WAVELET REGRESSION VIA BLOCK THRESHOLDING: ADAPTIVITY AND THE CHOICE OF BLOCK SIZE AND THRESHOLD LEVEL

by

T. Tony Cai

Technical Report #99–14

Department of Statistics
Purdue University
West Lafayette, IN USA

July 1999

◇

Wavelet Regression via Block Thresholding: Adaptivity and the Choice of Block Size and Threshold Level

T. Tony Cai
Department of Statistics
Purdue University

Abstract

We consider block thresholding rules for wavelet regression and derive an "optimal" block thresholding estimator that is fully specified and easy to implement, at a computational cost of $O(n)$.

We begin by studying the effect of block length on both the global and local adaptivity. The results show that there are conflicting requirements on block size for achieving the global and local adaptivity. We then consider block thresholding as a testing problem and discuss the choice of threshold level so that the resulting estimator enjoys a desirable denoising property, as well as achieving balance between variance and bias. These results lead us naturally to the optimal choice of block thresholding estimator.

Both the asymptotic and numerical properties of the estimator are investigated. We show that this estimator is indeed optimal in the sense that it achieves simultaneously the global and local adaptivity, while preserves the smoothing and denoising properties. Furthermore, numerical results show that the estimator performs excellently in comparisons with conventional methods.

Keywords: Block thresholding; Convergence rate; Global adaptivity; Local adaptivity; Minimax estimation; Nonparametric regression; Smoothing parameter; Wavelets.
AMS 1991 Subject Classification: Primary 62G07, Secondary 62G20.

1 Introduction

Consider the nonparametric regression model:

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

i = 1, 2, ..., n (= 2^J), $x_i = \frac{i}{n}$, ϵ 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 (MISE):

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

and locally by the expected loss at a point:

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

Wavelet bases offer efficient representations for functions in a wide range of function spaces and wavelet methods have demonstrated considerable successes in terms of adaptivity and computational efficiency in nonparametric regression. They enjoy excellent mean squared error properties when used to estimate functions that are only piecewise smooth and have near optimal convergence rates over large function classes. In contrast, traditional linear estimators typically achieve good performance only for relatively smooth functions.

Wavelet methods achieve their unusual 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 & Bruce (1997)), non-garrote shrinkage (Gao (1998)), and Bayesian shrinkage rules based on independent priors on empirical wavelet coefficients (see, e.g., Clyde, et al. (1998) and Abramovich, et al (1998)).

The main objective of VisuShrink is to produce "noise-free" reconstructions. VisuShrink achieves a degree of tradeoff between variance and bias contributions to the mean squared error. However, the tradeoff is not optimal. VisuShrink favors reducing variance over bias. As a result, it creates a logarithmic penalty in the MISE. The logarithmic penalty cannot be removed by simply fine tuning the threshold. In fact, the estimator is asymptotically optimal among all such universal term-by-term thresholding rules (see Donoho & Johnstone (1994); also see Section 3). The difficulty of term-by-term thresholding is due to the relative inaccuracy with which individual wavelet coefficients are estimated.

Hall, Kerkyacharian and Picard (1998 & 1999) introduced local block thresholding estimators which threshold 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. Hall, et al. (1998 & 1999) argued that block thresholding enjoys a number of advantages over the conventional term-by-term thresholding. See also Härdle, et al. (1998). The estimator of Hall, et al., however, has an obvious drawback. The smoothing parameters, block length

and threshold level, are not completely specified. No prescription is given for finite samples and the users thus need to select the parameters empirically.

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 block thresholding rules for wavelet regression and derive an "optimal" block thresholding estimator that is fully specified and easy to implement, at a computational cost of $O(n)$. Specifically, we have three goals. The first is to study the effect of block length and threshold level on the global as well as local adaptivity. The second is to determine the "optimal" choice for block size and threshold level and derive an estimator that achieves simultaneously the optimal global and local adaptivity while preserving the smoothing and denoising properties of the VisuShrink estimator. The third goal is to investigate both the asymptotic and numerical properties of the estimator.

As in any other smoothing method, the choice of smoothing parameters, in this case the block size and the threshold level, plays a critical role in the performance of the resulting estimator. After Section 2 in which basic notation and the block threshold method are introduced, we consider in Section 3 the effect of block length on both the global and local adaptivity of the estimator. The results show that there are conflicting demands on block size for achieving the global and local adaptivity. The block size must be at least of the order $\log n$ to achieve the optimal global adaptivity. On the other hand, to achieve the optimal local adaptivity, the block size must be no more than $\log n$ in order. Therefore no block thresholding estimator can achieve simultaneously the optimal global and local adaptivity, if the block size is larger or smaller than $\log n$ in order. Then, in Section 4, we consider block thresholding as a hypothesis testing problem and select threshold level so that the estimator enjoys a desirable denoising property, as well as achieving balance between variance and bias.

The results obtained in Sections 3 and 4 lead us naturally to consideration in Section 5 of a possible optimal choice of block thresholding estimator. The estimator, called *BlockShrink*, is completely specified, with explicit definition of both the block size and the threshold level. Asymptotic results show that the estimator is indeed optimal in the sense that it achieves simultaneously the exact global and local adaptivity, while preserves the smoothing and denoising properties. More specifically, we prove that *BlockShrink* 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. In addition, *BlockShrink* enjoys an interesting smoothness property which should offer high visual quality of the reconstruction.

We investigate the finite-sample performance of *BlockShrink* in Section 6. The estimator is compared both quantitatively and qualitatively with four conventional methods, VisuShrink, RiskShrink, SureShrink and Translation-Invariant (TI) de-noising. Simulation results show that the estimator has superior numerical performance in comparison to the other four estimators. It automatically adapts to subtle changes in the underlying functions, but do not contain the spurious fine-scale structure often contained in RiskShrink and SureShrink. Real and simulated data sets are also discussed. SPlus scripts implementing

the estimator and additional simulation results are provided on the web site [7]. The proofs are contained in Section 8.

2 Wavelet Thresholding

2.1 Wavelets

An orthonormal wavelet basis is generated from dilation and translation of two basic functions, a "father" wavelet ϕ and a "mother" wavelet ψ. The functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$. 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 the father wavelet ϕ and mother wavelet 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 $i = 1, ..., D-1$; and $\int x^i \psi(x)dx = 0$ for $i = 0, ..., D-1$. Denote the periodized wavelets

$$\phi_{jk}^p(x) = \sum_{l \in \mathcal{Z}} \phi_{jk}(x-l), \quad \psi_{jk}^p(x) = \sum_{l \in \mathcal{Z}} \phi_{jk}(x-l), \qquad \text{for } x \in [0,1].$$

where $\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k)$, and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. For simplicity in exposition, we use the periodized wavelet bases on [0, 1] in the present paper. The collection $\{\phi_{j_0k}^p, k = 1, ..., 2^{j_0}; \psi_{jk}^p, j \geq j_0 \geq 0, k = 1, ..., 2^j\}$ constitutes such an orthonormal basis of $L_2[0,1]$. The superscript "p" will be suppressed from the notations for convenience.

An orthonormal wavelet basis has an associated orthogonal Discrete Wavelet Trasnform (DWT) that transforms sampled data into wavelet coefficient domain in $O(n)$ operations. The DWT is norm-preserving and this enables one to transform the problem in the function domain into a problem in the sequence domain of the wavelet coefficients with isometry of risks. See Daubechies (1992) and Strang (1992) for more on wavelets and the DWT.

Wavelet bases have distinguished data compression and localization properties. A remarkable fact about wavelets is that full wavelet series (those having plenty of nonzero coefficients) represent really pathological functions, whereas "normal" functions have sparse wavelet series. Wavelet bases are well localized, i.e., local regularity properties of a function are determined by its local wavelet coefficients. Large wavelet coefficients cluster around the discontinuities and other irregularities of the function. See Meyer (1992).

Here is an example which depicts the data compression and localization properties of wavelets. Consider JumpSine, a sinusoid with three discontinuous jumps. Panels (a) and (b) of Figure 1 plot a sampled function of length 1024 and the DWT of the data, respectively. The vertical lines on panel (b) represents the values of the wavelet coefficients. Among the total of 1024 coefficients, there are only a few of them that are large enough to be visible on the plot. The large coefficients at high resolution levels occur only around the three jump points. Panel (c) shows the striking contrast between the energy concentration of the original data and the transformed data. The energy concentration function is defined by $e(k) \equiv \frac{\sum_{i \leq k} |\theta|_{(i)}^2}{\|\theta\|_2^2}$ where $|\theta|_{(i)}$ is the i-th largest absolute value in the vector θ (see Bruce & Gao (1997)). In panel (c), k is plotted on a log scale. The energy concentration function of the wavelet coefficients increases exponentially fast, whereas the energy concentration

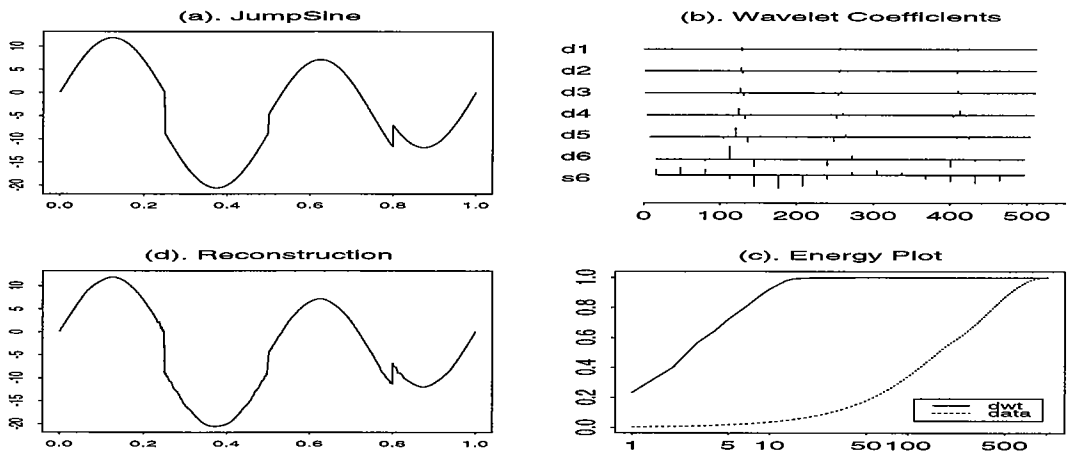

Figure 1: Compression and localization properties of wavelets

function of the original data increases very slowly. The information about the function is concentrated in a very small number of large wavelet coefficients. In fact, one has almost perfect reconstruction with only 50 largest coefficients (panel (d)).

2.2 Thresholding Estimators

The compression and localization properties of wavelets have important implications for statstical estimation. Efficient representation leads to efficient estimation. When the observations are contaminated with noise, a good strategy is to estimate the large coefficients accurately, and at the same time, kill the small coefficients. This can be achieved through thresholding.

