*Sankhyā: The Indian Journal of Statistics*
Special issue on Wavelets
2001, Volume 63, Series B, Pt. 2, pp. 127–148

# INCORPORATING INFORMATION ON NEIGHBOURING COEFFICIENTS INTO WAVELET ESTIMATION

*By* T. TONY CAI
*University of Pennsylvania, Philadelphia, U.S.A*
and
BERNARD W. SILVERMAN
*University of Bristol, Bristol, U.K*

*SUMMARY.* In standard wavelet methods, the empirical wavelet coefficients are thresholded term by term, on the basis of their individual magnitudes. Information on other coefficients has no influence on the treatment of particular coefficients. We propose and investigate a wavelet shrinkage method that incorporates information on neighbouring coefficients into the decision making. The coefficients are considered in overlapping blocks; the treatment of coefficients in the middle of each block depends on the data in the whole block. Both the asymptotic and numerical performances of two particular versions of the estimator are considered. In numerical comparisons with various methods, both versions of the estimator perform excellently; on the theoretical side, we show that one of the versions achieves the exact optimal rates of convergence over a range of Besov classes.

## 1. Introduction

Consider the nonparametric regression model

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

where $t_i = i/n$ for $i = 1, 2, \ldots n$, $\sigma$ is the noise level, and the $z_i$ are i.i.d. $N(0, 1)$. The function $f(\cdot)$ is an unknown function of interest.

Paper received June 2000; revised March 2001.
*AMS* (1991) *subject classification.* Primary 62G08; secondary 42C40, 62C20.
*Key words and phrases.* Adaptivity, Besov space, block thresholding, James-Stein estimator, local adaptivity, nonparametric regression, wavelets, white noise model.

Wavelet methods are attractive for nonparametric function estimation because of their spatial adaptivity, computational efficiency and asymptotic optimality properties. Standard wavelet methods achieve adaptivity through term-by-term thresholding of the empirical wavelet coefficients. Typically, to obtain the wavelet coefficients of the function estimate, each individual empirical wavelet coefficient $y$ is compared with a predetermined threshold $\tau$, and is processed taking account solely of its own magnitude. Other coefficients have no influence on the estimate. Examples of shrinkage functions applied to individual coefficients include the hard thresholding function $\eta_\tau^h(y) = y \cdot I(|y| > \tau)$ and the soft thresholding function $\eta_\tau^s(y) = \mathrm{sgn}(y) \cdot (|y| - \tau)_+$. For example, Donoho and Johnstone's (1994) VisuShrink estimates the true wavelet coefficients by soft thresholding with the *universal threshold* $\tau = \sigma(2 \log n)^{1/2}$.

Hall *et al.* (1999) and Cai (1996, 1999a and 1999b) studied local block thresholding rules for wavelet function estimation. These threshold the empirical wavelet coefficients in groups rather than individually, making simultaneous decisions to retain or to discard all the coefficients within a block. The aim is to increase estimation accuracy by utilizing information about neighbouring wavelet coefficients. These methods group coefficients in nonoverlapping blocks. The multiwavelet threshold estimators considered by Downie and Silverman (1998) also utilize block thresholding ideas.

In the present paper, we investigate wavelet shrinkage methods that incorporate information about neighbouring coefficients in a different way. The coefficients are considered in overlapping blocks. The basic motivation of block thresholding remains: if neighbouring coefficients contain some signal, then it is likely that the coefficients of current direct interest also do, and so a lower threshold should be used, essentially yielding a different local tradeoff between signal and noise. Two particular approaches are considered. One method, which we call *NeighCoeff*, chooses a threshold for each coefficient by reference not only to that coefficient but also to its neighbors. In the other approach, called *NeighBlock*, we aim to incorporate the advantages of the block thresholding method by estimating wavelet coefficients simultaneously in groups, but again use neighbouring coefficients outside the block of current interest in fixing the threshold. Both methods are specified completely, with explicit definition of both the block size and the threshold level.

After Section 2 in which basic notation and definitions are reviewed, the two estimators are defined in Section 3. We then investigate the two estimators both practically and theoretically. In Section 4, the estimators are applied both to simulated and real data, with good performance relative to other wavelet methods, with the NeighCoeff method performing particularly

well. Some theoretical results are derived in Section 5, where we show that a sequence space versions of the estimators enjoy a high degree of adaptivity. Specifically, we prove that the NeighBlock estimator simultaneously attains the exact optimal rate of convergence over a wide interval of the Besov classes with $p \geq 2$ without prior knowledge of the smoothness of the underlying functions. Over the Besov classes with $p < 2$, the estimator simultaneously achieves the optimal convergence rate within a logarithmic factor. We also prove that the NeighCoeff estimator is within a logarithmic factor of being minimax over a range of Besov classes. As shown in Cai (2000), this extra logarithmic factor is unavoidable for any estimator which uses a fixed number (independent of $n$) of empirical coefficients to estimate each wavelet coefficient.

The estimators are appealing visually as well as quantitatively. The reconstructions jump where the target function jump; the reconstruction is smooth where the target function is smooth. They do not contain the spurious fine-scale structure contained in some wavelet estimators, but adapt well to subtle changes in the underlying functions. The web site Cai and Silverman (1999) contains SPlus scripts implementing both our estimators. It also describes additional simulation results not included in this paper.

## 2. Wavelet Methods for Function Estimation

2.1 *Further background, notation and conventions.* We shall assume that we are working within an orthonormal wavelet basis generated by dilation and translation of a compactly supported scaling function $\phi$ and a mother wavelet $\psi$.

For simplicity in exposition, we work with periodized wavelet bases on $[0, 1]$, letting

$$\phi^p_{j,k}(t) = \sum_{l=-\infty}^{\infty} \phi_{j,k}(t-l), \quad \psi^p_{j,k}(t) = \sum_{l=-\infty}^{\infty} \psi_{j,k}(t-l), \qquad \text{for } t \in [0,1]$$

where

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k), \quad \psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k).$$

The collection $\{\phi^p_{j_0,k},\ k = 1, \ldots, 2^{j_0};\ \psi^p_{j,k},\ j \geq j_0 \geq 0, k = 1, ..., 2^j\}$ is then an orthonormal basis of $L^2[0, 1]$, provided the primary resolution level $j_0$ is large enough to ensure that the support of the scaling functions and wavelets at level $j_0$ is not the whole of $[0, 1]$. The superscript "$p$" will be suppressed from the notation for convenience.

An orthonormal wavelet basis has an associated exact orthogonal Discrete Wavelet Transform (DWT) that is norm-preserving and transforms sampled data into the wavelet coefficient domain in $O(n)$ steps. We use the standard device of transforming the problem in the function domain into a problem, in the sequence domain, of estimating the wavelet coefficients. See Daubechies (1992) and Strang (1992) for further details about the wavelets and the discrete wavelet transform.

Wavelets are known for their excellent compression and localization properties. In very many cases of interest, information about a function is essentially contained in relatively small number of large coefficients. Figure 1 displays the wavelet coefficients of the well-known test function Bumps (Donoho and Johnstone, 1994). It shows that large detail coefficients come as groups; they cluster around the areas where the function changes significantly.

