INCORPORATING INFORMATION ON NEIGHBORING COEFFICIENTS INTO WAVELET ESTIMATION

by

T. Tony Cai and Bernard W. Silverman
Purdue University University of Bristol

Technical Report #98–13

Department of Statistics
Purdue University
West Lafayette, IN USA

June 1998

◇

Incorporating Information on Neighboring Coefficients into Wavelet Estimation

T. Tony Cai
Department of Statistics
Purdue University
West Lafayette, IN 47907
U.S.A.

Bernard W. Silverman
Department of Mathematics
University of Bristol
Bristol BS8 1TW
U.K.

Abstract

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 a wavelet shrinkage method that incorporates information on neighboring 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. The asymptotic and numerical performances of two particular versions of the estimator are investigated. We show that, asymptotically, one version of the estimator achieves the exact optimal rates of convergence over a range of Besov classes for global estimation, and attains adaptive minimax rate for estimating functions at a point. In numerical comparisons with various methods, both versions of the estimator perform excellently.

Keywords: Block Thresholding; Wavelet; James-Stein Estimator; Adaptivity; Nonparametric function Estimation; Besov Space.
AMS 1991 Subject Classification: Primary 62G07, Secondary 62G20.

1 Introduction

Consider the nonparametric regression model

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

where $t_i = i/n$ for $i = 1, 2, \ldots n$, σ 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.

Wavelet methods have demonstrated success in nonparametric function estimation in terms of spatial adaptivity, computational efficiency and asymptotic optimality. Standard wavelet methods achieve adaptivity through term-by-term thresholding of the empirical wavelet coefficients. To obtain the wavelet coefficients of the function estimate, each individual empirical wavelet coefficient y is compared with a predetermined threshold τ, 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 VisuShrink [12] estimates the true wavelet coefficients by soft thresholding with the *universal threshold* $\tau = \sigma(2 \log n)^{1/2}$.

Hall, Kerkyacharian and Picard [15] and Cai ([4] and [5]) 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 neighboring wavelet coefficients. Estimators obtained by block thresholding enjoy a higher degree of spatial adaptivity than the standard term-by-term thresholding methods. The multiwavelet threshold estimators considered by Downie and Silverman [14] also utilize block thresholding ideas.

In the present paper, we propose a wavelet shrinkage method that incorporates into the thresholding decision information about neighboring coefficients outside the block of current interest. The basic motivation is that if neighboring 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. Two particular cases are considered. The NeighBlock method estimates wavelet coefficients simultaneously in groups, with the aim of gaining the advantages of the block thresholding method. The NeighCoeff approach is a special case that estimates coefficients individually.

After Section 2.1 in which basic notation and definitions are reviewed, the two estimators are defined in Section 2.2. We then investigate the two estimators both theoretically and by simulation. We show in Section 3 that the NeighBlock estimator enjoys a high degree of adaptivity and spatial adaptivity. Specifically, we prove that the 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. For estimating functions at a point, the estimator also attains the local adaptive minimax rate.

The theoretical properties of NeighCoeff are discussed in Section 4. The estimator is within a logarithmic factor of being minimax over a range of Besov classes, and shares the pointwise optimality properties of NeighBlock. Technical details and proofs are given in Section 6.

In Section 5, a simulation study of the two estimators is reported, together with a comparison on a data set collected in an anesthesiology study. Both estimators have excellent performance relative to conventional wavelet shrinkage methods; perhaps contrary to the indications provided by the theoretical discussion, that of the NeighCoeff method is, if anything, slightly superior. 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 [6] contains SPlus scripts implementing the estimators, and additional simulation results not included in the paper.

2 The estimation method

2.1 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 ϕ and a mother wavelet ψ. We call a wavelet ψ *r-regular* if ψ has r vanishing moments and r continuous derivatives.

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

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

where

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

The collection $\{\phi_{j_0k}^p,\ k = 1, \ldots, 2^{j_0};\ \psi_{jk}^p,\ j \geq j_0 \geq 0, k = 1, ..., 2^j\}$ 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 [9] and Strang [21] for further details about the wavelets and the discrete wavelet transform.

Suppose we observe the data $Y = \{y_i\}$ as in (1). We shall assume that the noise level σ is known. Let $\tilde{\Theta} = W \cdot Y$ be the discrete wavelet transform of Y. Then $\tilde{\Theta}$ is an n-vector with elements $\tilde{\xi}_{j_0k}$ $(k = 1, \cdots, 2^{j_0})$, which are the gross structure wavelet terms

at the lowest resolution level, and $\tilde{\theta}_{jk}$ $(j = 1, \cdots, J-1, k = 1, \cdots, 2^j)$, which are fine structure wavelet terms. Since the DWT is an orthogonal transform, the coefficients are independently normally distributed with variance σ^2.