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

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

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

$$\tilde{y}_{jk} = \theta'_{jk} + n^{-1/2} \epsilon z_{jk} \tag{5}$$

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

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

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

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

Hall, et al. (1999) introduce a block thresholding estimator which thresholds wavelet coefficients in groups instead of individually. At each resolution level j, the empirical wavelet coefficients $\tilde{y}_{jk}$ are divided into nonoverlapping blocks of length $L = C(\log n)^{1+\gamma}$ with $\gamma > 0$ and coefficients within a block are estimated simultaneously. Denote (jb) the indices of the coefficients in the b-th block at level j, i.e. $(jb) = \{(j,k) : (b-1)L + 1 \leq k \leq bL\}$. Let $S_{jb}^2 = \sum_{k \in (jb)} \tilde{y}_{jk}^2$ denote the sum of squares of the empirical coefficients in the block. A block (jb) is deemed important if S_{jb}^2 is larger than a threshold $T = \lambda L n^{-1} \epsilon^2$ with $\lambda \geq 48$ 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. For $(j,k) \in (jb)$,

$$\hat{\theta}_{jk} = \tilde{y}_{jk} \cdot I(S_{jb}^2 > \lambda L n^{-1} \epsilon^2). \tag{6}$$

The estimator of the whole function is given by

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

It is shown that, under the global risk measure (2), the estimator (7) attain the exact minimax rate of convergence without the logarithmic penalty over a wide range of function classes $\mathcal{H}$ considered in Section 5.2. A similar estimator was discussed in Hall, et al. (1998) in the case of density estimation. See also Härdle, et al. (1998).

Despite its virtues, this estimator has an obvious drawback. The smoothing parameters, block length and threshold level, are not completely specified by the theory. No prescription is given for finite samples and the users thus need to choose the parameters subjectively. Also as we will see in Section 3.2 that the estimator does not achieve optimal local adaptivity.

Throughout this paper, the term "block thresholding estimator" refers to an estimator of the form (7) with some $L > 0$ and $\lambda > 0$.

3 The Effect of Block Length on Adaptivity

We measure the performance of an estimator by its global as well as local adaptivity. An estimator that is global 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. 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.

We begin by investigating the effect of block length on global and local adaptivity. The results obtained in this section will lead us naturally to an "optimal" block thresholding estimator in Section 5 which achieves simultaneously the optimal global and local adaptivity, and at the same time, preserves the smoothing and denoising properties. The estimator is fully specified, with explicit definition of both the block size and the threshold level.

3.1 Effect on Global Adaptivity

We call an estimator achieving the optimal global adaptivity over certain function classes if, under the global risk measure (2), it attains the exact minimax rate of convergence

simultaneously over the function classes. The function classes of interest in this section is the traditional Hölder classes $\Lambda^\alpha(M)$ which are defined 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 $\alpha' = \alpha - \lfloor\alpha\rfloor$.

Denote the minimax risk over a function class $\mathcal{F}$ by $R(\mathcal{F}, n) = \inf_{\hat{f}_n} \sup_{\mathcal{F}} E\|\hat{f}_n - f\|_2^2$. It is well known that the minimax rate of convergence for global estimation over $\Lambda^\alpha(M)$ is $n^{-2\alpha/(1+2\alpha)}$. The results below shows the effect of block length on the global adaptivity in terms of convergence rate.

Theorem 1 *Suppose the wavelets $\{\phi, \psi\} \in W(D)$. Denote by $\hat{f}_n$ the estimator given by (7) with block size $L = (\log n)^\rho$ and thresholding constant λ.*

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

$$\overline{\lim_{n\to\infty}}\, n^{\frac{2\alpha}{1+2\alpha}} \cdot (\log n)^{-\frac{2\alpha(1-\rho)}{1+2\alpha}} \cdot \sup_{f\in\Lambda^\alpha(M)} E\|\hat{f}_n - f\|_2^2 > 0. \tag{8}$$

(ii). *On the other hand, if $\rho > 1$, then for any fixed $\lambda > 1$, and for all $0 < \alpha \le D$ and $0 < M < \infty$,*

$$0 < \overline{\lim_{n\to\infty}}\, n^{\frac{2\alpha}{1+2\alpha}} \cdot \sup_{f\in\Lambda^\alpha(M)} E\|\hat{f}_n - f\|_2^2 < \infty \tag{9}$$

Theorem 1 shows that, when $\rho < 1$, the rate of convergence for $\hat{f}_n$ over $\Lambda^\alpha(M)$ cannot exceed $(\log^{1-\rho} n/n)^{2\alpha/(1+2\alpha)}$. Therefore, it is impossible for a block thresholding estimator with $L = (\log n)^\rho$ and $\rho < 1$ to achieve the optimal global adaptivity. 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 $L = (\log n)^\rho$ and $\rho > 1$, and any fixed thresholding constant $\lambda > 1$, a block thresholding estimator is globally adaptive over the Hölder classes. In fact, it can be shown that the global adaptivity holds over much wider function classes.

Theorem 1(i) gives an upper bound for the global rate of convergence when $0 \le \rho < 1$. It can be shown that the upper bound is sharp. The estimator with thresholding constant derived in Section 4 attains the upper bound. A special case is $L = 1$. Theorem 1 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.

3.2 Effect on Local Adaptivity

We now consider local adaptivity of the estimators. 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 estimators locally. The local risk measure

$$R(\hat{f}(x_0),\ f(x_0)) = E(\hat{f}(x_0) - f(x_0))^2 \tag{10}$$

is used for 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 $\alpha' = \alpha - \lfloor\alpha\rfloor$. Denote the minimax risk for estimating functions at a point x_0 over $\Lambda^\alpha(M, x_0, \delta)$ by

$$R(\Lambda^\alpha(M, x_0, \delta), n) = \inf_{\hat{f}_n} \sup_{\Lambda^\alpha(M, x_0, \delta)} E(\hat{f}_n(x_0) - f(x_0))^2.$$

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 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, $R(\Lambda^\alpha(M, x_0, \delta), n)$ 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 & 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.

Theorem 2 *Suppose the wavelets $\{\phi, \psi\} \in W(D)$ and $x_0 \in (0, 1)$ is fixed. Denote by $\hat{f}_n$ the estimator given by (7), with block size $L = (\log n)^\rho$ and thresholding constant λ.*

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

$$0 < \overline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^{\frac{2\alpha}{1+2\alpha}} \cdot \sup_{f \in \Lambda^\alpha(M, x_0, \delta)} E(\hat{f}_n(x_0) - f(x_0))^2 < \infty, \tag{11}$$

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

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

In words, when $\rho > 1$, no block thresholding estimator $\hat{f}_n$ 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 could not be too large in order to well adapt to the local behavior of the underlying function. On the other hand, if $\rho < 1$, then, with an appropriate choice of λ, the optimal local adaptivity can be achieved. The choice of λ will be discussed in Section 4. Here we note that the block size in Hall, et al. (1999) is of the order $(\log n)^\rho$ with $\rho > 1$, so, in light of Theorem 2, it does not achieve the optimal local adaptivity.

It is revealing to put Theorems 1 and 2 together. One immediately sees that there are conflicting requirements on the block size for achieving the global and local adaptivity; and it is impossible to simultaneously achieve both by a block thresholding estimator with $L = (\log n)^\rho$ and $\rho \ne 1$.

These results lead us to consideration of the choice of $L = \log n$ as a possible optimal compromise. We will show in Section 5 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 over a range of function classes.

4 Block Thresholding as a Testing Problem and the Choice of the Thresholding Constant

The aim of block thresholding is to achieve better adaptivity while retaining the smoothing and denoising properties of the VisuShrink estimator. 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

$$x_i = \theta_i + z_i, \quad i = 1, 2, \cdots, n,$$

with $z_i \overset{iid}{\sim} N(0, 1)$. 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 the hypothesis

$$H_0 : \ \theta_1 = \theta_2 = \cdots = \theta_n = 0. \tag{13}$$

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 $b = 1, \cdots n/L$, and to estimate $\theta_{(b)}$ by 0 when the hypothesis $H_0^{(b)}$ is not rejected and by $x_{(b)}$ otherwise. The classical multivariate normal decision theory shows that, for each block, a uniformly most powerful test exists; and the best rejection region is of the form $\sum x_i^2 > T$, where T is a constant (see Lehmann & Casella (1998), pp. 351). Hence the shrinkage 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 p-value of the blockwise test under the null hypothesis is

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

where $Y_L \sim \chi_L^2$. It is reasonable to require that, under the null, the blockwise test asymptotically makes the correct decision with certainty, i.e., $p_L(\lambda) \to 0$, as $n \to \infty$.

Theorem 3 *Let $L = (\log n)^\rho$ and let $p_L(\lambda)$, as given in (14), be the p-value of the blockwise test. Denote $T_\rho = 2(\log n)^{1-\rho}$ for $0 < \rho < 1$ and $\delta_\rho = 2(\log n)^{-(\rho-1)/2}$ for $\rho > 1$. Let*

$$\text{(i).} \quad \lambda_L = 2 \log n, \ \textit{when } \rho = 0; \tag{15}$$

$$\text{(ii).} \quad \lambda_L = T_\rho + \log T_\rho + 1, \ \textit{when } 0 < \rho \leq 1/2; \tag{16}$$

$$\text{(iii).} \quad \lambda_L = T_\rho + \log T_\rho + 1 + (\log T_\rho + 1)/T_\rho, \ \textit{when } 1/2 < \rho < 1; \tag{17}$$

$$\text{(iv).} \quad \lambda_L = 4.5052 \ (\textit{the root of } \lambda - \log \lambda - 3 = 0), \ \textit{when } \rho = 1; \tag{18}$$

$$\text{(v).} \quad \lambda_L = 1 + \delta_\rho + \delta_\rho^2/3 + \delta_\rho^3/36, \ \textit{when } 1 < \rho < 2. \tag{19}$$

$$\text{(vi).} \quad \lambda_L = 1 + \delta_\rho + \delta_\rho^2/3, \quad \textit{when } 2 \leq \rho < 3. \tag{20}$$

$$\text{(vii).} \quad \lambda_L = 1 + \delta_\rho, \quad \textit{when } \rho \geq 3. \tag{21}$$

Then, for $\lambda \geq \lambda_L$, $p_L(\lambda) \to 0$ *as* $n \to \infty$. *Moreover, the bounds given above are sharp. For example, in the case of* $0 < \rho \leq 1/2$, *if* $\lambda \leq T_\rho + \log T_\rho + c$ *with a constant* $c < 1$, *then* $p_L(\lambda) \to 1$. *In particular, if* $\lim_{n\to\infty} \lambda/\lambda_L < 1$, *then* $p_L(\lambda) \to 1$.

It is interesting to note that in the case of $L = \log n$, λ_L is an absolute constant satisfying $\lambda - \log \lambda - 3 = 0$. In the special case of $L = 1$, the bound 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 & Johnstone (1994)).

Guided by Theorem 3, for a given block length $L = (\log n)^\rho$, we choose the thresholding constant λ_L as in (15)-(21). As a consequence of Theorem 3, with the selected λ_L, the resulting block thresholding estimator removes pure noise completely, with probability tending to 1. See Theorem 6 for the case of $L = \log n$.