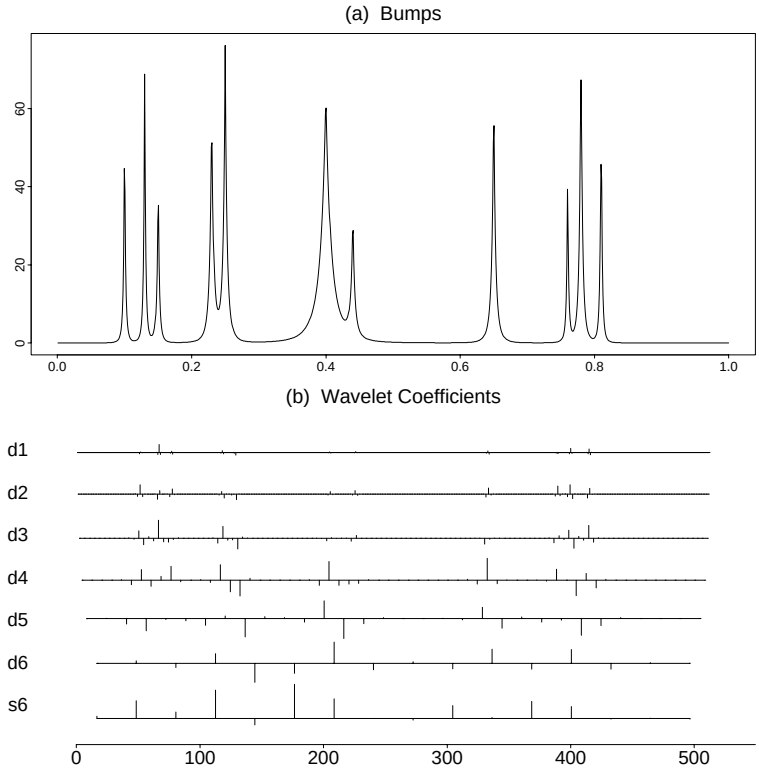


Figure 1. Wavelet coefficients of the Bumps function

This example illustrates the motivation for our methods—a coefficient is more likely to contain signal if neighbouring coefficients do also. There-

fore when the observations are contaminated with noise, estimation accuracy might be improved by incorporating information on neighbouring coefficients. Indeed, as we shall see, our estimators show significant numerical improvement over the conventional term-by-term thresholding estimators.

Suppose we observe the data $Y = \{y_i\}$ as in (1). We shall assume that the noise level $\sigma$ is known. Let $\tilde{\Theta} = W \cdot Y$ be the discrete wavelet transform of $Y$. Then $\tilde{\Theta}$ is an $n$-vector with elements $\tilde{\xi}_{j_0,k}$ $(k = 1, \cdots, 2^{j_0})$, which are the gross structure scaling function terms at the lowest resolution level, and $\tilde{\theta}_{j,k}$ $(j = j_0, \cdots, J-1, k = 1, \cdots, 2^j)$, which are fine structure wavelet terms. Since the DWT is an orthogonal transform, the coefficients are independently normally distributed with variance $\sigma^2$.

For any particular estimation procedure based on the wavelet coefficients, we use the notation $\hat{\Theta}$ for the estimate of the DWT $\Theta$ of the values of $f$ at the sample points. Up to the error involved in approximating $f$ at the finest level by a wavelet series, the mean integrated square error of the estimation satisfies

$$E\|\hat{f} - f\|_2^2 = n^{-1} E\|\hat{\Theta} - \Theta\|^2.$$

We therefore measure quality of recovery in terms of the mean square error in wavelet coefficient space.

## 3. The NeighBlock and NeighCoeff Procedures

We now define the estimates studied in this paper. We give a definition of the NeighBlock estimator first, because the NeighCoeff estimator can then be defined by reducing the basic block length to 1.

3.1 *The NeighBlock method.* The NeighBlock method has the following steps, aiming to build on the advantages previously found for block thresholding by incorporating information about neighbouring coefficients. The procedure is simple and easy to implement, and has a computational cost of $O(n)$.

STEP 1. Transform the data into the wavelet domain via the discrete wavelet transform: $\tilde{\Theta} = W \cdot Y$.

STEP 2. At each resolution level $j$, group the empirical wavelet coefficients into disjoint blocks $b_i^j$ of length $L_0 = [(\log n)/2]$. (If necessary, shorten one or both of the $b_i^j$ at the boundary to ensure that the blocks are nonoverlapping.)

STEP 3. Extend each block $b_i^j$ by an amount $L_1 = \max(1, [L_0/2])$ in each direction to form overlapping larger blocks $B_i^j$ of length $L = L_0 + 2L_1$. (If periodic boundary conditions are not being used, then the $b_i^j$ at the boundary are only extended in one direction to form $B_i^j$, again of length $L$.)

STEP 4. Within each block $b_i^j$, estimate the coefficients simultaneously via a shrinkage rule

$$\hat{\theta}_{j,k} = \beta_i^j \tilde{\theta}_{j,k}, \quad \text{for all } (j,k) \in b_i^j.$$

The shrinkage factor $\beta_i^j$ is chosen with reference to the coefficients in the *larger* block $B_i^j$:

$$\beta_i^j = (1 - \lambda_* L\sigma^2 / S_{j,i}^2)_+ \tag{2}$$

where

$$S_{j,i}^2 = \sum_{(j,k)\in B_i^j} \tilde{\theta}_{j,k}^2 \tag{3}$$

and $\lambda_* = 4.50524...$ is the solution of the equation $\lambda - \log\lambda = 3$. We can envision $B_i^j$ as a sliding window which moves $L_0$ positions each time and, for each given window, only the half of the coefficients in the center of the window are estimated.

STEP 5. Obtain the estimate of the function via the inverse discrete wavelet transform of the denoised wavelet coefficients.

The value of the thresholding coefficient $\lambda_*$ is derived from an oracle inequality introduced in Cai (1999a). Reasons for this choice will be discussed further when we consider the theoretical properties of the estimator. Note, in contrast to some other block thresholding methods, the various parameters are fully specified: the block length $L_0 = [(\log n)/2]$ depends on the sample size $n$ only and the thresholding constant $\lambda_*$ is an absolute constant.

The estimator can be modified by averaging over every possible position of the block centers. The resulting estimator sometimes has numerical advantages, at the cost of higher computational complexity.

3.2 *The NeighCoeff method.* The NeighCoeff procedure follows the same steps as the NeighBlock estimator, but with $L_0 = L_1 = 1$, $L = 3$, and $\lambda = \frac{2}{3}\log n$. The effect is that each individual coefficient is shrunk by an amount that depends on the coefficient and on its immediate neighbors.

NeighCoeff uses a lower threshold level than the VisuShrink method of Donoho and Johnstone (1994). In NeighCoeff, a coefficient is estimated by zero only when the sum of squares of the empirical coefficient and its

immediate neighbors is less than $2\sigma^2 \log n$, or the average of the squares is less than $\frac{2}{3}\sigma^2 \log n$.