For any particular estimation procedure based on the wavelet coefficients, use the notation $\hat{\Theta}$ for the estimate of the DWT Θ 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.

2.2 The NeighBlock and NeighCoeff procedures

We now define the estimates studied in this paper. The estimator has the following stages:

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

2. At each resolution level j, group the empirical wavelet coefficients into disjoint blocks b_i^j of length L_0. Each block b_i^j is extended by an amount $L_1 = \max(1, [L_0/2])$ in each direction to form overlapping larger blocks B_i^j of length $L = L_0 + 2L_1$.

3. 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 β_i^j is chosen with reference to the coefficients in the *larger* block B_i^j:

$$\beta_i^j = (1 - \lambda L \sigma^2 / S_{ji}^2)_+ \tag{2}$$

   where

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

   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. The choice of the block length L_0 and the threshold λ are discussed below.

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

The procedure is simple and easy to implement, at a computational cost of $O(n)$. We shall consider two special cases, as follows:

NeighBlock: Set $L_0 = [(\log n)/2]$ so that $L \approx \log n$. This method aims to combine the advantages previously found for block thresholding methods with those obtained by using information about neighboring coefficients. In this case the threshold λ is chosen by a James-Stein procedure; see Remark 2 below.

NeighCoeff: Set $L_0 = 1$ and $L = 3$, so that each individual coefficient is shrunk by an amount that may also depend on its immediate neighbors. The threshold $\lambda \cdot L$ is set to $2 \log n$, or $\lambda = (2/3) \log n$; see Section 4 for further details.

Remark 1: If L_0 is not a power of two, then one or both of the b_i^j at the boundary is shortened to ensure all the b_i^j are nonoverlapping. In the periodic case, the corresponding B_i^j are kept of length L with b_i^j at the center. 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.

Remark 2: In the NeighBlock procedure, the thresholding constant λ is set to $\lambda_* = 4.505...$, which is the solution of the equation $\lambda - \log \lambda = 3$. The value λ_* is derived from an oracle inequality introduced in Cai [5]. The estimator then enjoys superior numerical performance and asymptotic optimality, as our subsequent discussion shows. Instead of using a fixed block length, the block length can be allowed to increase with the resolution level; all the asymptotic results hold if we set the block length at level j to be $[(\log 2^j)/2]$.

Remark 3: 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 Optimality of the NeighBlock procedure

3.1 Global properties

As is traditional in the wavelet literature, we investigate the adaptivity of the NeighBlock procedure across Besov classes. Besov spaces are a very rich class of function spaces. They contain many traditional smoothness spaces such as Hölder and Sobolev Spaces. We shall show that NeighBlock enjoys excellent adaptivity across a wide range of Besov classes. Full details of Besov spaces are given, for example, in DeVore and Popov [10].

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

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

The function f can be expanded into a wavelet series:

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j_0 k} \phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} \theta_{jk} \psi_{jk}(x) \tag{4}$$

Let ξ be the vector of the scaling function coefficients, and for each j let θ_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 (α, p, q) quantifies the size in an L_p sense of the derivative of f of order α, with q giving a finer gradation; for a precise definition see DeVore and Popov [10].

Define $s = \alpha + 1/2 - 1/p$. For a given r-regular mother wavelet ψ with $r > \alpha$, the Besov sequence norm of the wavelet coefficients of a function f is then defined by

$$\|\xi\|_p + |\theta|_{b^s_{p,q}},$$

where

$$|\theta|^q_{b^s_{p,q}} = \sum_{j=j_0}^{\infty} 2^{jsq} \|\theta_j\|^q_p. \tag{5}$$

It is an important fact (see Meyer [19]) that the Besov function norm $\|f\|_{B^\alpha_{p,q}}$ is equivalent to the sequence norm of the wavelet coefficients of f. The Besov class $B^\alpha_{p,q}(M)$ is defined to be the set of all functions whose Besov norm is less than M.

Denote the minimax risk over a function class $\mathcal{F}$ by

$$R(\mathcal{F}, n) = \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} E\|\hat{f}_n - f\|_2^2$$

where $\hat{f}_n$ are estimators based on n observed data points. Donoho and Johnstone [11] showed that the minimax risk over a Besov class $B^\alpha_{p,q}(M)$ is given by

$$R(B^\alpha_{p,q}(M), n) \asymp n^{-2\alpha/(1+2\alpha)}, \quad n \to \infty$$

If attention is restricted to *linear* estimates, the corresponding minimax rate of convergence is $n^{-\rho'}$, with

$$\rho' = \frac{\alpha + (1/p_- - 1/p)}{\alpha + 1/2 + (1/p_- - 1/p)}, \quad \text{where } p_- = \max(p, 2). \tag{6}$$

Therefore the traditional linear methods such as kernel, spline and orthogonal series estimates are suboptimal for estimation over the Besov bodies with $p < 2$.