5 The BlockShrink Estimator and Its Optimality

5.1 The Estimator

The results in Section 3 show that it is impossible for a block thresholding estimator (7) with $L = (\log n)^\rho$ and $\rho \neq 1$ to achieve the optimal global and local adaptivity simultaneously. The discussion in the preceding sections lead naturally to consideration of the block thresholding estimator with block size $L = \log n$ and thresholding constant $\lambda = 4.5052$. In this section, we will discuss in detail the properties of this particular block thresholding estimator.

Denote by $L_* = \log n$ and $\lambda_* = 4.5052$, we define the block thresholding estimator $\hat{f}_n^*$ by

$$\hat{f}_n^*(x) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k} \phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1} \sum_{b} \sum_{k \in (jb)} \tilde{y}_{jk} I(S_{jb}^2 > \lambda_* L_* n^{-1} \epsilon^2)\, \psi_{jk}(x) \tag{22}$$

For specificity, we call this particular block thresholding estimator *BlockShrink* in the rest of the paper.

Often one is interested in estimating f at the sample points. In that case, the *BlockShrink* estimator can be easily implemented in three steps, at a computational cost of $O(n)$:

1. Transform the noisy data via the discrete wavelet transform.

2. At each resolution level, the empirical wavelet coefficients are grouped into nonoverlapping blocks of length L_*. If the sum of the squared empirical coefficients in a block is above a threshold $T = \lambda_* L_* \epsilon^2$, then the block is deemed to contain significant information about the signal and then all the coefficients in the block are retained, otherwise it is deemed insignificant and all the coefficients in the block are discarded.

3. Obtain the estimate of function f at the sample points, $(\widehat{f(x_i)})_{i=1}^n$, by the inverse discrete wavelet transform of the denoised wavelet coefficients.

We will show in Sections 5.2 – 5.5 that *BlockShrink* indeed enjoys simultaneously a high degree of global and local adaptivity as well as a desirable denoising property. Moreover,

a simulation study, summarized in Section 6, shows the estimator has excellent numerical performance. But first let us look at one example.

Consider the Doppler signal, a sinusoid with changing amplitude and frequency (Donoho & Johnstone (1994)). Random Gaussian noise is added to the signal. The signal-to-noise ratio (SNR) is 3. *BlockShrink* and four conventional wavelet methods, VisuShrink, RiskShrink, SureShrink, and TI de-noising, are used for recovering the true signal. (See Section 6 for a discussion of the four conventional methods.) Figure 2 displays the reconstructions.

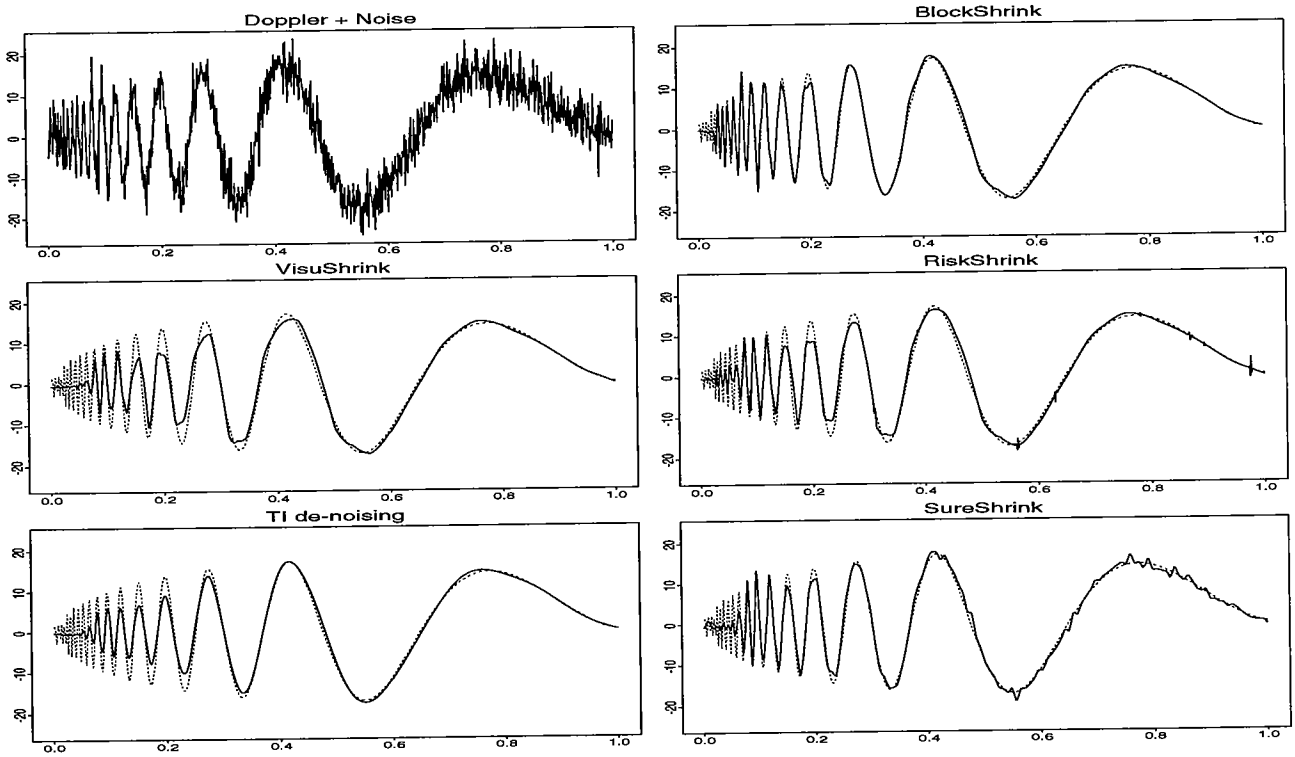

Figure 2: Comparison of the reconstructions. The dotted line is the true signal.

BlockShrink as well as the conventional methods recover the smooth lowest frequency part reasonably well. *BlockShrink* automatically adapts to the changing frequency of the target function. It estimates the smooth and low frequency part with a high degree of accuracy; at the same time, it also recovers well the rapidly oscillating area near the origin. In contrast, VisuShrink, RiskShrink and TI de-noising all significantly over-smooth the high frequency area. SureShrink does better than the other three conventional methods in recovering the high frequency part, but it contains a fair amount of local oscillation around the low frequency part and is visually unpleasant.

Quantitatively, the *BlockShrink* estimator is significantly more accurate than the other methods. In this case, the ratios of the mean squared error of *BlockShrink* to those of VisuShrink, RiskShrink, SureShrink, and TI de-noising are 0.280, 0.514, 0.628, and 0.368, respectively. See Section 6 for further numerical results.

An inspection of wavelet coefficients is also revealing. Figure 3 displays the wavelet coefficients used in the reconstructions. To show the detail coefficients more clearly, we use different scales at different levels. *BlockShrink* retains 9 blocks of size 4 with a total of 36

detail coefficients. The coefficients at high resolution levels cluster around the area near the origin where the function rapidly oscillates. SureShrink retains 64 detail coefficients with many around the low frequency area. As a result, the reconstruction contains spurious fine-scale structure in the low frequency area. RiskShrink keeps 40 detail coefficients, also with a few around the low frequency area which result in several unpleasant artifacts. VisuShrink keeps only 18 detail coefficients and the reconstruction is over-smoothed.

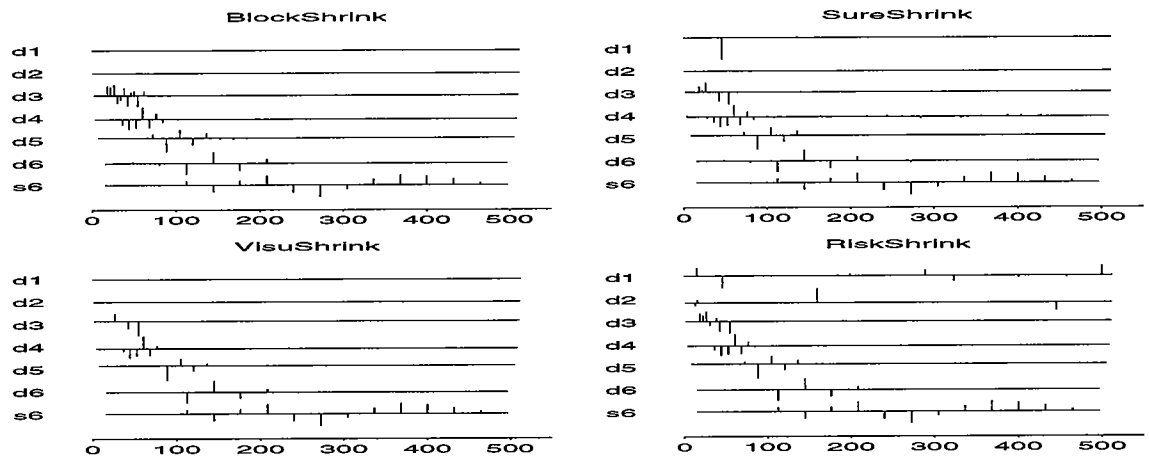

Figure 3: Wavelet coefficients of the reconstructions of Doppler.

5.2 The function classes $\mathcal{H}$

We consider the adaptivity of *BlockShrink* over a family of large function classes which was used in Hall, et al. (1999). These function 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 jump discontinuities and irregular perturbations.

Definition 1 *Let* $\mathcal{H} = \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$, *where* $0 \le \alpha_1 < \alpha \le D$, $0 \le \gamma < \frac{1+2\alpha_1}{1+2\alpha}$, *and* $M_1, M_2, M_3, \nu \ge 0$, *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* $a_0 = f(2^{-j}k), a_1, ..., a_{D-1}$ *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* $a_0 = f(2^{-j}k), a_1, ..., a_{D-1}$ *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 $\mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$ contains the traditional Besov class $B^{\alpha}_{\infty\infty}(M_2)$ as a subset for any given $\alpha_1, \gamma, M_1, M_3, D, \nu$. See Meyer (1992) for definitions and properties of Besov spaces.

A function $f \in \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$ can be regarded as the superposition of a regular smooth function f_s in a Besov class $B^{\alpha}_{\infty\infty}(M_2)$ 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: $\tau(x) = \sum_{k=1}^{K} a_k(x - x_k)^{\beta_k} \cos(x - x_k)^{-\omega_k}$. See Hall, et al. (1998 & 1999) for further discussions about the function classes $\mathcal{H}$.

5.3 Global Adaptivity

In this section, we investigate the global adaptivity of *BlockShrink* to unknown degree of inhomogeneous smoothness over the function classes $\mathcal{H} \equiv \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$. The optimal rate of convergence for global estimation over the Besov class $B^{\alpha}_{\infty\infty}(M)$ is $n^{2\alpha/(1+2\alpha)}$. Because the function class $\mathcal{H}$ contains $B^{\alpha}_{\infty\infty}(M_2)$ as a subset, the convergence rate over $\mathcal{H}$ can not exceed $n^{-2\alpha/(1+2\alpha)}$. Theorem 4 below shows that *BlockShrink* attains adaptively the optimal convergence rate of $n^{-2\alpha/(1+2\alpha)}$ across a wide interval of the function classes $\mathcal{H}$.