3.3 *Discussion.* In this paper, our main concern is with the nonparametric regression estimation of a function observed at regular intervals with independent homoscedastic noise. Nevertheless, the idea of the NeighBlock and NeighCoeff procedures can be generalized to treat other statistical function estimation problems. For instance, Johnstone and Silverman (1997) considered the case of data observed with stationary correlated noise. Such data lead to a wavelet transform that has level-dependent variance, but within each level the variance is constant, and their paper showed that thresholding such data as if they were independent would give good results. It is therefore straightforward to apply a block thresholding procedure in a case of this kind. Even though theoretical work remains to be done on the precise properties of such a procedure, the results of Johnstone and Silverman (1997) are encouraging.

Data with more general structure were considered by Kovac and Silverman (2000). Their work covers both the case of data observed at irregularly spaced design points and of data with more general covariance structure, and their paper provides efficient methods for finding the variances of all the empirical wavelet coefficients. A natural approach is then to rescale each coefficient by its own standard deviation, apply one of the block thresholding methods set out above, and then refer back to the original scale.

Thresholding of coefficients with unequal variances also arises in wavelet approaches to density estimation. For a discussion of the use of wavelets in density estimation, and for further references, see Herrick *et al.* (2001). Suppose we observe a random sample $X_1,\ X_2, \ldots, X_n$ from a density $f$ with wavelet expansion

$$f(t) = \sum_k \xi_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{\infty}\sum_k \theta_{jk}\psi_{jk}(t)$$

with the wavelet coefficients

$$\begin{aligned} \xi_{j_0 k} &= \int \phi_{j_0 k}(x) f(x)\, dx = E_f \phi_{j_0 k}(X) \\ \text{and } \theta_{jk} &= \int \psi_{jk}(x) f(x)\, dx = E_f \psi_{jk}(X). \end{aligned}$$

Denote the empirical wavelet coefficients by

$$\tilde{\xi}_{j_0 k} = \frac{1}{n}\sum_{i=1}^{n} \phi_{j_0 k}(X_i) \ \text{ and } \ \tilde{\theta}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \psi_{jk}(X_i).$$

At each level, the empirical wavelet coefficients will only be nonzero for a finite range of indices $k$. Herrick *et al.* consider ways of estimating the variances of the coefficients, and these estimates can then be used within a block thresholding procedure. A particular issue that requires some careful thought is the treatment of the non-normal distributions that arise at finer levels of the transform.

Detailed study of all these extensions of the NeighBlock and NeighCoeff estimators is an interesting topic for future work.

## 4. **Numerical Comparisons**

We first explore the performance of the estimators beginning with two illustrative examples, and then considering a more detailed simulation study. We implement the NeighBlock and NeighCoeff estimators in the software package S+Wavelets. The programs are available from the web site Cai and Silverman (1999).

The comparison methods include Donoho and Johnstone's Visu Shrink and SureShrink as well as Coifman and Donoho's Translation-Invariant (TI) denoising method. SureShrink selects the threshold at each resolution level by minimizing Stein's (1981) unbiased estimate of risk. In the simulation, we use the hybrid method proposed in Donoho and Johnstone (1995). The TI-denoising method was introduced by Coifman and Donoho (1995), and is equivalent to averaging over estimators based on all the shifts of the original data. This method has various advantages over the universal thresholding methods. For further details see the original papers. In the systematic simulation study in Section 4.3, we also consider the BlockJS estimator introduced in Cai (1999a). The BlockJS estimator has been shown to perform well both numerically and theoretically; see Cai (1999a) for further details.

4.1 *A simulated signal of varying frequency.* Figure 2 displays a noisy Doppler signal as well as reconstructions obtained using various methods. All the methods except SureShrink recover the smooth low frequency part reasonably well. Both NeighBlock and NeighCoeff automatically adapt to the changing frequency of the underlying signal. Both estimate the smooth and low frequency part accurately; at the same time, they also capture the more rapidly oscillating area between $t = 0.1$ and $t = 0.4$. In contrast, both VisuShrink and TI de-noising significantly over-smooth in this region. SureShrink does better than VisuShrink and TI de-noising in recovering the high frequency part, but it contains noticeable spurious local fluctuation and

is visually unpleasant. None of the estimators does a particularly good job in the region $t < 0.1$ of very high frequency oscillation, partly because of the

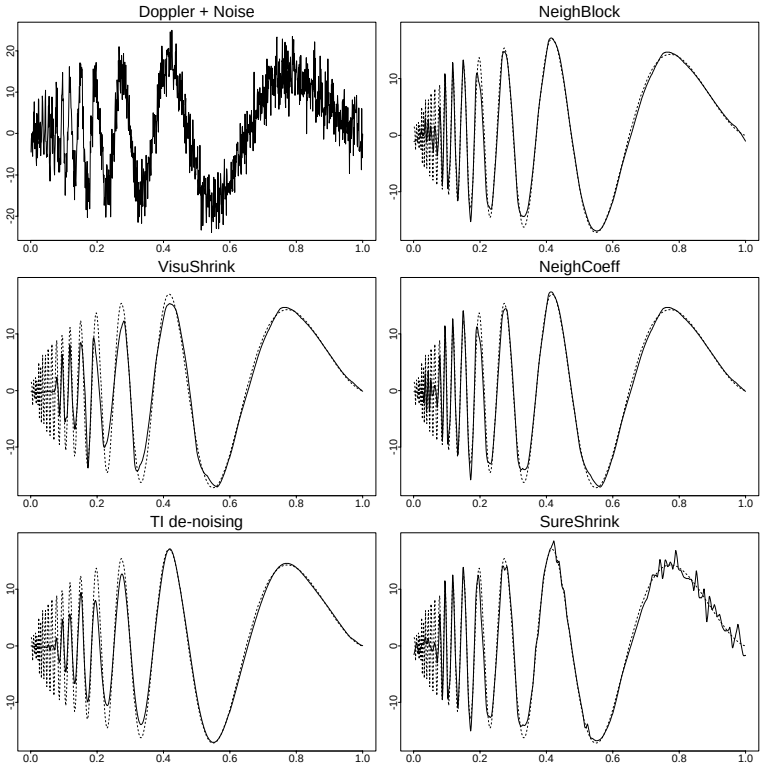


Figure 2. The noisy Doppler signal (top left panel) and the reconstructions (methods as labelled). The dotted line is the true signal.

low sampling rate relative to the rate of oscillation; however, in contrast to any of the other estimators, both NeighBlock and NeighCoeff do recover a little of the signal even in this region.

Quantitatively, NeighBlock and NeighCoeff are almost identical and both are significantly better than the other methods. In this particular example, the ratios of the mean squared error of NeighBlock and NeighCoeff to those of VisuShrink, SureShrink, and TI de-noising are 0.35, 0.72, and 0.45 respectively.

Inspection of wavelet coefficients shows that NeighBlock NeighCoeff, VisuShrink, and SureShrink use 33, 28, 15, and 61 detail coefficients in the reconstruction, respectively. SureShrink retains many detail coefficients in the

low frequency area and as a result, the reconstruction contains spurious oscillations. VisuShrink keeps only 15 detail coefficients and the reconstruction is over-smoothed. The additional smoothing inherent in the TI-denoising method has also led to over-smoothing.

