Adapting to Unknown Smoothness via Wavelet Shrinkage

David L. DONOHO and Iain M. JOHNSTONE

We attempt to recover a function of unknown smoothness from noisy sampled data. We introduce a procedure, *SureShrink,* that suppresses noise by thresholding the empirical wavelet coefficients. The thresholding is adaptive: A threshold level is assigned to each dyadic resolution level by the principle of minimizing the Stein unbiased estimate of risk (*Sure*) for threshold estimates. The computational effort of the overall procedure is order $N \cdot \log(N)$ as a function of the sample size N. *SureShrink* is smoothness adaptive: If the unknown function contains jumps, then the reconstruction (essentially) does also; if the unknown function has a smooth piece, then the reconstruction is (essentially) as smooth as the mother wavelet will allow. The procedure is in a sense optimally smoothness adaptive: It is near minimax simultaneously over a whole interval of the Besov scale; the size of this interval depends on the choice of mother wavelet. We know from a previous paper by the authors that traditional smoothing methods—kernels, splines, and orthogonal series estimates—even with optimal choices of the smoothing parameter, would be unable to perform in a near-minimax way over many spaces in the Besov scale. Examples of *SureShrink* are given. The advantages of the method are particularly evident when the underlying function has jump discontinuities on a smooth background.

KEY WORDS: Besov, Hölder, Sobolev, Triebel spaces; Compactly supported wavelets; Denoising; James–Stein estimator; Minimax decision theory; Nonparametric regression; Nonlinear estimation; Orthonormal bases; Stein unbiased risk estimate; Thresholding; White noise model.

1. INTRODUCTION

Suppose that we are given N noisy samples of a function f,

$$y_i = f(t_i) + z_i, \qquad i = 1, \ldots, N, \tag{1}$$

with $t_i = (i-1)/N$, z_i iid $N(0, \sigma^2)$. Our goal is to estimate the vector $\mathbf{f} = (f(t_i))_{i=1}^N$ with small mean squared error (MSE); that is, to find an estimate $\hat{\mathbf{f}}$ depending on $y_1, \ldots, y_N$ with small *risk* $R(\hat{\mathbf{f}}, \mathbf{f}) = N^{-1} \cdot E\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2 = E\ \mathrm{Ave}_i(\hat{f}(t_i) - f(t_i))^2$.

To develop a nontrivial theory, one usually specifies some fixed class $\mathcal{F}$ of functions to which f is supposed to belong. Then one may seek an estimator $\hat{f}$ attaining the *minimax risk* $R(N, \mathcal{F}) = \inf_{\hat{\mathbf{f}}} \sup_f R(\hat{\mathbf{f}}, \mathbf{f})$.

This approach has led to many theoretical developments of considerable interest (see, for example, Stone 1982, Nemirovskii, Polyak, and Tsybakov 1985; Nussbaum 1985). But from a practical point of view, it has the difficulty that it rarely corresponds with the usual situation where one is given data, but no knowledge of an a priori class $\mathcal{F}$.

To repair this difficulty, one may suppose that $\mathcal{F}$ is an unknown member of a *scale* of function classes and may attempt to behave in a way that is simultaneously near minimax across the entire scale. An example is the L^2 Sobolev scale, a set of function classes indexed by parameters m (degree of differentiability) and C (quantitative limit on the mth derivative):

$$W_2^m(C) = \left\{ f : \|f\|_2^2 + \left\| \frac{d^m}{dt^m} f \right\|_2^2 \le C^2 \right\}. \tag{2}$$

Here and later, $\|f\|_p^p = \int_0^1 |f(t)|^p\, dt$. Work of Efroimovich and Pinsker (1984) and Golubev and Nussbaum (1990), for example, shows how to construct estimates that are simultaneously minimax over a whole range of m and C. Those methods perform asymptotically as well when m and C are unknown as they would if these quantities were known.

Such results are limited to the case of L^2 smoothness measures. There are many other scales of function spaces, such as the Sobolev spaces,

$$W_p^m(C) = \left\{ f : \|f\|_p^p + \left\| \frac{d^m}{dt^m} f \right\|_p^p \le C^p \right\}. \tag{3}$$

If $p < 2$, then linear methods cannot attain the optimal rate of convergence over such a class when m and C are known (Nemirovskii 1985; Donoho and Johnstone 1994b). Thus adaptive linear methods cannot attain the optimal rate of convergence either. If one admits that not only the degree but also the type of smoothness are unknown, then it is not known how to estimate smooth functions adaptively.

In Section 2 we introduce a method, *SureShrink,* which is very simple to implement and attains much broader adaptivity properties than previously proposed methods. The properties apply over function classes measuring smoothness in traditional ways, such as (2), and also in less common but practically relevant ways, such as (3) and the Besov norms (see Sec. 3). The method is based on new results in multivariate normal decision theory that are interesting in their own right.

David L. Donoho and Iain M. Johnstone are Professors, Department of Statistics, Stanford University, CA 94305. The first author was supported at University of California Berkeley by National Science Foundation Grant DMS 88-10192, by NASA Contract NCA2-488, and by a grant from the ATT Foundation. The second author was supported in part by National Science Foundation Grants DMS 84-51750, 86-00235, and 92-09130 by National Institutes of Health, Public Health Service Grant CA 59039, and by a grant from the ATT Foundation. An early version of this article was presented as "Wavelets + Decision Theory = Optimal Smoothing" at the Wavelets and Applications Workshop, Luminy, France, March 10, 1991, and at the Workshop on Trends in the Analysis of Curve Data, University of Heidelberg, March 22, 1991. The authors thank P. Diaconis, T. Gasser, and R. Tibshirani for helpful comments and discussion.

Journal of the American Statistical Association
December 1995, Vol. 90, No. 432, Theory and Methods

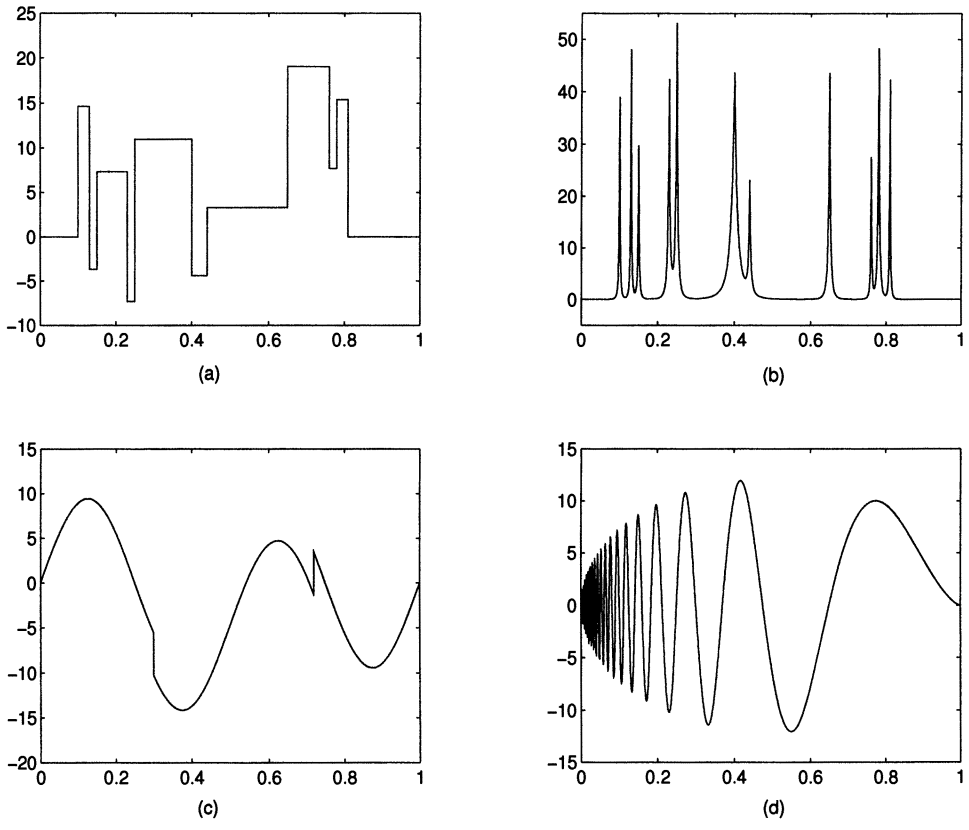

Figure 1. Four Spatially Variable Functions. (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler, N = 2,048. Formulas in Table 1.

SureShrink has the following ingredients (described in greater detail in Sec. 2):

1. *Discrete wavelet transform of noisy data.* The N noisy data are transformed via the discrete wavelet transform, to obtain N noisy wavelet coefficients $(y_{j,k})$.
2. *Thresholding of noisy wavelet coefficients.* Let

$$\eta_t(y) = \text{sgn}(y)(|y| - t)_+ \tag{4}$$

denote the *soft threshold,* which sets to zero data y below t in absolute value and pulls other data toward the origin by an amount t. The wavelet coefficients $y_{j,k}$ are subjected to soft thresholding with a level-dependent threshold level t_j^*.

Table 1. Formulas for Test Functions

Blocks

$$f(t) = \sum h_j K(t - t_j) \quad K(t) = (1 + \text{sgn}(t))/2.$$
$$(t_j) = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81)$$
$$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2)$$

Bumps

$$f(t) = \sum h_j K((t - t_j)/w_j) \quad K(t) = (1 + |t|)^{-4}.$$
$$(t_j) = t_{\text{Blocks}}$$
$$(h_j) = (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2)$$
$$(w_j) = (.005, .005, .006, .01, .01, .03, .01, .01, .005, .008, .005)$$

HeaviSine

$$f(t) = 4 \sin 4\pi t - \text{sgn}(t - .3) - \text{sgn}(.72 - t).$$

Doppler

$$f(t) = \sqrt{t(1 - t)} \sin(2\pi(1 + \varepsilon)/(t + \varepsilon)), \ \varepsilon = .05.$$

3. *Stein's unbiased estimate of risk (Sure) for threshold choice.* The level-dependent thresholds are found by regarding the different resolution levels (different j) of the wavelet transform as independent multivariate normal estimation problems. Within one level (fixed j), one has data $y_{j,k} = w_{j,k} + \varepsilon z_{j,k}, k = 0, \ldots, 2^j - 1$ and one wishes to estimate $(w_{j,k})_{k=0}^{2^j-1}$. Stein's unbiased estimate of risk for $\hat{\theta}_k^{(t)} = \eta_t(y_{j,k})$ gives an estimate of the risk for a particular threshold value t; minimizing this in t gives a selection of the threshold level for that level j. (A fixed threshold modification of this recipe is used in case the data vector has a very small l_2 norm.)

We briefly describe some examples of the method in action. Figure 1 depicts four specific functions f that we wavelet analyze repeatedly in this article:

- *Blocks.* A piecewise constant function, with jumps at $\{.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81\}$.
- *Bumps.* A sum of bumps $\sum_{j=1}^{11} h_j K((t - t_j)/w_j)$ with locations t_j at the same places as jumps in *Blocks;* the heights h_j and widths s_j vary, and the individual bumps are of the form $K(t) = 1/(1 + |t|)^4$.
- *HeaviSine.* A sinusoid of period 1 with two jumps, at $t_1 = .3$ and $t_2 = .72$.
- *Doppler.* The variable-frequency signal $f(t) = \sqrt{t(1-t)} \sin(2\pi \cdot 1.05/t + .05)$.

Precise formulas appear in Table 1. These examples have been chosen to represent various *spatially nonhomogeneous* phenomena. We regard *Blocks* as a caricature of the acoustic

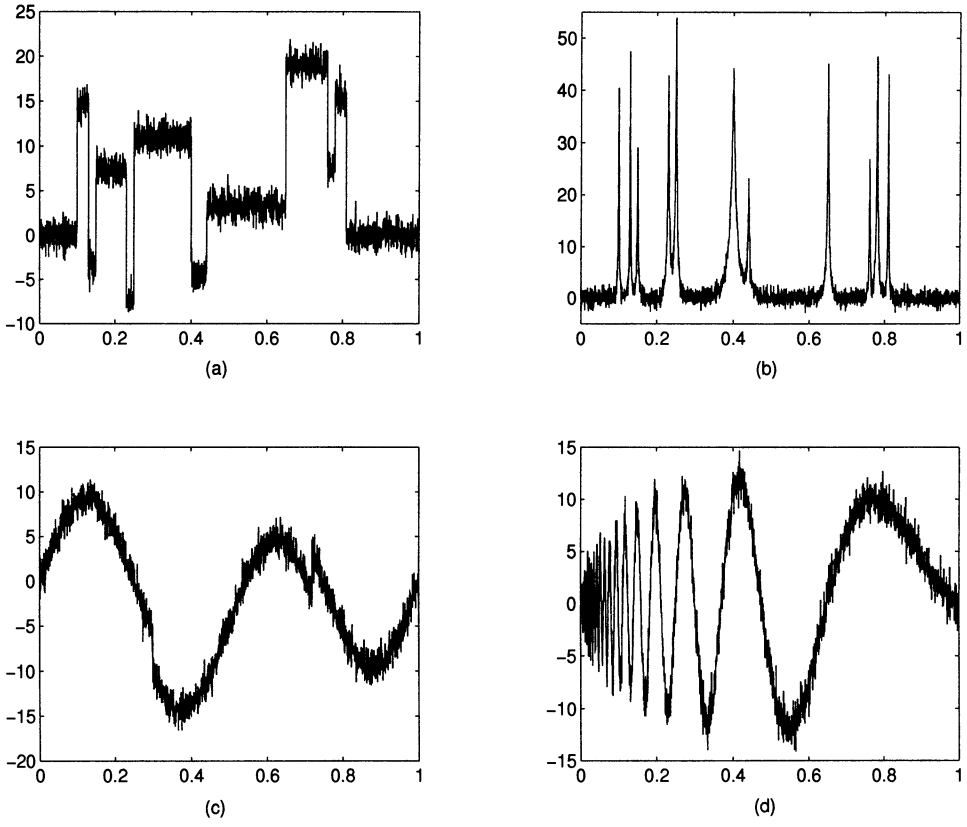

Figure 2. Four Functions With Gaussian White Noise, $\sigma = 1$, Rescaled to Have Signal-to-Noise Ratio SD(f)/$\sigma = 7$. (a) Noisy Blocks; (b) noisy Bumps; (c) noisy HeaviSine; (d) noisy Doppler.

impedance of a layered medium in geophysics and also of a 1-d profile along certain images arising in image-processing problems. We regard *Bumps* as a caricature of spectra arising, for example, in NMR, infrared, and absorption spectroscopy.

Figure 2 displays noisy versions of the same functions. The noise is independent $N(0, 1)$. Figure 3 displays the outcome of applying *SureShrink* (as described in Definition 1) in this case. The results are qualitatively appealing; the reconstructions jump where the true object jumps and are smooth where the true object is smooth. We emphasize that the same computer program, with the same parameters, produced all four reconstructions; no user intervention was permitted or required. *SureShrink* is automatically smoothness adaptive.

Section 3 gives a theoretical result showing that this smoothness adaptation is near optimal. *SureShrink* is asymptotically near minimax over large intervals of the Besov, Sobolev, and Triebel scales. Its speed of convergence is always the optimum one for the best smoothness condition obeyed by the true function, as long as the optimal rate is less than some "speed limit" set by the regularity of the wavelet basis. (By using increasingly higher-order wavelets, that is, wavelets with more vanishing moments and more smoothness, the "speed limit" may be expanded arbitrarily. The cost of such an expansion is a computational effort linearly proportional to the smoothness of the wavelet used.)

Linear methods like kernel, spline, and orthogonal series estimates, even with ideal choice of bandwidth, are unable to converge at the minimax speed over the members of the Besov, Sobolev, and Triebel scales involving L^p smoothness measures with $p < 2$. Thus *SureShrink* can achieve advantages over classical methods even at the level of rates. In fact such advantages are plainly visible in concrete problems where the object to be recovered exhibits significant spatial homogeneity. To illustrate this, Figure 4 shows an example of what can be accomplished by a representative adaptive linear method. This method applies the James–Stein shrinker (which may be interpreted as an adaptive linear shrinker; see Sec. 4.1) to the set of wavelet coefficients at each resolution level. It has a number of pleasant theoretical properties; it automatically achieves the minimax rate for linear estimates over large intervals of the Besov, Triebel, Sobolev, and Hölder scales. Nevertheless, Figure 4 shows that this adaptive linear method performs significantly worse than *SureShrink* in cases of significant spatial variability. A small simulation study described in Section 5 shows that for N in the range 10^3–10^4, *SureShrink* achieves the same level of performance with N samples that adaptive linear methods achieve for $2 \cdot N$ or $4 \cdot N$ samples.

