma^2\} \subseteq \{\sum_{(jb)} z_{j,k}^2 \ge \frac{1}{2}(1+\lambda)L\}.$$

So for sufficiently n, we have

$$\begin{aligned} E\hat{\theta}_{j,k}^2 &= E\tilde{y}_{jk}^2 I(S_{jb}^2 > \lambda L n^{-1}\sigma^2) \le 2n^{-1}\sigma^2 E z_{j,k}^2 I(S_{jb}^2 > \lambda L n^{-1}\sigma^2) + 2a_{j,k}^2 \\ &\le 2n^{-1}\sigma^2 EY(Y \ge \frac{1}{2}(1+\lambda)L) + 2a_{j,k}^2, \end{aligned}$$

where $Y = \sum_{(jb)} z_{j,k}^2 \sim \chi_L^2$. Denote $\lambda_1 = (1+\lambda)/2$. Lemma 4 now yields

$$E(\hat{\theta}_{j,k})^2 \le 2n^{-1}\sigma^2\lambda_1 L\ (\lambda_1^{-1}\ e^{\lambda_1 - 1})^{-L/2} + 2a_{j,k}^2 \le 2n^{-1}\sigma^2\lambda_1 L\ \beta^{-L} + Cn^{-2\alpha}2^{-j}$$

where $\beta = (\lambda_1^{-1}\ e^{\lambda_1 - 1})^{1/2} > 1$, since $\lambda_1 > 1$. Hence

$$T_{32} \equiv \sum_{(j,k)\in\mathcal{J}_2} (E\hat{\theta}_{j,k}^2)^{1/2} |\psi_{j,k}(x_0)| \le C\beta^{-L/2} L^{1/2} + Cn^{-\alpha} L^{1/2} = o(n^{-\alpha/(1+2\alpha)}). \tag{49}$$

It follows by combining (48) and (49),

$$T_3 = T_{31} + T_{32} = o(n^{-\alpha/(1+2\alpha)}). \tag{50}$$

Putting together (45), (46), and (50), we have $\mathcal{S} \ge T_1 - T_2 - T_3 \ge \frac{1}{\sqrt{2}} c_0 c_1 n^{-\alpha/(1+2\alpha)}$ $L^{\alpha/(1+2\alpha)}(1 + o(1))$. Now (12) follows by letting $L = (\log n)^s$ with $s > 1$.

8.5. Proof of Theorem 3

Let $\kappa(\lambda) \equiv (\lambda - \log\lambda - 1)/2$ and rewrite (28) accordingly. First consider $s = 1$. Since $\kappa(\lambda) \ge 1$ for $\lambda \ge \lambda_s = 4.5052$, for $L = \log n$ and $\lambda \ge \lambda_s$, one has

$$p_L(\lambda) = 1 - (1 - P(Y_L > \lambda L))^{n/L} \le 1 - (1 - (\lambda - 1)^{-1} \log^{-1/2} n / n^{\kappa(\lambda)})^{n/\log n} \to 0.$$

On the other hand, if λ is a constant less than λ_s, then $\kappa(\lambda) < 1$ and it is easy to see that $p_L(\lambda) \to 1$. The case of $s = 0$ is similar.

Now consider other cases. Suppose $\lambda = \beta + \delta$ with $\delta = o(\beta)$. Then, using Taylor expansion, one has for any $M > 1$,

$$\kappa(\lambda) = \beta + \delta - \log\beta - 1 + \sum_{m=1}^{M-1} (-1)^m \frac{\delta^m}{m\beta^m} + O(\frac{\delta^M}{\beta^M}). \tag{51}$$

Consider $0 < s \le 1/2$. Let $\lambda_s = 2(\log n)^{1-s} + \log(2(\log n)^{1-s}) + 1$. Applying (51) with $\beta = 2(\log n)^{1-s}$ and $\delta = \log(2(\log n)^{1-s}) + 1$, one has, for large n, $\kappa(\lambda_s) \ge (\log n)^{1-s} - \frac{\log(2(\log n)^{1-s})}{2(\log n)^{1-s}}$. Hence,

$$e^{-L\cdot\kappa(\lambda_s)} \le \sqrt{2}n^{-1}(\log n)^{(1-s)/2}. \tag{52}$$

Note that (52) also holds for any $\lambda \ge \lambda_s \ge 1$, since $\kappa(\lambda_s)$ is strictly increasing for $\lambda \ge 1$. Thus, for $\lambda \ge \lambda_s$, $p_L(\lambda) \le 1 - (1 - (\lambda - 1)^{-1}/n)^{n/(\log n)^s} \to 0$. The other cases can be verified similarly by using (51).