4.2 *An anesthesiology example.* Figure 3 shows a typical segment of the result of the same methods applied to the inductance plethysmography data analyzed, for example, by Abramovich *et al.* (1998). Because this is real data there is no 'right' answer, but both the VisuShrink and TI denoising estimates smooth out the broad features of the curve, while the SureShrink estimator contains high frequency effects near times 300 and 335, both of which are almost certainly spurious.

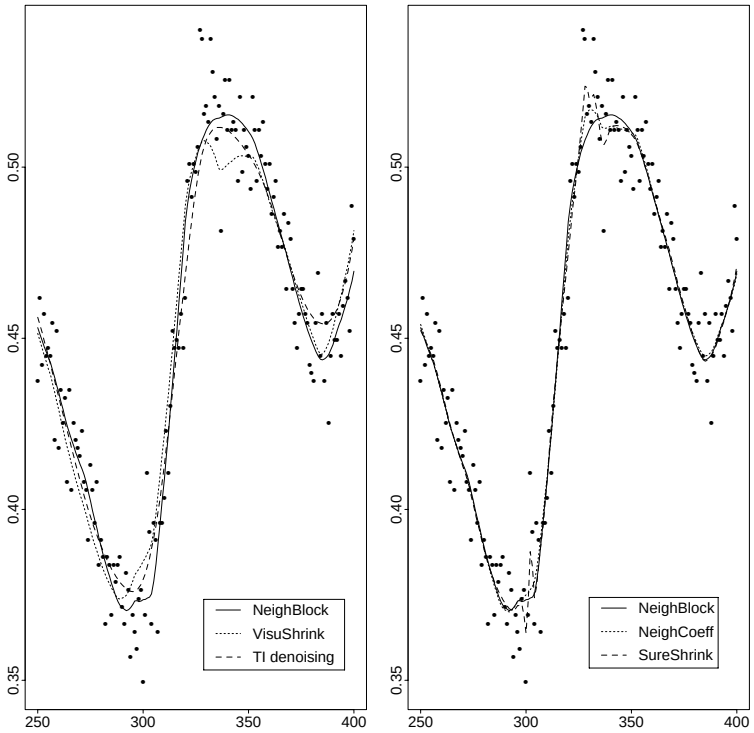


Figure 3. A segment of the data and curve estimates for the inductance plethysmography data. Left figure: NeighBlock (solid), VisuShrink (dotted), TI denoising (dashed). Right figure: NeighBlock (solid), NeighCoeff (dotted), SureShrink (dashed).

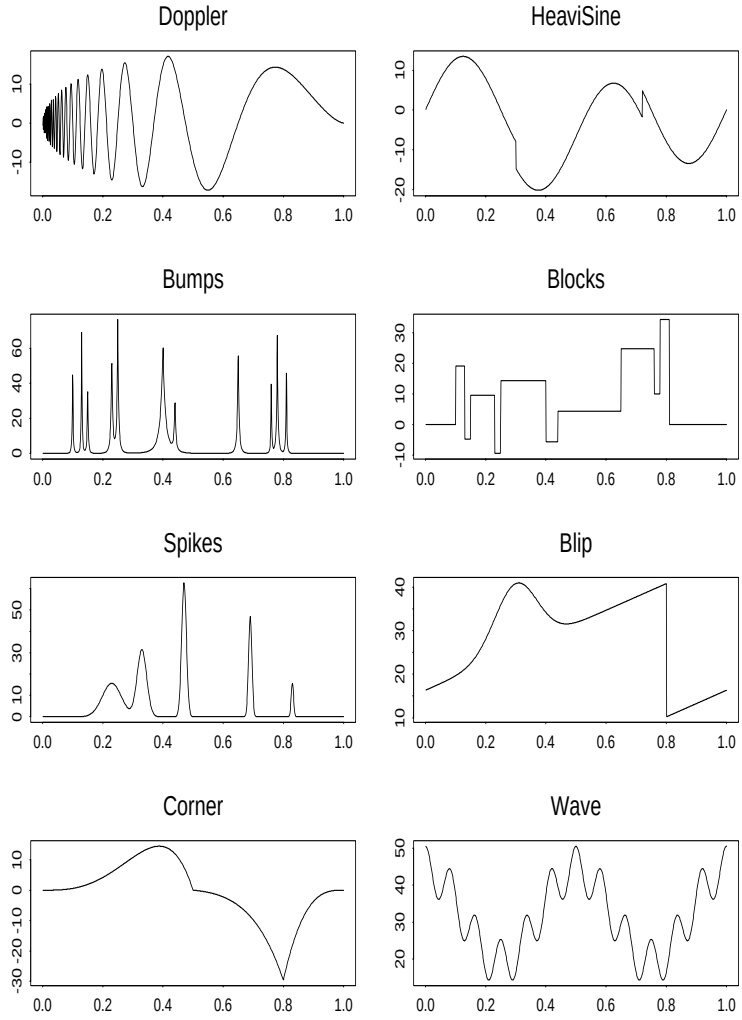


Figure 4. Test functions. Doppler, HeaviSine, Bumps and Blocks are from Donoho and Johnstone (1994). Blip and Wave are from Marron *et al.* (1998). The test functions are normalized so that every function has standard deviation 10. Formulae for Spikes and Corner are given in Cai (1999a).

4.3 *A simulation study.* To provide a more systematic comparison, we compared the numerical performance of the methods using eight test functions representing different level of spatial variability. The test functions are plotted in Figure 4. Sample sizes ranging from $n = 512$ to $n = 8192$ and root-signal-to-noise ratios (RSNR) from 3 to 7 were considered. The RSNR

is the ratio of the standard deviation of the function values to the standard deviation of the noise. Several different wavelets were used.

For reasons of space, we only report in detail the results for one particular case, using Daubechies' compactly supported wavelet *Symmlet* 8 and RSNR equal to 3. Table 1 reports the average squared errors over 60 replications with sample sizes ranging from $n = 512$ to $n = 8192$. A graphical presentation is given in Figure 5. Different combinations of wavelets and signal-to-noise ratios yield basically the same results; for details see the web site Cai and Silverman (1999).

