ADAPTIVE WAVELET ESTIMATION: A BLOCK THRESHOLDING AND ORACLE INEQUALITY APPROACH

by

T. Tony Cai
Purdue University

Technical Report #98-07

Department of Statistics
Purdue University

March 1998

△

Adaptive Wavelet Estimation: A Block Thresholding And Oracle Inequality Approach

T. Tony Cai
Department of Statistics
Purdue University

Abstract

We study wavelet function estimation via the approach of block thresholding and ideal adaptation with oracle. Oracle inequalities are derived to offer insights into the balance and tradeoff between block size and threshold level. Based on the oracle inequalities, an adaptive wavelet method for nonparametric regression is proposed and the optimality of the procedure is discussed.

We show that the estimator achieves simultaneously three objectives: adaptivity, spatial adaptivity and high visual quality. Specifically, we show that the estimator attains the exact optimal rates of convergence over a range of Besov classes and perturbed Besov classes and the estimator achieves adaptive local minimax rate for estimating functions at a point. Simulation results and generalizations of the method are also discussed.

Keywords: James-Stein Estimator; Adaptivity; Wavelet; Block Thresholding; Nonparametric function Estimation; Besov Space.

AMS 1991 Subject Classification: Primary 62G07, Secondary 62G20.

1 Introduction

Wavelet methods have demonstrated successes in nonparametric function estimation in terms of spatial adaptivity, computational efficiency and asymptotic optimality. In contrast to the traditional linear procedures, wavelet methods achieve (near) optimal convergence rates over large function classes and enjoy excellent mean squared error properties when used to estimate functions that are spatially inhomogeneous.

Standard wavelet methods achieve adaptivity through term-by-term thresholding of the empirical wavelet coefficients. There, each individual empirical wavelet coefficient is compared with a predetermined threshold. A wavelet coefficient is retained if its magnitude is above the threshold level and is discarded otherwise. A well-known example of the term-by-term thresholding procedures is the Donoho and Johnstone's VisuShrink (1994).

VisuShrink is spatially adaptive and easy to implement. The estimator is within a logarithmic factor of the optimal convergence rate over a wide range of Besov classes. VisuShrink achieves a degree of tradeoff between variance and bias contributions to the mean squared error. However, the tradeoff is not optimal. The VisuShrink estimator is often over-smoothed.

Hall, Kerkyacharian and Picard (1995) studied local block thresholding rules for wavelet function estimation which threshold the empirical wavelet coefficients in groups rather than individually. They showed that the estimator attain the minimax convergence rates for global estimation over a range of perturbed Besov classes. However, the estimator does not achieve optimal local adaptivity which the VisuShrink estimator enjoys. Also the estimator is difficult to be implemented. It uses a "near" unbiased estimator of the sum of the squares of the true coefficients in a block. It requires the evaluation of the wavelet functions at all the sample points and involves matrix multiplications which are computationally costly. Furthermore, simulation study showed that the estimator has little advantage over the VisuShrink estimators when the signal-to-noise ratio is high (see Hall, Penev, Kerkyacharian and Picard (1996)).

In the present paper, we study block thresholding via the approach of ideal adaptation with oracle. We study the performance of estimators by comparing the risk of the estimators with the risk of an "ideal estimator" in which case an oracle is available. The goal is to obtain estimators which can essentially mimic the performance of an oracle in the case of estimating a multivariate normal mean. We derive a risk inequality for block projection oracle. The block projection oracle inequality offers insights into the balance and tradeoff between threshold level and block length. We suggest that the oracle inequality serves as a guide for optimal selection of threshold level for a given block length including the special case of block length one which is the standard term-by-term thresholding. Similar to Donoho and Johnstone's diagonal projection oracle inequality, the block projection oracle inequality has important implications in wavelet function estimation.

Based on the Oracle Inequality, we propose a block thresholding estimator for nonparametric function estimation by using the classical James-Stein estimators and study the asymptotic and empirical performance of the estimator. We show in Section 5 that the estimator enjoys a high degree of adaptivity and spatial adaptivity. Specifically, we prove that the *BlockJS* estimator simultaneously attains the exact optimal rate of convergence over a wide interval of the Besov classes with $p \geq 2$ and over a range of perturbed Besov classes without prior knowledge of the smoothness of the underlying functions. Over the Besov classes with $p < 2$, the *BlockJS* 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 *BlockJS* estimator is not only quantitatively appealing but visually appealing as well. The reconstruction jumps where the target function jumps; the reconstruction is smooth where the target function is smooth. They do not contain spurious fine-scale structure that are contained in some wavelet estimators. The *BlockJS* adapts to the subtle changes of the underlying functions. We also show in Section 4 that the *BlockJS* has a similar smoothness property as the VisuShrink: if the underlying function is zero function, then, with high probability, the *BlockJS* is also zero function. In other words, the *BlockJS* removes pure noise completely.

In the present paper we also suggest that, through the example of the (positive-part) James-Stein estimator, block thresholding serves as a "bridge" between the traditional shrinkage estimators in multivariate normal decision theory and the more recent wavelet function estimation. This connection allows us to develop new classes of (near) optimal wavelet function estimators, and all of which may be useful in different estimation situations.

The paper is organized as follows. Section 2 describes the basics of wavelets. Section 3 derives an oracle inequality for block projection estimators. With preparation and motivations given in Sections 2 and 3, we introduce the ingredients of *BlockJS* procedure and show the "noise-free" feature of the estimator in Section 4. Section 5 presents the optimality results of the procedure. We consider the convergence rates of estimator uniformly over a wide scale of the Besov classes. Local adaptivity of the estimator is also presented. The choices of block size and threshold level are discussed in Section 6. A summary of the simulation results and discussions is presented in Section 7. The proofs are postponed to Section 9.

2 Wavelets

An orthonormal wavelet basis is generated from dilation and translation of two basic functions, a "father" wavelet ϕ and a "mother" wavelet ψ. In the present paper, the functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$. We call a wavelet ψ *r-regular* if ψ has r vanishing moments and r continuous derivatives.

Let

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

And denote the periodized wavelets

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

For simplicity in exposition, we use the periodized wavelet bases on [0, 1] in the present paper. The collection $\{\phi^p_{j_0k},\ k = 1,...,2^{j_0};\ \psi^p_{jk},\ j \geq j_0 \geq 0, k = 1,...,2^j\}$ constitutes such an orthonormal basis of $L_2[0,1]$. Note that the basis functions are periodized at the boundary. The superscript "p" will be suppressed from the notations for convenience.

An orthonormal wavelet basis has an associated exact orthogonal Discrete Wavelet Transform (DWT) that transforms sampled data into wavelet coefficient domain. A crucial point is that the transform is not implemented by matrix multiplication, but by a sequence of finite-length filtering which produce an order $O(n)$ transform. See Daubechies (1992) and Strang (1992) for further details about the wavelets and the discrete wavelet transform.

For a given square-integrable function f on [0, 1], denote $\xi_{jk} = \langle f, \phi_{jk}\rangle$, $\theta_{jk} = \langle f, \psi_{jk}\rangle$. So the function f can be expanded into a wavelet series:

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j_0k}\phi_{j_0k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{jk}\psi_{jk}(x) \tag{1}$$

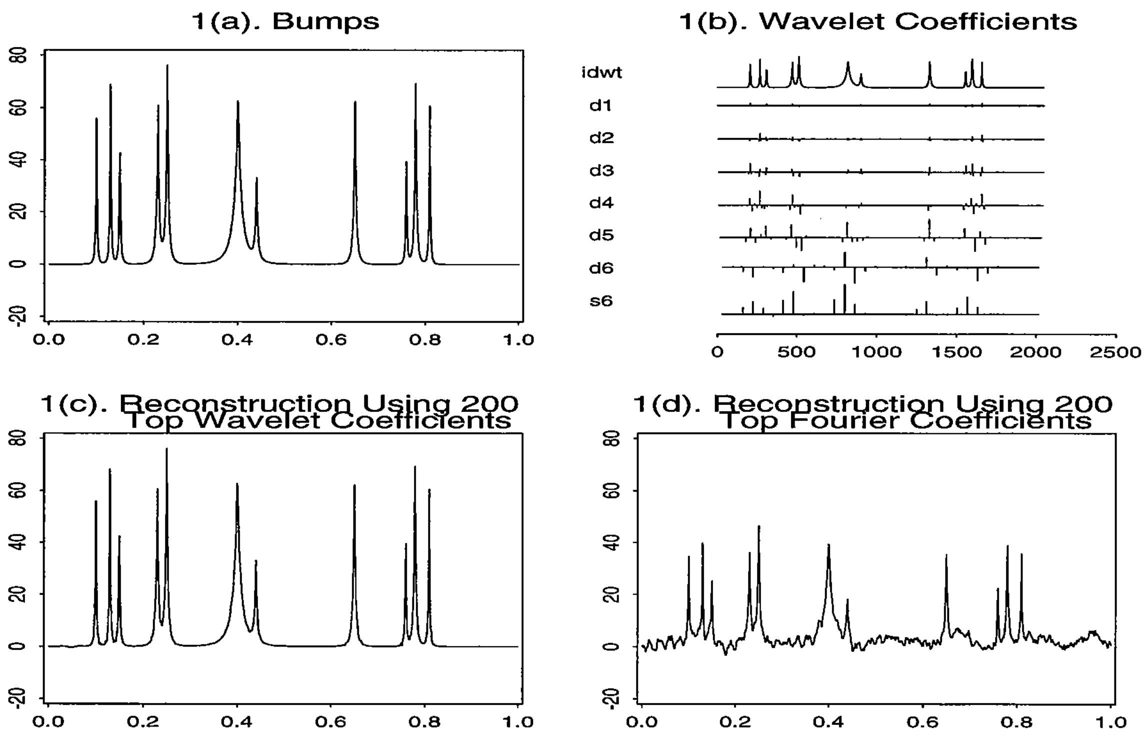
Wavelet transform decomposes a function into different resolution components. In (1), ξ_{j_0k} are the coefficients at the coarsest level. They represent the gross structure of the function f. And θ_{jk} are the wavelet coefficients. They represent finer and finer structures of the function f as the resolution level j increases.

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. In contrast, Fourier series of normal functions are full, whereas lacunary Fourier series represent pathological functions. (see Meyer (1992) pp. 113). Wavelet transform can compact the energy of a normal function into very few number of large wavelet coefficients.

Wavelet bases are well localized, i.e., local regularity properties of a function are determined by its local wavelet coefficients. In particular, a function is smooth at a point if and only if its local wavelet coefficients decay fast enough. The large wavelet coefficients of a function cluster around the discontinuities and other irregularities of the function. The wavelet coefficients at high resolution levels are small where the function is smooth.

The data compression and the localization properties of wavelets can be illustrated by the following example. Consider Donoho and Johnstone's Bumps function which is of significant spatial variability. Figure 1(a) plots the sampled Bumps function of length 2048. The data is then transformed into wavelet domain via DWT. We use Daubechies' *Symmelet 8* wavelet in the example. The wavelet coefficients plot (figure 1(b)) shows that among the 2048 wavelet coefficients there are only a small portion of coefficients that are large. And the large coefficients at high resolution level occur only around the spikes.

The ordered absolute values of the wavelet coefficients decay very fast. The information about the function is concentrated in a very small number of wavelet coefficients. One gets very good reconstruction with only 200 largest coefficients (figure 1(c)). That is fewer than 10% of the total number of coefficients. The relative error is less than 0.08%. As a comparison, we do the same with the traditional Fourier transform. Top 200 Fourier coefficients do not offer a reasonable reconstruction (figure 1(d)). It turns out that it takes more than 2000 largest Fourier coefficients to achieve the same performance as the 200 largest wavelet coefficients. The reason is that the Fourier basis is not localized in space. The example shows some heuristics, see DeVore, Jawerth and Popov (1992) and Meyer (1992) for theoretical treatment on the data compression property of the wavelets.



Based on these data compression and localization heuristics, one can intuitively envision that all but only a small number of wavelet coefficients of a "normal" function are negligible and large coefficients at high resolution levels cluster around irregularities of the function.

3 Oracle Inequality

In this section, we consider the problem of estimating a multivariate normal mean. Suppose we observe

$$y_i = \theta_i + \sigma z_i \qquad (2)$$

$i = 1, \cdots, n, \quad z_i \overset{iid}{\sim} N(0,1)$ with σ known. We wish to estimate $\theta = (\theta_1, \cdots, \theta_n)$ based on the observations $y = (y_1, \cdots, y_n)$ under the mean squared error:

$$R(\hat{\theta}, \theta) = \frac{1}{n} \sum E(\hat{\theta}_i - \theta_i)^2 \tag{3}$$

When $n = 1$, or 2, decision theory shows that the maximum likelihood estimator y is an admissible estimator of θ. When $n \geq 3$, it was shown by Stein(1955) that y is no longer a "good" estimator of θ in the sense that y is uniformly dominated by some other estimators. Multivariate normal decision theory shows that in order to do well according to the risk measure (3), some form of shrinkage is necessary (see, e.g. Lehmann (1983)).