8.6. Proof of Proposition 1

Denote $R(\hat\theta, \theta, \sigma_*) = E_{\sigma_*}\|\hat\theta - \theta\|_2^2$, and $\theta^* = \theta/\sigma_*$. Since $R(\hat\theta, \theta, \sigma_*) = \sigma_*^2 R(\hat\theta^*, \theta^*, 1)$, it suffices to consider only the case $\sigma_* = 1$. For brevity, we denote $R(\hat\theta, \theta, 1)$ by $R(\theta)$. It is easy to see that $R(\theta)$ is bounded from above by $(2\lambda+2)L$ since

$$R(\theta) = E\|xI(S^2 > \lambda L) - \theta\|_2^2 \le 2E\|x-\theta\|_2^2 + 2ES^2 I(S^2 \le \lambda L) \le (2\lambda+2)L. \tag{53}$$

On the other hand,

$$R(\theta) = E\|x-\theta\|_2^2 I(S^2 > \lambda L) + \|\theta\|_2^2 P_\theta(S^2 \le \lambda L) \le 2\|\theta\|_2^2 + 2ES^2 I(S^2 > \lambda L). \tag{54}$$

When $\|\theta\|_2^2 \ge L/2$, $ES^2 I(S^2 > \lambda L) \le ES^2 \le 3\|\theta\|_2^2$. So,

$$R(\theta) \le 8\|\theta\|_2^2, \quad \text{when } \|\theta\|_2^2 \ge L/2. \tag{55}$$

Now assume $\|\theta\|_2^2 < L/2$. Let $\mu = \|\theta\|_2^2$ and denote $g(\mu) = ES^2 I(S^2 > \lambda L)$. Denote by $f_{m,\mu}(y)$ the density of a noncentral χ^2-distribution with m degrees of freedom and noncentrality μ and denote $f_{m,0}(y)$ by $f_m(y)$. The pdf $f_{m,\mu}(y)$ has many representations (see, e.g., Johnson, Kotz and Balakrishnan (1995)). We need the Poisson form and the integral form:

$$f_{m,\mu}(y) = \sum_{k=0}^{\infty} \frac{(\mu/2)^k e^{-\mu/2}}{k!} f_{m+2k}(y), \tag{56}$$

$$f_{m,\mu}(y) = \frac{1}{2}\int_0^y [q(\sqrt{y-x} + \sqrt{\mu}) + q(\sqrt{y-x} - \sqrt{\mu})](y-x)^{-1/2} f_{m-1}(x)dx, \tag{57}$$

where $q(x)$ is the density of a standard normal distribution. Since S^2 has a noncentral χ^2 distribution with L degrees of freedom and noncentrality parameter μ, using (56), one has

$$g(\mu) = \sum_{k=0}^{\infty} \frac{(\mu/2)^k e^{-\mu/2}}{k!} EY_{L+2k} I(Y_{L+2k} > \lambda L), \tag{58}$$

where Y_m denotes a central χ^2 random variable with m degrees of freedom. Denote $a_k = EY_{L+2k}I(Y_{L+2k} > \lambda L)$ and differentiate both sides of (58): $g'(\mu) = \frac{1}{2}\sum_{k=0}^{\infty}\frac{(\mu/2)^k e^{-\mu}}{k!}(a_{k+1} - a_k)$. It is easy to verify that $a_{k+1} - a_k = 2P(Y_{L+2k+2} > \lambda L) + 2\lambda L f_{L+2k+2}(\lambda L)$. Therefore,

$$g'(\mu) = \sum_{k=0}^{\infty}\frac{(\mu/2)^k e^{-\mu/2}}{k!}\{P(Y_{L+2k+2} > \lambda L) + \lambda L f_{L+2k+2}(\lambda L)\} \le 1 + \lambda L f_{L+2,\mu}(\lambda L) \tag{59}$$

Now use the integral form (57) of $f_{L+2,\mu}$ to bound $\lambda L f_{L+2,\mu}(\lambda L)$.

$$\begin{aligned} f_{L+2,\mu}(\lambda L) &= \frac{1}{2}\int_0^{\lambda L}(q(\sqrt{\lambda L - x} + \sqrt{\mu}) + q(\sqrt{\lambda L - x} - \sqrt{\mu}))(\lambda L - x)^{-1/2} f_{L+1}(x)dx \\ &\le \frac{1}{2}\int_0^{(\lambda-2)L}(q(\sqrt{2L}) + q(\sqrt{2L} - \sqrt{L/2}))(2L)^{-1/2} f_{L+1}(x)dx \\ &\quad + \frac{1}{2}\int_{(\lambda-2)L}^{\lambda L}(\frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}}) f_{L+1}((\lambda-2)L)(\lambda L - x)^{-1/2}dx \\ &\le \frac{1}{4\sqrt{\pi}}L^{-1/2}(e^{-L} + e^{-L/4}) + \frac{1}{\sqrt{\pi}}L^{1/2} f_{L+1}((\lambda-2)L). \end{aligned}$$