Theorem 4 *Suppose the wavelets $\{\phi, \psi\} \in W(D)$ and $supp(\phi) = supp(\psi) = (0, N)$. Let $\mathcal{H} = \mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$. Then BlockShrink satisfies that for all $0 < \alpha \leq D$ and for all $\nu \geq N$,*

$$\sup_{f \in \mathcal{H}} E\|\hat{f}_n^* - f\|_2^2 \leq Cn^{-2\alpha/(1+2\alpha)} \tag{23}$$

and for the estimate at the sample points, $(\widehat{f(x_i)})_{i=1}^n$,

$$\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} E(\widehat{f(x_i)} - f(x_i))^2 \leq Cn^{-2\alpha/(1+2\alpha)} \tag{24}$$

Thus, *BlockShrink*, without knowing the a priori degree or amount of smoothness of the underlying function, attains the optimal convergence rate that one could achieve by knowing the regularity. That is, $\sup_{f \in \mathcal{H}} E\|\hat{f}_n^* - f\|^2 \asymp R(\mathcal{H}, n)$. In particular, *BlockShrink* attains the optimal rates over a range of the Hölder classes $\Lambda^{\alpha}(M)$.

Remark 1 (Use of Coiflets:) If the following local Lipschitz conditions are imposed on $\mathcal{H}$ when functions in $\mathcal{H}$ are relatively smooth, then there is no need for using Coiflets and the condition $\{\phi,\ \psi\} \in W(D)$ can be replaced by that ψ has D vanishing moments; and thus regular Daubechies' wavelets can be used.

(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 $|f(x) - f(2^{-j}k)| \leq M_5 2^{-j}$, for $x \in [2^{-j}k, 2^{-j}(k + v)]$.

5.4 Local Adaptivity

Again we use the pointwise risk (3) to measure the local adaptivity. As we noted in Section 3.2, the adaptive minimax rate for estimating a function at a point x_0 over the local Hölder class $\Lambda^{\alpha}(M, x_0, \delta)$ is $(\log n/n)^{2\alpha/(1+2\alpha)}$. Theorem 5 below shows that *BlockShrink* achieves optimal local adaptation with the minimal cost for estimating f at a point.

Theorem 5 *Let the wavelets* $\{\phi, \psi\} \in W(D)$ *and let* $x_0 \in (0,1)$ *be fixed. Then BlockShrink* $\hat{f}_n^*(x_0)$ *of* $f(x_0)$ *satisfies that for all* $0 < \alpha \leq D$, $\delta > 0$ *and* $0 < M < \infty$,

$$\sup_{f \in \Lambda^\alpha(M, x_0, \delta)} E(\hat{f}_n^*(x_0) - f(x_0))^2 \leq C \cdot (\log n/n)^{2\alpha/(1+2\alpha)} \tag{25}$$

Combining Theorems 4 and 5 and compare to Theorems 1 and 2, we can see that *BlockShrink* achieves both the global and local adaptivity which is impossible to achieve simultaneously for other block thresholding estimator with $L = (\log n)^\rho$ and $\rho \neq 1$.

5.5 Denoising Property

In addition to the global and local adaptivity, the *BlockShrink* estimator enjoys a smoothness property which should offer high visual quality of the reconstruction. The estimator, with high probability, removes pure noise completely.

Theorem 6 *If the underlying true function is the zero function* $f \equiv 0$, *then, with probability tending to 1, BlockShrink is also the zero function. That is, there exist universal constants* P_n *such that*

$$P(\hat{f}_n^* \equiv 0) \geq P_n \to 1, \quad as \ \ n \to \infty. \tag{26}$$

6 Numerical Results and Examples

A simulation study is carried out to investigate the finite-sample performance of the estimator. *BlockShrink* is compared to VisuShrink, RiskShrink, SureShrink as well as the Translation-Invariant (TI) de-noising method. RiskShrink, due to 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 wavelet coefficients by minimizing the Stein's unbiased risk estimate at each resolution level (see Donoho & Johnstone (1995)). Both RiskShrink and SureShrink usually have better mean squared error performance than VisuShrink, but the reconstructions often contain visually unpleasant spurious fine-structure. A TI de-noising estimator (Coifman and Donoho (1995)) is constructed by averaging over VisuShrink estimates based on all the shifts of the original data. For further details on these estimators the readers are referred to Donoho and Johnstone (1994 & 1995) and Coifman and Donoho (1995). For SureShrink, we use the hybrid method proposed in Donoho and Johnstone (1995) in the simulations.

Eight test functions representing different level of spatial variability were used in the simulations. For each of the eight objects under study, five different methods, *BlockShrink*, VisuShrink, RiskShrink, SureShrink and TI de-noising, are applied to noisy versions of the function. Sample sizes from $n = 512$ to $n = 8192$ and signal-to-noise ratios (SNR) from 3 to 7 are considered. And several different wavelets were used. Different combinations of wavelets and signal-to-noise ratios 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. For more details see the web site [7]

We implemented *BlockShrink* in S+Wavelets. There are total of 2^j empirical wavelet coefficients at a given resolution level j. For convenience one often wish to choose the block

size to be a dyadic integer and evenly divide the coefficients at each resolution level into nonoverlapping blocks. We suggest to take the block size to be the largest dyadic integer smaller than or equal to $\log n$, i.e. $L = 2^{[\log_2(\log n)]}$. This choice of block size is used in all simulations and works very well in our experience. Throughout, the lowest resolution level $j_0 = [\log_2 \log n] + 1$ was used for all methods. Table 1 reports the mean squared errors (MSE) over 500 replications. A graphical presentation is given in Figure 4.

Table 1: Mean Squared Error From 500 Replications (SNR=5)

n	Block	Visu	Risk	Sure	TI	n	Block	Visu	Risk	Sure	TI
Doppler						*HeaviSine*					
512	0.94	3.40	1.72	1.59	2.64	512	0.57	0.56	0.47	0.50	0.52
1024	0.59	2.22	1.17	0.89	1.66	1024	0.36	0.44	0.33	0.34	0.39
2048	0.34	1.42	0.77	0.54	1.02	2048	0.22	0.35	0.23	0.23	0.28
4096	0.16	0.80	0.45	0.34	0.56	4096	0.14	0.21	0.14	0.13	0.16
8192	0.09	0.52	0.29	0.18	0.34	8192	0.08	0.14	0.09	0.07	0.10
Bumps						*Blocks*					
512	2.19	10.92	4.99	2.23	7.53	512	2.15	6.29	3.05	2.61	5.35
1024	1.29	6.66	3.16	1.69	4.50	1024	1.33	4.20	2.08	1.59	3.61
2048	0.75	4.18	2.04	1.12	2.70	2048	0.85	2.85	1.48	1.04	2.40
4096	0.56	2.31	1.18	0.57	1.47	4096	0.71	1.72	0.94	0.71	1.39
8192	0.30	1.40	0.74	0.34	0.86	8192	0.42	1.16	0.64	0.44	0.89
Spikes						*Blip*					
512	0.92	2.85	1.45	1.05	2.11	512	0.47	0.94	0.56	0.63	0.75
1024	0.49	1.81	0.94	0.56	1.25	1024	0.27	0.69	0.40	0.42	0.51
2048	0.31	1.14	0.61	0.33	0.72	2048	0.18	0.45	0.27	0.24	0.32
4096	0.17	0.59	0.34	0.15	0.30	4096	0.09	0.25	0.16	0.15	0.19
8192	0.08	0.38	0.21	0.08	0.17	8192	0.06	0.16	0.10	0.09	0.11
Corner						*Wave*					
512	0.33	0.51	0.35	0.29	0.30	512	0.56	3.79	1.77	2.95	2.62
1024	0.17	0.32	0.21	0.17	0.20	1024	0.30	2.34	1.04	3.20	1.56
2048	0.10	0.20	0.13	0.09	0.12	2048	0.18	1.34	0.62	3.38	0.90
4096	0.04	0.08	0.06	0.05	0.06	4096	0.09	0.48	0.25	0.09	0.11
8192	0.03	0.05	0.04	0.03	0.03	8192	0.06	0.27	0.16	0.06	0.06

BlockShrink has smaller MSE than VisuShrink in all but one of cases, among the total of 40 combinations of signals and sample sizes (see Figure 4). For six of the eight test functions, Doppler, Bumps, Blocks, Spikes, Blip and Wave, *BlockShrink* has better precisions with sample size n than VisuShrink with sample size $2 \cdot n$ for all n from 512 to 8192 (see Table 1). *BlockShrink* outperforms the other methods as well. It yields better results than RiskShrink in 37 out of the 40 cases; and beats TI de-noising in 38 out of 40 cases. The differences are especially notable when the underlying function is of significant spatial variability. In terms of MSE, the only competitor among the conventional methods is SureShrink. Apart from being better than SureShrink in more than 75% of cases in mean square error, our estimator yields noticeably better results visually. The reconstruction is smooth where the underlying function is smooth. They do not contain spurious fine-scale structure that are often contained in RiskShrink and SureShrink. *BlockShrink* adapts well to the subtle changes of the target functions. See the web site [7] for more on simulation results.

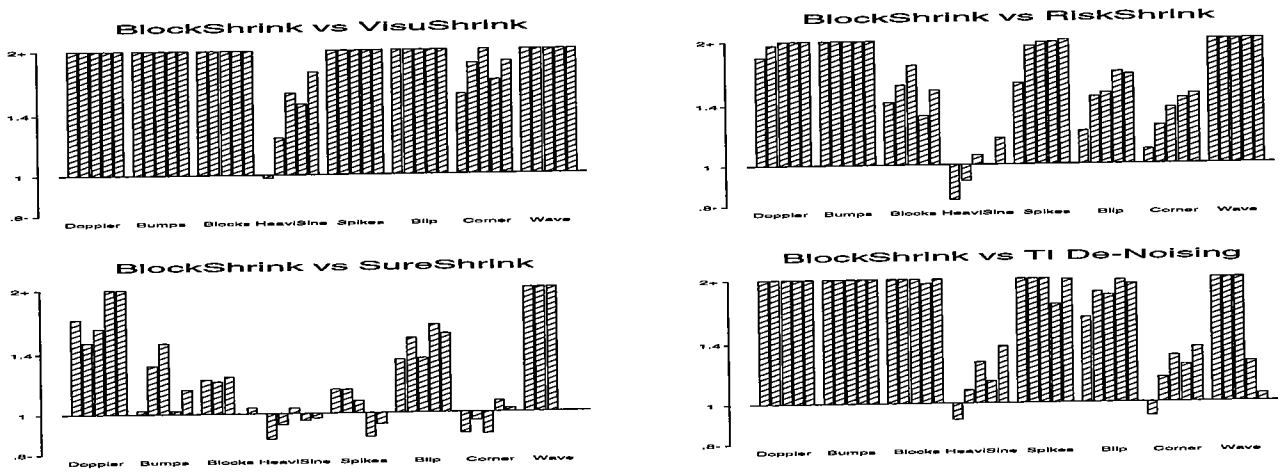

Figure 4: Comparison of MSEs (SNR=5). The vertical bars represent the ratios of the MSEs of the estimators to the corresponding MSE of *BlockShrink*. The higher the bar the better the relative performance of *BlockShrink*. The bars are plotted on a log scale and are truncated at the value 2. For each signal the bars are ordered from left to right by the sample sizes (n=512 to 8192).