To avoid possible confusion, we emphasize that the method *SureShrink* described in this article differs from variants *RiskShrink* and *VisuShrink* discussed by Donoho and Johnstone (1994a) and Donoho, Johnstone, Kerkyacharian, and Picard (1995) only in the choice of thresholds. Through use of a data based choice of threshold, *SureShrink* is more explicitly adaptive to unknown smoothness and has

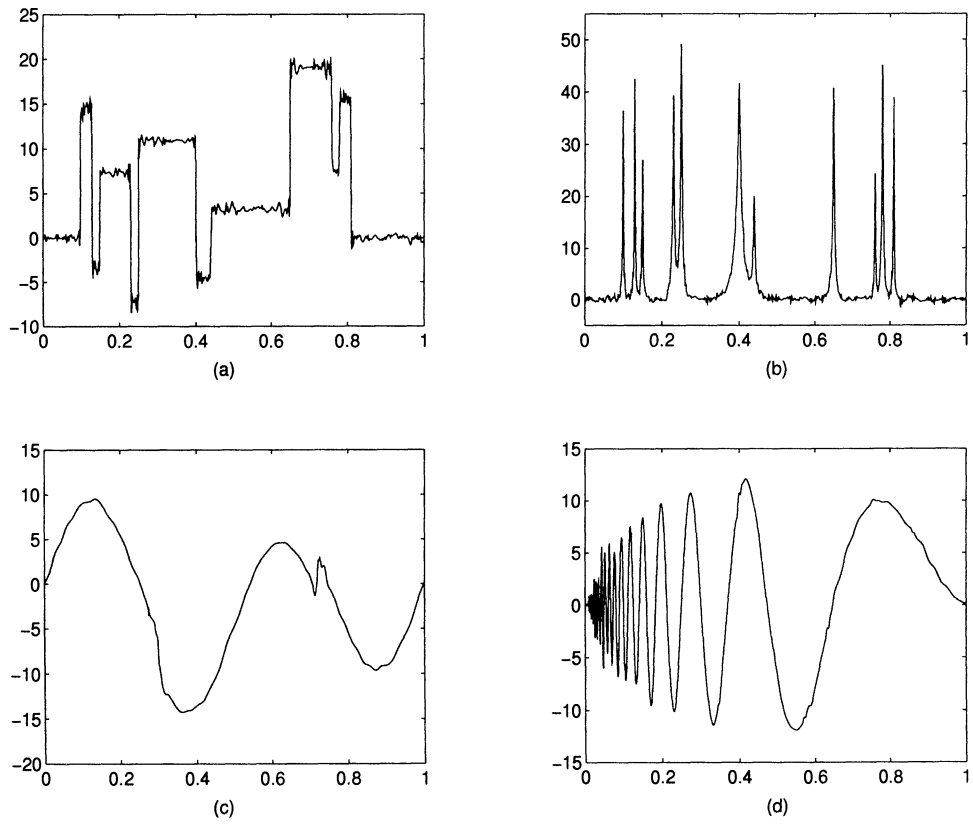

Figure 3. SureShrink Reconstruction Using Soft Thresholding, Most Nearly Symmetric Daubechies Wavelet with N = 8, and Cutoff L = 5. (a) SureShrink [Blocks]; (b) SureShrink [Bumps]; (c) SureShrink [HeaviSine]; (d) SureShrink [Doppler].

better large-sample MSE properties. For further comparative discussion, see Section 5.

2. SURESHRINK

We now describe in detail the ingredients of our procedure.

2.1 Discrete Wavelet Transform

Suppose we have data $y = (y_i)_{i=0}^{N-1}$, with $N = 2^J$. We consider here a family of discrete wavelet transforms, indexed by two integer parameters L and M, and one additional adjective "periodic" or "boundary adjusted." The parameter M represents the number of vanishing moments of the wavelet and L the coarsest resolution level considered; see (7). The construction relies heavily on concepts of Cohen, Daubechies, Jawerth, and Vial (1993), Daubechies (1992), and Meyer (1990, 1991). For a fixed value of M and L, we get a matrix $\mathcal{W}$; this matrix yields a vector $\mathbf{w}$ of the *wavelet coefficients* of $\mathbf{y}$ via

$$\mathbf{w} = \mathcal{W}\mathbf{y}.$$

For simplicity in exposition, we use the periodic version; in this case the transform is exactly orthogonal, so we have the inversion formula $\mathbf{y} = \mathcal{W}^T\mathbf{w}$. Brief comments on the minor changes needed for the boundary corrected version have been given by Donoho and Johnstone (1994a, sec. 4.6).

A crucial detail: The transform is implemented not by matrix multiplication, but by a sequence of special finite-length filtering steps that result in an order $O(N)$ transform. The choice of wavelet transform is essentially a choice of filter. (See Strang 1989 and especially Daubechies 1992 for the full story, and the appendix of Donoho, Johnstone, Kerkyacharian, and Picard 1995 for a brief summary relevant to our implementation.)

The vector $\mathbf{w}$ has $N = 2^J$ elements; it is convenient to index dyadically $N - 1 = 2^J - 1$ of the elements following the scheme

$$w_{j,k}: j = 0, \ldots, J-1; \qquad k = 0, \ldots, 2^j - 1;$$

we label the remaining element $w_{-1,0}$. To interpret these coefficients, let $\mathbf{W}_{j,k}$ denote the (j,k)th row of $\mathcal{W}$. The inversion formula $\mathbf{y} = \mathcal{W}^T\mathbf{w}$ becomes

$$y_i = \sum_{j,k} w_{j,k}\mathbf{W}_{j,k}(i),$$

expressing $\mathbf{y}$ as a sum of basis elements $\mathbf{W}_{j,k}$ with coefficients $w_{j,k}$.

In the special case $L = 0$ and $M = 0$, the transform reduces to the *discrete Haar transform.* Then, if $j \geq 0, \mathbf{W}_{j,k}(i)$ is proportional to 1 for $2^{-j}k \leq i/N < 2^{-j}(k + 1/2)$ and -1 for $2^{-j}(k + 1/2) \leq i/N < 2^{-j}(k + 1)$. $\mathbf{W}_{-1,0}$ is proportional to the constant function 1. Thus the wavelet coefficients measure the differences of the function across various scales, and the function is reconstructed from building blocks of zero-mean localized square waves. Figure 5 shows a schematic of W for the (artificially small) sample size $N = 16$.

In the case $M > 0$, the building blocks of the transform are smoother than square waves. In that case, the vector

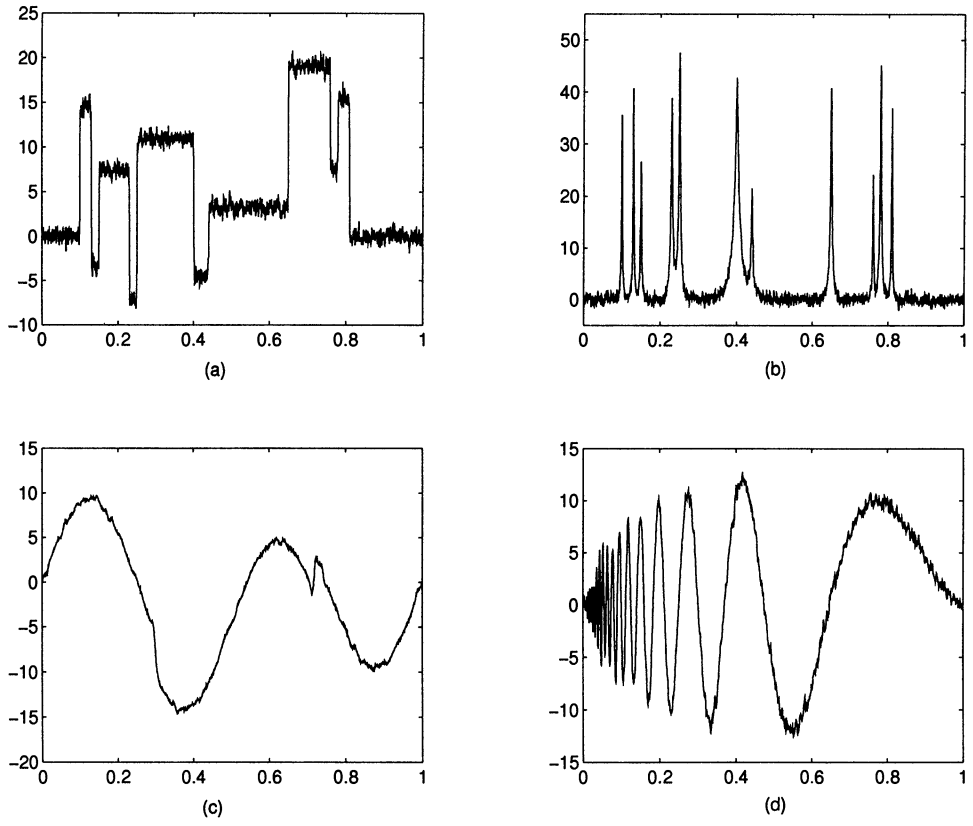

Figure 4. Reconstructions From Noisy Data Using WaveJS, Defined in Section 4.1. L = 5, S8 wavelet. (a) WaveJS [Blocks]; (b) WaveJS [Bumps]; (c) WaveJS [HeaviSine]; (d) WaveJS [Doppler].

$\mathbf{W}_{j,k}$, plotted as a function of i, has a continuous, wiggly, localized appearance that motivates the label "wavelet." For j and k bounded away from extreme cases by the condition

$$L < j \ll J, \qquad 0 \ll k \ll 2^j, \tag{5}$$

we have the approximation

$$\sqrt{N} \cdot \mathbf{W}_{j,k}(i) \approx 2^{j/2}\psi(2^j t) \qquad t = i/N - k2^{-j}, \tag{6}$$

where ψ is the mother wavelet arising in a wavelet transform on $\mathbb{R}$, as described by Daubechies (1988, 1992). This approximation improves with increasing N. ψ is an oscillating function of compact support. We thus speak of $\mathbf{W}_{j,k}$ as being localized to a spatial interval of size 2^{-j} and to have a frequency near 2^j. The basis element $\mathbf{W}_{j,k}$ has an increasingly smooth visual appearance, the larger the parameter M in the construction of the matrix $\mathcal{W}$. Daubechies (1988, 1992) has shown how the parameter M controls the smoothness (i.e., number of derivatives) of ψ; the smoothness is proportional to M.

The vectors $\mathbf{W}_{j,k}$ outside the range of (5) come in two types. First, there are those at $j < L$. These no longer resemble dilations of a mother wavelet ψ, and may no longer be localized. In fact, they may have support including all of $(0, 1)$. They are, qualitatively, low-frequency terms. Second, there are those terms at $j \geq L$ that have k near the boundaries 0 and 2^j. These cases fail to satisfy (6). If the transform is periodized, this is because $\mathbf{W}_{j,k}$ is actually approximated by dilation of a circularly wrapped version of ψ. If the transform is boundary-adjusted, this is because the boundary element $\mathbf{W}_{j,k}$ is actually approximated by a boundary wavelet as defined by Cohen et al. (1993).

Figure 6 displays $\mathbf{W}_{j,k}$ for $j = 6, k = 32$ (and $N = 2{,}048$), in four cases corresponding to specific wavelet filter sequences. The smoother wavelets have broader support.

The usual displays of wavelet transforms use S. Mallat's idea of multiresolution decomposition (Mallat 1989a,b). This adapts in the present situation as follows. Let $\mathbf{x} = (x_i)_{i=0}^{N-1}$ be the data, let

$$V_L\mathbf{x} = \sum_{j<L} w_{j,k}\mathbf{W}_{j,k} \tag{7}$$

denote the partial reconstruction from "gross structure" terms, and for $j \geq L$, let

$$W_j\mathbf{x} = \sum_{0\leq k<2^j} w_{j,k}\mathbf{W}_{j,k} \tag{8}$$

denote the partial reconstruction from terms at resolution level j, or scale 2^{-j}. Then $\mathbf{x}$ can be recovered from these components via $\mathbf{x} = V_L\mathbf{x} + \sum_{L\leq j<J} W_j\mathbf{x}$, and it is usual to examine the behavior of the components by displaying the graphs of $V_L\mathbf{x}$ and of $W_j\mathbf{x}$ for $j = L, L+1, \ldots, J-1$. In Figure 7 we do this for our four functions and the S8 wavelet. In Figure 8 we look just at the *Blocks* and *HeaviSine* functions to contrast the Haar and Daubechies D4 transforms.

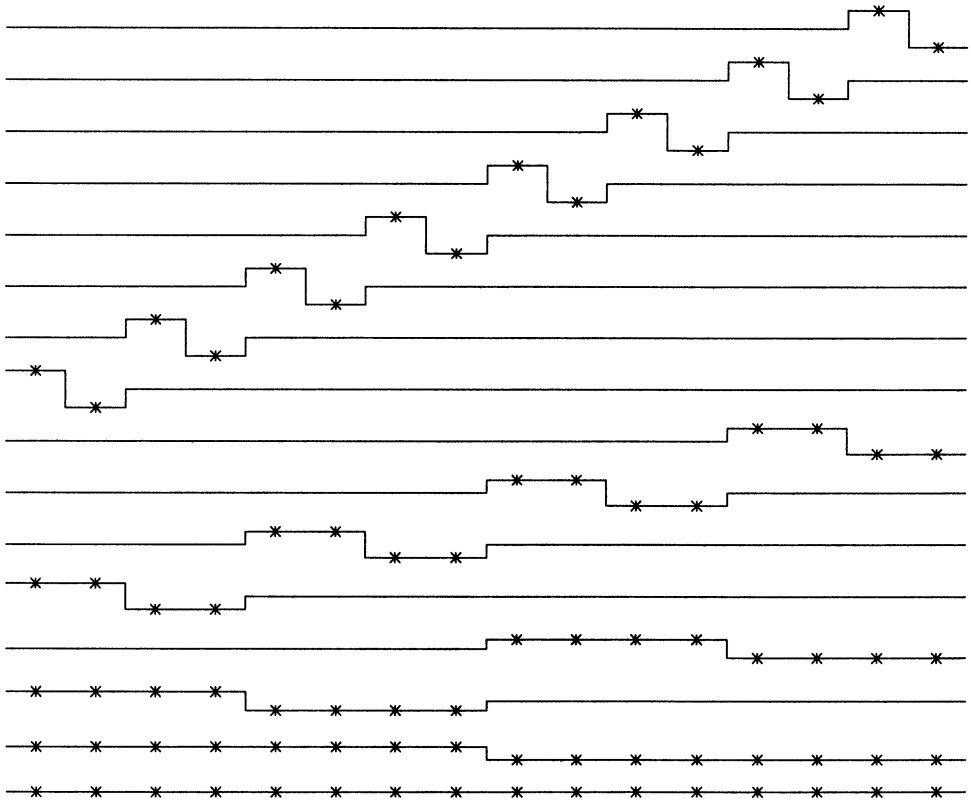

Figure 5. Schematic of Nonzero Entries of the Haar Wavelet Transform Matrix, M = L = 0, N = 16. Lines connect entries in rows $i \to W_{j,k}(i)$, with stars representing nonzero values (magnitudes equal $\sqrt{2^j/N}$.)

A less usual way to display wavelet transforms is to look at the wavelet coefficients directly. We do this in Figure 9. The display at level j depicts $w_{j,k}$ by a vertical line of height proportional to $w_{j,k}$ at horizontal position $k/2^j$. The low-resolution coefficients at $j < L$ are not displayed. The coefficients displayed are those of the S8 wavelet analysis of the four functions under consideration.

Note the considerable sparsity of the wavelet coefficient plots. In all of these plots more than 2,000 coefficients are displayed, but only a small fraction are nonzero at the resolution of the 300 dot-per-inch laser printer. It is also of interest to note the position of the nonzero coefficients, which at high-resolution number j cluster around the discontinuities and spatial nonhomogeneities of the function f. This is an instance of the data-compression properties of the wavelet transform. Indeed, the transform preserves the sum of squares, but in the wavelet coefficients this sum of squares is concentrated in a much smaller fraction of the components than in the raw data.

For comparison, we display in Figure 10 the Haar coefficients of the object; the compression is very pronounced for object *Blocks,* and in fact better than in the S8 case, but the compression is less effective for object *HeaviSine*—much less so than for the S8-based transform.

2.2 Thresholding of Noisy Wavelet Coefficients

The orthogonality of the (periodized) discrete wavelet transform has a fundamental statistical consequence: $\mathcal{W}$ transforms white noise into white noise. Hence if $(y_{j,k})$ are the wavelet coefficients of $(y_i)_{i=0}^{N-1}$ collected according to model (1) and $w_{j,k}$ are the wavelet coefficient of $(f(t_i))$, then

$$y_{j,k} = w_{j,k} + \sigma z_{j,k}, \tag{9}$$

where $z_{j,k}$ is an iid $N(0,1)$ noise sequence. Hence the wavelet coefficients of a noisy sample are themselves just noisy versions of the noiseless wavelet coefficients.

Moreover, $\mathcal{W}$ transforms estimators in one domain into estimators in the other domain, with isometry of risks. If $\hat{w}_{j,k}$ are estimates of the wavelet coefficients, then there is an estimate $\hat{\mathbf{f}}$ of $\mathbf{f} = (f(t_i))$ in the other domain obtained by

$$\hat{\mathbf{f}} = \mathcal{W}^T \hat{\mathbf{w}},$$

and the losses obey the Parseval relation

$$\|\hat{\mathbf{w}} - \mathbf{w}\|_2 = \|\hat{\mathbf{f}} - \mathbf{f}\|_2.$$

The connection also goes in the other direction: If $\hat{\mathbf{f}}$ is any estimator of $\mathbf{f}$, then $\hat{\mathbf{w}} = \mathcal{W}\hat{\mathbf{f}}$ defines an estimator with isometric risk.

The foregoing data-compression remarks were meant to suggest that most of the coefficients in a noiseless wavelet transform are effectively zero. Accepting this slogan, one reformulates the problem of recovering f as one of recovering those few coefficients of f that are significantly nonzero against a Gaussian white noise background.