3.1 Diagonal Projection Oracle

Donoho and Johnstone (1994) consider a special class of shrinkage estimators, diagonal projection estimators, in the context of wavelet function estimation. Denote by $\mathcal{H}$ a given subset of indices and consider

$$\hat{\theta}_i(\mathcal{H}) = \begin{cases} y_i & \text{if } j \in \mathcal{H} \\ 0 & \text{if } j \notin \mathcal{H} \end{cases}$$

Such diagonal projection estimator either keeps or omits a coordinate. For each individual coordinate, the expected loss is

$$E(\hat{\theta}_i(\mathcal{H}) - \theta_i)^2 = \sigma^2 I\{i \in \mathcal{H}\} + \theta_i^2 I\{i \notin \mathcal{H}\}$$

Ideally, one would estimate θ_i by y_i when $\theta_i^2 > \sigma^2$ and by 0 otherwise. An Diagonal Projection (DP) oracle would supply the extra side information $\mathcal{H}_{oracle}(\theta) = \{i : \theta_i^2 > \sigma^2\}$. The ideal diagonal projection consists in estimating only those θ_i larger than the noise level. Supplied with such an oracle, one would attain the ideal risk

$$R_{oracle}(\theta, \sigma, 1) = \frac{1}{n} \sum_{i=1}^{n} \min(\theta_i^2, \sigma^2).$$

The soft threshold estimator

$$\hat{\theta}_i^* = sgn(y_i)(|y_i| - \sigma\sqrt{2 \log n})_+, \tag{4}$$

introduced by Donoho and Johnstone (1994), can essentially "mimic" the performance of the DP oracle. The estimator comes (essentially) within a logarithmic factor of the ideal risk for all $\theta \in I\!R^n$. Specifically, they show the following DP oracle inequality

$$R(\hat{\theta}^*, \theta) \leq (2 \log n + 1)[R_{oracle}(\theta, \sigma, 1) + \sigma^2/n], \quad \text{for all } \theta \in I\!R^n. \tag{5}$$

Donoho and Johnstone develop the DP oracle inequality primarily for wavelet function estimation. They and co-authors show that the wavelet estimator, VisuShrink, achieves unusual adaptivity. The estimator attains within a logarithmic factor of the minimax rates over a wide range of Besov classes (Donoho, Johnstone, Kerkyacharian & Picard, 1995).

$$\sup_{f \in B^{\alpha}_{p,q}(M)} R(\hat{f}_{visu}, f) \le C(\log n/n)^{2\alpha/(1+2\alpha)}(1+o(1))$$

The DP oracle inequality is a key in showing the adaptivity of the VisuShrink estimator.

3.2 Block Projection Oracle

A DP estimator keeps or kills each coordinate individually without using information about other coordinates. On the contrary, a block projection (BP) estimator thresholds coordinates in groups, it uses information about neighboring coordinates. Simultaneous decisions are made to retain or to discard all the coordinates within the same group.

Suppose y_i are given as in (2). Let $B_1, B_2, \cdots, B_N$ be a partition of the index set $\{1, \cdots, n\}$ with each B_i of size L (For convenience, we assume that the sample size n is divisible by the block size L). Let $\mathcal{H}$ be a subset of the block indices $\{1, \cdots, N\}$. A Block Projection (BP) estimator associated with $\mathcal{H}$ is defined as

$$\hat{\theta}_{B_j}(\mathcal{H}) = \begin{cases} y_{B_j} & \text{if } j \in \mathcal{H} \\ 0 & \text{if } j \notin \mathcal{H} \end{cases}$$

where y_{B_j} denote the vector $(y_i)_{i \in B_j}$. For each given block the expected loss is

$$E\|\hat{\theta}_{B_j}(\mathcal{H}) - \theta_{B_j}\|^2 = L\sigma^2 I\{j \in \mathcal{H}\} + \|\theta_{B_j}\|_2^2 I\{j \notin \mathcal{H}\} \tag{6}$$

To minimize the risk (6), we would ideally like to choose $\mathcal{H}$ to consist of blocks with signal $>$ noise, i.e. $\|\theta_{B_j}\|_2^2 > L\sigma^2$. A BP oracle would supply exactly the side information

$$\mathcal{H}_* = \mathcal{H}_*(\theta) = \{j : \|\theta_{B_j}\|_2^2 > L\sigma^2\}$$

With the aid of the BP oracle, one has the ideal block projection "estimator"

$$\hat{\theta}_{B_j}(\mathcal{H}_*) = \begin{cases} y_{B_j} & \text{if } j \in \mathcal{H}_* \\ 0 & \text{if } j \notin \mathcal{H}_* \end{cases} \tag{7}$$

with the ideal risk

$$R_{oracle}(\theta, \sigma, L) = \inf_{\mathcal{H}} \frac{1}{n} E\|\hat{\theta}(\mathcal{H}) - \theta\|^2 = \frac{1}{n}\sum_{j=1}^{N}(\|\theta_{B_j}\|_2^2 \wedge L\sigma^2)$$

where $a \wedge b = \min(a,\ b)$. It is clear that the ideal risk is unattainable in general because it requires the knowledge of an oracle which is unavailable in most realistic situations. The ideal "estimator" (7) is not a true estimator in a statistical sense. Our first goal is to derive a practical estimator which can mimic the performance of the BP oracle. That is, we wish to construct an estimator whose risk is close to the risk of the ideal "estimator". Among the many traditional shrinkage estimators developed in the normal decision theory, the James-Stein estimator is perhaps the best-known and will be our primary focus in the present paper. Generalizations of the method is discussed in Section 8.

Suppose y_i are observed as in (2). James and Stein (1961) propose a particularly simple shrinkage estimator

$$\hat{\theta}^1 = (1 - (n-2)\sigma^2/S^2)\, y \tag{8}$$

where $S^2 = \sum y_i^2$ is the L_2 energy. James and Stein (1961) show that the estimate dominates the maximum likelihood estimator y when $n \geq 3$. It is easy to see that the estimator (8) is further dominated by

$$\hat{\theta}^2 = (1 - (n-2)\sigma^2/S^2)_+\, y \tag{9}$$

Efron and Morris (1973) showed that the (positive part) James-Stein estimator (9) does more than just demonstrate the inadequacy of the maximum likelihood estimator y . It is a member of a class of good shrinkage rules, all of which may be useful in different estimation problems. The class of estimators

$$\hat{\theta}^2 = (1 - c\sigma^2/S^2)_+\, y$$

can be regarded as truncated empirical Bayes rules. See Efron and Morris (1973).

Under the context of BP estimators, we consider a class of blockwise James-Stein estimators. Within each block B_j a James-Stein shrinkage rule is applied:

$$\hat{\theta}_{B_j}(L, \lambda) = (1 - \frac{\lambda L \sigma^2}{S_j^2})_+ y_{B_j}, \tag{10}$$

where $S_j^2 = \|y_{B_j}\|_2^2$. When the block size L and the threshold λ are properly chosen, the blockwise James-Stein rule can mimic the performance of a BP oracle. We first consider a special choice of block length and threshold level.

Theorem 1 (BP Oracle Inequality) *Assume that y_i and $\hat{\theta}_{B_j}(L, \lambda)$ are given as in (2) and (10) respectively. Suppose that the block size $L = \log n$ and the threshold $\lambda = 4.50524$. Then*

$$R(\hat{\theta}(\lambda, L), \theta) \leq \frac{1}{n}(\sum_{j=1}^{N}(\|\theta_{B_j}\|^2 \wedge \lambda L \sigma^2) + 2\sigma^2). \tag{11}$$

Written in "Oracular" form, we have

$$R(\hat{\theta}(\lambda, L), \theta) \leq 4.50524 R_{oracle}(\theta, \sigma, L) + \frac{2\sigma^2}{n}. \tag{12}$$

Therefore, with the given block size and threshold level, the estimator comes essentially within a constant factor of 4.50524 of the ideal risk.

Remark: The threshold $\lambda = 4.50524$ is the solution of the equation $\lambda - \log \lambda - 3 = 0$. This particular threshold is chosen so that the corresponding wavelet estimator is (near) optimal in function estimation problems. We focus on this estimator for the moment. Details on the choice of block size and threshold level are contained in Section 6.

4 Wavelet Shrinkage via the BP Oracle Inequality

Now let us turn attention to the function estimation problem and imagine that the mean vector μ is the wavelet coefficients of some function. According to the data compression and localization properties of wavelets, it is reasonable to think that most of the wavelet coefficients are small and negligible. On the other hand, it is also reasonable to believe that not all the coefficients are small. Large coefficients cluster around irregularities of the function. That is, we are aiming to estimate a high dimensional sparse normal mean vector. The objective is to estimate large coordinates accurately and to estimate the small coordinates by zero.

A natural approach is to estimate each coordinate individually by thresholding or shrinkage methods. VisuShrink is a good example of the term-by-term shrinkage methods. Our approach is to estimate the means in groups by putting the empirical wavelet coefficients into blocks and make simultaneous shrinkage decisions about all coefficients within a block. We estimate the true wavelet coefficients at each resolution level via a blockwise James-Stein rule. The optimality results in Section 5 show that it is more efficient than estimating coefficients individually. With these motivations in mind, we are now ready to formally describe the *BlockJS* procedure.

Suppose we observe a noisy sampled function f :

$$y_i = f(t_i) + \epsilon z_i, \quad i = 1, 2, \cdots, n \tag{13}$$

with $t_i = i/n, \quad n = 2^J$ and z_i i.i.d. $N(0,1)$. The noise level ϵ is assumed to be known. We are interested in recovering the unknown function f. The precision of an estimator is measured both globally and locally. The global quality of recovery is given by the expected integrated squared error:

$$R(\hat{f}, f) = E\|\hat{f} - f\|_2^2. \tag{14}$$

And local quality of recovery is measured by the expected loss at the point:

$$R(\hat{f}(t_0), f(t_0)) = E(\hat{f}(t_0) - f(t_0))^2. \tag{15}$$

Suppose we observe the data $Y = \{y_i\}$ as in (13). Let $\tilde{\Theta} = W \cdot n^{-1/2} Y$ be the discrete wavelet transform of $n^{-1/2}Y$. Write

$$\tilde{\Theta} = (\tilde{\xi}_{j_0 1}, \cdots, \tilde{\xi}_{j_0 2^{j_0}}, \tilde{\theta}_{j_0 1}, \cdots, \tilde{\theta}_{j_0 2^{j_0}}, \cdots, \tilde{\theta}_{J-1,1}, \cdots, \tilde{\theta}_{J-1,2^{J-1}})^T$$

Here $\tilde{\xi}_{j_0 k}$ are the gross structure terms at the lowest resolution level, and the coefficients $\tilde{\theta}_{jk}$ $(j = 1, \cdots, J-1, k = 1, \cdots, 2^j)$ are fine structure wavelet terms. One may write

$$\tilde{\theta}_{jk} = \theta'_{jk} + n^{-1/2} \epsilon z_{jk} \tag{16}$$

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

At each resolution level j, the empirical wavelet coefficients $\tilde{\theta}_{jk}$ are grouped into nonoverlapping blocks of length. Denote (jb) the set of indices of the coefficients in the b-th block at level j, i.e.

$$(jb) = \{(j,k) : (b-1)L + 1 \le k \le bL\}$$

Let $S_{(jb)} \equiv \sum_b \tilde{\theta}_{jk}^2$ denote the L_2 energy of the noisy signal in block (jb). We then apply the James-Stein shrinkage rule to each block (jb). For $jk \in (jb)$,

$$\hat{\theta}_{jk} = (1 - \lambda L \epsilon^2 / S_{(jb)}^2)_+ \tilde{\theta}_{jk}. \tag{17}$$

The estimate of the function f is given by

$$\hat{f}_n(t) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k} \phi_{j_0 k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{jk} \psi_{jk}(t) \tag{18}$$

If one is interested in estimating f at the sample points, then the fast Inverse Discrete Wavelet Transform (IDWT) is employed. And $\{f(t_i) : i = 1, \cdots, n\}$ is estimated by $\hat{f} = \{\widehat{f(t_i)} : i = 1, \cdots, n\}$ with

$$\hat{f} = W^{-1} \cdot n^{1/2} \hat{\Theta}$$

Based on the BP oracle inequality we derived in Section 3.2, we choose the block size $L = \log n$ and the threshold $\lambda = \lambda_* = 4.50524$. With these particular choice of block size and threshold level in (17), we call the estimator in (18) *BlockJS* and denote the estimator in (18) by $\hat{f}_n^*$.