It would be interesting to compare *BlockShrink* numerically with the estimator of Hall, et al. (1999). However, as mentioned earlier, their method requires to select block length and threshold level empirically and no specific prescription is given for choosing the parameters in finite sample cases. We therefore leave explicit numerical comparison for future work.

We now use the well-known sunspots data as an example to compare *BlockShrink* qualitatively with VisuShrink and SureShrink. Sunspots data has been analyzed 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 5.

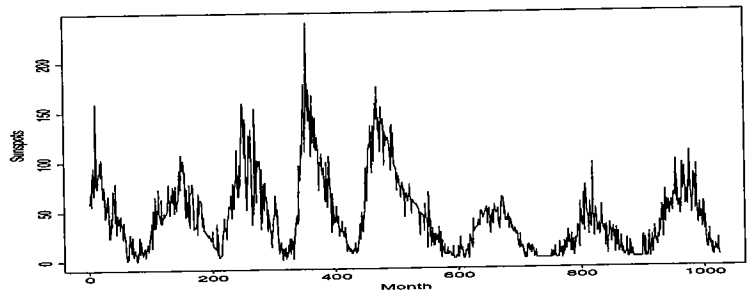

Figure 5: Monthly number of sunspots from January, 1749 to March, 1834.

Five wavelet methods, *BlockShrink*, SureShrink, VisuShrink, RiskShrink and TI denoising are applied to the data. Figure 6 displays the reconstructions and their residuals of *BlockShrink*, SureShrink and VisuShrink. Using the model in Section 5.3, we can conceptually 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 particular example, the smooth part f_s can be think of as the well-known periodic, seasonal component (with a

period of about 11 years).

The *BlockShrink* 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. *BlockShrink* permits the balance between variance and bias to be varied along the curve. It simultaneously retains the fine structures around the peaks and produces smooth reconstruction around the valleys. The reconstruction confirms the theoretical results derived in Section 5.

In comparison, VisuShrink grossly 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 VisuShrink. But SureShrink smoothes out some oscillations around the peaks, noticeably near the fourth and the seventh peaks, while 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. See the web site [7].

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 *BlockShrink* 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 *BlockShrink* keeps 63 blocks of size 4 with a total of 252 coefficients.

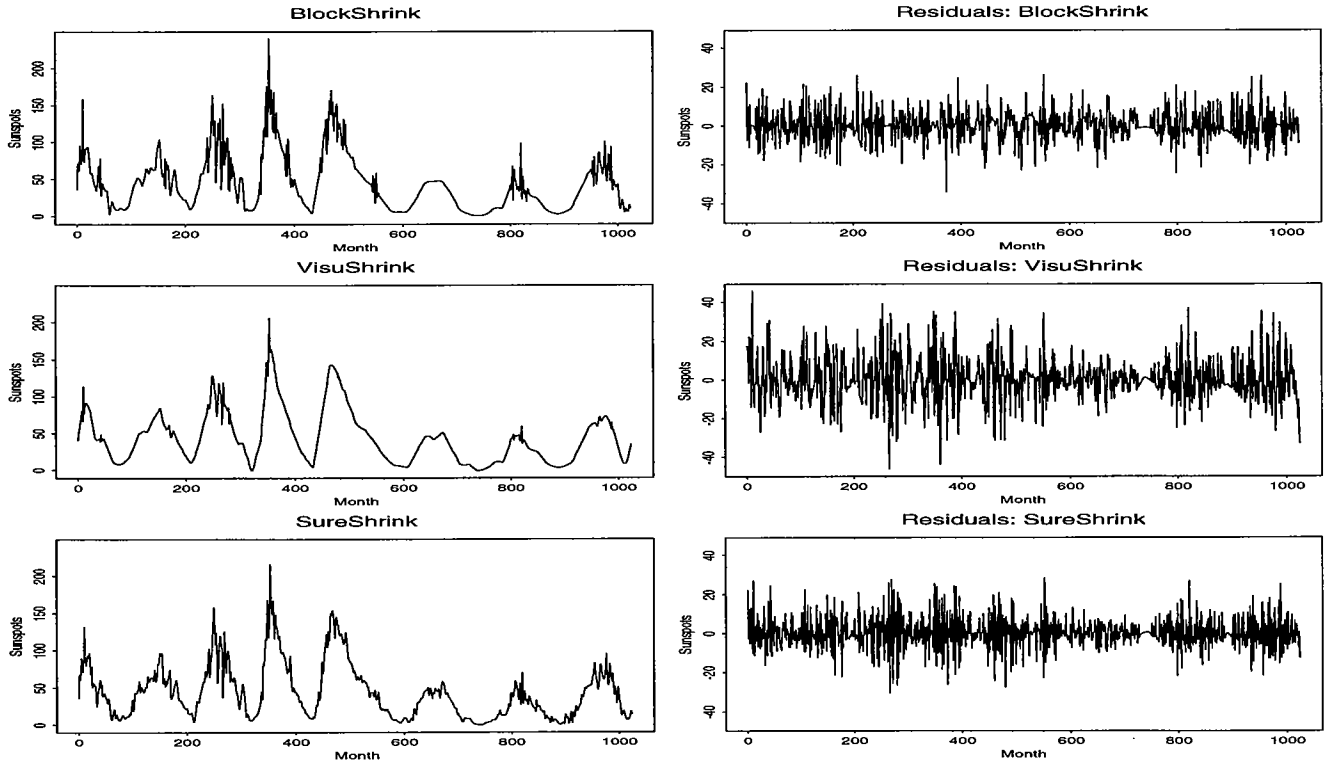

Figure 6: Comparison of reconstructions and residuals.

7 Concluding Remarks

We consider block thresholding rules for wavelet regression. It is shown that there are conflicting requirements on block size for achieving the global and local adaptivity. This and the results on the choice of threshold level lead naturally to a fully specified optimal choice of block thresholding estimator.

Asymptotic results show that the estimator, *BlockShrink*, is indeed optimal in the sense that it achieves simultaneously the exact global and local adaptivity, while preserves the smoothing and denoising properties. Numerical results also show that the estimator performs excellently in comparison with VisuShrink, RiskShrink, SureShrink and TI de-noising.

BlockShrink 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 *BlockShrink* is that it retains or deletes variables group-by-group rather than one-by-one.

Besides nonparametric regression, block thresholding techniques can be applied to other statistical problems such as linear inverse problems. For instance, block thresholding can be used to improve the asymptotic result obtained in Abramovich and Silverman (1998) for linear inverse problems. The extra logarithmic factor in the asymptotic risk bound can be removed. In other words, a block thresholding estimator will attain the exact minimax rate of convergence over a range of Besov classes for certain linear inverse problems.

8 Proofs

8.1 Preparatory Results

We will prove the main results in the order of Theorems 4, 5, 6, 1, 2, and 3. For simplicity, in the proofs we assume that n is divisible by L. 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* $x_i \overset{ind.}{\sim} N(\theta_i,\ \sigma^2)$, $i = 1, \cdots, L$. *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*

$$E\|\hat{\theta} - \theta\|_2^2 \leq (2\lambda + 2)(\|\theta\|_2^2 \wedge L\sigma^2) + 2\lambda L(\lambda^{-1}e^{\lambda-1})^{-L/2}\sigma^2. \tag{27}$$

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

$$E\|\hat{\theta} - \theta\|_2^2 \leq (2\lambda + 2)(\|\theta\|_2^2 \wedge L\sigma^2) + 2\lambda\epsilon^2 n^{-2}\log n. \tag{28}$$

The second term in (28) 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 within each block.

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

$$|\xi_{Jk} - n^{-\frac{1}{2}} f(k/n)| \leq M_1 \|\phi\|_1 n^{-(1/2+\alpha_1)} \quad \text{if } k \in A_j; \tag{29}$$

$$|\xi_{Jk} - n^{-\frac{1}{2}} f(k/n)| \leq M_2 \|\phi\|_1 n^{-(1/2+\alpha)} \text{ if } k \notin A_j; \tag{30}$$

$$|\theta_{jk}| \leq M_1 \|\psi\|_1 2^{-j(1/2+\alpha_1)} \text{ if } k \in A_j; \tag{31}$$

$$|\theta_{jk}| \leq M_2 \|\psi\|_1 2^{-j(1/2+\alpha)} \text{ if } k \notin A_j. \tag{32}$$

(ii). *for all functions $f \in \Lambda^\alpha(M)$, the wavelet coefficients of f satisfies*

$$|\theta_{jk}| \leq C' \cdot 2^{-j(1/2+\alpha)}$$

where the constant C' depends on the wavelets, α and M only.

Lemma 1 is a direct consequence of the vanishing moments conditions on the wavelets $\{\phi, \psi\}$.

Lemma 2 *If $\|u\|_{\ell_2}^2 \leq \gamma^2 t$ with $0 < \gamma < 1$, then*

$$\text{(i).} \quad \{x : \|x+u\|_{\ell_2}^2 \leq t\} \supseteq \{x : \|x\|_{\ell_2}^2 \leq (1-\gamma)^2 t\};$$

$$\text{(ii).} \quad \{x : \|x+u\|_{\ell_2}^2 \geq t\} \subseteq \{x : \|x\|_{\ell_2}^2 \geq (1-\gamma)^2 t\}.$$

Lemma 2 follows from the triangle inequality.

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

$$\text{(i).} \quad E(\sum X_i)^2 \leq (\sum (EX_i^2)^{1/2})^2; \tag{33}$$

$$\text{(ii).} \quad (E(Y + \sum X_i)^2)^{1/2} \geq (EY^2)^{1/2} - \sum (EX_i^2)^{1/2}. \tag{34}$$

The inequality (33) is a simple consequence of Cauchy-Schwartz inequality and

$$[E(Y + \sum X_i)^2]^{1/2} \geq (EY^2)^{1/2} - (E(\sum X_i)^2)^{1/2} \geq (EY^2)^{1/2} - \sum (EX_i^2)^{1/2},$$

where the first inequality follows from Minkowski's inequality and the second from (33).

Lemma 4 gives lower and upper bounds for the tail probability of the χ^2 distribution, and a bound for the expected value of a truncated χ^2 variable.