This motivates the use of a thresholding scheme that "kills" small $y_{j,k}$ and "keeps" large $y_{j,k}$. The particular soft

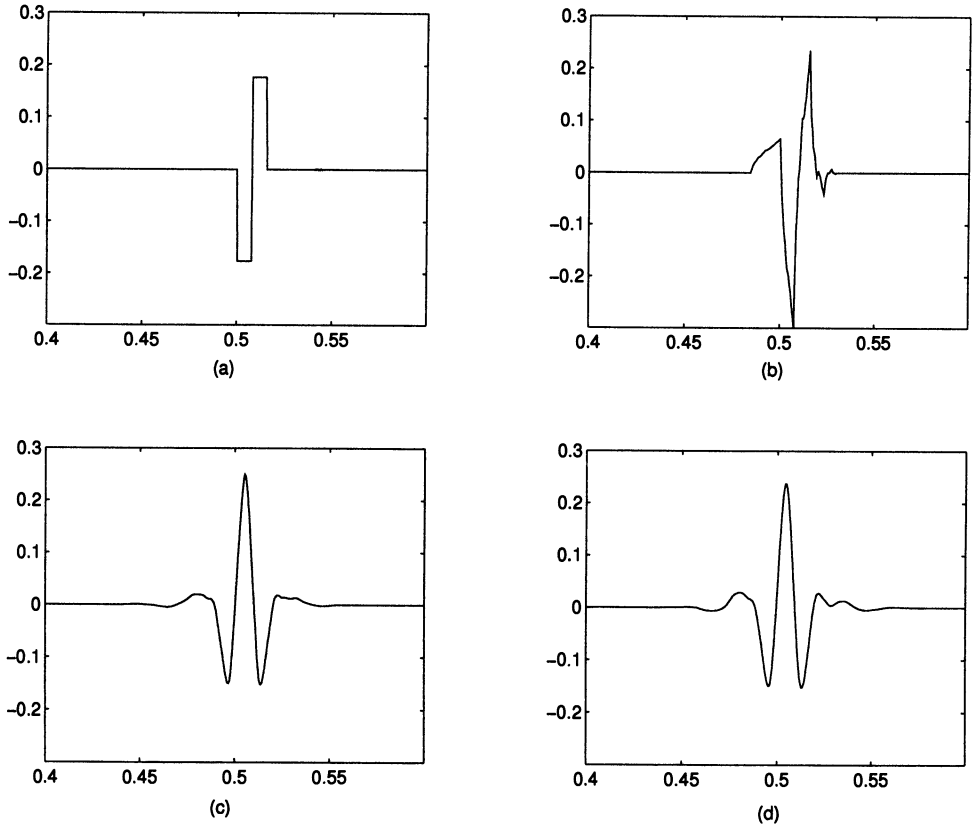

Figure 6. "Typical" Rows of the Wavelet Transform Matrix $\mathcal{W}$ Corresponding to j = 6, k = 32 in Four Cases. (a) Haar wavelet L = 0, M = 0; (b) Daubechies D4 wavelet L = 2, M = 2; (c) Coiflet C3 M = 9; (d) Daubechies "nearly linear phase" S8 wavelet M = 9.

thresholding scheme (4) that we introduced earlier is an instance of this.

Figure 3 has already shown the results such a scheme can provide in the domain of the original data (using S8). Figure 11 illustrates how this works in the wavelet domain using the Haar transform of a noisy version of *Blocks*.

The reconstruction obtained here is via the device of selecting from the noisy wavelet coefficients at level j a threshold t_j^* and applying this threshold to all the empirical wavelet coefficients at level j; the reconstruction is then $\hat{\mathbf{f}} = \mathcal{W}^T\hat{\mathbf{w}}$. Obviously, the choice of threshold t_j^* is crucial.

2.3 Threshold Selection by SURE

Let $\boldsymbol{\mu} = (\mu_i : i = 1, \ldots, d)$ be a d-dimensional vector, and let $x_i \sim N(\mu_i, 1)$ be multivariate normal observations with that mean vector. Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\mathbf{x})$ be a particular fixed estimator of $\boldsymbol{\mu}$. Stein (1981) introduced a method for estimating the loss $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ in an unbiased fashion. Stein showed that for a nearly arbitrary, nonlinear biased estimator, one can nevertheless estimate its loss unbiasedly.

Write $\hat{\boldsymbol{\mu}}(\mathbf{x}) = \mathbf{x} + \mathbf{g}(\mathbf{x})$, where $\mathbf{g} = (g_i)_{i=1}^d$ is a function from R^d into R^d. Stein (1981) showed that when $\mathbf{g}(\mathbf{x})$ is weakly differentiable, then

$$E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}(\mathbf{x}) - \boldsymbol{\mu}\|^2 = d + E_{\boldsymbol{\mu}}\{\|\mathbf{g}(\mathbf{x})\|^2 + 2\nabla \cdot \mathbf{g}(\mathbf{x})\}, \quad (10)$$

where

$$\nabla \cdot \mathbf{g} \equiv \sum_i \frac{\partial}{\partial x_i} g_i.$$

Now consider the soft threshold estimator $\hat{\mu}_i^{(t)}(\mathbf{x}) = \eta_t(x_i)$ and apply Stein's result. $\hat{\boldsymbol{\mu}}^{(t)}$ is weakly differentiable in Stein's sense, and so we get from (10) that the quantity

$$\mathrm{SURE}(t; \mathbf{x}) = d - 2 \cdot \#\{i : |x_i| \le t\} + \sum_{i=1}^{d} (|x_i| \wedge t)^2 \quad (11)$$

is an unbiased estimate of risk: $E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{(t)}(\mathbf{x}) - \boldsymbol{\mu}\|^2 = E_{\boldsymbol{\mu}}\mathrm{SURE}(t; \mathbf{x})$. Here $a \wedge b = \min(a, b)$.

Consider using this estimator of risk to *select* a threshold,

$$t^S = \mathrm{argmin}_{0 \le t \le \sqrt{2\log d}} \mathrm{SURE}(t; \mathbf{x}). \quad (12)$$

Arguing heuristically, one expects that for large dimension d, a sort of statistical regularity will set in, the law of large numbers will ensure that SURE is close to the true risk, and that t^S will be almost the optimal threshold for the case at hand. Theory developed later will show that this hope is justified (and explains the choice of upper bound at $\sqrt{2\log d}$.)

Computational evidence that t^S is a reasonable threshold selector is given in Figure 12. A vector $\boldsymbol{\mu}$ of dimension $d = 128$ consists of 16 consecutive 4's, followed by all zeros. White Gaussian noise of variance 1 was added (Fig. 12a). The profile of SURE(t) is displayed in Fig. 12c; it quite closely resembles the actual loss (Fig. 12d), which of course we know in this (artificial) example. The SURE principle was used to select a threshold that is applied to the data, resulting in an estimate of the mean vector (Fig. 12b). This

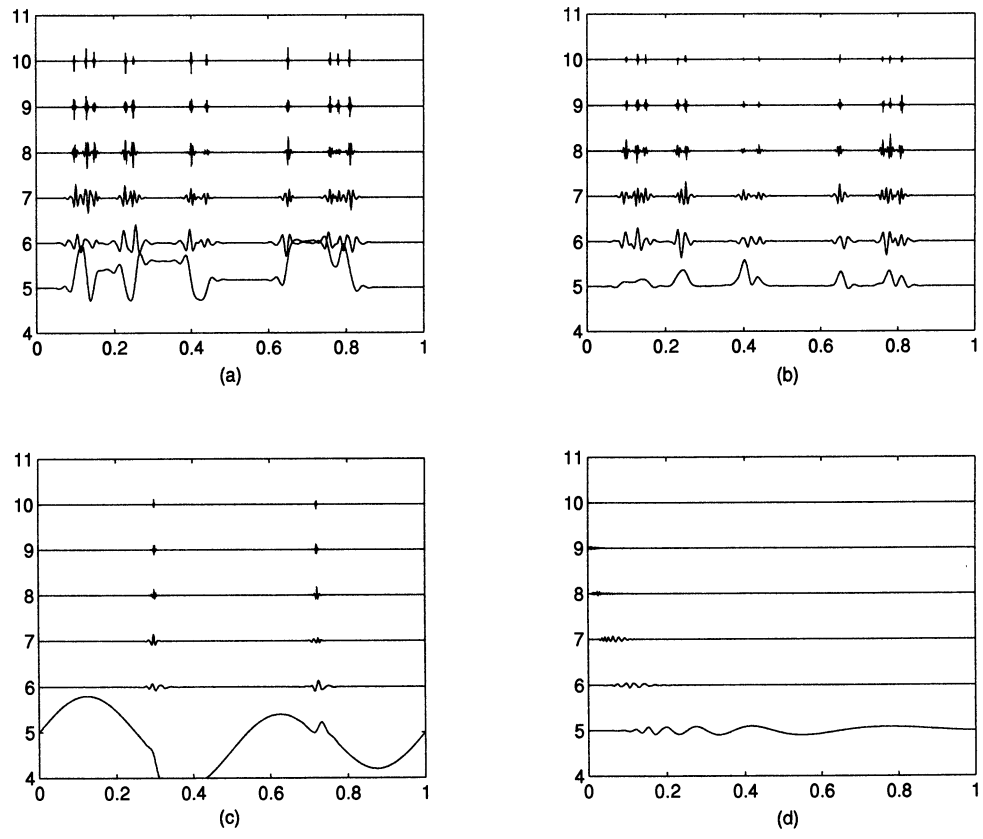

Figure 7. Mallat's Multiresolution Decomposition of the Four Basic Functions—(a) Blocks, (b) Bumps, (c) HeaviSine, (d) Doppler—S8 Wavelet. The line with baseline height j ($5 \le j \le 10$) plots the partial reconstruction $(W_j x)_i$ (cf. (8)) against i/N, $i = 0, \ldots N-1$, and the bottom line similarly shows the "gross structure" decomposition $V_L x$ (cf. (7); here $L = 4$.)

estimate is sparse and much less noisy than the raw data (Fig. 12a). Note also the shrinkage of the nonzero part of the signal.

The optimization problem (12) is computationally straightforward. Suppose, without any loss of generality, that the x_i have been reordered in order of increasing $|x_i|$. Then on intervals of t that lie between two values of $|x_i|$, SURE(t) is strictly increasing, indeed quadratic. Therefore, the minimum value t^S is one of the data values $|x_i|$. There are only d such values; when they have been already arranged in increasing order, the collection of all values SURE$(|x_i|)$ may be computed in order $O(d)$ additions and multiplications, with appropriate arrangement of the calculations. It may cost as much as order $O(d \log(d))$ calculations to arrange the $|x_i|$ in order; so the whole effort to calculate t^S is order $O(d \log(d))$. This is scarcely worse than the order $O(d)$ calculations required simply to apply either form of thresholding.

2.4 Threshold Selection in Sparse Cases

The SURE principle just described has a serious drawback in situations of extreme sparsity of the wavelet coefficients. In such cases the noise contributed to the SURE profile by the many coordinates at which the signal is zero swamps the information contributed to the SURE profile by the few coordinates where the signal is nonzero. Consequently, *SureShrink* uses a Hybrid scheme.

Figure 13a depicts the results of a small-scale simulation study. A vector μ of dimension $d = 1{,}024$ contained $\lfloor \varepsilon \cdot d \rfloor$ nonzero elements, all of size C. Independent $N(0,1)$ noise was added. The SURE estimator t^S was applied. Amplitudes $C = 3, 5$, and 7 were tried, and sparsities $\varepsilon = \{.005, .01, .02(.02).20, .25\}$ were studied. 25 replications were tried at each parameter combination, and the root MSE's are displayed in Figure 13. Evidently, the root MSE's do not tend to zero linearly as the sparsity tends to zero. For the theoretical results of Section 3, such behavior would be unacceptable.

In contrast, Figure 13b portrays the results of the same experiment, with a "fixed thresholding" estimator $\hat{\mu}^F$, where the threshold is set to $t_d^F = \sqrt{2 \log(d)}$ independent of the data. The losses tend to be larger than SURE for "dense" situations $\varepsilon \gg 0$, but much smaller for ε near zero. We have developed the rationale for the choice $\sqrt{2 \log(d)}$ in earlier work (Donoho and Johnstone 1994a). To summarize, first, the maximum of N iid standard Gaussian variates is smaller than $\sqrt{2 \log N}$, with probability increasing to 1 as N increases. Thus with high probability, a pure noise signal is correctly estimated as being identically zero. Second, the threshold t_d^F is also asymptotically optimal in a MSE sense for mimicking the MSE of an "oracle" that knows which coordinates of the mean vector are larger than the standard deviation of the noise.

Figure 13c displays the results of applying a hybrid method that we label $\hat{\mu}^*$, which is designed to behave like $\hat{\mu}^S$ in dense situations and like $\hat{\mu}^F$ in sparse ones. Its performance is roughly as desired.

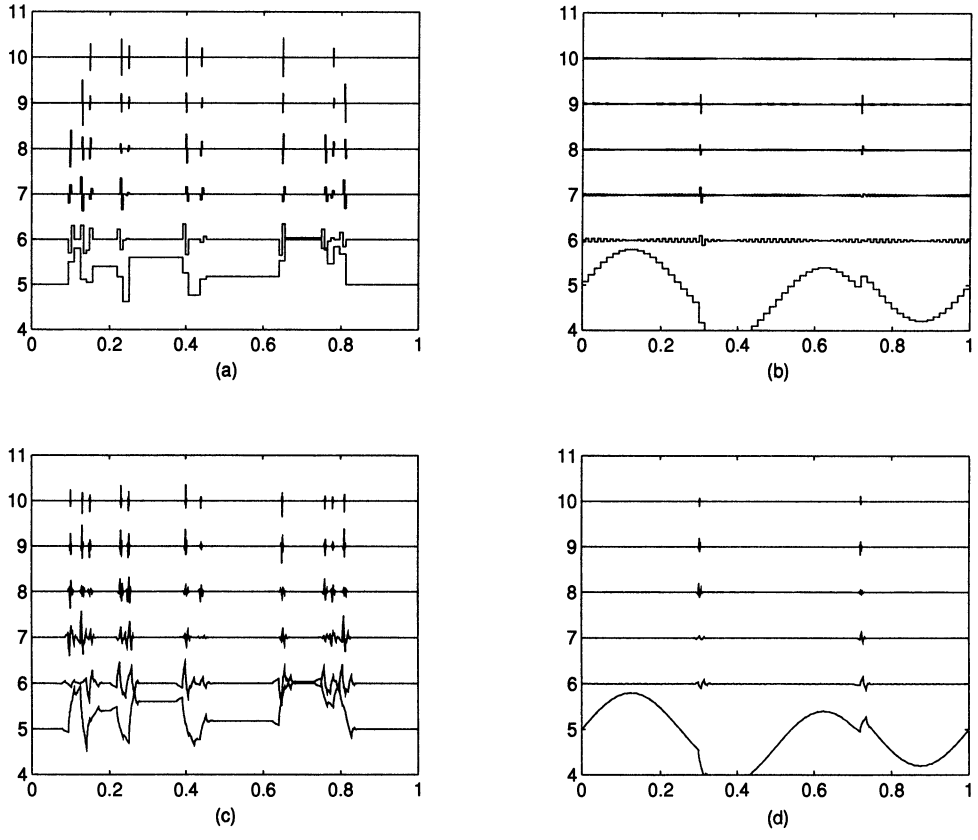

Figure 8. Mallat's Multiresolution Decomposition of Blocks and HeaviSine Using the Haar (a and b) and D4 (c and d) wavelets. Here L = 4.

In detail, the Hybrid method works as follows: Define $s_d^2 = d^{-1}\sum_i (x_i^2 - 1)$ and let γ_d be a critical value, which for the present we take as $\log_2^{3/2} d/\sqrt{d}$. Let I denote a random subset of half the indices in $\{1, \ldots, d\}$ and let I' denote its complement. Let t_I^S and $t_{I'}^S$ denote the minimizers of SURE for the given subsets of indices, only with an additional restriction on the search range,

$$t_I^S = \text{argmin}_{0 \le t \le t_d^F} \text{SURE}(t, (x_i)_{i \in I}),$$

and similarly for $t_{I'}^S$. Define the estimate

$$\begin{aligned} \hat{\boldsymbol{\mu}}^*(\mathbf{x})_i &= \eta_{t_d^F}(x_i) \qquad s_d^2 \le \gamma_d, \\ &= \eta_{t_I^S}(x_i) \qquad i \in I' \quad \text{and} \quad s_d^2 > \gamma_d, \\ &= \eta_{t_{I'}^S}(x_i) \qquad i \in I \quad \text{and} \quad s_d^2 > \gamma_d. \end{aligned} \tag{13}$$

In other words, we use one half-sample to estimate the threshold for use with the other half-sample; but unless there is convincing evidence that the signal is nonnegligible, we set the threshold to $\sqrt{2\log(d)}$.

This half-sample scheme was developed for the proof of Theorems 3 and 4 in Sections 3.2 and 3.3. In practice, the half-sample aspect of the estimate seems unnecessary; the simpler estimator $\hat{\mu}^+$ derived from

$$\begin{aligned} \hat{\mu}^+(\mathbf{x})_i &= \eta_{t_d^F}(x_i) \qquad s_d^2 \le \gamma_d \\ &= \eta_{t^S}(x_i) \qquad s_d^2 > \gamma_d, \end{aligned} \tag{14}$$

offers the same performance benefits in simulations and in fact is used in all the examples in this article; see Figure 13d.

Note Added in Proof. We have since developed a proof of Theorems 3 and 4 that applies to the more natural estimator $\hat{\mu}^+$, making the introduction of random half-sampling unnecessary. It is hoped to include details in the written version of our March 1995 lectures in Oberwolfach.

We now apply this multivariate normal theory in our wavelet setting.

Definition 1. The term *SureShrink* refers to the following estimator $\hat{\mathbf{f}}^*$ of $\mathbf{f}$. Assuming that $N = 2^J$ and that the noise is normalized so that it has standard deviation $\sigma = 1$, we set $\mathbf{x}_j = (y_{j,k})_{0 \le k < 2^j}$ and

$$\begin{aligned} \hat{w}_{j,k}^* &= y_{j,k}, \qquad j < L, \\ &= (\boldsymbol{\mu}^*(\mathbf{x}_j))_k, \qquad L \le j < L; \end{aligned}$$

the estimator $\hat{\mathbf{f}}^*$ derives from this via inverse discrete wavelet transform. We use $\hat{\mathbf{f}}^+$ to denote the variant using (14) used in practice.

Note that $\hat{\mathbf{f}}^*$ is fully automatic, modulo the choice of specific wavelet transform. Moreover, with appropriate arrangement of the work, the whole computational effort involved is order $O(N \log(N))$, scarcely worse than linear in the sample size N. Extensive experience with computations on a Macintosh show that performance is quite reasonable even on personal computers. The Matlab command *SureShrink* takes a few seconds to complete on an array of size $N = 4{,}096$.