We show in the following theorem that the NeighBlock method attains the exact optimal convergence rate over a wide range of the Besov scales. We denote by C a generic constant that may vary from place to place.

Theorem 1 *Suppose the wavelet ψ is r-regular. Then the NeighBlock estimator satisfies*

$$\sup_{f \in B^\alpha_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \leq C n^{-2\alpha/(1+2\alpha)} \tag{7}$$

for all $M \in (0, \infty), \alpha \in (0, r), q \in [1, \infty]$ and $p \in [2, \infty]$.

Thus, the NeighBlock estimator, without knowing the degree or amount of smoothness of the underlying function, attains the true optimal convergence rate that one could achieve knowing the regularity. The next theorem addresses the case of $p < 2$, and shows that the NeighBlock method achieves advantages over linear methods even at the level of rates.

Theorem 2 *Assume that the wavelet ψ is r-regular. Then the NeighBlock estimator is simultaneously within a logarithmic factor of minimax for $p < 2$:*

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \leq Cn^{-2\alpha/(1+2\alpha)}(\log n)^{(2-p)/p(1+2\alpha)} \tag{8}$$

for all $M \in (0, \infty), \alpha \in [1/p, r), q \in [1, \infty]$ and $p \in [1, 2)$.

The proofs of these theorems are given in Section 6.

3.2 Local adaptation

We now study the optimality of the NeighBlock procedure for estimating functions at a point. It is well known that for global estimation, it is possible to achieves complete adaptation for free in terms of convergence rate across a range of function classes. 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. But for estimation at a point, one must pay a price for not knowing the smoothness of the underlying function.

Denote the minimax risk for estimating functions at a point t_0 over a function class $\mathcal{F}$ by

$$R(\mathcal{F}, n, t_0) = \inf_{\hat{f}_n} \sup_{\mathcal{F}} E(\hat{f}_n(t_0) - f(t_0))^2$$

Consider the Hölder class $\Lambda^{\alpha}(M)$. The optimal rate of convergence for estimating $f(t_0)$ with α known is $n^{-\rho}$ where $\rho = 2\alpha/(1 + 2\alpha)$. Brown and Low [3] and Lepski [17] showed that even when α is known to be one of two values, one has to pay a price for adaptation of at least a logarithmic factor, They showed that the best one can do is $(\log n/n)^{\rho}$ when the smoothness parameter α is unknown. We call $(\log n/n)^{\rho}$ the local adaptive minimax rate over the Hölder class $\Lambda^{\alpha}(M)$. The following theorem shows that NeighBlock achieves the local adaptive minimax rate over a wide range of Hölder classes.

Theorem 3 *Suppose the wavelets $\{\phi, \psi\}$ are r-regular with $r \geq \alpha$. Let $t_0 \in (0, 1)$ be fixed. Then the NeighBlock estimator $\hat{f}^*_n(t_0)$ of $f(t_0)$ satisfies*

$$\sup_{f \in \Lambda^{\alpha}(M)} E\{\hat{f}^*_n(t_0) - f(t_0)\}^2 \leq C(n^{-1} \log n)^{2\alpha/(1+2\alpha)}. \tag{9}$$

3.3 Denoising property

In addition to the global and local estimation properties, the NeighBlock estimator enjoys an interesting smoothness property which should offer high visual quality of the reconstruction. The estimator, with high probability, removes pure noise completely.

Theorem 4 *If the target function is the zero function $f \equiv 0$, then, with probability tending to 1 as $n \to \infty$, the NeighBlock estimator is also the zero function, i.e., there exist universal constants P_n such that*

$$P(\hat{f}^*_n \equiv 0) \geq P_n \to 1, \quad as \ \ n \to \infty \tag{10}$$

4 Properties of the NeighCoeff estimator

The NeighCoeff estimator is intuitively appealing and easy to implement. It can be shown that the estimator is locally adaptive and is within a logarithmic factor of being minimax over a wide range of Besov classes. This is the same asymptotic performance as VisuShrink. We summarize the results without proof in the following theorems. We shall see subsequently that, in most cases, the estimator enjoys superior numerical performance to many classical wavelet estimators, such as VisuShrink, SureShrink and TI-denoising estimators in most cases. It even outperforms NeighBlock, even though that estimator apparently has better asymptotic properties.

The first result shows that the global performance of the NeighCoeff estimator is simultaneously within a logarithmic factor of minimax over a wide range of Besov classes, and the second result shows that NeighCoeff has the same good pointwise behavior as NeighBlock. Both results remain true if a 'hard thresholding' version of the estimator is used, replacing the shrinkage factor β_i^j as defined in (2) by the factor $I[S_{ji}^2 > \lambda\sigma^2]$.