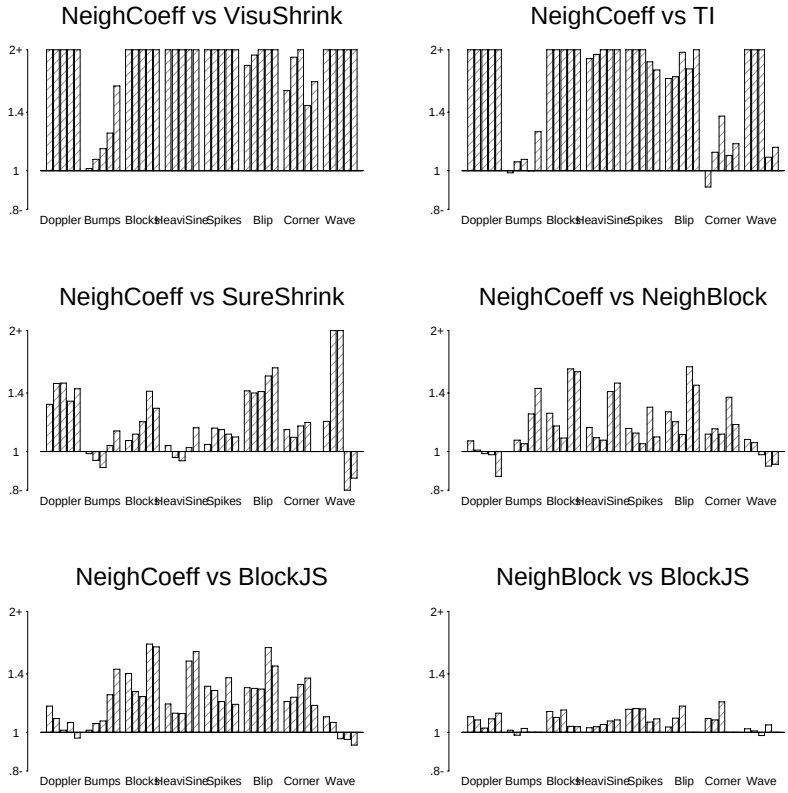


Figure 5. RSNR=3. The vertical bars represent the ratios of the MSEs of various estimators to the corresponding MSE of the NeighCoeff estimator. The higher the bar the better the relative performance of the NeighCoeff estimator, and a value of one means that the estimators have equal performance. The plotted ratios are truncated at a value of 2. For each signal the bars are ordered from left to right by the sample sizes $(n = 512, 1024, 2048, 4096, 8192)$.

Table 1. MEAN SQUARED ERROR FROM 60 REPLICATIONS (RSNR=3)

| $n$ | NeighCoeff | NeighBlock | SureShrink | TI-denoising | VisuShrink |
|---|---|---|---|---|---|
| *Doppler* | | | | | |
| 512 | 2.22 | 2.36 | 2.91 | 5.13 | 6.76 |
| 1024 | 1.34 | 1.35 | 1.98 | 3.36 | 4.49 |
| 2048 | 0.83 | 0.82 | 1.23 | 2.24 | 2.96 |
| 4096 | 0.51 | 0.50 | 0.68 | 1.25 | 1.61 |
| 8192 | 0.30 | 0.26 | 0.43 | 0.77 | 1.05 |
| *HeaviSine* | | | | | |
| 512 | 0.82 | 0.82 | 0.81 | 0.81 | 0.83 |
| 1024 | 0.59 | 0.63 | 0.56 | 0.62 | 0.63 |
| 2048 | 0.46 | 0.47 | 0.41 | 0.48 | 0.51 |
| 4096 | 0.28 | 0.36 | 0.30 | 0.29 | 0.36 |
| 8192 | 0.16 | 0.23 | 0.18 | 0.20 | 0.26 |
| *Bumps* | | | | | |
| 512 | 6.73 | 8.38 | 7.17 | 15.90 | 20.98 |
| 1024 | 3.66 | 4.24 | 4.04 | 10.08 | 13.63 |
| 2048 | 2.11 | 2.28 | 2.50 | 6.34 | 8.99 |
| 4096 | 1.08 | 1.75 | 1.54 | 3.42 | 5.09 |
| 8192 | 0.57 | 0.90 | 0.73 | 2.05 | 3.14 |
| *Blocks* | | | | | |
| 512 | 5.49 | 6.30 | 5.68 | 10.45 | 11.84 |
| 1024 | 3.78 | 4.09 | 3.65 | 7.37 | 8.29 |
| 2048 | 2.28 | 2.42 | 2.16 | 4.99 | 5.55 |
| 4096 | 1.39 | 1.96 | 1.42 | 2.92 | 3.38 |
| 8192 | 0.83 | 1.23 | 0.95 | 1.94 | 2.32 |
| *Spikes* | | | | | |
| 512 | 1.92 | 2.19 | 2.00 | 4.88 | 6.13 |
| 1024 | 1.18 | 1.31 | 1.35 | 3.11 | 4.00 |
| 2048 | 0.67 | 0.70 | 0.76 | 1.80 | 2.48 |
| 4096 | 0.38 | 0.49 | 0.42 | 0.71 | 1.19 |
| 8192 | 0.22 | 0.25 | 0.25 | 0.41 | 0.78 |
| *Blip* | | | | | |
| 512 | 1.06 | 1.33 | 1.50 | 1.80 | 1.94 |
| 1024 | 0.70 | 0.83 | 0.98 | 1.20 | 1.36 |
| 2048 | 0.39 | 0.43 | 0.55 | 0.77 | 0.93 |
| 4096 | 0.24 | 0.39 | 0.37 | 0.43 | 0.52 |
| 8192 | 0.13 | 0.19 | 0.21 | 0.28 | 0.34 |
| *Corner* | | | | | |
| 512 | 0.67 | 0.74 | 0.76 | 0.61 | 1.06 |
| 1024 | 0.36 | 0.41 | 0.40 | 0.40 | 0.69 |
| 2048 | 0.19 | 0.21 | 0.22 | 0.26 | 0.43 |
| 4096 | 0.11 | 0.15 | 0.13 | 0.12 | 0.16 |
| 8192 | 0.06 | 0.07 | 0.06 | 0.07 | 0.10 |
| *Wave* | | | | | |
| 512 | 2.65 | 2.84 | 3.15 | 5.75 | 7.14 |
| 1024 | 1.36 | 1.43 | 2.90 | 3.67 | 5.08 |
| 2048 | 0.55 | 0.54 | 3.18 | 2.22 | 3.27 |
| 4096 | 0.25 | 0.23 | 0.20 | 0.27 | 1.27 |
| 8192 | 0.14 | 0.13 | 0.12 | 0.16 | 0.70 |

The NeighBlock and NeighCoeff methods both uniformly outperform VisuShrink. For five of the eight test functions, Doppler, Bumps, Blocks, Spikes and Blip, our methods have better precision with sample size $n$ than VisuShrink with sample size $2n$ for all sample sizes where the comparison is possible. The NeighCoeff method is slightly better than NeighBlock in almost all cases, and outperforms the other methods as well. The NeighCoeff method is also better than TI-denoising in most cases, especially when the underlying function is of significant spatial variability. In terms of the mean square error criterion, conceivable competitors among the other methods are BlockJS and SureShrink. Both NeighCoeff and NeighBlock nearly always outperform BlockJS. Apart from being somewhat superior to SureShrink in mean square error, our methods yield noticeably better results visually; our estimates do not contain the spurious fine-scale effects that are often contained in the SureShrink estimator.

The curious behavior of some of the methods with the Waves signal calls for some explanation. Throughout, the primary resolution level $j_0 = [\log_2 \log n] + 1$ was used for all methods. Thus, $j_0 = 3$ for $n \leq 2048$, and $j_0 = 4$ for $n = 4096$ and 8192. This change in the value of $j_0$ affects whether or not the high frequency effect in the Waves signal is felt in the lowest level of wavelet coefficients. For $j_0 = 3$, the standard methods all smooth out the high frequency effect to some extent, because of applying a soft threshold with fixed threshold. An attractive feature of the NeighCoeff and NeighBlock methods is that they are not sensitive to the choice of primary resolution level in this way, because the threshold adapts to the presence of signal in all the coefficients.

4.4 *Summary of results.* Overall the two methods introduced in this paper have performed very well in comparison to other standard methods. If anything, the simple NeighCoeff procedure is the best of the estimators we have considered. Of course, there are many other approaches to the processing of wavelet coefficients now in the literature, but the simple message that could be applied more generally is that borrowing information from immediately neighbouring coefficients can make a substantial improvement.