3. MAIN ADAPTIVITY RESULT

In this section we investigate the adaptivity of *SureShrink*

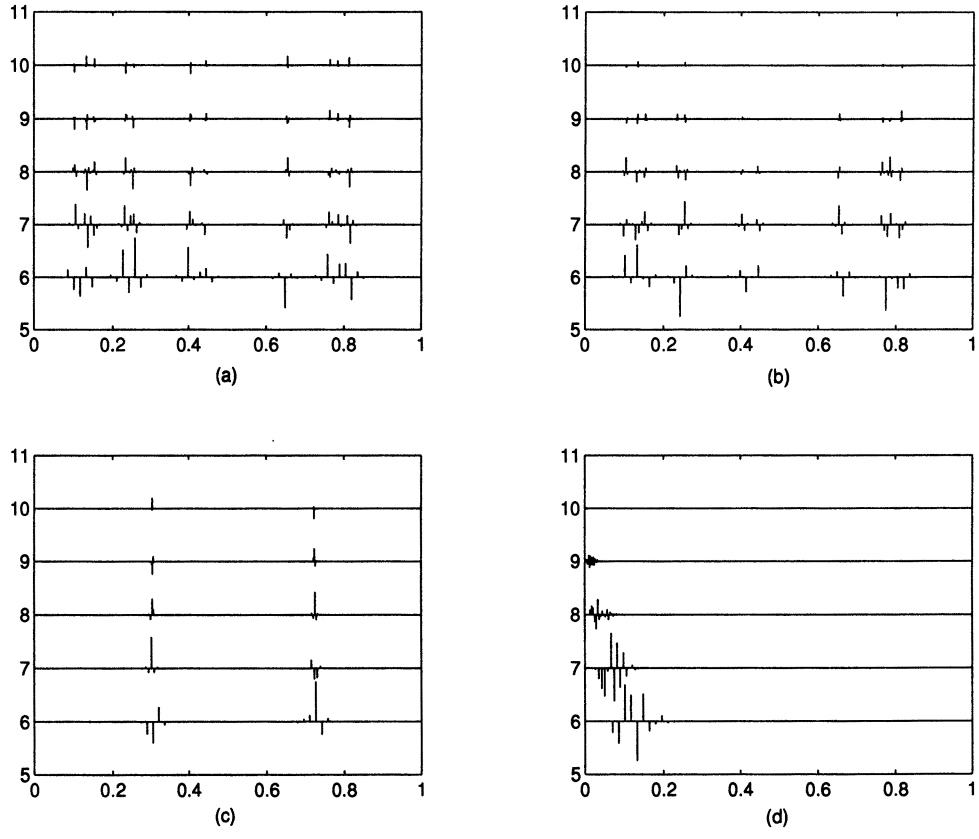

Figure 9. Plot of Wavelet Coefficients Using S8. (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler. The display at level j depicts w_{jk} by a vertical line of height proportional to $w_{j,k}$ at horizontal position $k2^{-j}$.

to unknown degree of smoothness. To state our result, we must define Besov spaces. We follow DeVore and Popov (1988). Let $\Delta_h^{(r)} f$ denote the rth difference $\sum_{k=0}^{r} \binom{r}{k} (-1)^k f(t + kh)$. The rth modulus of smoothness of f in $L^p[0,1]$ is

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

The *Besov* seminorm of index (σ, p, q) is derived for $r > \sigma$ by

$$|f|_{B^\sigma_{p,q}} = \left(\int_0^1 \left(\frac{w_{r,p}(f;h)}{h^\sigma} \right)^q \frac{dh}{h} \right)^{1/q}$$

if $q < \infty$, and by

$$|f|_{B^\sigma_{p,\infty}} = \sup_{0<h<1} \frac{w_{r,p}(f;h)}{h^\sigma}$$

if $q = \infty$. The *Besov ball* $B^\sigma_{p,q}(C)$ (resp. *space* $B^\sigma_{p,q}$) is then the class of functions $f : [0,1] \to \mathbb{R}$ satisfying $f \in L^p[0,1]$ and $|f|_{B^\sigma_{p,q}} \le C$ (resp. $|f|_{B^\sigma_{p,q}} < \infty$). Standard references on Besov spaces are works of Peetre (1975) and Triebel (1983, 1990).

This measure of smoothness includes, for various settings (σ, p, q), other commonly used measures. For example, let C^δ denote the *Hölder class* of functions with $|f(s) - f(t)| \le c|s - t|^\delta$ for some $c > 0$. Then f has for a given $m = 0, 1, \ldots$ a distributional derivative $f^{(m)}$ satisfying $f^{(m)} \in C^\delta$, $0 < \delta < 1$, if and only if $|f|_{B^{m+\delta}_{\infty,\infty}} < \infty$. Similarly, with W_2^m the L^2 Sobolev space as in Section 1, $f \in W_2^m$ iff $|f|_{B^m_{2,2}} < \infty$.

The Besov scale essentially includes other less traditional spaces as well. For example, recall the definition of total variation of a function,

$$TV(f) = \sup \Big\{ \sum |f(t_i) - f(t_{i-1})| : 0 = t_0 < t_1 < \cdots < t_{k-1} \le t_k = 1, k \in \mathcal{N} \Big\}$$

and note that the space of functions of bounded variation is a superset of $B^1_{1,1}$ and a subset of $B^1_{1,\infty}$. Similarly, all the L^p Sobolev spaces W_p^m contain $B^m_{p,1}$ and are contained in $B^m_{p\infty}$.

For the theoretical results, we permit alternate choices of cutoff γ_d for the pretest in *SureShrink;* for example, $\gamma_d = d^\gamma, 0 < \gamma < \frac{1}{2}$. We recall that $a_N \asymp b_N$ means $\liminf_N |a_N/b_N| > 0$ and $\limsup_N |a_N/b_N| < \infty$. Let the minimax risk be denoted by

$$R(N; B^\sigma_{p,q}(C)) = \inf_{\hat{\mathbf{f}}} \sup_{B^\sigma_{p,q}(C)} R(\hat{\mathbf{f}}, \mathbf{f}).$$

We note that for the ranges of (σ, p, q) considered later, $R(N; B^\sigma_{p,q}(C)) \asymp N^{-r}$, with $r = \sigma/(\sigma + \frac{1}{2})$.

Theorem 1. Let the discrete wavelet analysis correspond to a wavelet ψ having r null moments and r continuous derivatives, $r > \max(1, \sigma)$. Then, *SureShrink* is si-

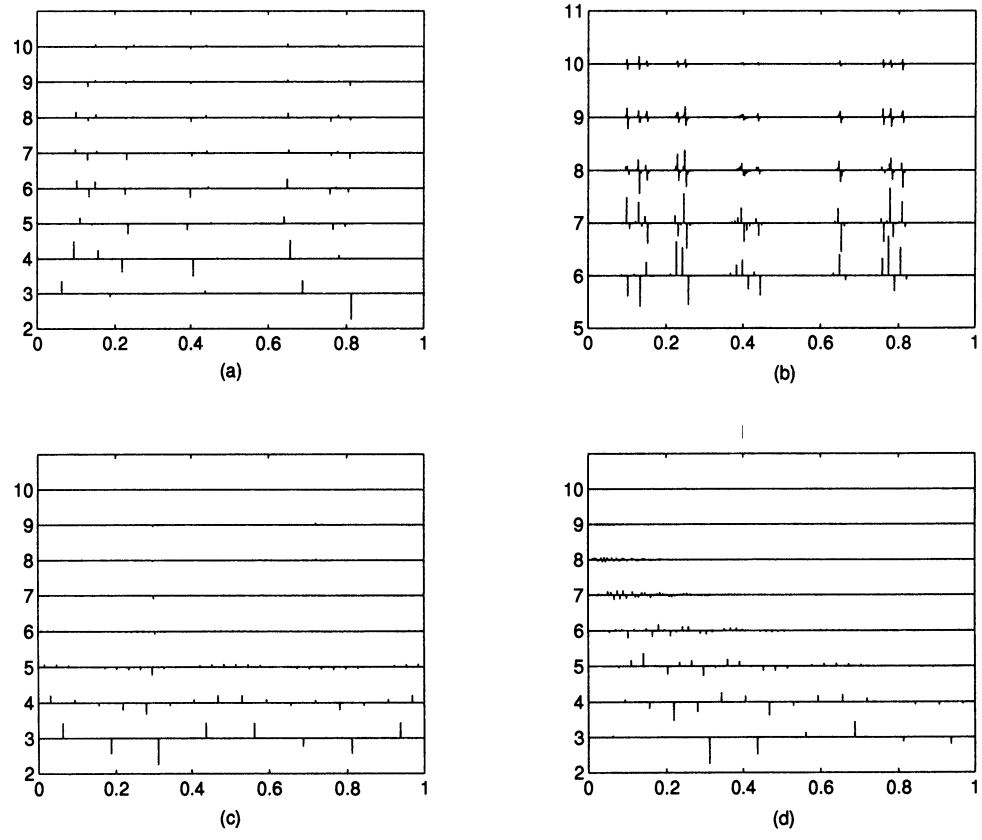

Figure 10. Wavelet Coefficients Using the Haar Wavelet. (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler. Compare amounts of compression with Figure 9.

multaneously nearly minimax,

$$\sup_{B^{\sigma}_{p,q}(C)} R(\hat{\mathbf{f}}^*, \mathbf{f}) \asymp R(N; B^{\sigma}_{p,q}(C)) \qquad N \to \infty,$$

for all $p, q \in [1, \infty]$, for all $C \in (0, \infty)$, and for all $\sigma_0 < \sigma < r$. In particular,

$$\gamma_d = \log^{3/2} d/\sqrt{d}, \Rightarrow$$
$$\sigma_0 = \max\left(\frac{1}{p}, 2\left(\frac{1}{p} - \frac{1}{2}\right)_+\right)$$

and

$$\gamma_d = d^{\gamma}, \qquad 0 < \gamma < \frac{1}{2}, \Rightarrow$$
$$\sigma_0 = \max\left(\frac{1}{p}, 2\left(\frac{1}{p} - \frac{1}{2}\right)_+ + \gamma - \frac{1}{2}\right).$$

In words, this estimator, which "knows nothing" about the a priori degree, type, or amount of regularity of the object, nevertheless achieves the optimal rate of convergence that one could attain by knowing such regularity. Over a Hölder class, it attains the optimal rate; over an L^2 Sobolev class, it achieves the optimal rate; and over Sobolev classes with $p < 2$, it also achieves the optimal rate.

We mentioned in Section 1 that no linear estimator achieves the optimal rate over all L^p Sobolev classes; as a result, the modification of *SureShrink* achieves something that usual estimates could not, even if the optimal bandwidth were known a priori.

Many other results along these lines could be proved, for other (σ, p, q). One particularly interesting result, because it refers to the Haar basis, is the following.

Theorem 2. Let $\mathcal{V}(C)$ denote the class of all functions on the unit interval of total variation $\leq C$. Let $\hat{f}^*$ denote the application of *SureShrink* in the Haar basis, with $\gamma_d = d^{\gamma}, 0 < \gamma < \frac{1}{2}$. This "*HaarShrink*" estimator is simultaneously nearly minimax,

$$\sup_{\mathcal{V}(C)} R(\hat{\mathbf{f}}^*, \mathbf{f}) \asymp R(N; \mathcal{V}(C)) \qquad N \to \infty$$

for all $C \in (0, \infty)$.

Again, without knowing any a priori limit on the total variation, the estimator behaves essentially as well as one could by knowing this limit. Figure 11 shows the plausibility of this result.

3.1 Estimation in Sequence Space

Our proof of Theorem 1 uses a method of sequence spaces described in earlier work (Donoho and Johnstone 1995). The key idea is to approximate the problem of estimating a function from finite noisy data by the problem of estimating an infinite sequence of wavelet coefficients contaminated with white noise.

The heuristic for this replacement is as follows. From (6) and (9), the empirical wavelet coefficient is $y_{j,k} = w_{j,k}$

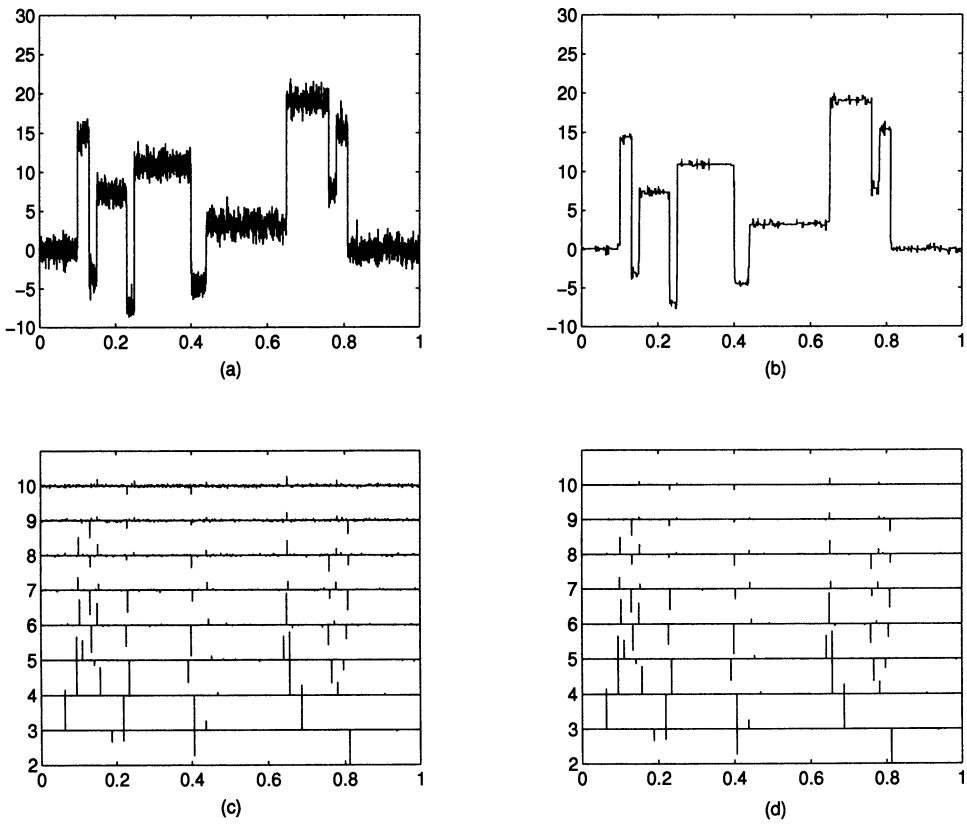

Figure 11. Component Steps of SureShrink Reconstruction. (a) Raw data: a noisy version of Blocks; (b) SureShrink reconstruction using Haar wavelet, and L = 2; (c) raw wavelet coefficients of the data; (d) the same coefficients after thresholding.

$+ \sigma z_{j,k}$, where the discrete $w_{j,k}$ obeys

$$w_{j,k} \approx \sqrt{N} \int f(t)\psi_{j,k}(t)\, dt$$

for a certain wavelet $\psi_{j,k}(t)$. In terms of the continuous wavelet coefficients $\theta_{j,k} = \int f(t)\psi_{j,k}(t)\, dt$, it is thus tempting to act as though our observations were actually

$$\sqrt{N} \cdot \theta_{j,k} + \sigma z_{j,k}$$

or, equivalently,

$$\theta_{j,k} + \varepsilon z_{j,k},$$

where $\varepsilon = \sigma/\sqrt{N}$ and $z_{j,k}$ is still a standard iid, $N(0,1)$ sequence. Moreover, due to the Parseval relation, $\|\hat{\mathbf{f}} - \mathbf{f}\|_2 = \|\hat{\mathbf{w}} - \mathbf{w}\|_2$, and the foregoing approximation, we are also tempted to act as if the loss $N^{-1}\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2$ were the same as $\|\hat{\theta} - \theta\|_2^2$.

These (admittedly vague) approximation heuristics lead to the study of the following sequence space problem. We observe an infinite sequence of data,

$$y_{j,k} = \theta_{j,k} + z_{j,k} \qquad j \geq 0, \qquad k = 0, \ldots, 2^j - 1, \quad (15)$$

where $z_{j,k}$ are iid. $N(0, \varepsilon^2)$ and $\theta = (\theta_{j,k})$ is unknown. We wish to estimate θ with small squared error loss $\|\hat{\theta} - \theta\|_2^2 = \sum(\hat{\theta}_{j,k} - \theta_{j,k})^2$. We let $\Theta(\sigma, p, q, C)$ denote the set of all wavelet coefficient sequences $\theta = (\theta_{j,k})$ arising from an $f \in B^{\sigma}_{p,q}(C)$. Finally, we search for a method $\hat{\theta}$ that is simultaneously nearly minimax over a range of $\Theta(\sigma, p, q, C)$.

Suppose that we can solve this sequence problem. Under certain conditions on σ, p, and q, this will imply Theorem 1. Specifically, if σ_0 is big enough and the wavelet is of regularity $r > \sigma_0$, an estimator that is simultaneously near minimax in the sequence space problem $\sigma_0 < \sigma < r$ may be applied to the empirical wavelet coefficients in the original problem under study and will also be simultaneously near minimax in the original function space problem. We have already discussed the approximation arguments necessary to establish this correspondence (Donoho 1992; Donoho and Johnstone 1994c), and for reasons of space we omit them. (See also Brown and Low 1990.)

3.2 Adaptive Estimation over Besov Bodies

The collections $\Theta(\sigma, p, q, C)$ of wavelet expansions $\theta = \theta(f)$ arising from functions $f \in B^{\sigma}_{p,q}(C)$ are related to certain simpler sets that we have called (Donoho and Johnstone 1994c) *Besov bodies*. These are sets $\|\theta\|_{\mathbf{b}^s_{p,q}} \leq C$, where

$$\|\theta\|^q_{\mathbf{b}^q_{p,q}} = \sum_{j \geq 0} \left(2^{js} \left(\sum_{0 \leq k < 2^j} |\theta_{j,k}|^p \right)^{1/p} \right)^q \quad (16)$$

and $s = \sigma + \frac{1}{2} - 1/p$.