Theorem 5 *Assume that the wavelet ψ is r-regular. Then the NeighCoeff estimator satisfies*

$$\sup_{f \in B_{p,q}^{\alpha}(M)} E\|\hat{f}_n^* - f\|^2 \leq C(n^{-1}\log n)^{2\alpha/(1+2\alpha)} \tag{11}$$

for all $M \in (0,\infty), \alpha \in (0,r), q \in [1,\infty]$ and $p \in [1,\infty]$.

Theorem 6 *Assume that the wavelet ψ is r-regular with $r \geq \alpha$. Let $t_0 \in (0,1)$ be fixed. Then the NeighCoeff estimator satisfies*

$$\sup_{f \in \Lambda^{\alpha}(M)} E(\hat{f}_n^*(t_0) - f(t_0))^2 \leq C(n^{-1}\log n)^{2\alpha/(1+2\alpha)}. \tag{12}$$

5 Numerical comparison

A simulation study was conducted to compare the numerical performance of the NeighBlock and NeighCoeff estimators with Donoho and Johnstone's VisuShrink 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 unbiased estimate of risk. In the simulation, we use the hybrid method proposed in Donoho and Johnstone [13]. The TI-denoising method was introduced by Coifman and Donoho [8], 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.

We implement the NeighBlock and NeighCoeff estimators in the software package S+Wavelets. The programs are available from the web site [6]. We compare the numerical performance of the methods using eight test functions representing different level of spatial variability. The test functions are plotted in Figure 1. 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 2. Different combinations of wavelets and signal-to-noise ratios yield basically the same results; for details see the web site [6].

The NeighBlock and NeighCoeff methods both uniformly outperform VisuShrink in all examples. 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, the only conceivable competitor among the standard methods is SureShrink. 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.

Though it would be interesting to include comparisons with the block thresholding estimator of Hall et al. [15], we do not include such comparisons for two reasons. Their estimator is not easy to implement, and furthermore simulation results by Hall et al. [16] 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 favorably over the estimator of Hall et al. in terms of mean squared error, at least in the case of high signal-to-noise-ratio.

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.

Figure 3 shows a typical segment of the result of the four methods applied to the inductance plethysmography data analyzed, for example, by Abramovich, Sapatinas and Silverman [1]. It can be seen that VisuShrink smooths out the broad features of the curve, while the SureShrink estimator allows through high frequency effects that are almost certainly spurious.

6 Proofs

6.1 Sequence space approximation

We shall prove Theorem 1 and 2 by using the sequence space approach introduced by Donoho and Johnstone in [11]. A key step is to use the asymptotic equivalence results presented by Brown and Low [2] and to approximate the problem of estimating f from the noisy observations in (1) by the problem of estimating the wavelet coefficient sequence of f contaminated with i.i.d. Gaussian noise.

Donoho and Johnstone [11] show a strong equivalence result on the nonparametric regression and the white noise model over the Besov classes $B^{\alpha}_{p,q}(M)$. When the wavelet ψ 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 [11] and [13] and Brown and Low [2]. For approximation results, see also Chambolle et al. [7].

Under the correspondence between the estimation problem in function space and the estimation problem in sequence space, it suffices to consider the following sequence estimation problem.

6.2 Estimation in sequence space by NeighBlock

Suppose we observe sequence data

$$y_{jk} = \theta_{jk} + n^{-1/2}\sigma z_{jk}, \quad j \geq 0,\ k = 1, 2, \cdots, 2^j \tag{13}$$

where z_{jk} are i.i.d. $N(0,1)$. The mean array θ is the object that we wish to estimate. We assume that θ is in some Besov Body $\Theta^s_{p,q}(M) = \{\theta : \|\theta\|_{b^s_{p,q}} \leq M\}$, where the norm is as defined in (5) above. Make the usual calibration $s = \alpha + 1/2 - 1/p$. Donoho and Johnstone [11] show that the minimax rate of convergence for estimating θ over the Besov body $\Theta^s_{p,q}(M)$ is $n^{-2\alpha/(1+2\alpha)}$ as $n \to \infty$. 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 now approach this sequence estimation problem using a procedure corresponding to NeighBlock. Let $J = [\log_2 n]$. Divide each resolution level $j_0 \leq j < J$ into nonoverlapping blocks of length $L_0 = [(\log n)/2]$. Again denote by b^j_i the i-th block at level j and similarly define B^j_i to be the larger block obtained by extending b^j_i by $[(\log n)/4]$ elements in each direction. Define S^2_{ji} to be the sum of the y^2_{jk} over B^j_i, by analogy with (3). Now estimate θ by $\hat{\theta}^*$ with

$$\hat{\theta}^*_{jk} = \begin{cases} y_{jk} & \text{for } j \leq j_0 \\ (1 - n^{-1}\lambda_* L\sigma^2/S^2_{ji})_+\, y_{jk} & \text{for } (j,k) \in b^j_i,\ \ j_0 \leq j < J \\ 0 & \text{for } j \geq J \end{cases} \tag{14}$$

For this estimator, we have the following minimax results, demonstrating that the estimator attains the exact minimax rate over all the Besov Bodies $\Theta^s_{p,q}(M)$ with $p \geq 2$, and the exact minimax rate up to a logarithmic term for $p < 2$.