One method we have not used in our comparisons is the block thresholding estimator of Hall *et al.* (1999). Their method requires the selection of smoothing parameters—block length and threshold level—neither of which is completely specified and no criterion is given for choosing the parameters objectively in finite sample cases. However, simulation results by Hall *et al.* (1997) show that even the translation-averaged version of the estimator has little advantage over VisuShrink when the signal to noise ratio is high. Our

simulation shows that NeighBlock uniformly outperforms VisuShrink in all examples, and indeed the relative performance of VisuShrink is even worse for values of RSNR higher than the one presented in detail. Therefore we expect our estimator to perform favourably over the estimator of Hall *et al.* in terms of mean squared error, at least in the case of high signal-to-noise-ratio.

## 5. **Theoretical Properties**

In the remainder of the paper, we consider the theoretical properties of our proposed estimators. In the Besov sequence space formulation that is by now classical for the analysis of wavelet regression methods, we find that both methods have excellent asymptotic properties. It should be noted that the Besov norms are invariant under permutation of the order of wavelet coefficients within each level of the transform, and it therefore may be the case that they do not completely capture the subtleties of inhomogeneous variability of functions actually arising in practice. This is an interesting topic for future work.

5.1 *Background.* Besov spaces are a very rich class of function spaces. They contain many traditional smoothness spaces such as Hölder and Sobolev Spaces. Full details of Besov spaces are given, for example, in DeVore and Popov (1988).

For a given square-integrable function $f$ on $[0,1]$, define the scaling function and wavelet coefficients of the wavelet expansion of $f$ by

$$\xi_{j,k} = \langle f, \phi_{j,k} \rangle, \qquad \theta_{j,k} = \langle f, \psi_{j,k} \rangle.$$

Let $\xi$ be the vector of the scaling function coefficients, and for each $j$ let $\theta_j$ be the vector of the wavelet coefficients at level $j$.

Suppose $\alpha > 0$, $0 < p \leq \infty$ and $0 < q \leq \infty$. Then, roughly speaking, the Besov function norm of index $(\alpha, p, q)$ quantifies the size in an $L_p$ sense of the derivative of $f$ of order $\alpha$, with $q$ giving a finer gradation; for a precise definition see DeVore and Popov (1988).

Define $s = \alpha + 1/2 - 1/p$. We call a wavelet $\psi$ *r-regular* if $\psi$ has $r$ vanishing moments and $r$ continuous derivatives. For a given $r$-regular mother wavelet $\psi$ with $r > \alpha$, the Besov sequence norm of the wavelet coefficients of a function $f$ is then defined by

$$\|\xi\|_p + \left( \sum_{j=j_0}^{\infty} 2^{jsq} \|\theta_j\|_p^q \right)^{1/q}. \tag{4}$$

It is an important fact (Meyer 1992) that the Besov function norm of index $(\alpha, p, q)$ of a function $f$ is equivalent to the sequence norm (4) of the wavelet coefficients of the function.

5.2 *Estimation in sequence space by NeighBlock and NeighCoeff* . In the present paper we shall confine our detailed theoretical discussion to a sequence space version of the NeighBlock and NeighCoeff estimators. Suppose $n = 2^J$ for some integer $J$ and that we observe sequence data

$$y_{j,k} = \theta_{j,k} + n^{-1/2}\sigma z_{j,k}, \quad j \geq 0,\ k = 1, 2, \cdots, 2^j \tag{5}$$

where $z_{j,k}$ are i.i.d. $N(0,1)$. The mean array $\theta$ is the object that we wish to estimate, and the accuracy of estimation is measured by the expected squared error

$$R(\hat{\theta}, \theta) = E \sum_{j,k} (\hat{\theta}_{j,k} - \theta_{j,k})^2.$$

We assume that $\theta$ is in some Besov Body $\Theta^s_{p,q}(M) = \{\theta : \sum_{j=j_0}^{\infty} 2^{jsq} \|\theta_j\|_p^q \leq M^q\}$. Make the usual calibration $s = \alpha + 1/2 - 1/p$. Donoho and Johnstone (1998) show that the minimax rate of convergence for estimating $\theta$ over the Besov body $\Theta^s_{p,q}(M)$ is $n^{-2\alpha/(1+2\alpha)}$ as $n \to \infty$.

We apply the NeighBlock procedure of Section 3.1 to the array of sample coefficients $\tilde{\theta}_{j,k}$ for $j < J$, to obtain estimated coefficients $\hat{\theta}_{j,k}$. For $j \geq J$ we set $\hat{\theta}_{j,k} = 0$. Similarly we denote by $\theta^*_{j,k}$ the result of applying the NeighCoeff procedure of Section 3.2, setting the estimate to zero for $j \geq J$.

We prove that both estimators attain the minimax rate up to logarithmic terms over all Besov Bodies $\Theta^s_{p,q}(M)$ with $\alpha p \geq 1$. For the NeighBlock estimator, our proofs yield the exact minimax rate for $p \geq 2$. The detailed results are as follows:

THEOREM **1** *Define $\hat{\theta}$ to be the NeighBlock estimator of the array $\theta$, as defined above. Then, as $n \to \infty$,*

$$\sup_{\theta \in \Theta^s_{p,q}(M)} E\|\hat{\theta} - \theta\|_2^2 \leq \begin{cases} Cn^{-2\alpha/(1+2\alpha)} & \text{for } p \geq 2 \\ Cn^{-2\alpha/(1+2\alpha)} (\log n)^{(2-p)/\{p(1+2\alpha)\}} & \text{for } p < 2 \\ & \text{and } \alpha p \geq 1. \end{cases} \tag{6}$$

THEOREM **2** *Define $\theta^*$ to be the NeighCoeff estimator of the array $\theta$. Then, for $\alpha p \geq 1$, as $n \to \infty$,*

$$\sup_{\theta \in \Theta^s_{p,q}(M)} E\|\theta^* - \theta\|_2^2 \leq C(\log n/n)^{2\alpha/(1+2\alpha)}.$$

Before proving these theorems, we remark that Donoho and Johnstone (1998) show a strong equivalence result between the nonparametric regression and the white noise models over Besov function classes of index $(\alpha, p, q)$. When the wavelet $\psi$ is $r$-regular with $r > \alpha$ and $p, q \geq 1$, then a simultaneously near-optimal estimator in the sequence estimation problem can be applied to the empirical wavelet coefficients in the function estimation problem in (1), and will be a simultaneously near-optimal estimator in the function estimation problem. For further details about the equivalence and approximation arguments, the readers are referred to Donoho and Johnstone (1995, 1998 and 1999) and Brown and Low (1996a). For approximation results, see also Chambolle *et al.* (1998).