Consider the problem of estimating θ when it is observed in a Gaussian white noise and is known a priori to lie in a

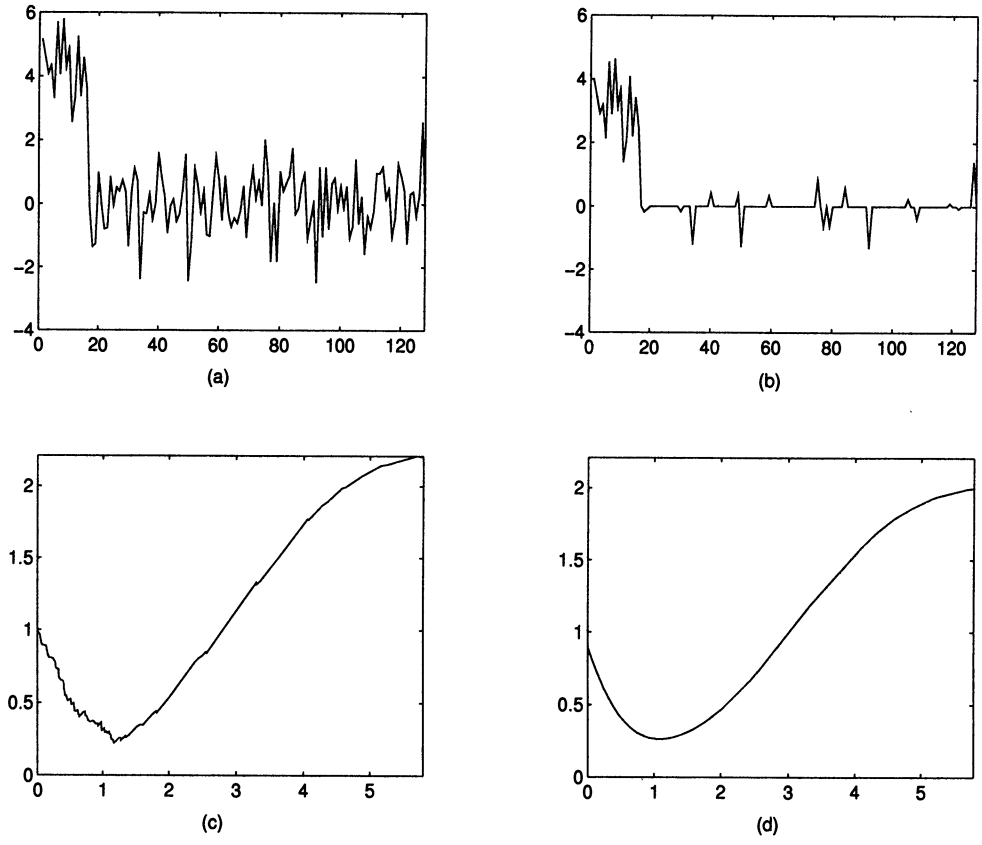

Figure 12. Illustration of Choice of Threshold Using SURE(t). See Section 2.3. (a) raw data; (b) estimate; (c) estimated risk versus lambda; (d) loss versus lambda.

certain convex set $\Theta^s_{p,q}(C) \equiv \{\theta : \|\theta\|_{\mathbf{b}^s_{p,q}} \le C\}$. We often put for short $\Theta^s_{p,q} = \Theta^s_{p,q}(C)$. The difficulty of estimation in this setting is measured by the *minimax risk*

$$R^*(\varepsilon; \Theta^s_{p,q}) = \inf_{\hat\theta} \sup_{\Theta^s_{p,q}} E\|\hat\theta - \theta\|_2^2, \tag{17}$$

and the minimax risk among threshold estimates is

$$R^*_T(\varepsilon; \Theta^s_{p,q}) = \inf_{(t_j)} \sup_{\Theta^s_{p,q}} E\|\hat\theta_{(t_j)} - \theta\|_2^2, \tag{18}$$

where $\hat\theta_{(t_j)}$ stands for the estimator $(\eta_t, (y_{j,k}))_{j,k}$. We have shown (Donoho and Johnstone 1995) that $R^*_T \le \Lambda(p) \cdot R^* \cdot (1 + o(1))$ with, for example, $\Lambda(1) \approx 1.6$. Hence threshold estimators are nearly minimax. Furthermore, they show that the minimax risk and minimax threshold risk over sets $\Theta^s_{p,q}(C)$ are equivalent, to within constants, to that over sets $\Theta(\sigma, p, q, C)$, provided that σ is large enough and we make the calibration $s = \sigma + \frac{1}{2} - 1/p$.

We may construct a *SureShrink*-style estimator in this problem by applying $\boldsymbol{\mu}^*$ level by level. Let $\mathbf{x}_j = (y_{j,k}/\varepsilon)_{k=0}^{2^j-1}$. Then set

$$\hat\theta^*_{j,k}(\mathbf{y}) = y_{j,k}, \qquad j < L \tag{19}$$

and

$$\hat\theta^*_{j,k}(\mathbf{y}) = \varepsilon \cdot \hat\mu^*(\mathbf{x}_j)_k \qquad j \ge L. \tag{20}$$

This is a particular adaptive threshold estimator.

Theorem 3. If either (a) $\gamma_d = d^{-1/2}\log^{3/2} d$ and $s > |\frac{1}{p} - \frac{1}{2}|$, or (b) $\gamma_d = d^{-\gamma}, 0 < \gamma < \frac{1}{2}$, and $s > |\frac{1}{p} - \frac{1}{2}| + \gamma - \frac{1}{2}$, then

$$\sup_{\Theta^*_{p,q}(C)} E_\theta\|\hat\theta^* - \theta\|_2^2 \le R^*_T(\varepsilon; \Theta^s_{p,q}(C))(1 + o(1))$$

as $\varepsilon \to 0$.

In short, without knowing s, p, q, or C, one obtains results as good asymptotically as if one did know those parameters. The result is effective across an infinite range of all the parameters in question. Because the minimax risk is close to the minimax threshold risk, this solves the problem of adapting across a scale of Besov bodies.

This theorem, together with the approximation arguments alluded to in Section 3.1, proves Theorems 1 and 2.

3.3 Adaptive Estimation at a Single Resolution Level

Theorem 3 depends on an analysis of adaptive threshold selection by the SURE principle. Return to the setup of Section 2.3.

Let $\tilde R(\boldsymbol{\mu})$ denote the ideal threshold risk, which we could achieve with information about the optimal threshold to use,

$$\tilde R(\boldsymbol{\mu}) = \inf_t d^{-1} \cdot \sum_i r(t, \mu_i),$$

where $r(t, \mu)$ is the risk $E(\eta_t(x) - \mu)^2$ in the scalar setting $x = \mu + z, z \sim N(0, 1)$. Of course, we can never hope to actually know the ideal threshold t attaining this expression; however, the following result says that the adaptive

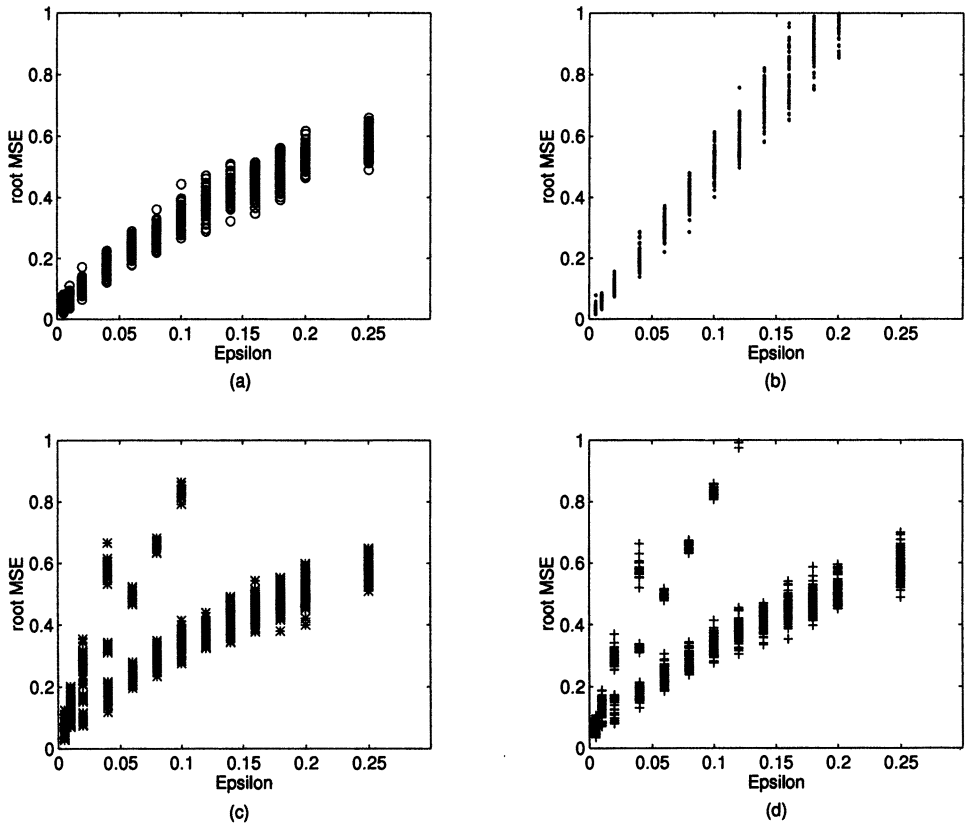

Figure 13. Root MSE's for Simulated Data at Varying Levels of Sparsity When Threshold is Chosen by (a) SURE, (b) $\sqrt{2 \log d}$, and (c) and (d) Two Variants of the Hybrid Method Described, Along With the Design, in Section 2.4.

estimator $\hat{\mu}^*$ performs almost as if we did know this ideal threshold. Let $\tau^2 = d^{-1}\sum \mu_i^2$; when τ^2 is small, the hybrid method switches to the fixed threshold t_d^F and leads to terms involving the fixed threshold risk

$$R_F(\boldsymbol{\mu}) = d^{-1}\sum_i r(t_d^F, \mu_i).$$

Theorem 4. Suppose that $\gamma_d \le 1$ and $\gamma_d^2 d/\log d \to \infty$. Then (a) uniformly in $\boldsymbol{\mu} \in \mathbb{R}^d$, we have

$$d^{-1}E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\|_2^2 \le \tilde{R}(\boldsymbol{\mu}) + R_F(\boldsymbol{\mu})I\{\tau^2 \le 3\gamma_d\} + c(\log d)^{3/2}d^{-1/2}, \text{ and}$$

(b) uniformly in $\tau^2 \le \frac{1}{3}\gamma_d$, we have

$$d^{-1}E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\|_2^2 \le R_F(\boldsymbol{\mu}) + O(d^{-1}(\log d)^{-3/2}).$$

These results can be thought of as asymptotic oracle inequalities; they describe the ability of the SURE-based estimator $\hat{\mu}^*$ to mimic the risk of an "ideal" estimator constructed with special knowledge of the optimal threshold $t = t(\boldsymbol{\mu})$.

4. COMPARISON WITH ADAPTIVE LINEAR SHRINKAGE

We now briefly explain in an informal fashion why *SureShrink* may be expected to compare favorably to adaptive linear shrinkage.

4.1 Adaptive Linear Estimation via James-Stein

In the multivariate normal setting of Section 2.3, the simplest linear shrinkage estimate $\hat{\boldsymbol{\mu}}^c = c\mathbf{x}$ has MSE

$$E\|\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu}\|^2 = c^2 d + (1-c)^2\|\boldsymbol{\mu}\|^2.$$

If $\boldsymbol{\mu}$ were known, we could choose

$$\tilde{c}(\boldsymbol{\mu}) = 1 - d/(d + \|\boldsymbol{\mu}\|^2)$$

to minimize the MSE at $\boldsymbol{\mu}$. Because $\boldsymbol{\mu}$ is unknown (it is, after all, the quantity we are trying to estimate), this linear shrinker represents an unattainable ideal. The James–Stein (positive part) estimate is $\hat{\mu}_i^{\mathrm{JS}} = c^{\mathrm{JS}}(\mathbf{x}) \cdot x_i, i = 1, \ldots, d$, where the shrinkage coefficient is

$$c^{\mathrm{JS}}(\mathbf{x}) = (1 - (d-2)/\|\mathbf{x}\|_2^2)_+.$$

From $E\|\mathbf{x}\|^2 = d + \|\boldsymbol{\mu}\|^2$, we see that the James–Stein shrinkage coefficient $c^{\mathrm{JS}}(\mathbf{x})$ is essentially an estimate of the ideal shrinkage coefficient $\tilde{c}$. This ideal estimator (not a statistic!) $\tilde{\boldsymbol{\mu}}^{\mathrm{IS}}(\mathbf{x}) = \tilde{c}(\boldsymbol{\mu})\mathbf{x}$ has MSE

$$E_{\boldsymbol{\mu}}\|\tilde{\boldsymbol{\mu}}^{\mathrm{IS}} - \boldsymbol{\mu}\|_2^2 = \frac{d\|\boldsymbol{\mu}\|^2}{d + \|\boldsymbol{\mu}\|^2}.$$

In fact, the James–Stein estimate does an extremely good job of approaching this ideal.

Theorem 5. For all $d > 2$, and for all $\boldsymbol{\mu} \in R^d$,

$$E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\|_2^2 \le 2 + E_{\boldsymbol{\mu}}\|\tilde{\boldsymbol{\mu}}^{\mathrm{IS}} - \boldsymbol{\mu}\|_2^2. \tag{21}$$

We pay a price of at most 2 for using the James–Stein shrinker rather than the ideal shrinker. In high dimensions

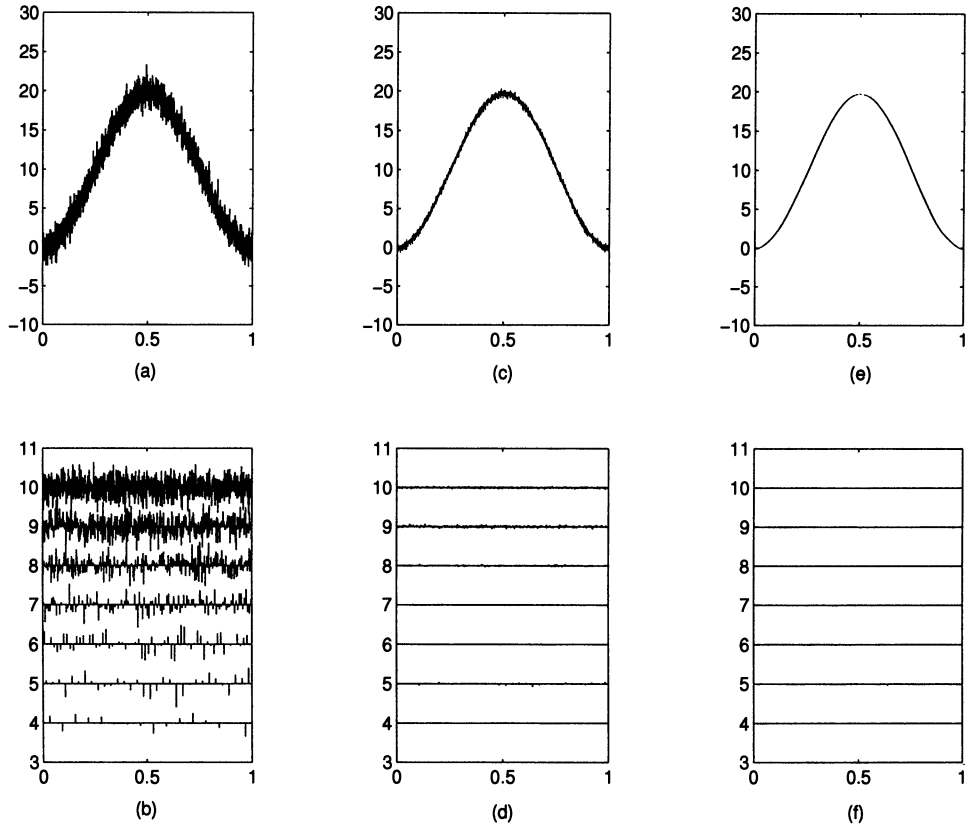

Figure 14. Comparison of WaveJS (c) and SureShrink (e) Reconstructions on Noisy Version (a) of f(t) = c sin $^2\pi$ t, and the Concomitant Action on Noisy Wavelet Coefficients (b) of WaveJS (d) and SureShrink (f).

d, this price is negligible. Because (21) bounds the risk of a genuine estimator in terms of an ideal estimator, it is an example of an "oracle inequality" (cf. Donoho and Johnstone 1994).

Now return to the function estimation setting setting of (1) and Definition 1 with observed wavelet coefficients $\mathbf{x}_j = (y_{j,k})_{0\le k<2^j}, 0 \le j < n$. We apply James–Stein shrinkage separately on each resolution level in the wavelet domain:

$$\hat{w}_j^{\mathrm{JS}} = \begin{cases} \mathbf{x}_j & j < L \\ \hat{\boldsymbol{\mu}}^{\mathrm{JS}}(\mathbf{x}_j) = \left(1 - \frac{2^j-2}{\|\mathbf{x}_j\|^2}\right)_+ \mathbf{x}_j & j \ge L. \end{cases}$$

Inverting the wavelet transform via

$$\hat{f}^{\mathrm{WJS}}(t_i) = \sum_{j,k} \hat{w}_{j,k}^{\mathrm{JS}} \mathcal{W}_{jk}(t_i)$$

gives an estimate that we call *WaveJS*.

Figures 14 and 15 show *WaveJS* in action on two spatially homogeneous functions: a low-frequency sinusoid $f(t) = c\sin^2 \pi t$ (with $L = 4$) and a high-frequency sinusoid $f(t) = c\sin 50\pi t$ (with $L = 6$), scaled in each case to have signal-to-noise ratio 7. In the first case, there is essentially no signal in levels $j \ge 4$, and so all coefficients are shrunk heavily. (Whether they are shrunk exactly to zero depends on whether $|\mathbf{x}_j|^2 < 2^j - 2$, a threshold that lies in the central part of the $\chi^2_{2^j}$, distribution.) For the high-frequency sinusoid, in levels $j \ge 6$ the signal is concentrated at level 6, so little shrinkage occurs there. The low-frequency oscillation in the reconstruction reflects the fact that no shrinkage is applied at levels $j < 6$. (Note also that *SureShrink* performs essentially as well as *WaveJS*, both visually and in MSE terms, on these homogeneous examples. At the higher levels, the pretest leads to the use of a $\sqrt{2\log 2^j}$ threshold, which in turn shrinks *almost* all coefficients to zero.)