Lemma 4 *Let $Y_L \sim \chi_L^2$ and $\lambda > 1$. Then*

$$\text{(i).} \quad \frac{2}{5} \lambda^{-1} L^{-1/2} (\lambda^{-1} e^{\lambda - 1})^{-L/2} \leq P(Y_L > \lambda L) \leq \pi^{-1/2} (\lambda - 1)^{-1} L^{-1/2} (\lambda^{-1} e^{\lambda - 1})^{-L/2}; \tag{35}$$

$$\text{(ii).} \quad E Y_L I(Y_L \geq \lambda L) \leq \lambda L \, (\lambda^{-1} e^{\lambda - 1})^{-L/2}. \tag{36}$$

Proof: Denote by $f_m(y)$ the pdf of a χ^2_m variable. Then integration by parts yields

$$P(Y_m > x) = 2f_m(x) + P(Y_{m-2} > x). \tag{37}$$

Applying (37) recursively, one gets

$$P(Y_L > \lambda L) \leq 2 \sum_{k=0}^{[(L-1)/2]} f_{L-2k}(\lambda L). \tag{38}$$

It is easy to see that, for $m \leq L$,

$$f_m(\lambda L) = \frac{m}{\lambda L} f_{m+2}(\lambda L) \leq \lambda^{-1} f_{m+2}(\lambda L). \tag{39}$$

Combining (38) and (39), one has

$$P(Y_L > \lambda L) \leq 2 \sum_{k=0}^{[(L-1)/2]} \lambda^{-k} f_L(\lambda L) \leq \frac{2\lambda}{\lambda - 1} \cdot \frac{1}{2^{L/2}\Gamma(L/2)} (\lambda L)^{L/2-1} e^{-\lambda L/2}. \tag{40}$$

Now Stirling's formula,

$$\Gamma(x+1) = \sqrt{2\pi}\, x^{x+1/2}\, e^{-x+\theta/(12x)}, \quad \text{with } 0 < \theta < 1, \tag{41}$$

yields

$$P(Y_L > \lambda L) \leq \pi^{-1/2} (\lambda - 1)^{-1} L^{-1/2} (\lambda^{-1} e^{\lambda - 1})^{-L/2}.$$

On the other hand,

$$P(Y_L \geq \lambda L) = \frac{1}{2^{L/2}\, \Gamma(L/2)} \int_{\lambda L}^{\infty} x^{L/2-1} e^{-\frac{x}{2}}\, dx \geq \frac{(\lambda L)^{L/2-1}\, 2e^{-\lambda L/2}}{2^{L/2}\, \Gamma(L/2)}. \tag{42}$$

Again, it follows from Stirling's formula (41), after some simple algebra,

$$P(Y_L \geq \lambda L) \geq \frac{2}{5}\, \lambda^{-1}\, L^{-1/2}\, (\lambda^{-1}\, e^{\lambda - 1})^{-L/2}.$$

The proof of (36) is straightforward:

$$\begin{aligned} EY_L I(Y_L \geq \lambda L) &= \frac{1}{2^{L/2}\Gamma(L/2)} \int_{\lambda L}^{\infty} x^{L/2} e^{-x/2}\, dx \\ &= \frac{1}{2^{L/2}\Gamma(L/2)} \int_{L}^{\infty} x^{L/2} e^{-x/2} (\lambda^{L/2+1} e^{-x(\lambda-1)/2})\, dx \leq \lambda L\, (\lambda^{-1}\, e^{\lambda-1})^{-L/2}. \end{aligned}$$

8.2 Proof of Theorem 4

Let $\tilde{Y}$ be the discrete wavelet transform of $\{n^{-1/2}Y\}$ and be written as in (4). One may write

$$\tilde{y}_{jk} = \theta_{jk} + a_{jk} + n^{-1/2} \epsilon z_{jk} \tag{43}$$

where θ_{jk} is the true wavelet coefficients of f, a_{jk} is some approximation error which is considered "small" by the results of Lemma 1(i), and z_{jk}'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

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

Here, ξ_{j_0k} and θ_{jk} are the orthogonal transform of $\{\xi_{Ji}\}$ via W, likewise $\tilde{a}_{j_0k}$ and a_{jk} the transform of $\{n^{-1/2}f(x_i) - \xi_{Ji}\}$, and $\tilde{z}_{j_0k}$ and z_{jk} the transform of $\{z_i\}$. Thus $\tilde{z}_{j_0k}$ and z_{jk} are i.i.d. $N(0,1)$. Let $\tilde{\xi}_{j_0k} = \xi_{j_0k} + \tilde{a}_{j_0k} + n^{-1/2}\epsilon\tilde{z}_{j_0k}$ and $\tilde{y}_{jk} = \theta_{jk} + a_{jk} + n^{-1/2}\epsilon z_{jk}$. Lemma 1(i) and the orthogonality of the discrete wavelet transform yield that

$$\sum_{k=1}^{2^{j_0}}\tilde{a}_{j_0k}^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}a_{jk}^2 = \sum_{i=1}^{n}(n^{-1/2}f(x_i) - \xi_{Ji})^2 = o(n^{-2\alpha/(1+2\alpha)}). \tag{44}$$

Let $\hat{\xi}_{j_0k} = \tilde{\xi}_{j_0k}$ and $\hat{\theta}_{jk} = \tilde{y}_{jk}I(S_{jb}^2 > \lambda_* L_* n^{-1}\epsilon^2)$, for $(j,k) \in (jb)$. By the isometry of the function norm and the sequence norm, the risk of the *BlockShrink* estimator $\hat{f}_n^*$ can be written as

$$E\|\hat{f}_n^* - f\|_2^2 = \sum_k E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 + \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{jk} - \theta_{jk})^2 + \sum_{j=J}^{\infty}\sum_k \theta_{jk}^2. \tag{45}$$

Lemma 1(i) and (44) yield that

$$\sum_k E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2 + \sum_{j=J}^{\infty}\sum_k \theta_{jk}^2 = o(n^{-2\alpha/(1+2\alpha)}). \tag{46}$$

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

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

The term $S \equiv \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{jk} - \theta_{jk})^2$ can be bounded by using Proposition 1 and (44).

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

Denote

$$S_1 = \sum_{j=j_0}^{J-1}\sum_{(jb)\in G_j}\sum_{(j,k)\in(jb)}\theta_{jk}^2 \wedge L_* n^{-1}; \quad S_2 = \sum_{j=j_0}^{J-1}\sum_{(jb)\in G_j'}\sum_{(j,k)\in(jb)}\theta_{jk}^2 \wedge L_* n^{-1}.$$

Note that $card(G_j) \leq M_3 2^{j\gamma}$ and let J_1 and J_2 be two integers satisfying $2^{J_1} \asymp n^{1/(1+\alpha_1)}$ and $2^{J_2} \asymp n^{1/(1+\alpha)}$ respectively. Then

$$\begin{aligned} S_1 &\leq \sum_{j=J_0}^{J_1-1} \sum_{(jb)\in G_j} L_* n^{-1} + \sum_{j=J_1}^{J-1} \sum_{(jb)\in G_j} \sum_{(j,k)\in(jb)} \theta_{jk}^2 \leq L_* n^{-1} 2^{J_1\gamma} + C L_* 2^{-J_1(1+2\alpha_1-\gamma)} \\ &= o(n^{-2\alpha/(1+2\alpha)}), \end{aligned} \tag{47}$$

and

$$S_2 \leq \sum_{j=J_0}^{J_2-1} \sum_{(jb)\in G'_j} L_* n^{-1} + \sum_{j=J_2}^{J-1} \sum_{(jb)\in G'_j} \sum_{(j,k)\in(jb)} \theta_{jk}^2 \leq C n^{-2\alpha/(1+2\alpha)}. \tag{48}$$

Now (23) follows from (46), (47) and (48). The proof of (24) is similar. ∎

Remark 2 Under the conditions of Remark 1 following Theorem 4, Lemma 1 still holds with (29) and (30) replaced, respectively, by

$$|\xi_{Jk} - n^{-\frac{1}{2}} f(k/n)| \leq M_4 \|\phi\|_1 n^{-(1/2+\alpha_1\wedge 1)} \quad \text{if } k \in A_j$$

and

$$|\xi_{Jk} - n^{-\frac{1}{2}} f(k/n)| \leq M_5 \|\phi\|_1 n^{-(1/2+\alpha\wedge 1)} \quad \text{if } k \notin A_j.$$

These ensure that the approximation error is of higher order than the estimation error,

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

which is the same as in (44). The rest is identical to the proof of Theorem 4.

8.3 Proof of Theorem 5

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

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

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

$$Q_1 = \sum_{k=1}^{2^{j_0}} (E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2)^{1/2} |\phi_{j_0k}(x_0)| \leq 2^{j_0/2} \|\phi\|_\infty N n^{-1/2} \epsilon = o(n^{-\alpha/(1+2\alpha)}). \tag{49}$$

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

$$Q_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{jk}||\psi_{jk}(x_0)| \leq \sum_{j=J}^{\infty} N\|\psi\|_\infty 2^{j/2} C 2^{-j(1/2+\alpha)} \leq Cn^{-\alpha}. \tag{50}$$

Now consider the second term Q_2. First note that for function $f \in \Lambda^\alpha(M)$, the approximation error a_{jk} satisfies $|a_{jk}| \leq Cn^{-\alpha}2^{-j/2}$. By applying Lemma 1(ii) and Proposition 1, we have

$$\begin{aligned} Q_2 &\leq \sum_{j=j_0}^{J-1} \sum_{k \in K(j,x_0)} 2^{j/2}\|\psi\|_\infty (E(\hat{\theta}_{jk} - \theta_{jk})^2)^{1/2} \\ &\leq C \sum_{j=j_0}^{J-1} 2^{j/2}[(2^{-j(1+2\alpha)} + 2^{-j}n^{-2\alpha}) \wedge L_* n^{-1}\epsilon^2 + L_* n^{-2}\epsilon^2]^{1/2} \\ &= C(\log n/n)^{\alpha/(1+2\alpha)}. \end{aligned} \tag{51}$$

Combining (49), (50) and (51), we have $E(\hat{f}_n^*(x_0) - f(x_0))^2 \leq C(\log n/n)^{2\alpha/(1+2\alpha)}$. ∎

8.4 Proof of Theorem 6

The function is estimated by zero if and only if all the coefficients are estimated by zero. When $\theta_{jk} \equiv 0$, then the probability that a block is estimated by zero is $P(\sum_{k\in(jb)} z_{jk}^2 \leq \lambda_* L_*)$. Since z_{jk} are i.i.d. $N(0,1)$, $Y_{L_*} = \sum_{k\in(jb)} z_{jk}^2$ has a χ^2 distribution with L_* degrees of freedom. Lemma 4, together with the facts that $L_* = \log n$ and $\lambda_* = 3 + \log \lambda_*$, yield

$$P(Y_{L_*} \geq \lambda_* L_*) \leq e^{-L_* \cdot (\lambda_* - \log \lambda_* - 1)/2} = 1/n. \tag{52}$$

The total number of blocks is n/L_*, so it follows from (52) that

$$P(\hat{f}^* \equiv 0) = [1 - P(Y_{L_*} \geq \lambda_* L_*)]^{n/L_*} \geq [(1 - 1/n)^n]^{1/L_*} \tag{53}$$

Let $P_n = [(1-1/n)^n]^{1/L_*}$. Since $(1-1/n)^n \to e^{-1}$ and $1/L_* \to 0$, so $P_n \to 1$ as $n \to \infty$.

8.5 Proof of Theorem 1

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

$$f_n(t) = \sum_k \theta_{J_1 k}\psi_{J_1 k}(t)$$

where $\theta_{J_1 k} = c_0(\log n)^{\gamma/2} n^{-1/2} \asymp 2^{-J_1(1/2+\alpha)}$ 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 (43),

$$\tilde{y}_{jk} = \theta_{jk} + a_{jk} + n^{-1/2}\epsilon z_{jk} \tag{54}$$

Since $f_n \in \Lambda^\alpha(M)$, Lemma 1(ii) yields that the approximation error satisfies $|a_{jk}| \le Cn^{-\alpha}2^{-j/2}$.