Theorem 7 *Define $\hat{\theta}^*$ as in (14). Then, as $n \to \infty$,*

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

The results of Theorem 1 and 2 follow from this theorem and the equivalence and the approximation arguments discussed in Section 6.1.

6.3 Proof of the main results

We will prove Theorem 7. The proof of Theorem 4 is straightforward and the proof of Theorem 3 is similar to a corresponding result in Cai [5]. A key result used in the proof of Theorem 7 is the following oracle inequality.

Lemma 1 *Assume that $y_{j,k}$ and $\hat{\theta}^*_{j,k}$ are given as in (13) and (14) respectively. Then, defining λ_* by $\lambda_* - \log \lambda_* = 3$, for each j and i*

$$\sum_{(j,k)\in b_i^j} E(\hat{\theta}^*_{j,k} - \theta_{j,k})^2 \leq \lambda_*(\sigma^2 n^{-1} \log n \wedge \sum_{(j,k)\in B_i^j} \theta^2_{j,k}) + 2n^{-2}\sigma^2. \tag{15}$$

The proof of this lemma is an extension of the proof of Theorem 1 of Cai [5]. For j, k in B_i^j define

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

Since $\hat{\theta}^\dagger_{j,k} = \hat{\theta}^*_{j,k}$ for (j,k) in b_i^j, extending the sum from b_i^j to B_i^j, and replacing $\hat{\theta}^*$ by $\hat{\theta}^\dagger$, can only increase the left hand side of (15). The argument of Theorem 9 and Lemma 2 of Cai [5] shows that the inequality holds with these changes, completing the proof. ∎

We also recall two elementary inequalities between two different ℓ_p norms, and a bound for a certain sum.

Lemma 2 *Let $x \in \mathbb{R}^m$, and $0 < p_1 \leq p_2 \leq \infty$. Then the following inequalities hold:*

$$\|x\|_{p_2} \leq \|x\|_{p_1} \leq m^{\frac{1}{p_1}-\frac{1}{p_2}}\|x\|_{p_2} \tag{16}$$

Lemma 3 *Let $0 < a < 1$ and $S = \{x \in \mathbb{R}^k : \sum_{i=1}^k x_i^a \leq B,\ x_i \geq 0,\ i = 1, \cdots, k\}$. 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 return to the proof of Theorem 7. Let y and $\hat{\theta}^*$ be given as in (13) and (14) respectively. Then,

$$E\|\hat{\theta}^* - \theta\|_2^2 = \sum_{j<j_0}\sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 + \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 + \sum_{j=J}^{\infty}\sum_k \theta_{jk}^2 \equiv S_1 + S_2 + S_3, \quad (17)$$

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

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

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 (15). Since $L \approx \log n$ and the number of blocks is definitely less than n, this yields

$$S_2 = \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 \le C\sum_{j=j_0}^{J-1}\sum_i (A_i^j \wedge \sigma^2 n^{-1} L) + 2n^{-1}\sigma^2. \quad (18)$$

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

The case $p \ge 2$ For $\theta \in \Theta^s_{p,q}(M)$, Lemma 2 implies that

$$\|\theta_j\|_2^2 \le (2^j)^{2(\frac{1}{2}-\frac{1}{p})}\|\theta_j\|_p^2 \le M^2 2^{2j(\frac{1}{2}-\frac{1}{p}-s)} = M^2 2^{-2\alpha j}. \quad (19)$$

It follows that

$$S_1 + S_3 \le 2^{j_0} n^{-1}\sigma^2 + \sum_{j=J}^{\infty} M^2 2^{-2\alpha j} = o(n^{-2\alpha/(1+2\alpha)}), \quad (20)$$

so that $S_1 + S_3$ can be neglected.

We divide the sum in (18) into two parts. Choose J_1 such that $2^{J_1} \approx n^{1/(1+2\alpha)}$. Then,

$$\sum_{j=j_0}^{J_1-1}\sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \le \sum_{j=j_0}^{J_1-1}\sum_i \sigma^2 n^{-1} L \le C 2^{J_1} n^{-1} \le C n^{-2\alpha/(1+2\alpha)}, \quad (21)$$

and, making use of the bound (19),

$$\sum_{j=J_1}^{J-1}\sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \le \sum_{j=J_1}^{J-1}\sum_i A_i^j \le 2\sum_{j=J_1}^{J-1}\|\theta_j\|_2^2 \le C n^{-2\alpha/(1+2\alpha)}. \quad (22)$$

Combining (21) and (22) demonstrates that $S_2 \le C n^{-2\alpha/(1+2\alpha)}$, completing the proof for this case.