Corresponding to this pleasant visual performance, a number of nice adaptivity properties of *WaveJS* follow immediately from Theorem 5. We state one such in the sequence-space setting of Sections 3.1 and 3.2. Consider model (15) and, as before, set $\mathbf{x}_j = (y_{j,k}/\varepsilon)$. Make the calibration $J = \log_2 \varepsilon^{-2}$ (so that $N = 2^J = \varepsilon^{-2}$). The corresponding WaveJS estimator is defined by

$$\begin{aligned} \hat{\theta}_{j,k}^{\mathrm{WJS}}(y) &= y_{jk} && j < L, \\ &= \varepsilon\hat{\boldsymbol{\mu}}^{\mathrm{JS}}(\mathbf{x}_j) && L \le j \le J, \\ &= 0 && j > J. \end{aligned}$$

Let $\hat{\theta}_L$ denote an estimator that is linear in the data y; the minimax risk among linear estimators is

$$R_L^*(\varepsilon;\Theta) = \inf_{\hat{\theta}_L} \sup_{\Theta} E\|\hat{\theta} - \theta\|^2.$$

We now state a linear adaptivity result for *WaveJS* that is analogous in form to that given for *SureShrink* in Theorem 3.

Theorem 6. If $\sigma > 1/p$, then

$$\sup_{\Theta^s_{p,q}(C)} E_\theta\|\hat{\theta}^{\mathrm{WJS}} - \theta\|_2^2 \le R_L^*(\varepsilon;\Theta^s_{p,q}(C))(1 + o(1))$$

as $\varepsilon \to 0$.

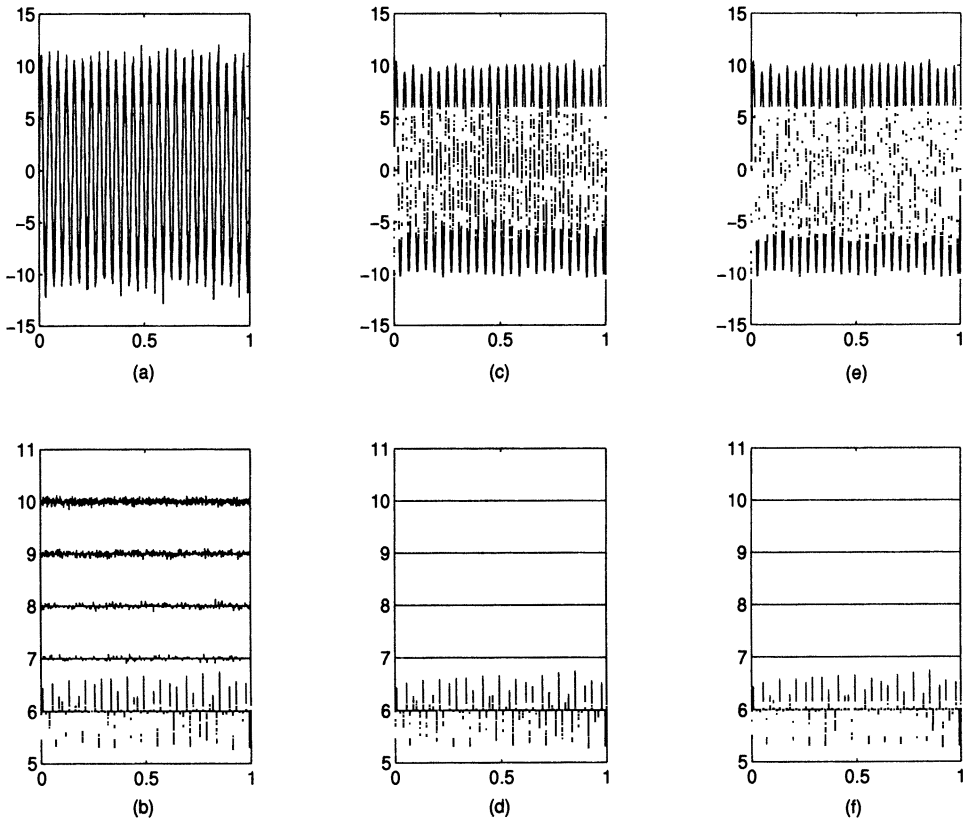

Figure 15. Comparison of WaveJS (c) and SureShrink (e) Reconstructions on Noisy Version (a) of f(t) = c sin 50πt, and the Concomitant Action on Noisy Wavelet Coefficients (b) of WaveJS (d) and SureShrink (f).

In the same way that sequence-space Theorem 3 leads to function-space Theorem 1, one may deduce a function-space version of Theorem 6 for the *WaveJS* estimator $\hat{\mathbf{f}}$. We indicate some of the features of this without going into full detail. Consider the ideal linear shrinkage estimator (again not a statistic)

$$(\tilde{w}_{j,k})_{0\le k<2^j} = \tilde{\boldsymbol{\mu}}^{\mathrm{IS}}(\mathbf{x}_j),$$

with inverse wavelet transform $\tilde{\mathbf{f}}^{\mathrm{ID}}$. Then, as an immediate corollary of Theorem 5, for all $N = 2^J$, and for all f,

$$R(\hat{\mathbf{f}}^{\mathrm{WJS}}, \mathbf{f}) \le R(\tilde{\mathbf{f}}^{\mathrm{ID}}, \mathbf{f}) + \frac{2^L + 2\log_2(N)}{N}.$$

Now the ideal estimator achieves within a constant factor of the minimax risk for *linear* estimators. Moreover, the minimax linear risk measured as earlier behaves like $N^{-r'}$, where $r' = \sigma/(\sigma + \frac{1}{2})$ if $p \ge 2$ and $s/(s + \frac{1}{2})$ if $p \le 2$. The ideal estimator is not, however, a statistic, whereas the James–Stein estimate *is* a statistic; and because $(2^L + 2\log_2(N))/N = o(N^{-r})$, it follows that $\hat{\mathbf{f}}^{\mathrm{WJS}}$ achieves the optimal rate of convergence for *linear* estimates over the whole Besov scale. This is in fact a better adaptivity result than previously established for adaptive linear schemes, because it holds over a very broad scale of spaces. Note, however, that the linear rate is slower than the nonlinear rate $N^{-\sigma/(\sigma+1/2)}$ if $p < 2$. This is one of the theoretical reasons for preferring the nonlinear *SureShrink* to *WaveJS*.

Moreover, such an estimate is not very good in practice, as we have seen in Figure 4. The *WaveJS* reconstruction is much noisier than *SureShrink*. This can be seen in Figure 16, which compares the action of *WaveJS* and *SureShrink* on wavelet coefficients of the Doppler signal: if in one resolution level there are significant coefficients that need to be kept, then the James–Stein estimate keeps all the coefficients, incurring a large variance penalty.

To obtain estimators with acceptable performance on spatially variable functions, one must, like *SureShrink*, adaptively keep large coordinates and kill small ones. An adaptive linear estimator does not do this, because it operates on coordinates at each level by the same multiplicative factor.

4.2 Other Adaptive Linear Estimation Methods

To a considerable degree, *WaveJS* serves as a representative for other, more familiar adaptive linear smoothing regimens. Examples include (Priestly–Chao) kernel smoothers, smoothing splines, and truncated Fourier inversion. Each case can be thought of as applying a linear shrinkage of Fourier coefficients. In each case there is a smoothing parameter (i.e., bandwidth, regularization weight, and truncation point) that controls the degree of shrinkage. This smoothing parameter might be chosen to minimize an unbiased estimate of MSE (just as the James–Stein shrinkage factor does approximately). This was done with smoothing splines and Fourier inversion (Donoho et al. 1995, figs. 5 and 6); these two figures are qualitatively quite similar to Figure 4 for *WaveJS* in this article. This and the structural similarities of all these adaptive linear methods underly our

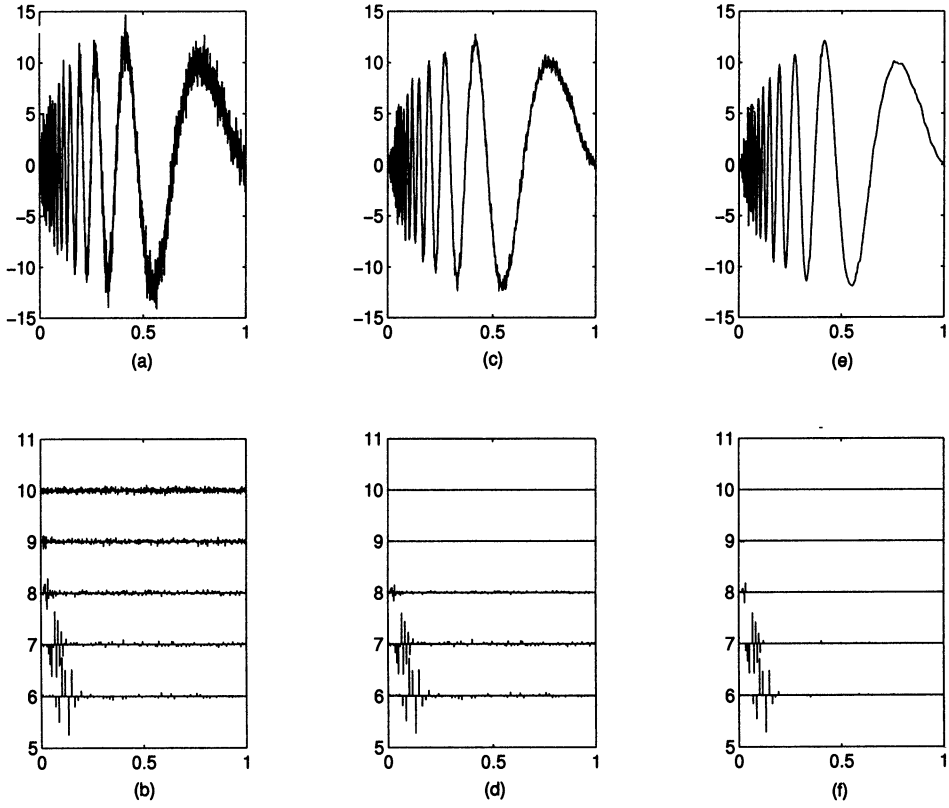

Figure 16. Comparison of WaveJS (c) and SureShrink (e) Reconstructions on Noisy Version (a) of the Doppler Signal, and the Concomitant Action on Noisy Wavelet Coefficients (b) of WaveJS (d) and SureShrink (f).

belief that the advantages of *SureShrink* over *WaveJS* apply quite generically to other adaptive linear methods. In the following subsection we briefly discuss a refined form of Fourier inversion for which impressive theoretical results exist.

4.3 Linear Adaptation Using Fourier Coronae

Suppose that we identify zero with 1, so that $[0, 1]$ has a circular interpretation. Work by Efroimovich and Pinsker (1984), and other recent Russian literature, would consider the use of adaptive linear estimators based on empirical Fourier coefficients $(\hat{v}_l)$. One divides the frequency domain into coronas, $l_i \le l < l_{i+1}$, and within each corona one uses a linear shrinker,

$$\hat{f}_l = c_i \cdot \hat{v}_l \qquad l_i \le l < l_{i+1}.$$

The weights are chosen adaptively by an analysis of the Fourier coefficients in the corresponding coronas. Letting $\mathbf{v}_i$ denote the vector of coefficients belonging to the ith corona, the choice used by Efroimovich and Pinsker (1984) is essentially

$$c_i = c^{\mathrm{EP}}(\mathbf{v}_i) = (\|\mathbf{v}_i\|_2^2 - d)/\|\mathbf{v}_i\|_2^2.$$

We propose an adaptive linear scheme that differs from the Efroimovich–Pinsker choice in two ways. First, we propose to use dyadic coronas $l_i = 2^{i+L}$. Such dyadic Fourier coronas occur frequently in Littlewood–Paley theory: (Frazier, Jawerth, Weiss 1991; Peetre 1975; Triebel 1983). Second, within each corona, we shrink via the James–Stein estimate $c_i = c^{\mathrm{JS}}(\mathbf{v}_i)$, which has nicer theoretical properties than the Efroimovich–Pinsker choice. The estimator that we get in this way we shall label LPJS.

LPJS is an adaptive linear estimator. Indeed, from Theorem 5, its risk is at most a term $[4\log_2(N)]/N$ worse than an ideal linear estimator $\tilde{\mathbf{f}}^{LPIS}$ defined in the obvious way. This ideal linear estimator, based on constant shrinkage in dyadic coronas, has performance not worse than a constant factor times the performance of so-called Pinsker weights, and hence we conclude that, except for constants, LPJS replicates the adaptive-rate advantages of the Efroimovich–Pinsker choice of coronas. LPJS offers advantages the Efroimovich–Pinsker choice does not. It achieves the optimal rate of *linear* estimators over a whole range of L^2 Sobolev, Hölder, and Besov spaces. Theoretically, LPJS is a very good adaptive linear estimator.

But in practice LPJS is disappointing; Figure 17 shows the LPJS reconstructions of our basic examples. The answers are significantly noisier than what can be obtained by *SureShrink*. Instead, the result is comparable to the (disappointing) performance of WaveJS. There is a deeper reason for the similarity between the LPJS and WaveJS, which derives from the Littlewood–Paley theory (Frazier et al. 1991).

5. DISCUSSION

5.1 Simulation Results

A small-scale simulation experiment was conducted to investigate the performance of the methods we have discussed. For each of the four objects under study, we applied nine different methods to noisy versions of the data:

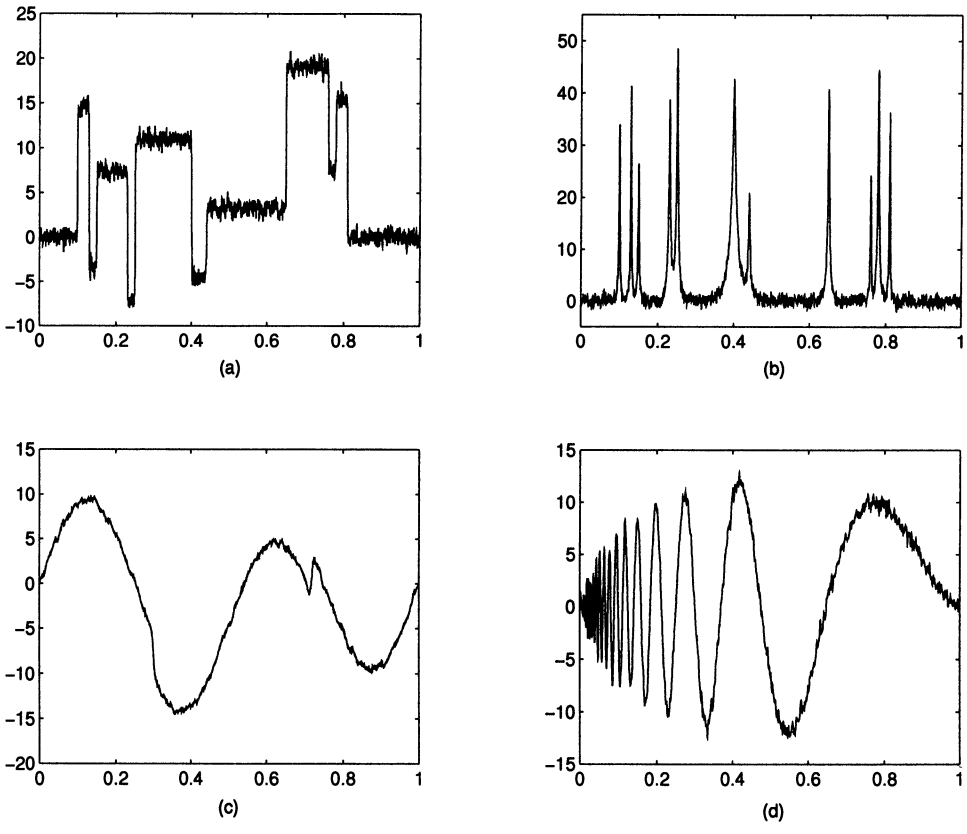

Figure 17. LPJS Reconstruction With Cutoff L = 5. (a) Blocks; (b) Bumps; (c) Heavisine; (d) Doppler.

SureShrink in the Haar, Db4, C3, and S8 wavelet bases, *VisuShrink* (Donoho et al. 1995 and Sec. 5.2 herein) in the S8 basis, *WaveJS* in the S8 wavelet basis, LPJS; and finally, the procedure *RiskShrink* (Donoho and Johnstone 1994a) using the C3 and S8 wavelet bases (denoted as ThrC3 and ThrS8). *RiskShrink* uses a fixed threshold chosen to yield minimax performance for MSE against an "oracle." These threshold values have been tabulated (Donoho and Johnstone 1994a). Dyadic sample sizes $N = 2^J$, from $N = 128$ to $N = 16{,}384$, were studied, and in all cases $L = 6$ was used ($L = 5$ for LPJS).

Sample results are given in Table 2, which reports the root mean square over 20 replications (10 if $N = 8{,}192, 1$ if $N = 16{,}384$) of the root loss $N^{-1/2}\|\hat{\mathbf{f}} - \mathbf{f}\|_2$ (not its square). The data are plotted in Figure 18.

In all examples, the *SureShrink* variants are generally competitive or close to best, with the partial exceptions of *Blocks* (where Haar *SureShrink* is clearly more appropriate) and *Doppler* (where the short and relatively rough filter D4 is noticeably inferior to C3 and S8). The adaptive linear methods *WaveJS* and LPJS are (predictably) less successful on the very nonhomogeneous examples, sometimes needing two to four times the sample size to obtain similar risks. The most extreme case is object *Blocks,* where the performance of shrinkage in the Haar basis at sample size 1,024 is comparable to the performance of LPJS at sample size 8,192 and 16,384. The results for *SureShrink* and *RiskShrink* are remarkably similar for *Doppler* and for *HeaviSine. VisuShrink* quite generally pays a high MSE price for its visually smooth appearance, with the high threshold incurring a larger bias.