5.3 *The choice of the thresholding constant $\lambda_*$ in NeighBlock.* In the NeighBlock procedure, the thresholding constant $\lambda_*$ is set to $\lambda_* = 4.505...$, which is the solution of the equation $\lambda - \log \lambda = 3$. The reasons for choosing this value is analogous to those for the choice of $(2 \log n)^{1/2}$ in term by term thresholding. Donoho and Johnstone (1994) use $(2 \log n)^{1/2}$ as thresholding constant in their VisuShrink estimator based on an oracle inequality and the following fact which makes the VisuShrink estimator almost "noise free". For $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,\ 1)$

$$P\left\{\max_i |Z_i| > (2 \log n)^{1/2}\right\} \to 0, \quad \text{as } n \to \infty.$$

In NeighBlock, the choice of $\lambda_*$ is also based an oracle inequality ( See Theorem 1 in Cai (1999a)) and the following smoothness property. Let $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,\ 1)$ and $L = \log n$. Divide $Z_i$ into blocks of size $L$, then the sums of squares $S_b^2 = \sum_{i=b(L-1)+1}^{bL} Z_i^2$ of the blocks satisfies

$$P\left\{\max_b\ S_b^2 > \lambda_* L\right\} \to 0, \quad \text{as } n \to \infty. \tag{7}$$

The value of $\lambda_* = 4.50524...$ is the smallest constant satisfying (7). With this choice of $\lambda_*$, the NeighBlock estimator, with high probability, removes pure noise completely. This smoothness property offers high visual quality of the reconstruction. The choice of $\lambda_*$ can also be motivated by a hypothesis testing formulation. See Cai (1999b) for further details.

Finally, we note that the theoretical results in Theorem 1 remain valid for any constant $\lambda \geq \lambda_*$. (Similarly, term by term thresholding estimators attain the same convergence rate of $(\log n/n)^{2\alpha/(1+2\alpha)}$ with the threshold $(a \log n)^{1/2}$ for any $a \geq 2$.)

5.4 *Proofs.* Our proofs depend on three lemmas. The first contains two key oracle inequalities for the estimators we are considering.

LEMMA **1** *Assume that $y_{j,k}$, $\hat{\theta}_{j,k}$ and $\theta^*_{j,k}$ are as defined in Section 5.2. Then, defining $\lambda_* > 1$ by $\lambda_* - \log \lambda_* = 3$, for each $i$ and $j_0 \le j < J$*

$$\sum_{(j,k)\in b_i^j} E(\hat{\theta}_{j,k} - \theta_{j,k})^2 \le \lambda_*(\sigma^2 n^{-1} \log n \wedge \sum_{(j,k)\in B_i^j} \theta_{j,k}^2) + 2n^{-2}\sigma^2. \tag{8}$$

$$E(\theta^*_{j,i} - \theta_{j,i})^2 \le (2\sigma^2 n^{-1} \log n) \wedge \sum_{k=i-1}^{i+1} \theta_{j,k}^2 + 2\sigma^2 n^{-2} (\log n)^{1/2} \tag{9}$$

At the boundary, the sum in (9) is taken over the block of length 3 containing $(j,i)$. The proof of this lemma is an extension of the proof of Theorem 1 of Cai (1999a), but with certain essential modifications. First consider (8). For $j,k$ in $B_i^j$ define

$$\theta^\dagger_{j,k} = (1 - n^{-1}\lambda_* L\sigma^2 / S^2_{j,i})_+ \, y_{j,k}.$$

Then $\theta^\dagger_{j,k} = \hat{\theta}_{j,k}$ for $(j,k)$ in $b_i^j$, so extending the sum from $b_i^j$ to $B_i^j$, and replacing $\hat{\theta}$ by $\theta^\dagger$, can only increase the left hand side of (8). The argument of Theorem 1 and Lemma 2 of Cai (1999a) shows that the inequality holds with these changes, completing the proof of (8). The proof of (9) follows from Theorem 1 in Cai (1999a) and the following upper bound for the tail probability of the $\chi^2_m$ for integers $m$:

$$P(\chi^2_m > \lambda m) \le \pi^{-1/2}(\lambda - 1)^{-1} m^{-1/2} \, e^{-\frac{m}{2}(\lambda - \log\lambda - 1)} \text{ for } \lambda > 1. \tag{10}$$

To prove (10), denote by $f_m(y)$ the pdf of a $\chi^2_m$ variable, and let $\tilde{F}_m(x)$ be the tail probability $\int_x^\infty f_m(y)dy$. Then, by exercise 16.7 of Stuart and Ord (1994) and elementary calculations for the $\chi^2_1$ distribution,

$$\tilde{F}_m(\lambda m) \le 2 \sum_{k=0}^{[(m-1)/2]} f_{m-2k}(\lambda m). \tag{11}$$

It is easy to see that, for $\ell \le m$,

$$f_\ell(\lambda m) = \frac{\ell}{\lambda m} f_{\ell+2}(\lambda m) \le \lambda^{-1} f_{\ell+2}(\lambda m). \tag{12}$$

Combining (11) and (12), one has

$$\tilde{F}(\lambda m) \le 2 \sum_{k=0}^{[(m-1)/2]} \lambda^{-k} f_m(\lambda m) \le \frac{2\lambda}{\lambda - 1} \cdot \frac{1}{2^{m/2}\Gamma(m/2)} (\lambda m)^{m/2-1} e^{-\lambda m/2}. \tag{13}$$

Now by Stirling's formula, $\Gamma(x+1) \geq \sqrt{2\pi}\, x^{x+1/2} e^{-x}$ for all $x > 0$, and so

$$\tilde{F}_m(\lambda m) \leq \pi^{-1/2}(\lambda - 1)^{-1} m^{-1/2} e^{-\frac{m}{2}(\lambda - \log \lambda - 1)}$$

as required, completing the proof of Lemma 1.

We now recall two elementary inequalities between two different $L_p$ norms, and a bound for a certain sum.

LEMMA **2** *Let $x \in I\!R^m$, and $0 < p_1 \leq p_2 \leq \infty$. Then the following inequalities hold:*

$$\|x\|_{p_2} \leq \|x\|_{p_1} \leq m^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2}. \tag{14}$$

LEMMA **3** *Let $0 < a < 1$ and $S = \{x \in I\!R^k : \sum_{i=1}^k x_i^a \leq B,\ x_i \geq 0,\ i = 1, \cdots, k\}$. Then for $\tau > 0$,*

$$\sup_{x \in S} \sum_{i=1}^{k} (x_i \wedge \tau) \leq B \cdot \tau^{1-a}.$$

We can now proceed to the proofs of the theorems themselves. We have

$$\begin{aligned} E\|\hat{\theta} - \theta\|_2^2 &= \sum_{j<j_0} \sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2 + \sum_{j=j_0}^{J-1} \sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2 + \sum_{j=J}^{\infty} \sum_k \theta_{j,k}^2 \\ &\equiv S_1 + S_2 + S_3, \end{aligned} \tag{15}$$

say. We bound the term $S_2$ by using Lemma 1. Let

$$A_i^j = \sum_{(j,k) \in B_i^j} \theta_{j,k}^2,$$

the sum of squared coefficients within the block $B_i^j$. We then split up the sum defining $S_2$ into sums over the individual blocks $b_i^j$, and apply the oracle inequality (8). Since $L = \log n$ and the number of blocks is definitely less than $n$, this yields

$$S_2 = \sum_{j=j_0}^{J-1} \sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2 \leq C \sum_{j=j_0}^{J-1} \sum_i (A_i^j \wedge \sigma^2 n^{-1} L) + 2n^{-1}\sigma^2. \tag{16}$$

Note also that, since $\theta \in \Theta_{p,q}^s(M)$, we have $2^{js}\|\theta_j\|_p \leq M$ for each $j$. We now complete the proof for the two cases separately.