Using Stirling's formula, after some algebra, one has $f_{L+1}((\lambda-2)L) \le \frac{1}{2\sqrt{2(\lambda-2)\pi}}\left(\frac{\lambda-2}{e^{\lambda-3}}\right)^{L/2}$. So,

$$\lambda L f_{L+2,\mu}(\lambda L) \le \frac{\lambda}{4\sqrt{\pi}}L^{1/2}(e^{-L} + e^{-L/4}) + \frac{\lambda}{2\pi\sqrt{2(\lambda-2)}}L^{3/2}\left(\frac{\lambda-2}{e^{\lambda-3}}\right)^{L/2}. \tag{60}$$

Some calculus shows that for $a, b > 0$,

$$L^{1/2}e^{aL} \le \sup_{x>0} xe^{-ax^2} = (2ae)^{-1/2}, \quad \text{and} \quad L^{3/2}b^{-L} \le \sup_{x>0} x^3 b^{-x^2} = (3/(2e\log b))^{3/2}. \tag{61}$$

Set $a = 1$ and $a = 1/4$, and let $b = (e^{\lambda-3}/(\lambda-2))^{1/2}$. It follows from (60) and (61) that

$$\begin{aligned} \lambda L f_{L+2,\mu}(\lambda L) &\le \frac{\lambda}{4\sqrt{\pi}}((2e)^{-1/2} + (2/e)^{1/2}) \\ &\quad + \frac{\lambda}{2\pi\sqrt{2(\lambda-2)}}\left(\frac{3}{e(\lambda - 3 - \log(\lambda-2))}\right)^{3/2} \le \lambda - 1, \end{aligned}$$

for $\lambda \ge 4$. Now (59) yields $g'(\mu) \le \lambda$ and consequently $g(\mu) \le \lambda\mu + g(0) = \lambda\mu + EY_L I(Y_L > \lambda L)$. It now follows from Lemma 4 and (54) that

$$R(\theta) \le (2\lambda + 2)\|\theta\|_2^2 + 2\lambda L(\lambda^{-1}e^{\lambda-1})^{-L/2}, \quad \text{when } \|\theta\|_2^2 < L/2. \tag{62}$$

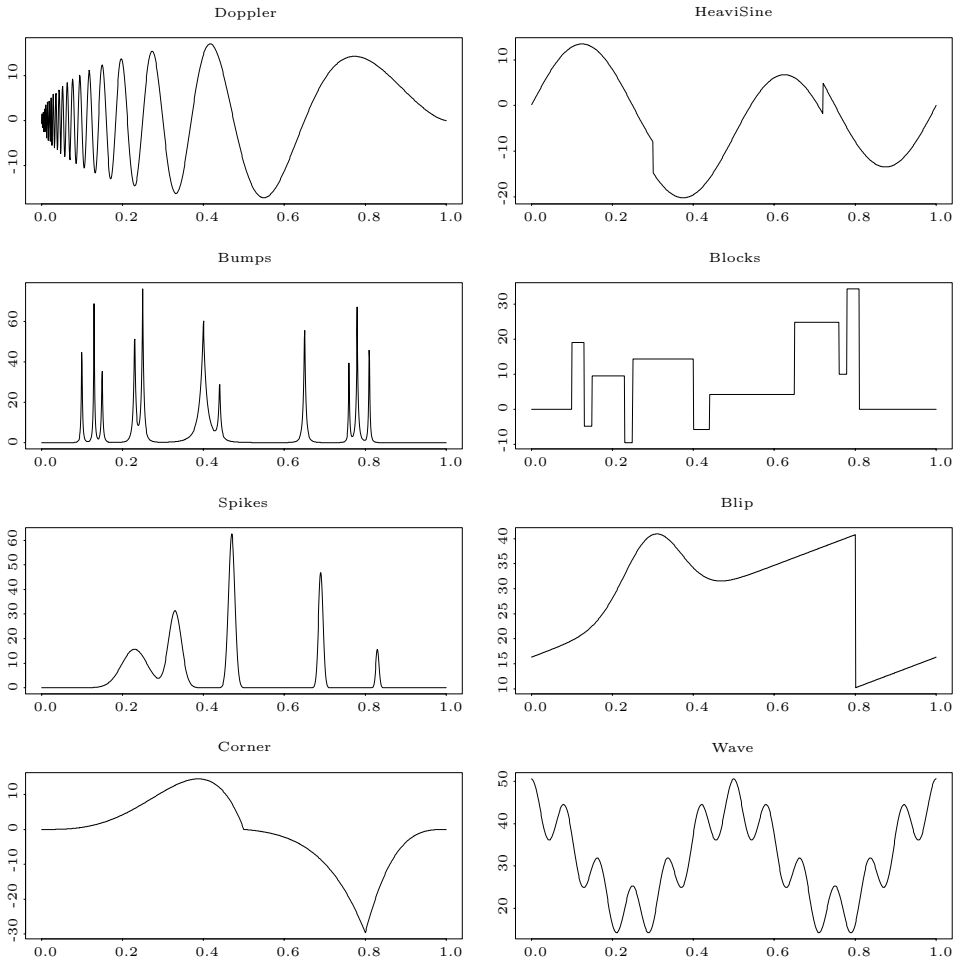
The inequality (24) follows by putting together (53), (55), and (62).