5.2 Visual Quality of Reconstruction

The reader may have noticed that *SureShrink* reconstructions contain structure at all scales. This is inherent in the method, which has no a priori knowledge about the smoothness (or lack of smoothness) of the object. Occasional spurious fine-scale structure must sneak into the reconstruction; otherwise, the method would not be able to adapt spatially to true fine-scale structure.

Some readers may be actually annoyed at the tendency of *SureShrink* to show a small amount of spurious fine-scale structure and will demand a more thorough explanation. The presence of this fine-scale structure is demanded by the task of minimizing the l^2 norm loss, which always involves a trade-off between noise and bias. The l^2 norm balances these in equilibrium, which ensures that some noise artifacts will be visible.

Enhanced visual quality can be obtained by keeping the noise term in the trade-off to a minimum. This may be obtained by uniformly applying the threshold $\sqrt{2\log(N)}$ without any adaptive selection. This ensures that essentially all "pure noise" wavelet coefficients (i.e., coefficients where $w_{j,k} = 0$) are set to zero by the thresholding. The resulting curve shows most of the structure and very little noise. Further discussion of threshold selection by the $\sqrt{2\log(N)}$

Table 2. Root Mean Square Errors of Estimation Using Various Threshold Methods

	Haar	DB4	Coif3	Symm8	WvJS	LPJS	ThrC3	ThrS8	VisS8	SD's
Blocks										
128	.81	.88	.87	.89	.94	.93	.89	.91	1.18	.062
256	.68	.78	.80	.80	.92	.91	.87	.88	1.20	.039
512	.59	.76	.77	.78	.85	.85	.78	.80	1.12	.034
1024	.47	.64	.63	.64	.77	.76	.69	.72	1.03	.023
2048	.41	.50	.54	.56	.67	.68	.60	.60	.85	.016
4096	.29	.41	.43	.44	.58	.58	.50	.50	.72	.011
8192	.24	.33	.34	.37	.50	.50	.41	.42	.59	.008
16384	.22	.27	.29	.30	.43	.42	.34	.35	.49	*NA*
Bumps										
128	.93	.94	.94	.94	.99	.99	1.00	.99	1.35	.062
256	.85	.85	.85	.85	.99	.98	.96	.97	1.41	.039
512	.76	.74	.73	.74	.94	.94	.90	.92	1.38	.034
1024	.70	.63	.70	.70	.84	.84	.78	.79	1.16	.023
2048	.67	.55	.51	.52	.70	.69	.65	.66	.96	.016
4096	.58	.43	.42	.41	.54	.53	.53	.54	.77	.011
8192	.51	.32	.31	.32	.42	.42	.42	.43	.61	.008
16384	.38	.25	.22	.22	.33	.33	.33	.34	.48	*NA*
Heavi										
128	.77	.73	.74	.73	.74	.61	.75	.75	.73	.062
256	.62	.55	.55	.56	.56	.49	.56	.55	.56	.039
512	.55	.42	.43	.44	.45	.40	.42	.43	.45	.034
1024	.46	.32	.33	.33	.34	.34	.33	.33	.34	.023
2048	.35	.26	.26	.27	.28	.28	.25	.26	.28	.016
4096	.28	.20	.20	.20	.23	.23	.20	.20	.23	.011
8192	.22	.15	.17	.17	.20	.20	.17	.17	.20	.008
16384	.18	.12	.11	.12	.16	.16	.12	.13	.16	*NA*
Doppler										
128	.93	.88	.83	.82	.93	.94	.82	.80	.92	.062
256	.88	.83	.73	.74	.92	.90	.74	.75	.98	.039
512	.89	.73	.65	.63	.77	.75	.61	.58	.80	.034
1024	.72	.64	.54	.50	.59	.61	.50	.49	.65	.023
2048	.61	.45	.35	.38	.47	.47	.39	.38	.51	.016
4096	.50	.36	.28	.28	.38	.36	.31	.30	.42	.011
8192	.38	.28	.20	.19	.27	.28	.24	.23	.31	.008
16384	.32	.18	.15	.15	.21	.21	.18	.18	.25	*NA*

NOTE: Haar, DB4, Coif3, Symm8 = *SureShrink* using the indicated wavelet filter; WvJS = *WaveJS* using S8; LPJS = LPJS (Sec. 4.3); ThrC3 and ThrS8 = *RiskShrink* using C3 and S8 filters; and VisS8 = *VisuShrink* using S8. Sample sizes $N = 2^j$ for $j = 7(1)14$, with 20 replications for each N, except for 10 replications at $N = 8{,}192$ and 1 replication only at $N = 16{,}384$. Standard deviations shown in column SD are average over estimators and target function of the 36 individual SD's obtained for each sample size N.

rule (called *VisuShrink*) and the connection with optimum "visual quality" may be found in other work (Donoho et al. 1995).

5.3 Hard Thresholding

Could one use "hard thresholding,"

$$\begin{aligned} \xi_t(y) &= y \qquad |y| \geq t, \\ &= 0 \qquad |y| < t, \end{aligned}$$

in place of soft thresholding η_t? Indeed, hard thresholding seems more natural to nonstatisticians. We prefer soft thresholding because of various statistical advantages (e.g., continuity of the rule; simplicity of the SURE formula). But in principle, the foregoing results could have equally well been derived for hard thresholding. Because hard thresholding is not continuous, let alone weakly differentiable, a more complicated SURE formula would be required to implement the idea on data. The proofs would also be more complicated.

5.4 Estimated Noise Level

For practical use, it is important to estimate the noise level σ from the data rather than to assume that the noise level is known. In practice we derive an estimate from the finest scale empirical wavelet coefficients: $\hat{\sigma} = \text{median}(|y_{J-1,k}| : 0 \leq k < 2^{J-1})/.6745$. We believe that it is important to use a robust estimator like the median, in case the fine-scale wavelet coefficients contain a small proportion of strong "signals" mixed in with "noise."

5.5 Other Literature

There are a number of interesting comparisons or extensions that we have not discussed for lack of space:

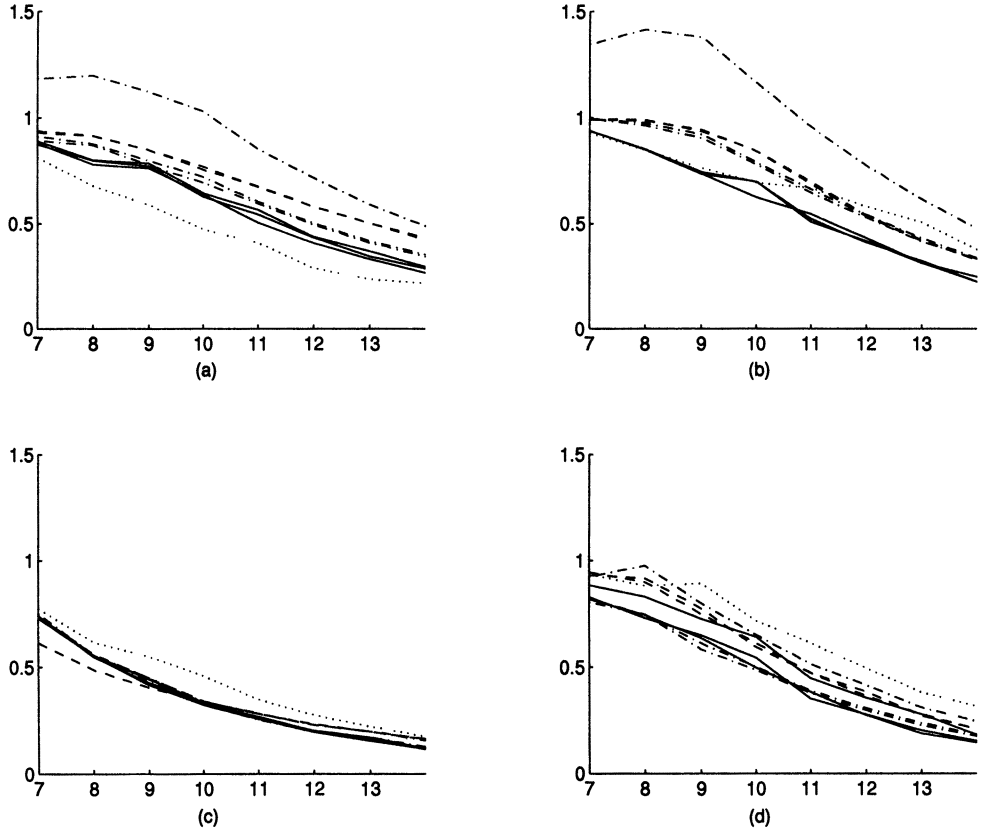

Figure 18. Root MSE's (y Axis) Plotted Against $\log_2 N$ for SureShrink in the Haar basis (dotted line) and Db4, C3, and S8 Bases (Solid Line); WaveJS and LPJS in the S8 Basis (Dashed Line); RiskShrink in C3 and S8 (Dash-Dot line); and VisuShrink in S8 (Dash-Dot Line). (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler.

- We have not compared our results here with the considerable literature on the use of cross-validation to select bandwidth of fixed-kernel smoothers (compare Johnstone and Hall 1992 and the many references therein).
- Neither have we considered how some of the recent more stable bandwidth selection methods discussed and cited there might be applied in the present context.
- Nor have we discussed in detail earlier applications of SURE with linear estimates (compare Li 1985 and references therein).
- Other spatially adaptive, nonlinear regression methods based on variable bandwidth kernels (Brockmann, Gasser, and Herrmann 1993), local polynomials (Cleveland 1979; Fan and Gijbels 1995) or adaptive splines (Friedman 1991).
- Finally we have not discussed applications of wavelet thresholding in density estimation (Johnstone, Kerkyacharian, and Picard 1992).

5.6 Software

Software for the methods described in this article forms part of the WaveLab package of MATLAB M-files, data sets, demonstrations, and documentation. It and further technical reports are available either by anonymous ftp to `playfair.stanford.edu` in directories `pub/donoho` and `pub/johnstone`, or via WWW and URL `http://stat.stanford.edu/`. In particular, the figures in this paper may be reproduced (or modified) by using the M-files in the directory `~/Papers/Adapt` in the WaveLab release. The *SureShrink* method is also implemented in the S+WAVELETS toolkit, by A. G. Bruce and H.-Y. Gao (1995), which runs within S-PLUS.

6. CONCLUSION

A curious divergence has recently developed between minimax adaptation theory and the heuristics guiding algorithm development. As we have seen, the former established adaptivity properties for essentially linear methods (up to the data-based choice of linear shrinkage or bandwidth parameter). The latter was concerned in part with spatial adaptivity, using local polynomials or splines and strongly nonlinear methods. The results in this article can be thought of as a step toward reversing this divergence. First, wavelet bases have spatial adaptation properties that are essentially equivalent with those of local polynomials and splines— these are certainly reflected in *SureShrink,* which has a fast and at least potentially practical algorithm. Second, we establish rate-optimal adaptation results for *SureShrink. SureShrink* is (necessarily) strongly nonlinear, but fortunately in the wavelet domain, the nonlinearity can take the relatively simple form of coordinatewise thresholding.

APPENDIX: PROOFS OF THEOREMS 3–6

We proceed in reverse: first collecting tools, then establishing Theorem 4, and finally returning to Theorem 3.

Exponential Inequalities

We first recall two basic exponential inequalities for bounded variates from Hoeffding (1962) and note a corresponding inequality for chi-squared variates:

(A) Let $Y_1, \dots, Y_n$ be independent, $a_i \le Y_i \le b_i$, and $\bar{Y}_n = n^{-1}\sum_1^n Y_i$ and $\mu = E\bar{Y}_n$. For $t > 0$,

$$P\{|\bar{Y}_n - \mu| \ge t\} \le 2\exp\left\{-2n^2t^2 \Big/ \sum_1^n (b_i - a_i)^2\right\}. \quad \text{(A.1)}$$

(B) Let $X_1, \dots, X_m$ be sampled *without replacement* from $\{c_1, \dots, c_n\}$. Suppose that $a \le c_i \le b$ for all i. Set $\bar{X}_m = m^{-1}\sum_1^m X_i$ and $\mu = n^{-1}\sum_1^n c_i$. For $t > 0$,

$$P\{|\bar{X}_m - \mu| \ge t\} \le 2\exp\{-2nt^2/(b-a)^2\}. \quad \text{(A.2)}$$

(C1) Let $Z_1, \dots, Z_n$ be iid $N(0,1)$. Then by elementary arguments,

$$P\{|\Sigma\alpha_j(z_j^2 - 1)| > t\} \le 2e^{2s^2\Sigma\alpha_j^2 - |s|t}$$

for

$$|s| \le 1/(4\max(|\alpha_j|)).$$

(C2) If all $\alpha_j = n^{-1}$, then, by optimizing over s,

$$P\left\{\left|n^{-1}\sum (z_j^2 - 1)\right| > t\right\} \le 2e^{-nt(t\wedge 1)/8}. \quad \text{(A.3)}$$

Preparatory Propositions

We use (A) to bound the deviation of the unbiased risk estimate (11) from its expectation. To recapitulate the setting of Section 2.3, suppose that $x_i \sim N(\mu_i, 1), i = 1, \dots, d$ are independent. Let F_d denote the empirical distribution function of $\{\mu_i\}$. As earlier, let $r(t, \mu_i) = E[\eta_t(x_i) - \mu_i]^2$ denote the MSE of the soft threshold estimate of a single coordinate and define

$$r(t, F) = \int r(t, \mu)F(d\mu).$$

In particular,

$$r(t, F_d) = d^{-1}\Sigma r(t, \mu_i) = d^{-1}E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{(t)} - \boldsymbol{\mu}\|^2. \quad \text{(A.4)}$$

Stein's unbiased risk estimate (11),

$$\begin{aligned} U_d(t) &= d^{-1}\mathrm{SURE}(t, \mathbf{x}), \\ &= 1 - 2d^{-1}\sum_i I\{x_i^2 \le t^2\} + d^{-1}\sum_i x_i^2 \wedge t^2, \\ &= d^{-1}\sum_i 1 - 2I\{x_i^2 \le t^2\} + x_i^2 \wedge t^2, \end{aligned} \quad \text{(A.5)}$$

has expectation $r(t, F_d)$. We study the deviation

$$Z_d(t) = U_d(t) - r(t, F_d)$$

uniformly for $0 \le t \le t_d = \sqrt{2\log d}$.

Proposition 1. Uniformly in $\boldsymbol{\mu} \in \mathbb{R}^d$,

$$E_{\boldsymbol{\mu}} \sup_{0 \le t \le t_d} |U_d(t) - r(t, F_d)| = O\left(\frac{\log^{3/2} d}{d^{1/2}}\right).$$

Proof. Combining (A.4) and (A.5) with the bound $r(t, \mu_i) \le 1 + t^2$, we can write $Z_d(t) = d^{-1}\sum_1^d Y_i(t)$ with zero mean summands that are uniformly bounded: $|Y_i(t)| \le 2 + t^2$. Hoeffding's inequality (A.1) gives, for a fixed t and (for now) arbitrary $r_d > 1$,

$$P\{|Z_d(t)| > r_d d^{-1/2}\} \le 2\exp\{-r_d^2/2(t^2+2)^2\}. \quad \text{(A.6)}$$

For distinct $t < t'$, let $N_d(t, t') = \#\{i : t < |x_i| \le t'\}$ and

$$\begin{aligned} U_d(t) - U_d(t') &= 2d^{-1}\Sigma I\{t^2 < x_i^2 \le t'^2\} \\ &\quad + d^{-1}\sum_i x_i^2 \wedge t^2 - x_i^2 \wedge t'^2 \\ &\le d^{-1}(2 + t'^2 - t^2)N_d(t, t'). \end{aligned}$$

We may bound $r(t, F_d) - r(t', F_d)$ by recalling that for $t \le t_d$, $(\partial/\partial t)r(t, F_d) \le 5t_d$. Then, so long as $|t - t'| < \delta_d$,

$$|Z_d(t) - Z_d(t')| \le 2d^{-1}(1 + \delta_d t_d)N_d(t, t') + 5\delta_d t_d.$$

Now choose $t_j = j\delta_d \in [0, t_d]$; clearly,

$$A_d = \{\sup_{[0,t_d]} |Z_d(t)| \ge 3r_d d^{-1/2}\} \subset D_d \cup E_d,$$

where $D_d = \{\sup_j |Z_d(t_j)| \ge r_d d^{-1/2}\}$ and

$$E_d = \{\sup_j \sup_{|t - t_j| \le \delta_d} |Z(t) - Z(t_j)| \ge 2r_d d^{-1/2}\}.$$

Choose δ_d so that $\delta_d t_d = o(d^{-1/2})$; then E_d is contained in

$$\begin{aligned} E_d' &= \{\sup_j 2d^{-1}N_d(t_j, t_j - \delta_d) \ge r_d d^{-1/2}\} \\ &\subset \{\sup_j d^{-1}|N_d(t_j, t_j + \delta_d) - EN_d| \ge r_d d^{-1/2}/3\} \\ &= E_d'', \end{aligned}$$

say, for large d where we used $EN_d(t_j, t_j + \delta_d) \le c_0 d\delta_d = O(r_d d^{1/2})$. Again, from Hoeffding's inequality (A.1),

$$P\{d^{-1}|N_d(t_j, t_j + \delta_d) - EN_d| \ge r_d d^{-1/2}/3\} \le e^{-2r_d^2/9}. \quad \text{(A.7)}$$

Finally, using (A.6), (A.7), and the cardinality of $\{t_j\}$,

$$\begin{aligned} P(A_d) &\le P(D_d) + P(E_d'') \\ &\le 2t_d\delta_d^{-1}(\exp\{-r_d^2/2(t_d^2+2)^2\} + \exp\{-2r_d^2/9\}). \end{aligned}$$

Set $r_d = (2b\log d)^{1/2}(t_d^2 + 2) = O(\log^{3/2} d)$. Then

$$P(A_d) \le \frac{3t_d}{\delta_d d^b}. \quad \text{(A.8)}$$

Let $\|Z_d\| = \sup\{|Z_d(t)| : 0 \le t \le t_d\}$ and $r_d^0 = (2\log d)^{1/2}(t_d^2 + 2)$. We may rephrase (A.8) as

$$P_{\boldsymbol{\mu}}\{\sqrt{d}\|Z_d\|/r_d^0 > s\} \le 3t_d\delta_d^{-1}e^{-s^2\log d},$$

which suffices for the L_1 convergence claimed.