The case $p < 2$ with $\alpha p \geq 1$: For $\theta \in \Theta^s_{p,q}(M)$, Lemma 2 now yields $\|\theta_j\|_2^2 \leq \|\theta_j\|_p^2 \leq M^2 2^{-2js}$. The assumption $\alpha p \geq 1$ implies that $s \geq \frac{1}{2}$, so that

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

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 (21),

$$\sum_{j=j_0}^{J_2-1} \sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \leq \sum_{j=j_0}^{J_2-1} \sum_i \sigma^2 n^{-1} L \leq Cn^{-2\alpha/(1+2\alpha)} (\log n)^{-(2-p)/p(1+2\alpha)}. \tag{23}$$

Turning to the other part of S_2, it follows by convexity that, for each j,

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

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

$$\sum_{j=J_2}^{J-1} \sum_i (A_i^j \wedge \sigma^2 n^{-1} L) \leq Cn^{-2\alpha/(1+2\alpha)} (\log n)^{-(2-p)/p(1+2\alpha)}. \tag{24}$$

We complete the proof by combining the bounds (23) and (24), as in the case $p \geq 2$. ∎

Acknowledgment

This work was carried out while Bernard Silverman was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford. He is grateful for financial support provided by National Science Foundation Grant number SBR-9601236.

References

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

[2] Brown, L.D. & Low, M.G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384-2398.

[3] Brown, L.D. & Low, M.G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24**, 2524-2335.

[4] Cai, T. (1996). Minimax wavelet estimation via block thresholding. Technical Report #96-41, Department of Statistics, Purdue University. (Revised 1998.)

[5] Cai, T. (1998). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. Technical Report #98-07, Department of Statistics, Purdue University.

[6] Cai, T. & Silverman, B. W. (1998). Incorporating information on neighboring coefficients into wavelet estimation. Web page available at
`www.stat.purdue.edu/~tcai/neighblock.html`

[7] Chambolle, A., DeVore, R., Lee, N. & Lucier, B. (1996). Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing,* to appear.

[8] 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.

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

[10] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.,* **305**, 397-414.

[11] Donoho, D.L. & Johnstone, I.M. (1997). Minimax estimation via wavelet shrinkage. *Ann. Statist.,* **25**, (In press).

[12] Donoho, D.L. & Johnstone, I.M. (1995). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425-455.

[13] Donoho, D.L. & Johnstone, I.M. (1994). Adapting to unknown smoothness via wavelet shrinkage. Technical Report, Stanford University.

[14] Downie, T.R. & Silverman, B.W. (1998). The discrete multiple wavelet transform and thresholding methods. *IEEE Transactions in Signal Processing*, to appear.

[15] Hall, P., Kerkyacharian, G. & Picard, D. (1995). On the minimax optimality of block thresholded wavelet estimators. Manuscript.

[16] Hall, P., Penev, S., Kerkyacharian, G. & Picard, D. (1996). Numerical performance of block thresholded wavelet estimators. Manuscript.

[17] Lepski, O.V. (1990). On a problem of adaptive estimation on white gaussian noise. *Theory of Probability and Appl.* **35**, 454-466.

[18] Marron, J.S., Adak, S., Johnstone, I.M., Neumann, M.H. and Patil, P. (1995). Exact risk analysis of wavelet regression. Technical Report, Stanford University.

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

[20] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135-1151.