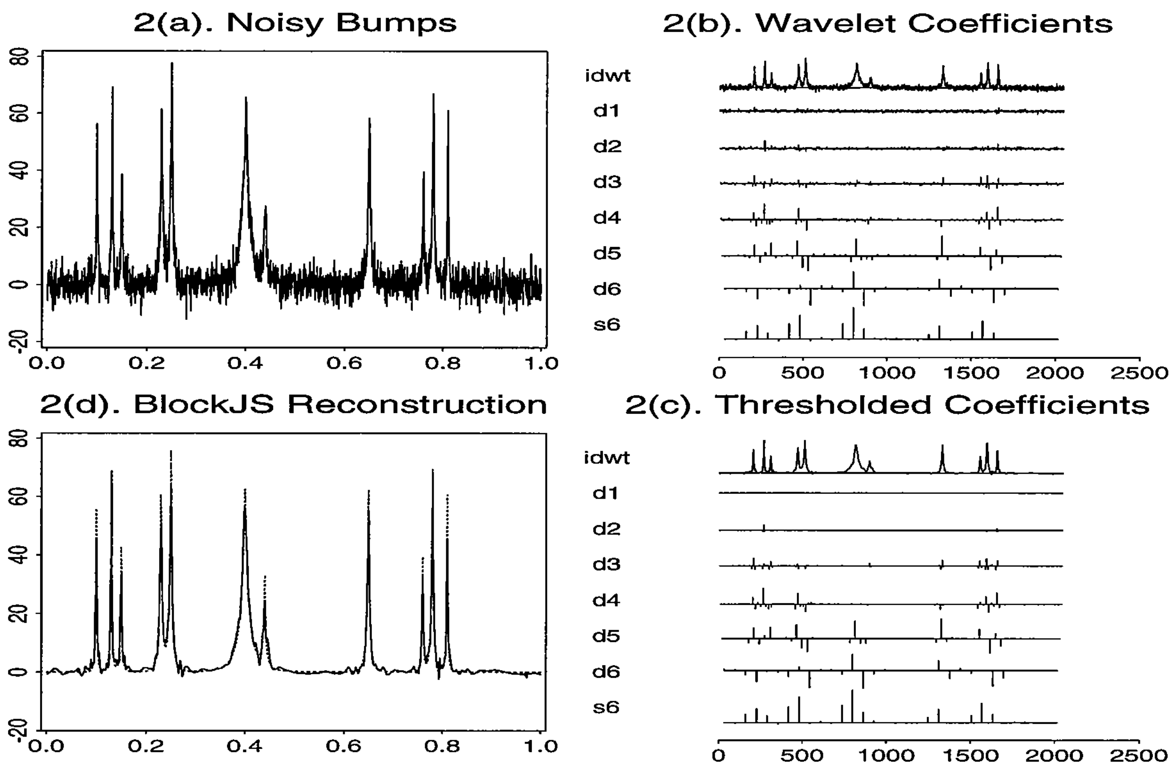
Remark: Here we use the same block length L at all resolution levels. In general the block length can be allowed to increase with the resolution level. All the results remain valid if we set the block length at level j to be $L_j = \log 2^j$.

The *BlockJS* procedure is simple and easy to implement. The computational cost is of the order $O(n)$. The *BlockJS* reconstruction is appealing both quantitatively and qualitatively. In particular, the *BlockJS*, with high probability, removes pure noise completely.

Theorem 2 *If the target function is the zero function $f \equiv 0$, then with probability tending to 1 the BlockJS estimator is also the zero function, i.e., there exist universal constants P_n such that*

$$P(\hat{f}_n^* \equiv 0 | f \equiv 0) \ge P_n \to 1, \quad \text{as } n \to \infty \tag{19}$$

We use Donoho and Johnstone's test function Bumps as an example. Figures 2(a)–2(d) show the *BlockJS* procedure in action. The sample size is 2048 and the signal-to-noise ratio is 3. Again we use Daubechies' *Symmelet 8* wavelet. We can see from figure 2(d) that the *BlockJS* estimator captures both the smooth and the jump features of the function well. For better comparison, the true function, Bumps, is superimposed on the estimator as a dotted line. See Section 7 for more simulation results.



5 Optimality Of The BlockJS Procedure

5.1 Global Properties

In this section, we investigate the adaptivity of the *BlockJS* procedure across the Besov classes and the perturbed Besov classes. The reason for studying adaptivity over Besov spaces is that they are very rich function spaces. They contain many traditional smoothness spaces such as Hölder and Sobolev Spaces. Besov spaces also include function classes of significant spatial inhomogeneity such as the Bump Algebra and the Bounded Variation Classes.

Testing adaptivity over the Besov classes is now becoming a standard procedure for wavelet methods. The *BlockJS* enjoys excellent adaptivity across a wide range of Besov classes. Before we state the results, we must first define the Besov spaces.

Let $\Delta_h^0 f(t) \equiv f(t)$ and $\Delta_h^{r+1} f(t) = \Delta_h^r f(t+h) - \Delta_h^r f(t), \quad r = 1, 2, \cdots$. The $L_p[0,1]$-modulus of smoothness is defined as

$$\omega_r(f;h) = \|\Delta_h^r f\|_{L^p[0,1-rh]}.$$

Given $\alpha > 0$, $0 < p \le \infty$ and $0 < q \le \infty$, choose $r > \alpha$. Then the Besov seminorm of index (α, p, q) is defined as

$$|f|_{B^\alpha_{p,q}} = (\int [h^{-\alpha}\omega_r(f;h)]^q \frac{dh}{h})^{1/q}$$

with usual change to a supremum when $q=\infty$. The Besov Space norm is

$$\|f\|_{B^\alpha_{p,q}} = \|f\|_p + |f|_{B^\alpha_{p,q}}$$

And the Besov space $B^\alpha_{p,q}$ is the set of functions $f:[0,1]\to I\!R$ satisfying $\|f\|_{B^\alpha_{p,q}} \le \infty$. See DeVore and Popov (1988).

For a given r-regular mother wavelet ψ with $r>\alpha$, define the sequence seminorm of the wavelet coefficients of a function f by

$$|\theta|_{\tilde{b}^\alpha_{p,q}} = (\sum_{j=j_0}^{\infty} (2^{js}(\sum_k |\theta_{jk}|^p)^{1/p})^q)^{1/q}$$

where $s=\alpha+1/2-1/p$. The wavelet basis provides smoothness characterization of the Besov spaces. It is an important fact that the Besov function norm $\|f\|_{B^\alpha_{p,q}}$ is equivalent to the sequence norm of the wavelet coefficients of f. See Meyer (1992).

$$\|f\|_{B^\alpha_{p,q}} \asymp \|\xi_{j_0 k}\|_p + |\theta|_{\tilde{b}^\alpha_{p,q}}.$$

We will always use the equivalent sequence norm in our calculations with $\|f\|_{B^\alpha_{p,q}}$.

We now investigate the adaptivity of the *BlockJS* procedure over Besov classes. The *BlockJS* utilizes information about neighboring wavelet coefficients. The block length increases slowly as the sample size increases. As a result, the amount of information available from the data to estimate the energy of the function within a block, and making a decision about keeping or omitting all the coefficients in the block, would be more than in the case of the term-by-term threshold rule. The *BlockJS* increases the estimation accuracy of the wavelet coefficients and so it allows convergence rates to be improved. In this section, we show that this is in fact true.

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

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

The minimax risk over Besov classes has been studied by Donoho and Johnstone (1997). They showed that the minimax risk over a Besov class $B^\alpha_{p,q}(M)$ is of the order n^{-r} with $r=2\alpha/(1+2\alpha)$, i.e.

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

And the minimax linear rate of convergence is $n^{-r'}$ as $n\to\infty$ with

$$r = \frac{\alpha + (1/p_- - 1/p)}{\alpha + 1/2 + (1/p_- - 1/p)}, \quad \text{where } p_- = max(p,2). \tag{20}$$

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 simple block thresholding rule attain the exact optimal convergence rate over a wide range of the Besov scales.

Theorem 3 *Suppose the wavelet ψ is r-regular. Then the BlockJS estimator satisfies*

$$\sup_{f \in B^\alpha_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{21}$$

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

Thus, the *BlockJS* estimator, without knowing the a priori degree or amount of smoothness of the underlying function, attains the true optimal convergence rate that one could achieve by knowing the regularity.

$$\sup_{f \in B^\alpha_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \asymp R(B^\alpha_{p,q}(M), n), \qquad \text{for p} \ge 2$$

Theorem 4 *Assume that the wavelet ψ is r-regular. Then the BlockJS estimator is simultaneously within a logarithmic factor from being minimax for $p < 2$:*

$$\sup_{f \in B^\alpha_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2/p-1}{1+2\alpha}}(1 + o(1)) \tag{22}$$

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

Therefore, the *BlockJS* achieves advantages over the traditional methods even at the level of rates.

Hall, Kerkyacharian and Picard (1995) and Cai (1996) study adaptivity of wavelet procedures over perturbed instead of standard Besov classes. We finally note that the *BlockJS* estimator also achieves optimal performance over a wide range of perturbed Besov classes. We state the results in the following theorem. The readers are referred to Cai (1996) and Hall, Kerkyacharian and Picard (1995) for the definition of perturbed Besov classes considered in the theorem.

Theorem 5 *Suppose the wavelet ψ is r-regular. The BlockJS estimator is simultaneously rate-optimal over an interval of perturbed Besov classes $\mathcal{F} = \mathcal{F}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, D, \nu)$.*

$$\sup_{f \in \mathcal{F}} E\|\hat{f}^*_n - f\|^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{23}$$

for all $0 < \alpha \le r$ and for all $\nu \ge N$.

5.2 Asymptotic Equivalence And Approximation

Brown and Low (1996) obtain an important result on the asymptotic equivalence between the nonparametric regression and the white noise model. Specifically, they show that, under conditions, observing the noisy sampled data as in (13) is asymptotically equivalent to observing the stochastic process $Y(t), \ t \in [0, 1]$ where the process Y is characterized by

$$dY(t) = f(t)dt + n^{-1/2}\epsilon dW(t) \tag{24}$$

with W a standard Wiener process. The two experiments cannot be distinguished asymptotically by any statistical tests. Furthermore, for any procedure in one experiment, we can construct an equivalent procedure in another experiment. Because the wavelet bases we use are orthogonal bases, observing the white-noise-with-drift process (24) is in turn equivalent to observing an infinite sequence of wavelet coefficients of f contaminated with i.i.d. Gaussian noise of noise level $n^{-1/2}\epsilon$.

We shall prove Theorem 3 and 4 by using a method of sequence spaces introduced by Donoho and Johnstone (1997). A key step is to use the equivalence idea and to approximate the problem of estimating f from the noisy observations in (13) by the problem of estimating the wavelet coefficient sequence of f contaminated with i.i.d. Gaussian noise.

The approximation arguments are given in Donoho and Johnstone (1997). They show a strong equivalence result on the white noise model and the nonparametric regression 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 employed to the empirical wavelet coefficients in the function estimation problem in (13), 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 (1997 and 1995) and Brown and Low (1996). For approximation results, see also Chambolle, DeVore, Lee and Lucier (1996).

Under the correspondence between the estimation problem in function spaces and the estimation problem in sequence spaces, it suffices to solve the sequence estimation problem.

5.3 Estimation in Sequence Space by BlockJS

Suppose we observe sequence data:

$$y_{jk} = \theta_{jk} + n^{-1/2}\epsilon z_{jk}, \quad j \geq 0,\ k = 1, 2, \cdots, 2^j \tag{25}$$

where z_{jk} are i.i.d. $N(0,1)$. The mean vector θ is the object that we wish to estimate. The accuracy of estimation is measured by the expected squared error $R(\hat{\theta}, \theta) = E\sum_{j,k}(\hat{\theta} - \theta)^2$. We assume that θ is known to be in some Besov Body $\Theta^{s}_{p,q}(M) = \{\theta : \|\theta\|_{b^s_{p,q}} \leq M\}$, where

$$\|\theta\|_{b^s_{p,q}} = (\sum_{j=0}^{\infty}(2^{js}(\sum_{k}|\theta_{jk}|^p)^{1/p})^q)^{1/q}$$

The minimax risk of estimating θ over the Besov Body is defined as

$$R(\sigma, \Theta^{s}_{p,q}(M)) = \inf_{\hat{\theta}} \sup_{\Theta^{s}_{p,q}(M)} E\|\hat{\theta} - \theta\|_2^2$$

The minimax rate of estimation over Besov Body has been derived by Donoho and Johnstone (1997). First let us make the usual calibration $s = \alpha + 1/2 - 1/p$. Donoho and Johnstone 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 $n \to \infty$.