**Case $p \geq 2$:** For $\theta \in \Theta_{p,q}^s(M)$, Lemma 2 implies that

$$\|\theta_j\|_2^2 \leq (2^j)^{2(\frac{1}{2} - \frac{1}{p})} \|\theta_j\|_p^2 \leq M^2 2^{2j(\frac{1}{2} - \frac{1}{p} - s)} = M^2 2^{-2\alpha j}. \tag{17}$$

It follows that

$$S_1 + S_3 \le 2^{j_0} n^{-1}\sigma^2 + \sum_{j=J}^{\infty} M^2 2^{-2\alpha j} = o(n^{-2\alpha/(1+2\alpha)}), \tag{18}$$

so that $S_1 + S_3$ can be neglected.

We divide the sum in (16) into two parts. Choose $J_1$ such that $2^{J_1} \asymp n^{1/(1+2\alpha)}$. Then,

$$\sum_{j=j_0}^{J_1-1}\sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \le \sum_{j=j_0}^{J_1-1}\sum_i \sigma^2 n^{-1} L \le C 2^{J_1} n^{-1} \le C n^{-2\alpha/(1+2\alpha)}, \tag{19}$$

and, making use of the bound (17),

$$\sum_{j=J_1}^{J-1}\sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \le \sum_{j=J_1}^{J-1}\sum_i A_i^j \le 2\sum_{j=J_1}^{J-1} \|\theta_j\|_2^2 \le C n^{-2\alpha/(1+2\alpha)}. \tag{20}$$

Combining (19) and (20) demonstrates that $S_2 \le C n^{-2\alpha/(1+2\alpha)}$, completing the proof for this case.

**Case $p < 2$ with $\alpha p \ge 1$:** For $\theta \in \Theta_{p,q}^s(M)$, Lemma 2 now yields $\|\theta_j\|_2^2 \le \|\theta_j\|_p^2 \le M^2 2^{-2js}$. The assumption $\alpha p \ge 1$ implies that $s \ge \frac{1}{2}$, so that

$$S_3 \le C\sum_{j=J}^{\infty} 2^{-2js} \le C n^{-2s} \le C n^{-1}.$$

Thus $S_1 + S_3 = o(n^{-2\alpha/(1+2\alpha)})$ as before.

Now let $J_2$ be an integer satisfying $2^{J_2} \asymp n^{1/(1+2\alpha)} (\log n)^{-(2-p)/p(1+2\alpha)}$. Then, by an argument analogous to that leading to (19),

$$\sum_{j=j_0}^{J_2-1}\sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \le \sum_{j=j_0}^{J_2-1}\sum_i \sigma^2 n^{-1} L \le C n^{-2\alpha/(1+2\alpha)} (\log n)^{(2-p)/p(1+2\alpha)}. \tag{21}$$

Turning to the other part of $S_2$, it follows from Lemma 2 that, for each $j$,

$$\sum_i (A_i^j)^{p/2} \le \sum_i \sum_{(j,k)\in B_i^j} (\theta_{j,k}^2)^{p/2} \le 2\sum_k (\theta_{j,k}^2)^{p/2} \le 2M^p 2^{-jsp}.$$

Applying Lemma 3 with $a = p/2$, we have, after some algebra,

$$\sum_{j=J_2}^{J-1}\sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \le C n^{-2\alpha/(1+2\alpha)} (\log n)^{(2-p)/p(1+2\alpha)}. \tag{22}$$

We complete the proof of Theorem 1 by combining the bounds (21) and (22), as in the case $p \geq 2$.

The proof of Theorem 2 is similar, using the oracle inequality (9) instead of (8).

*Acknowledgment.* Part of this work was carried out while Bernard Silverman was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, supported by National Science Foundation Grant number SBR-9601236. He is also grateful for the financial support of the British Engineering and Physical Sciences Research Council.

## References

Abramovich, F., Sapatinas, T. and Silverman, B.W. (1998). Wavelet thresholding via a Bayesian approach, *Journal of the Royal Statistical Society,* Series B, **60**, 725–749.

Brown, L.D. and Low, M.G. (1996a). Asymptotic equivalence of nonparametric regression and white noise, *Annals of Statistics*, **24**, 2384–2398.

– – – – (1996b). A constrained risk inequality with applications to nonparametric functional estimation, *Annals of Statistics*, **24**, 2524–2335.

Cai, T. (1996). Minimax wavelet estimation via block thresholding. *Technical Report* #96-41, Department of Statistics, Purdue University.

– – – – (1999a). Adaptive wavelet estimation: a block thresholding and oracle inequality approach, *Annals of Statistics* **27**, 898-924.

– – – – (1999b). On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Technical Report*, Department of Statistics, Purdue University.

– – – – (2000). On adaptability and information-pooling in nonparametric function estimation. *Technical Report*, Department of Statistics, University of Pennsylvania.

Cai, T. and Silverman, B. W. (1999). Incorporating information on neighbouring coefficients into wavelet estimation. Web page available at `http://www-stat.wharton.upenn.edu/~tcai/ neighblock.html`

Chambolle, A., DeVore, R., Lee, N. and Lucier, B. (1998). Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage, *IEEE Transactions on Image Processing,* **70**, 319–335.

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

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

DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces, *Transaction of the American Mathematical Society*, **305**, 397–414.

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

– – – – (1995). Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association*, **90**, 1200–1224.

– – – – (1998). Minimax estimation via Wavelet shrinkage, *Annals of Statistics*, **26**, 879–921.

– – – – (1999). Asymptotic minimaxity of wavelet estimators with sampled data, *Statistica Sinica*, **9**, 1–32.

Downie, T.R. and Silverman, B.W. (1998). The discrete multiple wavelet transform and thresholding methods, *IEEE Transactions in Signal Processing*, **46**, 2558–2561.

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

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

Herrick, D.R.M., Nason, G.P. and Silverman, B.W. (2001). Some new methods for wavelet density estimation, *Sankhyā* Series A, **63**.

Johnstone, I.M. and Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise, *Journal of the Royal Statistical Society*, Series B, **59**, 319–351.

Kovac, A. and Silverman, B.W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding, *Journal of the American Statistical Association*, **95**, 172–183.

Lepski, O.V. (1990). On a problem of adaptive estimation in white Gaussian noise, *Theory of Probability and its Application*, **35**, 454–466.

Marron, J.S., Adak, S., Johnstone, I.M., Neumann, M.H. and Patil, P. (1998). Exact risk analysis of wavelet regression, *Journal of Computational and Graphical Statistics*, **7**, 278–309.

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

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *Annals of Statistics*, **9**, 1135–1151.

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

Stuart, A. and Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics. Volume 1: Distribution Theory*. Edward Arnold, London.

T. Tony Cai
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104, U.S.A.
Email: tcai@wharton.upenn.edu

Bernard W. Silverman
Institute for Advanced Studies
University of Bristol
Royal Fort House
Bristol, BS8 1UJ, U.K.
Email: b.w.silverman@bristol.ac.uk