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

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

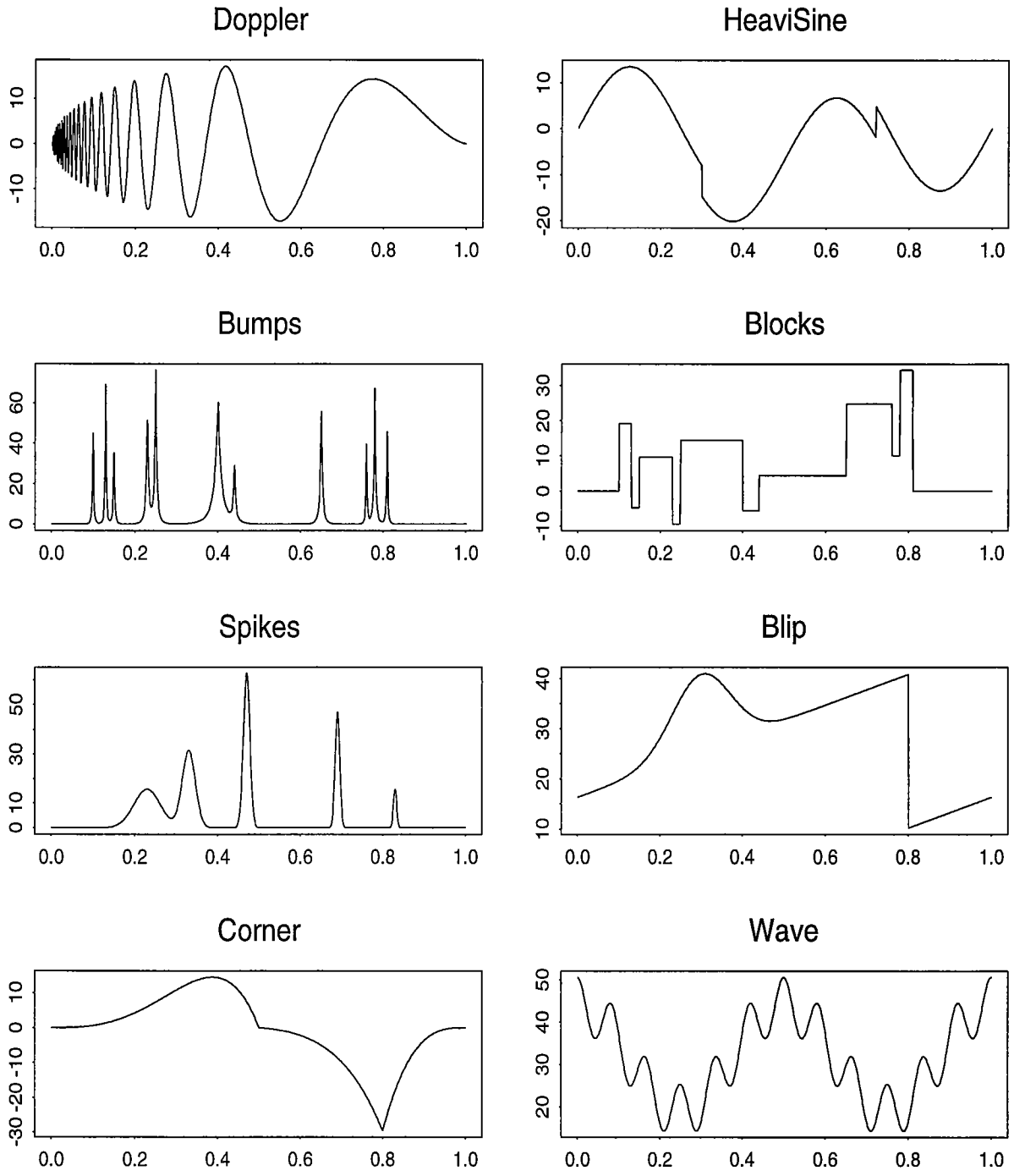

Figure 1: Test functions. Doppler, HeaviSine, Bumps and Blocks are from Donoho and Johnstone [12]. Blip and Wave are from Marron et al. [18]. The test functions are normalized so that every function has standard deviation 10. Formulae for all the functions are given in Cai [5]