We now apply a *BlockJS*-type procedure to this sequence estimation problem. Let $J = [\log_2 n]$. Divide each resolution level $j_0 \le j < J$ into nonoverlapping blocks of length $L = [\log n]$. Again denote *(jb)* the *b*-th block at level *j*. Now estimate θ by $\hat{\theta}^*$ with

$$\hat{\theta}^*_{jk} = \begin{cases} y_{jk} & \text{for } j \le j_0 \\ (1 - \lambda_* L\epsilon^2 / S^2_{(jb)})_+ \, y_{jk} & \text{for } jk \in (jb), \;\; j_0 \le j < J \\ 0 & \text{for } j \ge J \end{cases} \tag{26}$$

This estimator enjoys a high degree of adaptivity. Specifically, we have

Theorem 6 *Let $p \ge 2$. Then*

$$\sup_{\Theta^s_{p,q}(M)} E\|\hat{\theta}^* - \theta\|_2^2 \le C n^{-\frac{2\alpha}{1+2\alpha}}(1 + o(1)), \quad as \;\; n \to \infty \tag{27}$$

That is, the estimator attains the exact minimax rate over all the Besov Bodies $\Theta^s_{p,q}(M)$ with $p \ge 2$. For $p < 2$, we have the following result.

Theorem 7 *Let $p < 2$ and $\alpha p \ge 1$. Then*

$$\sup_{\Theta^s_{p,q}(M)} E\|\hat{\theta}^* - \theta\|_2^2 \le C n^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}} (1 + o(1)), \quad as \;\; n \to \infty \tag{28}$$

The results of Theorem 3 and 4 follow from these two theorems and the equivalence and the approximation arguments we mention in Section 5.2.

5.4 Local Adaptation

A distinguished property of many wavelet procedures is their high degree of spatial adaptivity. For functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point. Global risk measures such as (14) cannot wholly reflect the local adaptivity of the estimators. It is more appropriate to use local risk measure (15) for spatial adaptivity.

In this section, we study the optimality of the *BlockJS* 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. Brown and Low (1996) and Lepski (1990) showed that local adaptation cannot be achieved for free.

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 local Hölder class $\Lambda^\alpha(M, t_0, \delta)$ defined as follows. If $\alpha \le 1$,

$$\Lambda^\alpha(M, t_0, \delta) = \{f : |f(t) - f(t_0)| \le M\,|t - t_0|^\alpha, \text{ for } t \in (t_0 - \delta,\ t_0 + \delta)\}$$

If $\alpha > 1$, then

$$\Lambda^\alpha(M, t_0, \delta) = \{f : |f^{(\lfloor\alpha\rfloor)}(t) - f^{(\lfloor\alpha\rfloor)}(t_0)| \le M\,|t - t_0|^{\alpha'} \text{ for } t \in (t_0 - \delta,\ t_0 + \delta)\}$$

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

The optimal rate of convergence for estimating $f(t_0)$ over function class $\Lambda^\alpha(M, t_0, \delta)$ with α known is $n^{-2\alpha/(1+2\alpha)}$. Brown and Low (1996) and Lepski (1990) showed that one has to pay a price for adaptation of at least a logarithmic factor even when α is known to be one of two values. They showed that the best one can do is $(\log n/n)^r$ when the smoothness parameter α is unknown. We call $(\log n/n)^r$ the local adaptive minimax rate over the Hölder class $\Lambda^\alpha(M, t_0, \delta)$.

The *BlockJS* achieves optimal local adaptation with the minimal cost:

Theorem 8 *Suppose the wavelets $\{\phi, \psi\}$ are r-regular with $r \ge \alpha$. Let $t_0 \in (0, 1)$ be fixed. Then the BlockJS estimator $\hat{f}_n^*$ satisfies*

$$\sup_{f \in \Lambda^\alpha(M, t_0, \delta)} E(\hat{f}_n^*(t_0) - f(t_0))^2 \le C \cdot (\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}(1 + o(1)) \tag{29}$$

Remark: The choice of block length $L \sim \log n$ is important for achieving the optimal local adaptivity. The result does not hold if $L = (\log n)^{1+\delta}$, $\delta > 0$. The estimator proposed by Hall, Kerkyacharian and Picard (1995) uses block length $L = (\log n)^{1+\delta}$ and thus does not achieve optimal local adaptivity.

6 Choices of Block Size and Threshold Level

In the function estimation problem (13), We have three objectives in mind: *adaptivity, spatial adaptivity,* and *visual quality* of the estimator. The first objective, adaptivity, requires to minimize the global risk measure (14) adaptively over some unknown smoothness class. The second objective, spatial adaptivity, needs the estimator be well localized so irregularities at one point will not affect the estimator at locations away from the point. The third is to have a "noise-free" reconstruction.

We have shown in the previous sections that with block size $L = \log n$ and $\lambda = 4.50524$, the *BlockJS* estimator achieves the three objectives simultaneously. In particular, Theorems 2, 3 and 8 hold. Naturally, one would ask from where the block size and the threshold come? To answer the question, we first look at the risk of block thresholding estimators of the form (10). The following oracle inequality provides a generalization of (12).

Theorem 9 *Assume that y_i and $\hat{\theta}_{B_j}(L,\lambda)$ are given as in (2) and (10) respectively. Then*

$$R(\hat{\theta}(\lambda, L), \theta) \leq \lambda R_{oracle}(\theta, \sigma, L) + 4\sigma^2 \cdot P(\chi_L^2 > \lambda L). \tag{30}$$

For the purpose of selecting block size and threshold level we regard the RHS of the oracle inequality (30) as the true risk instead of an upper bound. The second term on the RHS of (30) is important. It determines the balance between the block length L and the threshold level λ. For a chosen block size L, we select the corresponding threshold based on the minimax quantity

$$\lambda_L = \arg\min_{\lambda} \sup_{\theta} \frac{\lambda(\|\theta\|_2^2 \wedge L\sigma^2) + 4\sigma^2 \cdot P(\chi_L^2 > \lambda L)}{(\|\theta\|_2^2 \wedge L\sigma^2) + L\sigma^2/n} \tag{31}$$

Based on this criterion, the threshold λ decreases as the block size L increases.We consider here three special cases of block size.

Proposition 1 *Let the threshold λ_L be defined as in (31), then*
(i). with block size $L = \log n$,

$$\lambda_L \sim 4.50524, \qquad as \;\; n \to \infty; \tag{32}$$

(ii). with $L = 1$,

$$\lambda_L \sim \sqrt{2\log n}, \qquad as \;\; n \to \infty; \tag{33}$$

(iii). with $L = (\log n)^{1+\delta}$, $\delta > 0$,

$$\lambda_L \sim 1 \qquad as \;\; n \to \infty. \tag{34}$$

Our choice of threshold $\lambda_L = 4.50524$ used in the *BlockJS* estimator is based on (32). The choice of block size $L = \log n$ is aiming for achieving a high degree of both global and local adaptivity. The proof of Proposition 1 uses Lemma 2 on chi-square tail probabilities in Section 9. We now consider other choices of block sizes discussed in Proposition 1 and summarize the results in the following theorems.

Theorem 10 *With $L = 1$ and $\lambda = \sqrt{2\log n}$ in (17), denote $\hat{f}_n^{(1)}$ the estimator given by (18). Then Theorems 2 and 8, but not Theorem 3, hold for $\hat{f}_n^{(1)}$. Under the conditions of Theorem 3, $\hat{f}_n^{(1)}$ satisfies*

$$\sup_{f \in B_{p,q}^{\alpha}(M)} E\|\hat{f}_n^* - f\|^2 \leq C n^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2\alpha}{1+2\alpha}} (1 + o(1)) \tag{35}$$

Therefore, the estimator $\hat{f}_n^{(1)}$ has noise-free feature and optimal local adaptivity, but not optimal global adaptivity. The extra logarithmic factor in (35) is unavoidable because this is a term-by-term thresholding estimator. The shrinkage function $\eta_\lambda^{JS}(x) = (1 - \lambda/x^2)_+ x$ is bounded between the hard threshold $\eta_\lambda^h(x) = x \cdot I(|x| > \lambda)$ and the soft threshold $\eta_\lambda^s(x) = sgn(x)(|x| - \lambda)_+$. The estimator enjoys essentially the same properties as the VisuShrink estimator. This estimator has also been studied by Gao (1997).

Theorem 11 *With block size $L = (\log n)^{1+\delta}$, $\delta > 0$ and a constant threshold $\lambda > 1$ in (17), denote $\hat{f}_n^{(2)}$ the estimator given by (18). Then Theorems 2 and 3, but not Theorem 8, hold for $\hat{f}_n^{(2)}$. Under the conditions of Theorem 8, $\hat{f}_n^{(2)}$ satisfies*

$$\sup_{f\in\Lambda^\alpha(M)} E(\hat{f}_n^{(2)}(t_0) - f(t_0))^2 \le C \cdot (\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}} \cdot (\log n)^{\frac{2\alpha\delta}{1+2\alpha}}(1 + o(1)) \tag{36}$$

In words, the estimator $\hat{f}_n^{(2)}$ achieves global adaptivity, but not optimal local adaptivity. The extra logarithmic factor in (36) is due to the fact that the estimator is not well localized. The block size $L = (\log n)^{1+\delta}$ is too large. Intuitively it is also clear that the block size should be relatively small in order to achieve a high degree of spatial adaptivity.

In comparison, it is clear that the choice of block size $L = \log n$ and threshold $\lambda = 4.50524$ is superior than $L = 1$ and $\lambda = \sqrt{2\log n}$, or $L = (\log n)^{1+\delta}$ and $\lambda \sim 1$.

7 Simulation Results

A simulation study is conducted to investigate the empirical performance of the *BlockJS* estimator. We compare the *BlockJS* with Donoho and Johnstone's VisuShrink and SureShrink as well as Coifman and Donoho's Translation-Invariant (TI) De-Noising method. SureShrink selects threshold at each resolution level by minimizing Stein's unbiased estimate of risk at each resolution level. In the simulation, we use the hybrid method proposed in Donoho and Johnstone (1995). The TI De-Noising method was introduced by Coifman and Donoho (1995). The reconstruction was obtained by 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 on SureShrink and TI De-Noising the readers are referred to Donoho and Johnstone (1995) and Coifman and Donoho (1995).

We implement the *BlockJS* estimator in the statistical software package S+Wavelets. The complexity of the algorithm is of the order $O(n)$. We compare the numerical performance of the methods using eight test functions representing different level of spatial variability. The *BlockJS*, the VisuShrink , the SureShrink, and the TI De-Noising are applied to noisy versions of the test functions. Sample sizes from $n = 512$ to $n = 8192$ and signal-to-noise ratios (SNR) from 3 to 7 are considered. And several different wavelets are used.

The *BlockJS* uniformly outperforms the VisuShrink in all examples. Among the eight test functions, five of them, Doppler, Bumps, Blocks, Spikes and Blip, *BlockJS* has better precisions with sample size n than the VisuShrink with sample size $2 \cdot n$ for all n from 512 to 819 (see Table 1). The *BlockJS* also yields better results than the TI De-Noising in most cases, especially when the underlying function is of significant spatial variability. The *BlockJS* is comparable to SureShrink numerically. See Table 1 and Figure 10.

The *BlockJS* estimator is visually appealing. The reconstruction is smooth where the underlying function is smooth. They do not contain spurious fine-scale structure that are often contained in the SureShrink estimator (see Figures 4 and 5).

Different combinations of wavelets and signal-to-noise ratios yield basically the same results. For the reasons of space, we report in Table 1 only the numerical results for $SNR = 7$ using Daubechies compactly supported wavelet *Symmlet* 8. Table 1 reports the average squared error over 20 replications with sample sizes ranging from $n = 512$ to $n = 8192$. Figures 4 - 7 compare the visual quality of the four different estimation methods. All the figures were produced with the sample size 2048, the wavelet *Symmlet* 8 and the SNR 3.

It would have been very interesting to compare the mean squared error performance of *BlockJS* with that of Hall, Kerkyacharian and Picard's block thresholding estimator. But as we pointed out earlier, the block thresholding estimator proposed by Hall, Kerkyacharian and Picard is difficult to be implemented. It requires the evaluation of the wavelet functions at all the sample points and involves matrix multiplications which are computationally costly. Also simulation results in Hall, Penev, Kerkyacharian, and Picard (1996) show that even the translation-averaged version of their estimator has little advantage over the VisuShrink when the signal to noise ratio is high. Since our simulation shows that *BlockJS* uniformly outperforms the VisuShrink in all examples, we expect the *BlockJS* estimator performs favorably over HKP's estimator in terms of the mean squared error, at least in the case of high signal-to-noise-ratio.

8 Discussion

The *BlockJS* estimator can be modified by averaging over different block centers. This technique has also been used in Hall, Penev, Kerkyacharian and Picard (1996). Specifically, for each given $0 \le i \le L-1$, partition the indices at each resolution level j into blocks

$$(jb) = \{(j,k) : \ (b-1)L + i + 1 \le k \le bL + i\}.$$

In the original *BlockJS* estimator, we take $i = 0$. Defining $\hat{f}_n^{(i)}$ to be the version of $\hat{f}_n^*$ for a given i and set

$$\hat{f}_n^{**} = \sum_{i=0}^{L-1} \hat{f}_n^{(i)} / L.$$

The estimator $\hat{f}_n^{**}$ often has better numerical advantage.