For a given block $(J_1 b)$ at level J_1,

$$\begin{aligned}
\sum_{(j,k)\in(J_1 b)} E(\hat{\theta}_{jk}-\theta_{jk})^2 &\ge \sum_{(j,k)\in(J_1 b)} (\frac{1}{2}E[\hat{\theta}_{jk}-(\theta_{jk}+a_{jk})]^2 - a_{jk}^2) \\
&= \sum_{(j,k)\in(J_1 b)} [\frac{1}{2}E(\tilde{y}_{jk}-\theta_{jk})^2 I(S_{J_1 b}^2 > \lambda_n L) + \frac{1}{2}(\theta_{jk}+a_{jk})^2 P(S_{J_1 b}^2 \le \lambda_n L) - a_{jk}^2] \\
&\ge \frac{1}{4}\sum_{(j,k)\in(J_1 b)} \theta_{jk}^2 P(S_{J_1 b}^2 \le \lambda_n L) - 2\sum_{(j,k)\in(J_1 b)} a_{jk}^2 \qquad (55)
\end{aligned}$$

To get a lower bound for $P(S_{J_1 b}^2 \le \lambda_n L)$, we will apply Lemma 2 to $S_{J_1 b}^2$.

Since $|a_{jk}| \le Cn^{-\alpha}2^{-j/2}$, $\sum_{(J_1 b)} a_{jk}^2 \le Cn^{-1-4\alpha^2/(1+2\alpha)}(\log n)^{(1-2\alpha\gamma)/(1+2\alpha)}$. Hence there exists $N > 0$ such that for $n > N$, $\sum_{(J_1 b)} a_{jk}^2 \le \frac{1}{16}\lambda_n L n^{-1}\epsilon^2$. Now $\sum_{(J_1 b)} \theta_{jk}^2 = c_0^2 n^{-1}\log n$, so for small $c_0 > 0$, $\sum_{(J_1 b)} \theta_{jk}^2 \le \frac{1}{16}\lambda_n L n^{-1}\epsilon^2$. Choosing the constant $c_0 > 0$ small enough, we have, for $n > N$,

$$\sum_{(J_1 b)} (\theta_{jk}+a_{jk})^2 \le 2\sum_{(J_1 b)} \theta_{jk}^2 + 2\sum_{(J_1 b)} a_{jk}^2 \le \frac{1}{4}\lambda_n L n^{-1}\epsilon^2.$$

Then it follows from Lemma 2 that

$$\{S_{J_1 b}^2 \le \lambda L n^{-1}\epsilon^2\} = \{\sum_{(J_1 b)} (\theta_{jk}+a_{jk}+n^{-1/2}\epsilon z_{jk})^2 \le \lambda_n L n^{-1}\epsilon^2\} \supseteq \{\sum_{(J_1 b)} z_{jk}^2 \le \frac{1}{4}\lambda_n L\}.$$

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

$$P(S_{J_1 b}^2 \le \lambda_n L n^{-1}\epsilon^2) \ge P(\sum_{(J' b)} z_{jk}^2 \le 2L) \ge 1/2. \qquad (56)$$

Combining (55) and (56), we have, for large n,

$$E\|\hat{f}_n - f_n\|^2 \ge \sum_k E(\hat{\theta}_{J_1 k}-\theta_{J_1 k})^2 \ge \frac{1}{8}\sum_k \theta_{J_1 k}^2 - 2\sum_k a_{J_1 k}^2 = (c_0^2/8)(n/\log^\gamma n)^{-2\alpha/(1+2\alpha)}(1+o(1)).$$

So in this case,

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

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

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

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

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

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

8.6 Proof of Theorem 2

We give the proof of part (ii) in detail. With the thresholding constant λ_L chosen as in Section 4, the proof of part (i) is similar to that of Theorem 5.

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 small enough.

Noting that $\xi_{j_0k} = \langle f_n^*, \phi_{jk}\rangle = 0$ for all k and $\theta_{jk} = \langle f_n^*, \psi_{jk}\rangle = 0$ for all $(j,k) \neq (J',k')$, we have

$$\begin{aligned}
\mathcal{S} &\equiv \{\sup_{f\in\Lambda^\alpha(M)} E_f(\hat{f}_n(x_0) - f(x_0))^2\}^{1/2} \geq (E_{f_n^*}(\hat{f}_n(x_0) - f_n^*(x_0))^2)^{1/2} \\
&= \left\{E[(\hat{\theta}_{J'k'} - \theta_{J'k'})\psi_{J'k'}(x_0) + \sum_{k=1}^{2^{j_0}} \hat{\xi}_{j_0k}\phi_{j_0k}(x_0) + \sum_{(j,k)\in\mathcal{J}} \hat{\theta}_{jk}\psi_{jk}(x_0)]^2\right\}^{1/2} \qquad (57)
\end{aligned}$$

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 (57), we have

$$\begin{aligned}
\mathcal{S} &\geq (E(\hat{\theta}_{J'k'} - \theta_{J'k'})^2)^{1/2}|\psi_{J'k'}(x_0)| - \sum_{k=1}^{2^{j_0}} (E\hat{\xi}_{j_0k}^2)^{1/2}|\phi_{j_0k}(x_0)| - \sum_{(j,k)\in\mathcal{J}} (E\hat{\theta}_{jk}^2)^{1/2}|\psi_{jk}(x_0)| \\
&\equiv T_1 - T_2 - T_3. \qquad (58)
\end{aligned}$$

We will show that the first term T_1 is dominating and 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

$$\begin{aligned}
E(\hat{\theta}_{J'k'} - \theta_{J'k'})^2 &= E(\tilde{y}_{J'k'} - \theta_{J'k'})^2 I(S_{J'b}^2 > \lambda L n^{-1}\epsilon^2) + \theta_{J'k'}^2 P(S_{J'b}^2 \leq \lambda L n^{-1}\epsilon^2) \\
&\geq \theta_{J'k'}^2 P(S_{J'b}^2 \leq \lambda L n^{-1}\epsilon^2) \qquad (59)
\end{aligned}$$

Same as in the proof of Theorem 1,we will apply Lemma 2 to $S_{J_1b}^2$ to get a lower bound for $P(S_{J_1b}^2 \leq \lambda_n L)$. Note again that the approximation error a_{jk} satisfies $|a_{jk}| \leq Cn^{-\alpha}2^{-j/2}$, so $\sum_{(J'b)} a_{jk}^2 \leq Cn^{-1-4\alpha^2/(1+2\alpha)}L^{(2+2\alpha)/(1+2\alpha)}$. Hence there exists a constant $N_* > 0$ such that for $n > N_*$

$$\sum_{(J'b)} a_{jk}^2 \leq \frac{1}{4}(1 - \lambda^{-1/2})^2 \lambda L n^{-1}\epsilon^2. \qquad (60)$$

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

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

It follows from Lemma 2

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

So, $P(S^2_{J'b} \le \lambda L n^{-1}\epsilon^2) \ge P(\sum_{(J'b)} z^2_{jk} \le L) \ge 1/2$. Now (59) 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

$$T_1 = (E(\hat{\theta}_{J'k'} - \theta_{J'k'})^2)^{1/2} 2^{J'/2}|\psi(2^{J'}x_0 - k')| \ge \frac{1}{\sqrt{2}} c_0 c_1 n^{-\alpha/(1+2\alpha)} L^{\alpha/(1+2\alpha)}. \tag{61}$$

For T_2, same as in the proof of Theorem 5, we have

$$T_2 = \sum_{k=1}^{2^{j_0}} (E\hat{\xi}^2_{j_0k})^{1/2}|\phi_{j_0k}(x_0)| = o(n^{-\alpha/(1+2\alpha)}). \tag{62}$$

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

$$E\hat{\theta}^2_{jk} = E\tilde{y}^2_{jk} I(S^2_{jb} > \lambda L n^{-1}\epsilon^2) \le E\tilde{y}^2_{jk} = a^2_{jk} + n^{-1}\epsilon^2 \le C n^{-2\alpha} 2^{-j} + n^{-1}\epsilon^2. \tag{63}$$

So,

$$T_{31} \equiv \sum_{(j,k)\in\mathcal{J}_1} (E\hat{\theta}^2_{jk})^{1/2}|\psi_{jk}(x_0)| \le C n^{-\alpha}\log n + C n^{-1/2} 2^{j_1/2}\log n = o(n^{-\alpha/(1+2\alpha)}). \tag{64}$$

Now consider $(j,k) \in \mathcal{J}_2$. In this case, similar to (60), for large n, we have

$$\sum_{(jb)} a^2_{jk} \le (1 - (\frac{1+\lambda}{2\lambda})^{1/2})^2 \lambda L n^{-1}\epsilon^2.$$

It then follows from Lemma 2 that

$$\{S^2_{jb} \ge \lambda L n^{-1}\epsilon^2\} = \{\sum_{(jb)}(a_{jk} + n^{-1/2}\epsilon z_{jk})^2 \ge \lambda L n^{-1}\epsilon^2\} \subseteq \{\sum_{(jb)} z^2_{jk} \ge \frac{1}{2}(1+\lambda)L\}.$$

So for sufficiently n, we have

$$\begin{aligned} E\hat{\theta}^2_{jk} &= E\tilde{y}^2_{jk} I(S^2_{jb} > \lambda L n^{-1}\epsilon^2) \le 2n^{-1}\epsilon^2 E z^2_{jk} I(S^2_{jb} > \lambda L n^{-1}\epsilon^2) + 2a^2_{jk} \\ &\le 2n^{-1}\epsilon^2 EY(Y \ge \frac{1}{2}(1+\lambda)L) + 2a^2_{jk}, \end{aligned}$$

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

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

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

$$T_{32} \equiv \sum_{(j,k)\in\mathcal{J}_2} (E\hat{\theta}_{jk}^2)^{1/2}|\psi_{jk}(x_0)| \le C\beta^{-L/2}L^{1/2} + Cn^{-\alpha}L^{1/2} = o(n^{-\alpha/(1+2\alpha)}). \tag{65}$$

It follows by combining (64) and (65),

$$T_3 = T_{31} + T_{32} = o(n^{-\alpha/(1+2\alpha)}). \tag{66}$$

Putting together (61), (62), and (66), we have

$$\mathcal{S} \ge T_1 - T_2 - T_3 \ge \frac{1}{\sqrt{2}}c_0c_1n^{-\alpha/(1+2\alpha)}L^{\alpha/(1+2\alpha)}(1+o(1)). \tag{67}$$

Now (12) follows by letting $L = (\log n)^\rho$ with $\rho > 1$. ∎

8.7 Proof of Theorem 3

Let $\kappa(\lambda) \equiv (\lambda - \log\lambda - 1)/2$. Lemma 4 shows

$$\frac{2}{5}\lambda^{-1}L^{-1/2}e^{-L\cdot\kappa(\lambda)} \le P(Y_L > \lambda L) \le \pi^{-1/2}(\lambda-1)^{-1}L^{-1/2}e^{-L\cdot\kappa(\lambda)}.$$

First consider $\rho = 1$. Since $\kappa(\lambda) \ge 1$ for $\lambda \ge \lambda_L = 4.5052$, for $L = \log n$ and $\lambda \ge \lambda_L$, 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_L, then $\kappa(\lambda) < 1$ and it is easy to see that $p_L(\lambda) \to 1$. The case of $\rho = 0$ is similar.