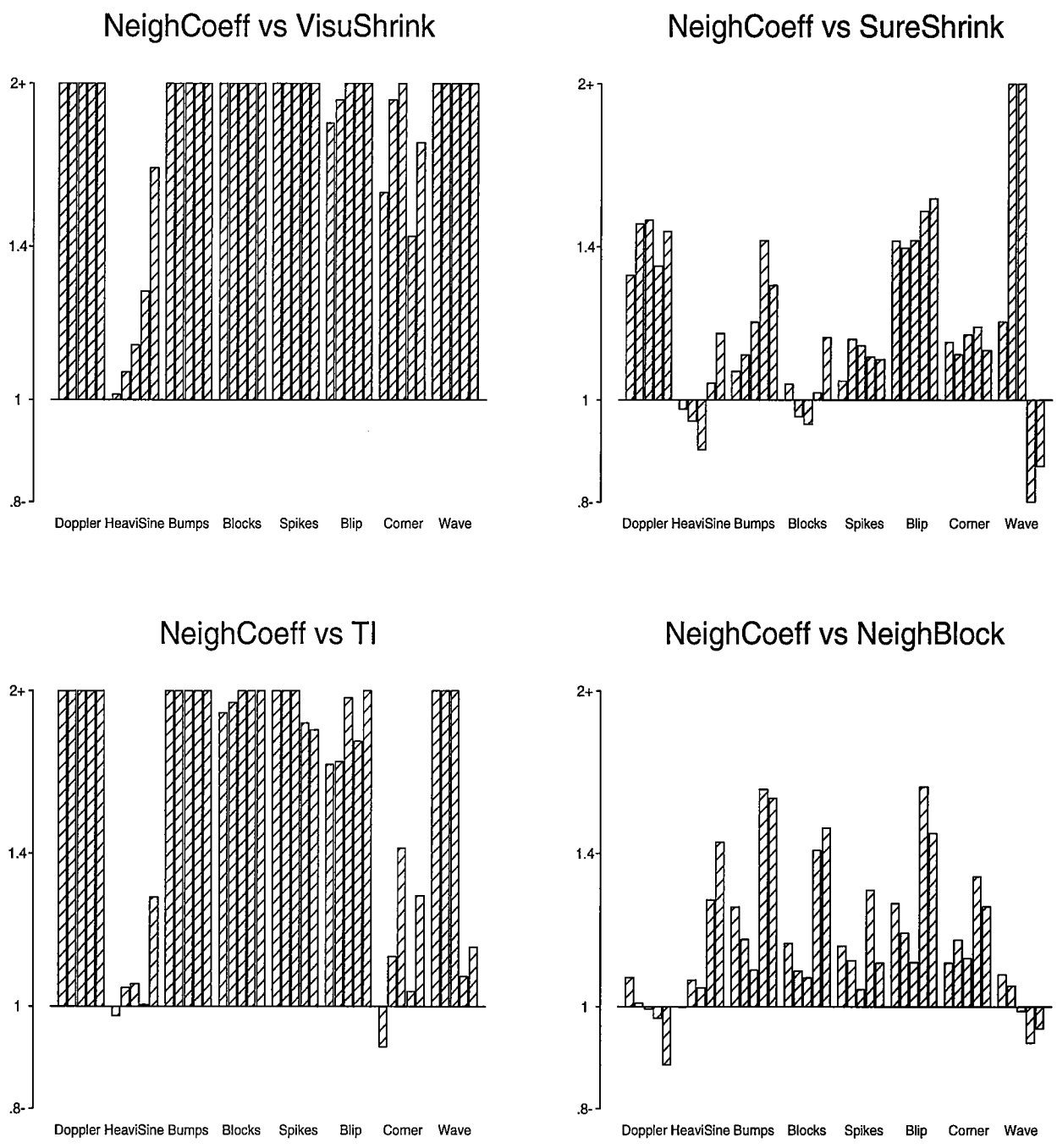

Figure 2: RSNR=3. The vertical bars represent the ratios of the MSEs of various estimators to the corresponding MSE of the NeighCoeff estimator. The higher the bar the better the relative performance of the NeighCoeff estimator. The bars are plotted on a log scale and are truncated at the value 2 of the original ratio. For each signal the bars are ordered from left to right by the sample sizes (n=512 to 8192).

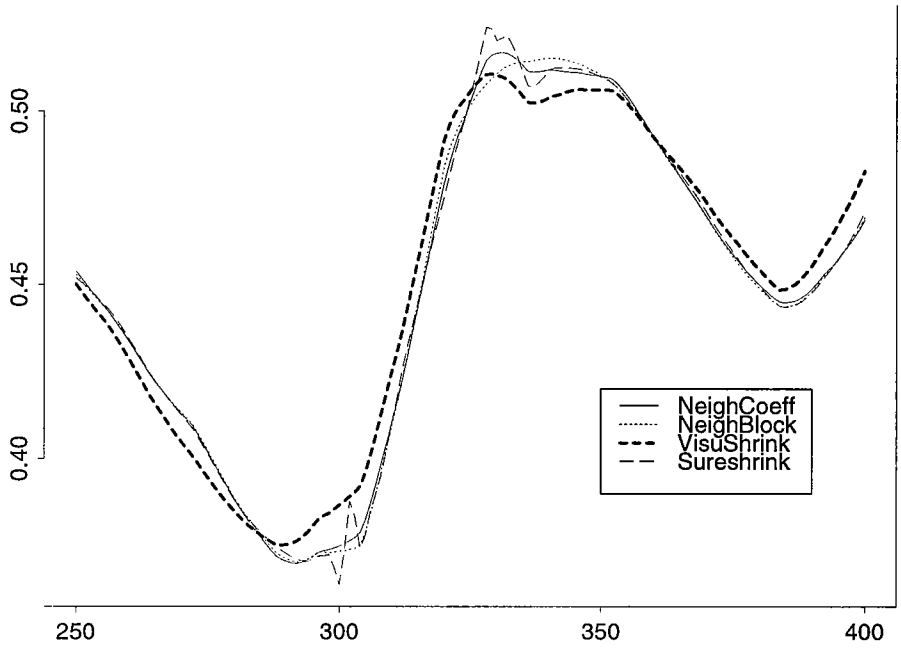

Figure 3: Curve estimates for a segment of the inductance plethysmography data. —— — NeighCoeff; ····· NeighBlock; − − − VisuShrink; — — SureShrink. The VisuShrink estimate smooths out the broad features, while the SureShrink estimate contains high frequency effects near times 300 and 335, both of which are presumably spurious.