James-Stein estimator has been used in wavelet function estimation by Donoho and Johnstone (1995). The estimator, WaveJS, is constructed by applying the James-Stein estimator resolution-level-wise, so it is not local and does not have the spatial adaptivity enjoyed by the *BlockJS* estimator introduced in the present paper. Indeed, the main purpose of Donoho and Johnstone's introduction of the WaveJS procedure is to show that linear estimator does not perform well even it is an adaptive and nearly-ideal linear estimator(See Donoho and Johnstone (1995)).

In the present paper, we focus on the James-Stein estimators. The block thresholding method can be used with other type of shrinkage estimators in normal decision theory, for example, estimators of the forms $\hat{\theta} = (1 - \lambda_1\sigma^2/(\lambda_2 + S^2))_+ y$ or $\hat{\theta} = (1 - c(S^2)/S^2)_+ y$. In this sense, block thresholding serves as a "bridge" between traditional normal decision theory and the recent adaptive wavelet estimation. This bridge enables us to utilize the rich results developed in the decision theory for wavelet function estimation.

On the other hand, block thresholding methods can also be used in other statistical function estimation problems such as density estimation and linear inverse problems. We will address these applications elsewhere.

9 Proofs

Proof of Theorem 9. Let $x_i = \mu_i + \sigma z_i,\ \ i = 1, \cdots, L$, and let

$$\hat{\mu}_i = (1 - \frac{\lambda L \sigma^2}{S^2})_+ x_i. \tag{37}$$

where $S^2 = \|x\|^2$ and $\lambda \geq 1$ is a constant. Denote $R(\hat{\mu}, \mu, \sigma) = E_\sigma\|\hat{\mu} - \mu\|_2^2$, and $\mu^* = \mu/\sigma$. Since $R(\hat{\mu}, \mu, \sigma) = \sigma^2 R(\hat{\mu^*}, \mu^*, 1)$. It suffices to consider only the case $\sigma = 1$ and to show

$$E\|\hat{\theta} - \theta\|_2^2 \leq \|\mu\|^2 \wedge \lambda L + 4P(\chi_L^2 > \lambda L). \tag{38}$$

The (positive part) James-Stein estimator is weakly differentiable, Stein's formula for unbiased estimate of risk yields

$$E\|\hat{\mu} - \mu\|_2^2 = E\ [Sure(x, \mu, \lambda, L)] \tag{39}$$

where

$$Sure(x, \mu, \lambda, L) = L + \frac{\lambda^2 L^2 - 2\lambda L(L-2)}{S^2} \cdot I(S^2 > \lambda L) + (S^2 - 2L) \cdot I(S^2 \leq \lambda L) \tag{40}$$

is Stein's unbiased risk estimate. Simple algebra yields

$$Sure(x, \mu, \lambda, L) \leq \max\{\lambda L - L + 4,\ L\} \tag{41}$$

And it follows from (41) trivially

$$E\|\hat{\mu} - \mu\|_2^2 \leq \lambda L + 4P(\chi_L^2 > \lambda L) \tag{42}$$

for $\lambda \geq 1$ and $L \geq 4$. The inequality (42) can be verified directly for the cases of $L = 1,\ 2$ and 3 using the specific noncentral chi-square distributions. For the sake of brevity we omit the proof. It remains to be shown

$$E\|\hat{\mu} - \mu\|_2^2 \leq \|\mu\|^2 + 4P(\chi_L^2 > \lambda L) \tag{43}$$

It follows from (39) and (40) that

$$E\|\hat{\mu}-\mu\|_2^2 = \|\mu\|^2 + E\left[\frac{\lambda^2L^2 - 2\lambda L^2 + 4\lambda L}{S^2} - S^2 + 2L\right] I(S^2 > \lambda L). \tag{44}$$

Let $\mu_* = \|\mu\|^2/2$, then $E\|\hat{\mu}-\mu\|_2^2$ is a function of μ_*, λ and L. $S^2 = \sum x_i^2$ has a noncentral χ^2-distribution with density

$$f(y) = \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!} f_{L+2k}(y) \tag{45}$$

where $f_m(y) = \frac{1}{2^{m/2}\Gamma(m/2)} y^{m/2-1} e^{-y/2}$ is the density of a central χ^2-distribution. Write

$$G(\mu_*, \lambda, L) = E\left[\frac{\lambda^2L^2 - 2\lambda L^2 + 4\lambda L}{S^2} - S^2 + 2L\right] I(S^2 > \lambda L).$$

It is easy to see that $G(0, \lambda, L) \leq 4P(\chi_L^2 > \lambda L)$. So it suffices to show that $G(\mu_*, \lambda, L)$ is decreasing in μ_*. Denote Y_m a central χ^2 variable with degrees of freedom m and use (45) as the density of S^2, we have

$$\begin{aligned} G(\mu_*, \lambda, L) &= \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!} E\left[\frac{\lambda^2L^2 - 2\lambda L^2 + 4\lambda L}{Y_{L+2k}} - Y_{L+2k} + 2L\right] I(Y_{L+2k} > \lambda L) \\ &\equiv \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!} g_k \end{aligned}$$

Since,

$$\frac{\partial G(\mu_*, \lambda, L)}{\partial \mu_*} = \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!} [g_{k+1} - g_k],$$

it is thus sufficient to show $g_{k+1} - g_k \leq 0$ for all $k \geq 0$. Some algebra yields that for $L > 2$

$$g_{k+1} - g_k = (\frac{2\lambda L}{L+2k} - 2) P(Y_{L+2k} > \lambda L) - \frac{2\lambda L(\lambda L - L + 2k + 2)}{(L+2k)(L+2k-2)} P(Y_{L+2k-2} > \lambda L) \tag{46}$$

It is easy to see that $g_{k+1} - g_k \leq 0$ when $\lambda \leq L + 2k$. For the case of $\lambda > L + 2k$, we appeal to the following lemma on two chi-square tail probabilities.

Lemma 1

$$P(\chi_{n+2}^2 \geq T) \leq \frac{T}{n}(1 + \frac{2}{T-n}) P(\chi_n^2 \geq T) \quad \textit{if} \;\; T > n. \tag{47}$$

Applying (47) to (46), it yields $g_{k+1} - g_k \leq 0$ when $\lambda > L + 2k$. Therefore, when $L > 2$, $G(\mu_*, \lambda, L)$ is decreasing in μ_*. Hence

$$G(\mu_*, \lambda, L) \leq G(0, \lambda, L) \leq 4P(\chi_L^2 > \lambda L),$$

and

$$E\|\hat{\mu}-\mu\|_2^2 = \|\mu\|^2 + G(\mu_*,\lambda,L) \le \|\mu\|^2 + 4P(\chi_L^2 > \lambda L).$$

The cases of $L=1$ and $L=2$ can be verified directly, we omit the proof here. ∎

Proof of Theorem 1 : Theorem 1 follows from Theorem 9 and the following lemma on the bounds of the tail probability of a central chi-square distribution.

Lemma 2 *The tail probability of χ_L^2 has the following lower and upper bounds.*

$$\frac{2}{5}\lambda^{-1}\, L^{-\frac{1}{2}}\, [\lambda^{-1}\, e^{\lambda-1}]^{-\frac{L}{2}} \le P(\chi_L^2 \ge \lambda L) \le \frac{1}{2}\, [\lambda^{-1}\, e^{\lambda-1}]^{-\frac{L}{2}} \tag{48}$$

With $L=\log n$ and $\lambda_* = 4.50524$, Lemma 2 yields

$$P(\chi_L^2 > \lambda L) \le \frac{1}{2}[\lambda^{-1}e^{\lambda-1}]^{-L/2} \le \frac{1}{2n}.$$

Theorem 1 now follows from Theorem 9. ∎

The following elementary inequalities between two different ℓ_p norms are needed in the proofs of Theorems 6 and 7.

Lemma 3 *Let $x \in I\!\!R^m$, and $0 < p_1 \le p_2 \le \infty$. Then the following inequalities hold:*

$$\|x\|_{p_2} \le \|x\|_{p_1} \le m^{\frac{1}{p_1}-\frac{1}{p_2}}\|x\|_{p_2} \tag{49}$$

Proof of Theorem 6 : Let y and $\hat{\theta}^*$ be given as in (25) and (26) respectively. Then,

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

Denote by C a generic constant that may vary from place to place. Since $\theta \in \Theta_{p,q}^{\alpha}(M)$, so $2^{js}(\sum_{k=1}^{2^j}|\theta_{jk}|^p)^{1/p} \le M$. It follows from Lemma 3 that $p \ge 2$ implies

$$\sum_{k=1}^{2^j}|\theta_{jk}|^2 \le M^2 2^{-j2\alpha}$$

It is clear that both S_1 and S_3 are "small".

$$S_1 = 2^{j_0}n^{-1}\epsilon^2 = o(n^{-2\alpha/(1+2\alpha)}) \tag{50}$$

$$S_3 = \sum_{j=J}^{\infty}\sum_{k=1}^{2^j}\theta_{jk}^2 \le \sum_{j=J}^{\infty} M^2 2^{-j2\alpha} \le Cn^{-2\alpha} = o(n^{-2\alpha/(1+2\alpha)}) \tag{51}$$

Now let us consider the term S_2. Denote by $\beta^2_{(jb)} = \sum_{k\in(jb)} \theta^2_{jk}$ the sum of squared coefficients within the block (jb). From the BP oracle inequality (1) we can bound the term S_2 by the following.

$$S_2 = \sum_{j=j_0}^{J-1} \sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 \le \lambda_* \sum_{j=j_0}^{J-1} \sum_b (\beta^2_{(jb)} \wedge Ln^{-1}\epsilon^2) + 2n^{-1}\epsilon^2 \tag{52}$$

Let $J_1 = [\frac{1}{1+2\alpha}\log_2 n]$. So, $2^{J_1} \approx n^{1/(1+2\alpha)}$. Then

$$\lambda_* \sum_{j=j_0}^{J_1-1} \sum_b (\beta^2_{(jb)} \wedge Ln^{-1}\epsilon^2) \le \sum_{j=j_0}^{J_1-1} \sum_b \lambda_* Ln^{-1}\epsilon^2 \le Cn^{-2\alpha/(1+2\alpha)} \tag{53}$$

$$\lambda_* \sum_{j=J_1}^{J-1} \sum_b (\beta^2_{(jb)} \wedge Ln^{-1}\epsilon^2) \le \lambda_* \sum_{j=J_1}^{J-1} \sum_b \beta^2_{(jb)} \le Cn^{-2\alpha/(1+2\alpha)} \tag{54}$$

Combining (53) and (54) together with (50) and (51), we prove the theorem.

$$E\|\hat{\theta}^* - \theta\|_2^2 \le Cn^{-2\alpha/(1+2\alpha)}(1 + o(1)) \quad \blacksquare$$

Proof of Theorem 7: Same as in the proof of Theorem 6, we first separate the risk into three components.

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

Since $\theta \in \Theta^\alpha_{p,q}(M)$ and $p < 2$, Lemma 3 yields $\sum_{k=1}^{2^j} |\theta_{jk}|^2 \le M^2 2^{-j2s}$. The first term S_1 is "small" and the assumption $\alpha p \ge 1$ implies that S_3 is of higher order:

$$S_1 = 2^{j_0} n^{-1}\epsilon^2 = o(n^{-2\alpha/(1+2\alpha)}) \tag{55}$$

$$S_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} \theta^2_{jk} \le \sum_{j=J}^{\infty} M^2 2^{-j2s} \le Cn^{-2\alpha-1+2/p} = o(n^{-2\alpha/(1+2\alpha)}) \tag{56}$$

Now we consider the term S_2. First we state the following lemma without proof.

Lemma 4 *Let $0 < p < 1$ and $S = \{x \in \mathbb{R}^k : \sum_{i=1}^k x_i^p \le B,\ x_i \ge 0,\ i = 1, \cdots, k\}$. Then for $A > 0$,*

$$\sup_{x\in S} \sum_{i=1}^k (x_i \wedge A) \le B \cdot A^{1-p}.$$

Again denote $\beta^2_{(jb)} = \sum_{k\in(jb)} \theta^2_{jk}$. Applying the BP oracle inequality (1), we have

$$S_2 = \sum_{j=j_0}^{J-1} \sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 \le \lambda_* \sum_{j=j_0}^{J-1} \sum_b (\beta^2_{(jb)} \wedge Ln^{-1}\epsilon^2) + 2n^{-1}\epsilon^2 \tag{57}$$

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

$$\lambda_* \sum_{j=j_0}^{J_2-1} \sum_b (\beta_{(jb)}^2 \wedge Ln^{-1}\epsilon^2) \leq \sum_{j=j_0}^{J_2-1} \sum_b \lambda_* Ln^{-1}\epsilon^2 \leq Cn^{-2\alpha/(1+2\alpha)}(\log n)^{-(2-p)/p(1+2\alpha)} \quad (58)$$

Note that $\sum_b (\beta_{(jb)}^2)^{p/2} \leq \sum_k (\theta_{j,k}^2)^{p/2} \leq M2^{-jsp}$. Applying Lemma 4, we have

$$\lambda_* \sum_{j=J_2}^{J-1} \sum_b (\beta_{(jb)}^2 \wedge Ln^{-1}\epsilon^2) \leq Cn^{-2\alpha/(1+2\alpha)}(\log n)^{-(2-p)/p(1+2\alpha)} \quad (59)$$

We finish the proof by putting (55)—(59) together.

$$E\|\hat{\theta}^* - \theta\|_2^2 \leq Cn^{-2\alpha/(1+2\alpha)}(\log n)^{-(2-p)/p(1+2\alpha)}(1 + o(1)) \quad \blacksquare$$

Proof of Theorem 8 : For simplicity, we give the proof for Hölder classes $\Lambda^\alpha(M)$ instead of local Hölder classes $\Lambda^\alpha(M, t_0, \delta)$. First note that for Hölder classes $\Lambda^\alpha(M)$ there exists a constant $C > 0$ such that for all $f \in \Lambda^\alpha(M)$,

$$|\theta_{j,k}| = |\langle f, \psi_{j,k}\rangle| \leq C2^{-j(1/2+\alpha)}. \quad (60)$$

The proof of the theorem makes use of the following elementary inequality.