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

$$\kappa(\lambda) = \beta + \delta - \log\beta - 1 + \sum_{m=1}^{M-1}(-1)^m\frac{\delta^m}{m\beta^m} + O(\frac{\delta^M}{\beta^M}). \tag{68}$$

Consider $0 < \rho \le 1/2$. Let $\lambda_L = 2(\log n)^{1-\rho} + \log(2(\log n)^{1-\rho}) + 1$. Applying (68) with $\beta = 2(\log n)^{1-\rho}$ and $\delta = \log(2(\log n)^{1-\rho}) + 1$, one has, for large n,

$$\kappa(\lambda_L) \ge (\log n)^{1-\rho} - \frac{\log(2(\log n)^{1-\rho})}{2(\log n)^{1-\rho}}.$$

Hence,

$$e^{-L\cdot\kappa(\lambda_L)} \le \sqrt{2}n^{-1}(\log n)^{(1-\rho)/2}. \tag{69}$$

Note that (69) also holds for any $\lambda \ge \lambda_L \ge 1$, since $\kappa(\lambda_L)$ is strictly increasing for $\lambda \ge 1$. Thus, for $\lambda \ge \lambda_L$,

$$p_L(\lambda) \le 1 - (1 - (\lambda-1)^{-1}/n)^{n/(\log n)^\rho} \to 0.$$

The other cases can be verified in the same way by using (68). We omit the details here for brevity. ∎

8.8 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 above by $(2\lambda + 2)L$ since

$$R(\theta) = E\|xI(S^2 > \lambda L) - \theta\|_2^2 \leq 2E\|x - \theta\|_2^2 + 2ES^2 I(S^2 \leq \lambda L) \leq (2\lambda + 2)L. \tag{70}$$

On the other hand,

$$R(\theta) = E\|x - \theta\|_2^2 I(S^2 > \lambda L) + \|\theta\|_2^2 P_\theta(S^2 \leq \lambda L) \leq 2\|\theta\|_2^2 + 2ES^2 I(S^2 > \lambda L). \tag{71}$$

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

$$R(\theta) \leq 8\|\theta\|_2^2, \quad \text{when } \|\theta\|_2^2 \geq L/2. \tag{72}$$

Now assume $\|\theta\|_2^2 < L/2$. Let $\mu = \|\theta\|_2^2$ and denote

$$g(\mu) = ES^2 I(S^2 > \lambda L). \tag{73}$$

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, et al. (1995)). We will need the Poisson form and the integral form:

$$f_{m,\mu}(y) = \sum_{k=0}^{\infty} \frac{(\mu/2)^k e^{-\mu/2}}{k!} f_{m+2k}(y) \tag{74}$$

$$f_{m,\mu}(y) = \frac{1}{2}\int_0^y [q(\sqrt{y-x} + \sqrt{\mu}) + q(\sqrt{y-x} - \sqrt{\mu})](y-x)^{-1/2} f_{m-1}(x)\, dx \tag{75}$$

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 (74), one has

$$g(\mu) = \sum_{k=0}^{\infty} \frac{(\mu/2)^k e^{-\mu/2}}{k!} EY_{L+2k} I(Y_{L+2k} > \lambda L) \tag{76}$$

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 (76),

$$g'(\mu) = \frac{1}{2}\sum_{k=0}^{\infty} \frac{(\mu/2)^k e^{-\mu}}{k!}(a_{k+1} - a_k) \tag{77}$$

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,

$$g'(\mu) = \sum_{k=0}^{\infty} \frac{(\mu/2)^k e^{-\mu/2}}{k!}\{P(Y_{L+2k+2} > \lambda L) + \lambda L f_{L+2k+2}(\lambda L)\} \leq 1 + \lambda L f_{L+2,\mu}(\lambda L) \tag{78}$$

Now use the integral form (75) 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 (41), 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,

$$\lambda L f_{L+2,\mu}(\lambda L) \le \frac{\lambda}{4\sqrt{\pi}}L^{1/2}(e^{-L}+e^{-L/4}) + \frac{\lambda}{2\pi\sqrt{2(\lambda-2)}}L^{3/2}\left(\frac{\lambda-2}{e^{\lambda-3}}\right)^{L/2}. \tag{79}$$

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

$$L^{1/2}e^{-aL} \le \sup_{x>0} xe^{-ax^2} = (2ae)^{-1/2}, \text{ and } L^{3/2}b^{-L} \le \sup_{x>0} x^3 b^{-x^2} = (3/(2e\log b))^{3/2}. \tag{80}$$

Set $a = 1$ and $a = 1/4$, and let $b = (e^{\lambda-3}/(\lambda-2))^{1/2}$, it follows from (79) and (80) that, for $\lambda \ge 4$,

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

Now (78) yields $g'(\mu) \le \lambda$ and hence $g(\mu) \le \lambda\mu + g(0) = \lambda\mu + EY_L I(Y_L > \lambda L)$. It now follows from Lemma 4 and (71) that

$$R(\theta) \le (2\lambda+2)\|\theta\|_2^2 + 2\lambda L(\lambda^{-1}e^{\lambda-1})^{-L/2}, \quad \text{when } \|\theta\|_2^2 < L/2. \tag{81}$$

The inequality (27) follows by putting together (70), (72), and (81). ∎

Acknowledgment

It is a pleasure to acknowledge helpful comments by Anirban DasGupta.

References

[1] Abramovich, F., Sapatinas, T. & Silverman, B.W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Stat. Soc. Ser. B,* **60**, 725-749.

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

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

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

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

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

[7] Cai, T. (1998). Wavelet regression using block thresholding. Web page available at `www.stat.purdue.edu/~tcai/blockshrink.html`

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

[9] Coifman, R.R. & Donoho, D.L. (1995). Translation invariant denoising. In A. Antoniadis and G. Oppenheim (eds), *Wavelets and Statistics*, Lecture Notes in Statistics **103**. New York: Springer-Verlag, pp. 125–150.

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

[11] Donoho, D.L. & Johnstone, I.M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika,* **81**, 425–455.

[12] Donoho, D.L. & Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Stat. Assoc.* **90**, 1200-24.

[13] Efroimovich, S.Y. (1999). Quasi-linear wavelet estimation. *J. Amer. Stat. Assoc.* **94**, 189-204.

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

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

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

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

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

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

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

[21] Lehmann, E.L. & Casella, G. (1998). *Theory of Point Estimation.* Springer, New York.

[22] Lepski, O.V. (1990). On a problem of adaptive estimation in white Gaussian noise. *Theory of Probability and Appl.* **35**, 3, 454-466

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

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

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

9 Appendix

The test functions are normalized so that all of the functions have the same $s.d.(f) = 100$. Doppler, HeaviSine, Bumps and Blocks are from Donoho & Johnstone (1994). Blip and Wave are from Marron, et al. (1998). Formulae for Spikes and Corner are given below.

Spikes:

$$f(x) = 15.6676 \cdot \left[e^{-500(x-0.23)^2} + 2\, e^{-2000(x-0.33)^2} + 4\, e^{-8000(x-0.47)^2} + 4\, e^{-8000(x-0.47)^2} + 3\, e^{-16000(x-0.69)^2} + e^{-32000(x-0.83)^2} \right]$$

Corner:

$$f(x) = 62.3865 \cdot [10x^3(1-4x^2) I_{(0,.5]}(x) + 3(0.125 - x^3)x^4 I_{(.5,.8]}(x) + 59.4432(x-1)^3 I_{(.8,1]}(x)]$$

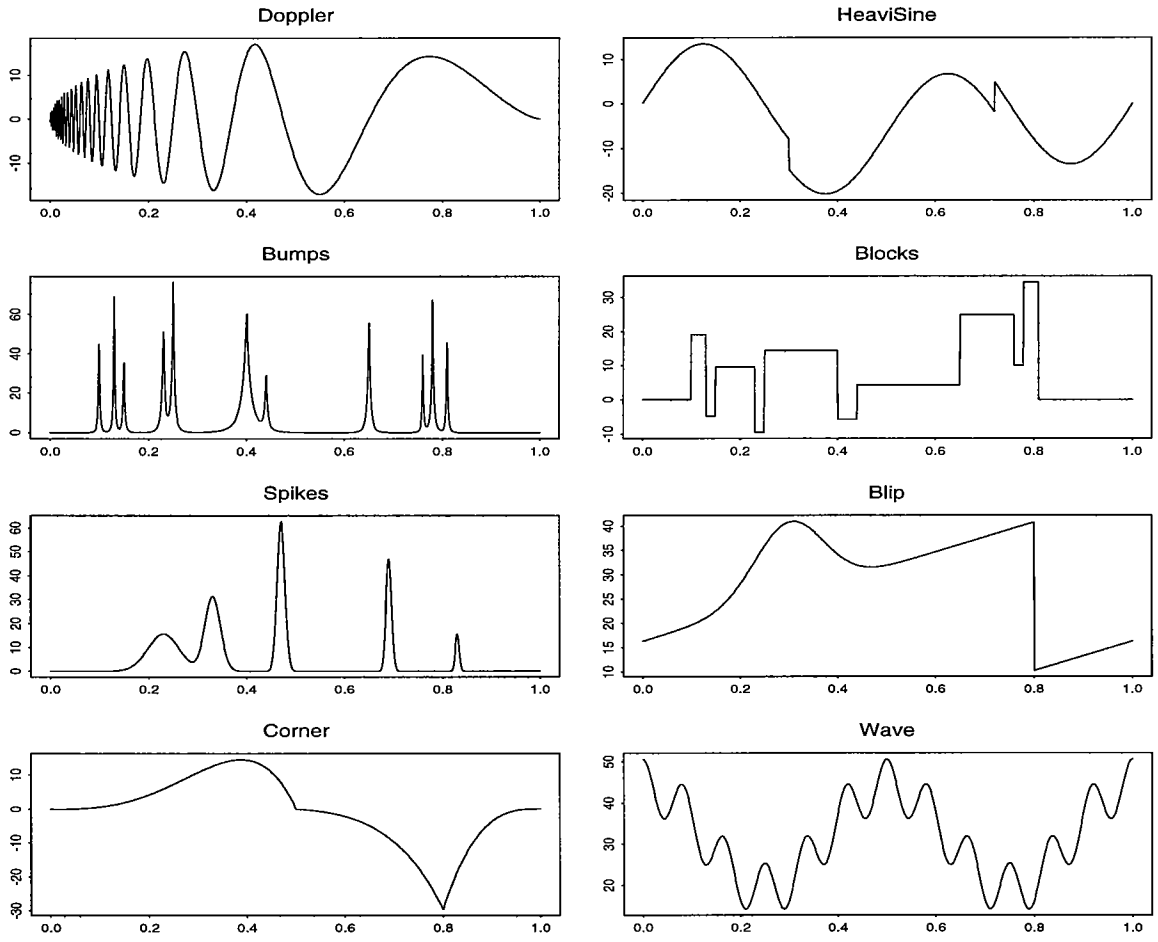

Figure 7: The test functions.