Acknowledgements

I thank all the referees, the Associate Editor, and the Editor for the constructive comments which have led to an improved presentation.

This research supported in part by NSF Grant DMS-0072578.

9. Appendix. The Test Functions



References

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. Ser. B* **60**, 725-749.

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115-129.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

Brockwell, P. and Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer-Verlag, New York.

Brown, L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimations. *Ann. Statist.* **24**, 2524-2535.

Bruce, A. and Gao, H-Y. (1997). *Applied Wavelet Analysis with S-PLUS*, Springer, New York.

Cai, T. (1999a). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898-924.

Cai, T. (1999b). Quantitative and qualitative comparisons of the block thresholding estimators and conventional wavelet methods. Manuscript.

Cai, T. (2000a). On adaptive estimation of a derivative and other related linear inverse problems. *J. Statist. Plann. Inference*, in press.

Cai, T. (2000b). On adaptability and information pooling in nonparametric function estimation. Technical Report. Department of Statistics, University of Pennsylvania.

Cai, T. and Silverman, B.W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhyā Ser. B* **63**, 127-148.

Cavalier, L. and Tsybakov, A. (2000). Sharp adaptation for inverse problems with random noise. Preprint 559, CNRS (UMR 7599).

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391-402.

Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993). Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris* (A). **316**, 417-421.

Coifman, R. R. and Donoho, D. L. (1995). Translation invariant denoising. In *Wavelets and Statistics* (Edited by A. Antoniadis and G. Oppenheim) **103**, 125-150, *Lecture Notes in Statist.*, Springer-Verlag, New York.

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

Daubechies, I. (1994). Two recent results on wavelets: wavelet bases for the interval, and biorthogonal wavelets diagonalizing the derivative operator. In *Recent Advances in Wavelet Analysis* (Edited by Schumaker L. L. and Webb G.), 237-258. Academic Press.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Statist. Assoc.* **90**, 1200-24.

Efromovich, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30**, 557-661.

Efromovich, S. Y. (1999). Quasi-linear wavelet estimation. *J. Amer. Statist. Assoc.* **94**, 189-204.

Efromovich, S. Y. (2000a). Sharp linear and block shrinkage wavelet estimation. *Statist. Probab. Lett.* **49**, 323-329.

Efromovich, S. Y. (2000b). Can adaptive estimators for Fourier series be of interest to wavelets? *Bernoulli* **6**, 699-708.

Gao, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.* **7**, 469-488.

Gao, H.-Y. and Bruce, A. G. (1997). WaveShrink with firm shrinkage. *Statist. Sinica* **7**, 855-874.

Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26**, 922-942.

Hall, P., Kerkyacharian, G. and Picard, D. (1999a). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9**, 33-50.

Hall, P.Kerkyacharian, G. and Picard, D. (1999b). A note on the wavelet oracle. *Statist. Probab. Lett.* **43**, 415-420.

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

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications.* Springer, New York.

Johnson, N. L., Kotz. S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Vol. 2. Wiley, New York.

Kerkyacharian, G., Picard, D. and Tribouley, K (1996). L_p adaptive density estimation. *Bernoulli* **2**, 229-247.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer, New York.

Lepski, O. V. (1990). On a problem of adaptive estimation on white Gaussian noise. *Theory Probab. Appl.* **35**, 3, 454-466

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H. and Patil, P. (1998). Exact risk analysis of wavelet regression. *J. Comput. Graph. Statist.* **7**, 278-309.

Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

Strang, G. (1992). Wavelet and dilation equations: a brief introduction. *SIAM Rev.* **31**, 614-627.

Triebel, H. (1983). *Theory of Function Spaces.* Birkhäuser Verlag, Basel.

Department of Statistics, the Wharton School, University of Pennsylvania.

E-mail: tcai@wharton.upenn.edu

(Received October 2000; accepted June 2002)