Lemma 5 *Let X_i be random variables, $i = 1, \cdots, n$. Then*

$$E(\sum_{i=1}^n X_i)^2 \leq (\sum_{i=1}^n (EX_i^2)^{1/2})^2 \quad (61)$$

Apply the inequality (61), we have

$$\begin{aligned} E(f_n^*(t_0) - f(t_0))^2 &= E\left[\sum_{k=1}^{2^{j_0}} (\hat{\xi}_{j_0k} - \xi_{j_0k})\phi_{j_0k}(t_0) + \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} (\hat{\theta}_{jk} - \theta_{jk})\psi_{jk}(t_0)\right]^2 \\ &\leq \left[\sum_{k=1}^{2^{j_0}} (E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2\phi_{j_0k}^2(t_0))^{1/2} + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} (E(\hat{\theta}_{jk} - \theta_{jk})^2\psi_{jk}^2(t_0))^{1/2} + \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{jk}\psi_{jk}(t_0)|\right]^2 \\ &\equiv (Q_1 + Q_2 + Q_3)^2 \end{aligned}$$

Since we are using wavelets of compact support, so there are at most N basis functions ψ_{jk} at each resolution level j that are nonvanishing at t_0, where N is the length of the support of the wavelets ϕ and ψ. Denote $K(t_0, j) = \{k : \psi_{j,k}(t_0) \neq 0\}$. Then $|K(t_0, j)| \leq N$. It is easy to see that both Q_1 and Q_3 are small:

$$Q_1 = \sum_{k=1}^{2^{j_0}} (E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2)^{1/2} |\phi_{j_0k}(t_0)| = O(n^{-1}) \quad (62)$$

$$Q_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{jk}||\psi_{jk}(t_0)| \leq \sum_{j=J}^{\infty} N\|\psi\|_\infty 2^{j/2} C2^{-j(1/2+\alpha)} \leq Cn^{-\alpha} \quad (63)$$

We now consider the second term Q_2. Apply Lemma 3 and Theorem 12, and use (60), we have

$$\begin{aligned} Q_2 &\le \sum_{j=j_0}^{J-1} \sum_{k\in K(t_0,\, j)} 2^{j/2}\|\psi\|_\infty (E(\hat\theta_{jk}-\theta_{jk})^2)^{1/2} \le C\sum_{j=j_0}^{J-1} 2^{j/2}[(2^{-j(1+2\alpha)} \wedge Ln^{-1}\epsilon^2) + Ln^{-2}\epsilon^2]^{1/2} \\ &\le C(\log n/n)^{\alpha/(1+2\alpha)}(1+o(1)) \qquad (64) \end{aligned}$$

Combining (62), (63) and (64), we have

$$E(f_n^*(t_0)-f(t_0))^2 \le C(\log n/n)^{2\alpha/(1+2\alpha)}(1+o(1)) \quad \blacksquare$$

Acknowledgments: It is a pleasure to acknowledge helpful comments by Anirban Dasgupta.

References

[1] Brown, L.D. & Cai, T. (1997). Wavelet Shrinkage For Nonequispaced Samples. Technical Report #97-06, Department of Statistics, Purdue University. To appear *Ann. Statist.*

[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 Estimations. *Ann. Statist.* **24**, 2524 - 2335.

[4] Cai, T. (1996). Minimax Wavelet Estimation Via Block Thresholding. Technical Report # 96-41, Department of Statistics, Purdue University.

[5] Cai, T. (1997). On Adaptivity of BlockShrink Wavelet Estimator Over Besov Spaces. Technical Report #97-05, Department of Statistics, Purdue University.

[6] Chambolle, A., DeVore, R., Lee, N. & Lucier, B. (1996). Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal through Wavelet Shrinkage. *IEEE Trans. on Image Processing,* to appear.

[7] Coifman, R.R. & Donoho, D.L. (1995). Translation Invariant De-Noising. *Wavelets and Statistics*, Ed. Anestis Antoniadis, Springer-Verlag, Lecture Notes, 1995.

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

[9] DeVore, R., Jawerth, B. and Popov, V. (1992). Compression of Wavelet Decompositions, *Amer. J. Math.,* 114, 737-785.

[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. To appear *Ann. Statist.*

[12] Donoho, D.L. & Johnstone, I.M. (1994). Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika,* ***81***, 425–455.

[13] Donoho, D.L. & Johnstone, I.M. (1995). Adapt to Unknown Smoothness via Wavelet Shrinkage, *J. Amer. Statist. Assoc.* 90, pp.1200-1224.

[14] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet Shrinkage: Asymptopia?, *J. Roy. Stat. Soc. Ser. B,* ***57***, 301–369.

[15] Efron, B. & Morris, C. (1973). Stein's Estimation Rule and Its Competitors— An Empirical Bayes Approach. *J. Amer. Statist. Assoc.* **68**, 117 –130.

[16] Gao, H.-Y.(1997). Wavelet Shrinkage Denoising Using the Non-Negative Garrote. *J. Comput. Graph. Statist.*, to appear.

[17] Hall, P., Kerkyacharian, G. & Picard, D. (1995). On The Minimax Optimality of Block Thresholded Wavelet Estimators, manuscript.

[18] Hall, P., Penev, S., Kerkyacharian, G. & Picard, D. (1996). Numerical Performance of Block Thresholded Wavelet Estimators, manuscript.

[19] James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361-380. Univ. California Press.

[20] Lehmann, E.L. (1983). *Theory of Point Estimation.* John Wiley & Sons, New York.

[21] Lepski, O.V. (1990). On a Problem of Adaptive Estimation on White Gaussian Noise. *Theory of Probability and Appl.* **35**, 3, 454-466

[22] 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.

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

[24] Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Ann. Statist.* ***9***, 1135 –1151.

[25] Strang, G. (1992). Wavelet And Dilation Equations: A Brief Introduction. *SIAM Review,* 31(4), 614 - 627.

10 Appendix

The formulae of the test functions. (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 and Johnstone (1994). Blip and Wave are from Marron, Adak, Johnstone, Neumann and Patil (1995).

1. *Doppler.*

$$f(x) = 34.5856 \cdot \sqrt{x(1-x)}\sin(2.1\pi/(x+.05))$$

2. *HeaviSine.*

$$f(x) = 3.3662 \cdot [4\sin 4\pi x - sgn(x-.3) - sgn(.72-x)]$$

3. *Bumps.*

$$f(x) = 15.0769 \cdot \sum h_j K((x-x_j)/w_j) \qquad K(x) = (1+|x|)^{-4}.$$

$$\begin{array}{lllllllllllll}
(x_j) & = (.1, & .13, & .15, & .23, & .25, & .40, & .44, & .65, & .76, & .78, & .81) \\
(h_j) & = (4, & 5, & 3, & 4, & 5, & 4.2, & 2.1, & 4.3, & 3.1, & 5.1, & 4.2) \\
(w_j) & = (.005, & .005, & .006, & .01, & .01, & .03, & .01, & .01, & .005, & .008, & .005)
\end{array}$$

4. *Blocks.*

$$f(x) = 4.7606 \cdot \sum h_j K(x-x_j) \qquad K(x) = (1+sgn(x))/2.$$

$$\begin{array}{lllllllllllll}
(x_j) & = (.1, & .13, & .15, & .23, & .25, & .40, & .44, & .65, & .76, & .78, & .81) \\
(h_j) & = (4, & -5, & 3, & -4, & 5, & -4.2, & 2.1, & 4.3, & -3.1, & 5.1, & -4.2)
\end{array}$$

5. *Spikes.*

$$\begin{aligned} f(x) \quad = \quad & 15.6676 \cdot \Big[e^{-500(x-0.23)^2} + 2\,e^{-2000(x-0.33)^2 + 4\,e^{-8000(x-0.47)^2}} \\ & +4\,e^{-8000(x-0.47)^2} + 3\,e^{-16000(x-0.69)^2} + e^{-32000(x-0.83)^2} \Big] \end{aligned}$$

6. *Blip.*

$$f(x) = 50.9859 \cdot [(0.32+0.6x+0.3e^{-100(x-0.3)^2})I_{(0,.8]}(x) + (-0.28+0.6x+0.3e^{-100(x-0.3)^2})I_{[.8,1]}(x)]$$

7. *Corner.*

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

8. *Wave.*

$$f(x) = 63.2301 \cdot [.5 + .2\cos(4\pi x) + .1\cos(24\pi x)]$$

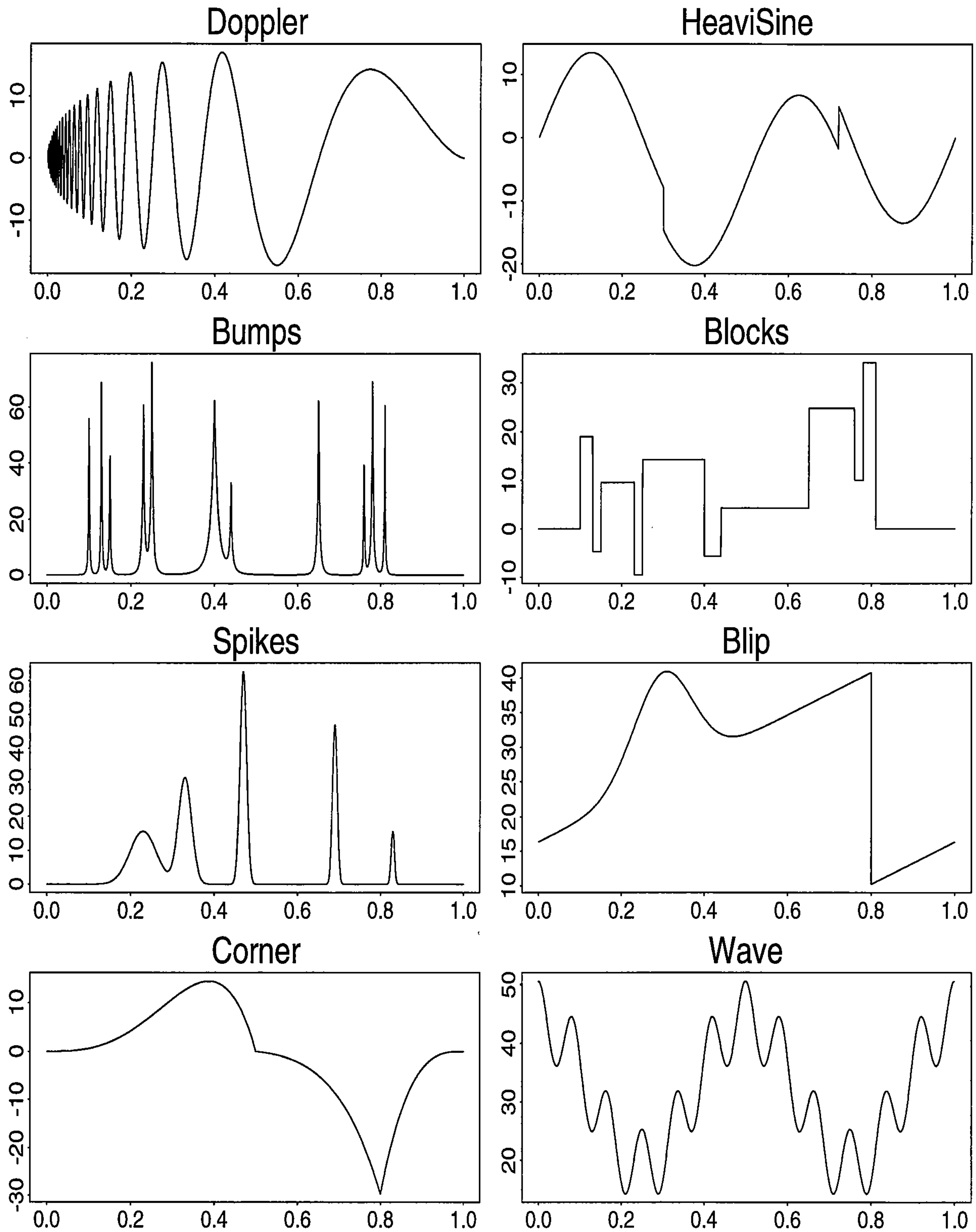


Figure 1: Test Functions

Table 1: Mean Squared Error From 20 Replications (SNR=7)

n	BlockJS	VisuShrink	SureShrink	TI De-Noising
Doppler				
512	0.756	1.838	0.984	1.438
1024	0.424	1.188	0.564	0.886
2048	0.236	0.781	0.352	0.541
4096	0.121	0.424	0.182	0.292
8192	0.060	0.259	0.105	0.169
HeaviSine				
512	0.370	0.395	0.361	0.323
1024	0.217	0.290	0.236	0.223
2048	0.129	0.204	0.138	0.154
4096	0.099	0.117	0.080	0.091
8192	0.059	0.078	0.051	0.062
Bumps				
512	1.758	5.835	1.187	4.034
1024	0.929	3.610	0.977	2.342
2048	0.528	2.211	0.547	1.354
4096	0.391	1.160	0.343	0.712
8192	0.210	0.707	0.210	0.418
Blocks				
512	1.562	3.569	1.335	2.746
1024	0.949	2.290	0.836	1.847
2048	0.584	1.615	0.648	1.253
4096	0.501	0.883	0.367	0.696
8192	0.290	0.620	0.268	0.461
Spikes				
512	0.274	0.502	0.256	0.268
1024	0.149	0.339	0.114	0.155
2048	0.106	0.265	0.086	0.110
4096	0.068	0.191	0.060	0.075
8192	0.053	0.151	0.046	0.055
Blip				
512	0.258	0.455	0.364	0.369
1024	0.150	0.335	0.235	0.253
2048	0.090	0.229	0.132	0.161
4096	0.069	0.139	0.095	0.096
8192	0.038	0.085	0.053	0.061
Corner				
512	0.170	0.208	0.187	0.152
1024	0.077	0.114	0.086	0.086
2048	0.040	0.072	0.045	0.054
4096	0.036	0.036	0.036	0.035
8192	0.018	0.018	0.018	0.018
Wave				
512	0.395	1.402	0.277	0.339
1024	0.178	0.782	0.155	0.203
2048	0.098	0.461	0.092	0.117
4096	0.060	0.060	0.060	0.028
8192	0.044	0.044	0.037	0.014

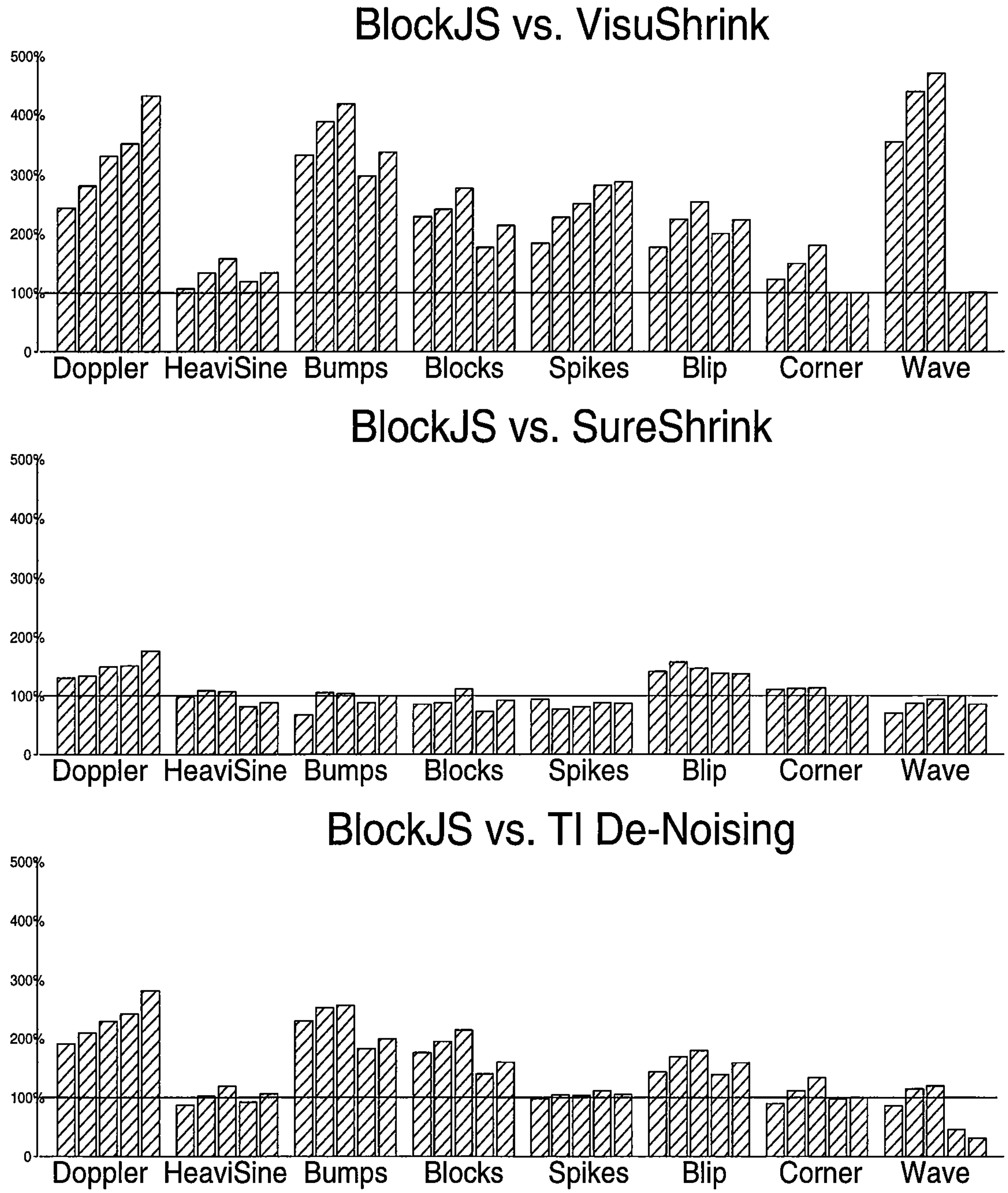


Figure 2: Comparison of MSEs (SNR=7). The average squared errors of VisuShrink, SureShrink, and TI De-Noising as the percentages of the corresponding average squared errors of BlockJS. For each test function, the bars are order from left to right by the sample sizes from 512 to 8192.

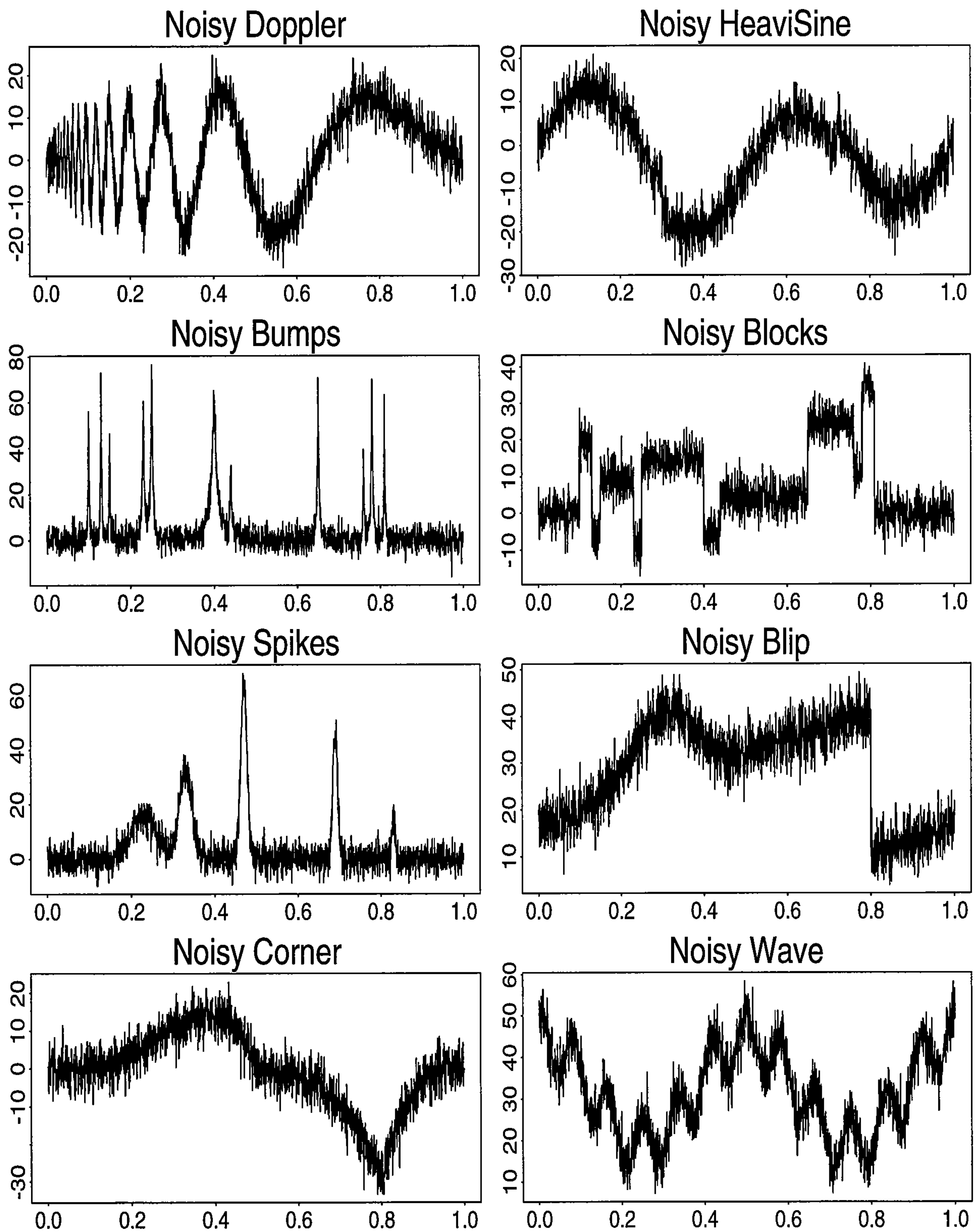


Figure 3: Test Functions + Noise (SNR=3)

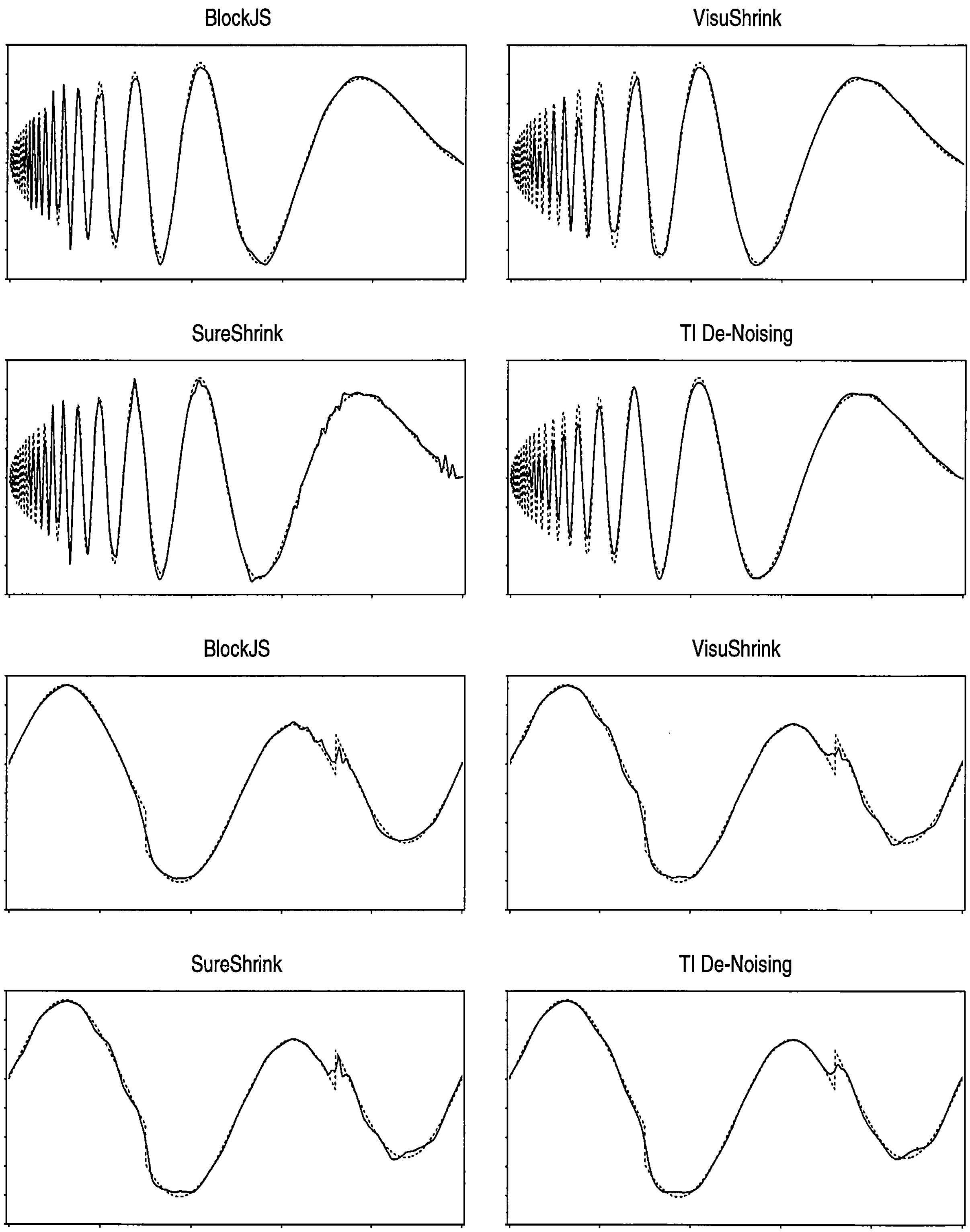


Figure 4: Comparison of Reconstructions (I)

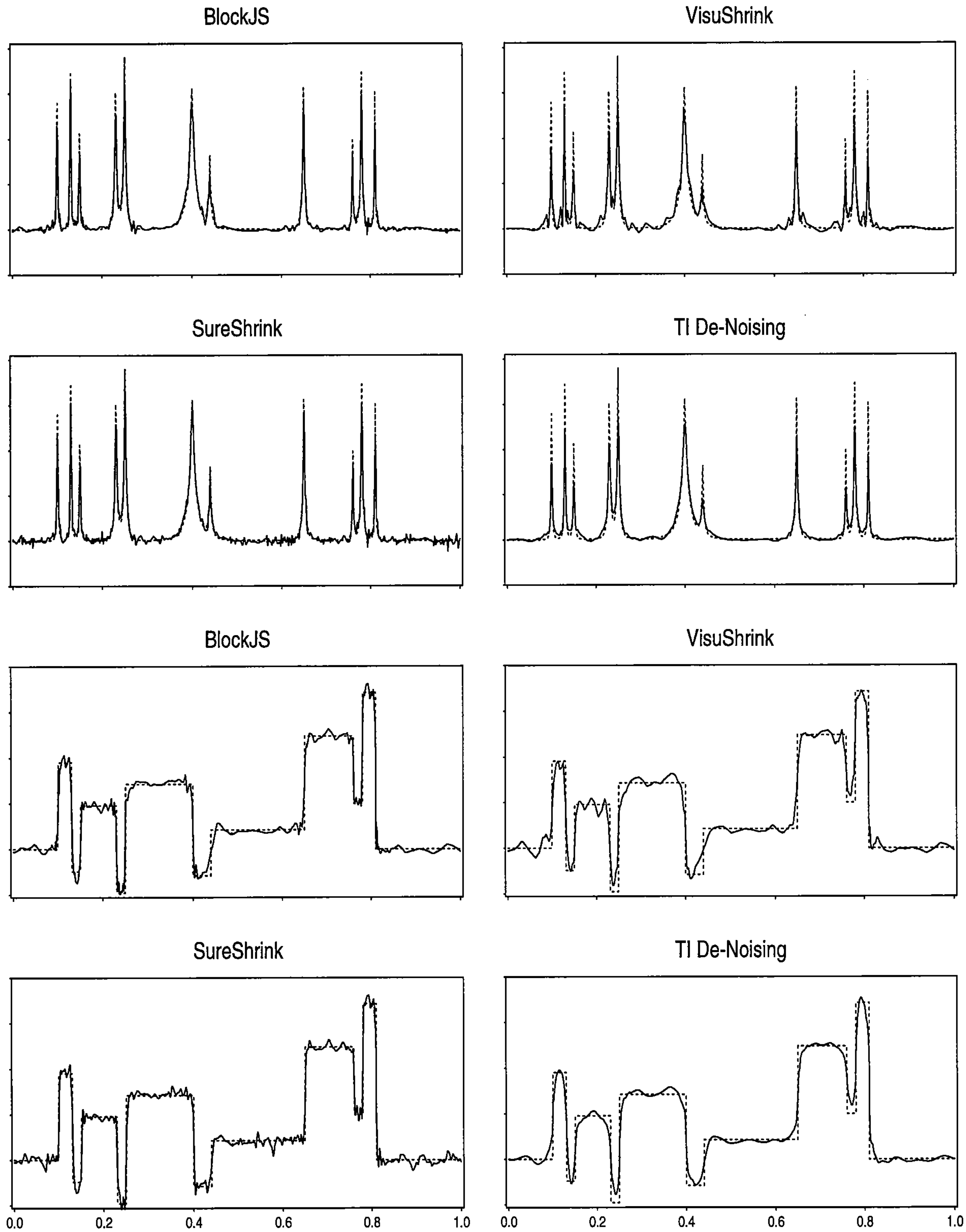


Figure 5: Comparison of Reconstructions (II)

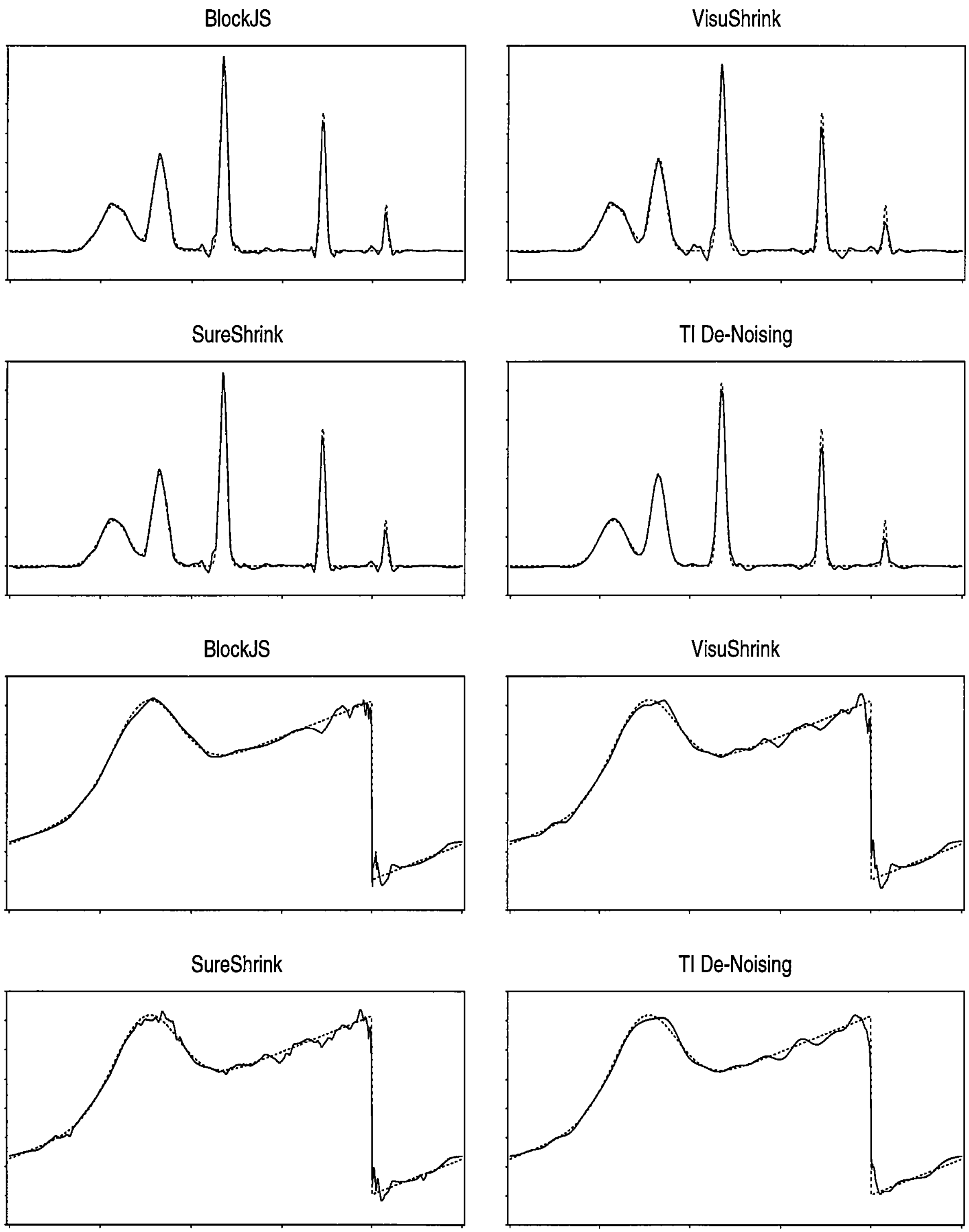


Figure 6: Comparison of Reconstructions (III)

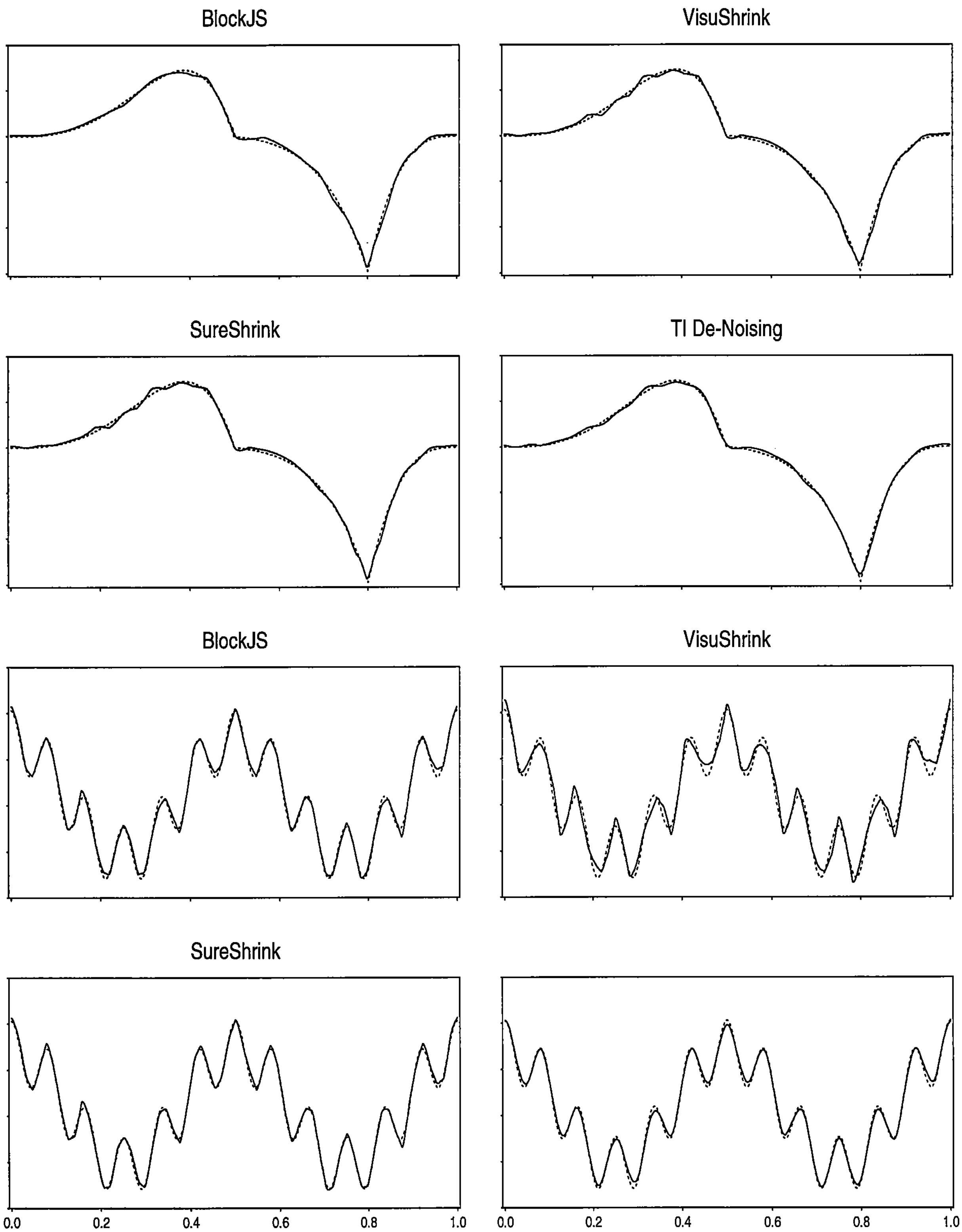


Figure 7: Comparison of Reconstructions (IV)