Proposition 2. Uniformly in $\boldsymbol{\mu} \in \mathbb{R}^d$,

$$E_I\|r(\cdot, F_I) - r(\cdot, F)\|_\infty = O\left(\frac{\log^{3/2} d}{d^{1/2}}\right).$$

Proof. This is similar to that of Proposition 1, but is simpler and uses Hoeffding's inequality (A.2). In the notation of (A.2), set $n = d, c_i = r(t, \mu_i) \le 1 + t_d^2$, and $m = d/2$, so that $\mu = r(t, F_d), \bar{X}_m = r(t, F_I)$, and

$$Z(t) := r(t, F_I) - r(t, F_d) = \bar{X}_m - \mu,$$

$$P\{|Z(t)| > r_d d^{-1/2}\} \le 2\exp\{-2r_d^2/(1+t_d^2)^2\}.$$

Because $|(\partial/\partial t)r(t,F)| \le 5t_d$ for any F, it follows that for $|t'-t| < \delta_d$,

$$|Z(t') - Z(t)| \le 10\delta_d t_d.$$

Thus if δ_d is small enough so that $10\delta_d t_d \le r_d d^{-1/2}$ and $r_d = (2b\log d)^{1/2}(t_d^2+1)$, then

$$P\{\|Z_d\| \ge 2r_d d^{-1/2}\} \le P\{\sup_{j:j\delta_d \le t_d} |Z_d(j\delta_d)| \ge r_d d^{-1/2}\} \le \frac{2t_d}{\delta_d}\frac{1}{d^{4b}}.$$

As for Proposition 1, this yields the result.

Lemma 1. Let $x_i \sim N(\mu_i, 1), i = 1, \dots, d$, be independent, $s_d^2 = d^{-1}\Sigma(x_i^2 - 1)$, and $\tau^2 = d^{-1}\Sigma\mu_i^2$. Then, if $\gamma_d^2 d/\log d \to \infty$,

$$\sup_{\tau^2 \ge 3\gamma_d} (1+\tau^2)P\{s_d^2 \le \gamma_d\} = o(d^{-1/2}). \tag{A.9}$$

Proof. This is a simple statement about the tails of the noncentral chi-squared distribution. Write $x_i = z_i + \mu_i$, where $z_i \sim N(0,1)$. The event in (A.9) may be rewritten as

$$\begin{aligned} A_d &= \{d^{-1}\Sigma(z_i^2 - 1) + d^{-1}\Sigma 2\mu_i z_i \le -(\tau^2 - \gamma_d)\} \\ &\subset \{d^{-1}\Sigma(z_i^2 - 1) \le -\tau^2/3\} \cup \{d^{-1}\Sigma 2\mu_i z_i \le -\tau^2/3\} \\ &= B_d \cup C_d, \end{aligned} \tag{A.10}$$

from the lower bound on τ^2 in (A.9).

By elementary inequalities, for $\tilde{\Phi}(z) = \int_z^\infty e^{-s^2/2}\, ds/\sqrt{2\pi}$,

$$P(C_d) = \tilde{\Phi}\left(\frac{\tau^2}{3}\frac{d^{1/2}}{2\tau}\right) \le c_1 e^{-c_2 d\tau^2}, \tag{A.11}$$

and it is easily verified that $(1+\tau^2)e^{-c_2 d\tau^2} \le 2e^{-3c_2 d\gamma_d} = o(d^{-1/2})$ for $\tau^2 \ge 3\gamma_d$ and d large.

For B_d, apply the exponential inequality (A.3) to obtain

$$\begin{aligned} (1+\tau^2)P(B_d) &\le 2(1+\tau^2)\exp\{-d\tau^2(\tau^2 \wedge 3)/72\} \\ &\le c_3 \exp\{-\gamma_d^2 d/8\} = o(d^{-1/2}), \end{aligned}$$

because $\gamma_d^2 d/\log d \to \infty$.

Proof of Theorem 4(a)

Decompose the risk of $\hat{\boldsymbol{\mu}}^*$ according to the outcome of the pretest event $A_d = \{s_d^2 \le \gamma_d\}$, with the goal of showing that

$$\begin{aligned} R_{1d}(\boldsymbol{\mu}) &= d^{-1}E[\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\|^2, A_d] \\ &\le R_F(\boldsymbol{\mu})I\{\tau^2 \le 3\gamma_d\} + o(d^{-1/2}) \end{aligned} \tag{A.12}$$

and

$$R_{2d}(\boldsymbol{\mu}) = d^{-1}E[\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\|^2, A_d^c] \le \tilde{R}(\boldsymbol{\mu}) + c(\log d)^{3/2}d^{-1/2}. \tag{A.13}$$

On event A_d, fixed thresholding is used:

$$R_{1d} = d^{-1}E\left[\sum_i (\eta(x_i, t_d^F) - \mu_i)^2, A_d\right] \le R_F(\boldsymbol{\mu}).$$

If $\tau^2 \ge 3\gamma_d$, then we first note that on event A_d,

$$d^{-1}\sum_i \eta(x_i, t_d^F)^2 \le d^{-1}\Sigma x_i^2 \le 1 + \gamma_d.$$

Using Lemma 1, (A.12) follows from

$$R_{1d} \le 2(1 + \gamma_d + \tau^2)P(A_d) = o(d^{-1/2}).$$

On event A_d^c, the adaptive, half-sample–based thresholding applies. Let $E_{\boldsymbol{\mu}}$ denote expectation over the distribution of (x_i) and let E_I denote expectation over the random choice of half-sample I. Then

$$dR_{2d} \le E_I\left\{\sum_{i\in I} E_{\boldsymbol{\mu}}[\eta(X_i, \hat{t}_{I'}) - \mu_i]^2 + \sum_{i\in I'} E_{\boldsymbol{\mu}}[\eta(X_i, \hat{t}_I) - \mu_i]^2.\right\}$$

Let F_I, (resp. $F_{I'}, F_d$) denote the empirical distribution functions of $\{\mu_i : i \in I\}$ (resp. of $\mu_i, i \in I', \{1, \dots, d\}$), and set $r(t,F) = \int r(t,\boldsymbol{\mu})F(d\boldsymbol{\mu})$. Then, using the symmetry between I and I', we have

$$\begin{aligned} R_{2d} &= \frac{1}{2}E_I E_{\boldsymbol{\mu}}\{r(\hat{t}_{I'}, F_I) + r(\hat{t}_I, F_{I'})\} \\ &= E_I E_{\boldsymbol{\mu}} r(\hat{t}_I, F_{I'}). \end{aligned}$$

Thus to complete the proof of (A.13), it suffices to show that

$$R_{3d}(\boldsymbol{\mu}) := E_I E_{\boldsymbol{\mu}} r(\hat{t}_I, F_{I'}) - \tilde{R}(\boldsymbol{\mu}) \le c(\log d)^{3/2}d^{-1/2}. \tag{A.14}$$

There is a natural decomposition,

$$\begin{aligned} R_{3d} &= E_{\boldsymbol{\mu}}E_I[r(\hat{t}_I, F_{I'}) - r(\hat{t}_I, F_I)] \\ &\quad + E_{\boldsymbol{\mu}}E_I[r(\hat{t}_I, F_I) - r_{\min}(F_I)] \\ &\quad + E_I[r_{\min}(F_I) - r_{\min}(F_d)], \\ &= S_{1d} + S_{2d} + S_{3d}, \end{aligned}$$

where we have set $r_{\min}(F) = \inf\{r(t,F), 0 \le t \le t_d^F\}$ and note that $r_{\min}(F_d) = \tilde{R}(\boldsymbol{\mu})$. We use

$$r(t, F_I) - r(t, F_d) = \frac{1}{2}[r(t, F_I) - r(t, F_{I'})] \tag{A.15}$$

together with the simple observation that $|r_{\min}(F) - r_{\min}(G)| \le \|r(\cdot, F) - r(\cdot, G)\|_\infty$ to conclude that

$$S_{1d} + S_{3d} \le 3E_I\|r(\cdot, F_I) - r(\cdot, F_d)\|_\infty = O\left(\frac{\log^{3/2} d}{d^{1/2}}\right)$$

using Proposition 2.

Finally, let $U_{d/2}(t, I)$ denote the unbiased risk estimate derived from subset I. Then, using Proposition 1,

$$\begin{aligned} S_{2d} &\le E_{\boldsymbol{\mu}}E_I|r(\hat{t}_I, F_I) - U_{d/2}(\hat{t}_I, I)| + |U_{d/2}(\hat{t}_I, I) - r_{\min}(F_I)| \\ &\le 2E_I E_{\boldsymbol{\mu}}\|r(\cdot, F_I) - U_{d/2}(\cdot, I)\|_\infty = O\left(\frac{\log^{3/2} d}{d^{1/2}}\right). \end{aligned}$$

Putting all together, we obtain (A.14).

Proof of Theorem 4(b)

When $\|\boldsymbol{\mu}\|$ is small, the pretest of $s_d^2 \le \gamma_d$ will with high probability lead to use of the fixed threshold t_d^F. The $O(d^{-1/2}\log^{3/2} d)$ error term in Theorem 4, which arises from empirical process fluctuations connected with minimization of SURE, is then not germane and can be improved to $O(d^{-1})$.

We decompose the risk of $\boldsymbol{\mu}^*$ as in (A.12) and (A.13), but now $P(A_d) \nearrow 1$ as $d \nearrow \infty$. On A_d, fixed thresholding is used and so

$$dR_{1d} \le \sum_1^d r(t_d^F, \mu_i).$$

We use large derivation inequalities to bound the remaining term R_{2d}. Using symmetry between I and I',

$$dR_{2d} \le 2E_I \sum_{i\in I'} E_{\boldsymbol{\mu}}\{(\eta(X_i, \hat{t}_I) - \mu_i)^2, A_d^c\}.$$

Using the Cauchy–Schwartz inequality and noting from the limited translation structure of $\eta(\cdot,t)$ that $E_\mu(\eta(X,t)-\mu)^4 \le c(t_d^F)^4$ for $t \le t_d^F$, we get

$$dR_{2d} \le cd(t_d^F)^2 P_{\boldsymbol{\mu}}(A_d^c)^{1/2}.$$

Arguing similarly to (A.10), we note that the hypothesized bound on $\boldsymbol{\mu}$ implies that

$$A_d^c \subset \left\{d^{-1}\sum(z_i^2-1) > \gamma_d/3\right\} \cup \left\{d^{-1}\sum 2\mu_i z_i > \gamma_d/3\right\} = B_d \cup C_d.$$

The chi-squared exponential inequality (A.3) and standard Gaussian inequalities give

$$P(B_d) \le \exp\{-d\gamma_d^2/72\}$$

and

$$P(C_d) \le \exp\{-\gamma_d d/24\},$$

which imply that $d\log d \cdot P_d(A_d^c)^{1/2} = o(\log^{-3/2} d)$, which shows negligibility of R_{2d} and completes the proof.

Proof of Theorem 3

The idea behind the proof is as follows. For a given $\Theta_{p,q}^s(C)$ and noise level ε, there is a certain level at which the least favorable cases are found; that is, j_* at (A.26). For j near j_*, the unbiased risk estimate picks the threshold in an asymptotically efficient manner. At levels where total signal is negligible, the pretest picks t_j^F with high probability, so that risk and modulus of continuity bounds for $\sqrt{2\log d}$ thresholding can be used to show that the remaining terms are of lower order.

We make use of the definitions (19) and (20) to write

$$E_\theta\|\hat\theta^* - \theta\|^2 = 2^L\varepsilon^2 + \varepsilon^2\sum_{j\ge L} E\|\hat{\boldsymbol{\mu}}^*(x_j) - \boldsymbol{\mu}_j\|^2,$$

where $\boldsymbol{\mu}_j = (\mu_{jk}) = (\theta_{jk}/\varepsilon)$ and $\theta_j = (\theta_{jk})$. We use the abbreviations $t_j^F = t_{2^j}^F = \sqrt{2\log 2^j}$ and $\gamma_j = \gamma_{2^j}$. For a $L = j_0(\varepsilon,\sigma,p,q) \nearrow \infty$ to be specified later, we use Theorem 4(a) for levels $j \le j_0$ and Theorem 4(b) for $j > j_0$:

$$E_\theta\|\hat\theta^* - \theta\|^2 \le O(\varepsilon^2) + S_{1\varepsilon} + S_{2\varepsilon}, \tag{A.16}$$

$$S_{1\varepsilon} \le \varepsilon^2\sum_{j\le j_0}\left\{\inf_{t_j}\sum_k r(t_j,\mu_{jk}) + I\{\tau_j^2 \le 3\gamma_j\} \times \sum_k r(t_j^F,\mu_{jk}) + cj^{3/2}2^{j/2}\right\} \tag{A.17}$$

$$= S_{11\varepsilon} + S_{12\varepsilon} + S_{13\varepsilon}, \tag{A.18}$$

$$S_{2\varepsilon} \le \varepsilon^2\sum_{j>j_0}\left\{\sum_k r(t_j^F,\mu_{jk}) + cj^{-3/2}\right\} \tag{A.19}$$

$$= S_{21\varepsilon} + S_{22\varepsilon}, \tag{A.20}$$

say. Maximizing now over $\Theta_{p,q}^s$,

$$\sup_\Theta S_{11\varepsilon} \le \inf_{(t_j)}\sup_\Theta E\|\hat\theta_{(t_j)} - \theta\|^2 \tag{A.21}$$

$$= R_T^*(\varepsilon,\Theta_{p,q}^s) \asymp \varepsilon^{2r}, \tag{A.22}$$

where $r = 2\sigma/(2\sigma+1)$. It remains to verify that j_0 can be chosen so that all other terms are $o(\varepsilon^{2r})$. Define $j_\varepsilon = (1-r)\log C^2\varepsilon^{-2}$. The term $S_{13\varepsilon}$ is negligible if

$$j_0 + 3\log_2 j_0 - 2j_\varepsilon \to -\infty. \tag{A.23}$$

Because $S_{22\varepsilon} = O(\varepsilon^2)$, it remains to consider $S_{12\varepsilon}$ and $S_{21\varepsilon}$.

We begin by borrowing some notation and results from Donoho et al. (1994). Suppose that $y \sim N_d(\xi,\varepsilon^2 I)$ and $\xi \in \Xi \subset \mathbb{R}^d$. Define a modulus of continuity,

$$\Omega(\delta;\Xi) = \sup\{\|\xi\|_2 : \xi \in \Xi, \|\xi\|_\infty \le \delta\}.$$

Let $\hat\xi_{F,i}(y) = \eta(y_i, t_d^F\varepsilon)$ denote the fixed threshold estimator; using arguments similar to those of section 5 of Donoho et al. (1995), we have

$$\sup_\Xi E\|\hat\xi_F - \xi\|^2 \le c\Omega^2(2t_d^F\varepsilon;\Xi). \tag{A.24}$$

We shall denote an l_p-ball in $\mathcal{R}^n$ by $B_{p,n}(r) = \{\xi \in \mathbb{R}^n : \|\xi\|_p \le r\}$. Return to the sequence-space setting of Section 3. If $\theta = (\theta_{jk})$, then define $\theta^{(j)}$ to be the same vector with all coefficients set to zero that are not at resolution level j. With a slight abuse of notation, we define

$$\Theta^{(j)} \equiv \{\theta^{(j)} : \theta \in \Theta_{p,q}^s(C)\} \equiv B_{p,2^j}(C2^{-sj}). \tag{A.25}$$

The constraint $\tau^2 \le 3\gamma_j$ in (A.17) corresponds to a set

$$\bar\Theta^{(j)} \equiv B_{2,2^j}(c_{j\varepsilon}), \qquad c_{j\varepsilon}^2 = 3\varepsilon^2 2^j\gamma_j.$$

From Donoho et al. (1995), we borrow the inequality

$$\Omega(\delta;\Theta^{(j)}) \le \delta^r C^{1-r}2^{-\eta|j-j_*|}, \tag{A.26}$$

where $\eta = \eta(\sigma,p) > 0, r = 2\sigma/(2\sigma+1)$, and $j_* = 2(1-r)\log_2 C/\delta$. Again, note from Donoho et al. (1995) that because $2^j(2\varepsilon t_j^F)^2 \ge c_{j\varepsilon}^2$,

$$\Omega(2\varepsilon t_j^F, \bar\Theta^{(j)}) = c_{j\varepsilon}. \tag{A.27}$$

The bound (A.24) shows that

$$S_{12\varepsilon} \le c\sum_{j\le j_0}\Omega^2(2t_j^F\varepsilon;\Theta^{(j)}\cap\bar\Theta^{(j)}),$$

and we plan to use (A.26) and (A.27) in conjunction with $\Omega(\delta,\cap\Theta_i) \le \min_i\Omega(\delta,\Theta_i)$.

For small j, we note that $\Omega(2\varepsilon t_j^F,\bar\Theta^{(j)}) = o(\varepsilon^{2r})$ if

$$j + \log_2\gamma_j - j_\varepsilon \to -\infty. \tag{A.28}$$

Choose $\alpha > 0$ small. If we choose γ_j so that $\gamma_j \le 2^{-\alpha j}$ for all j, then (A.28) will certainly hold for all $j \le j_{1\varepsilon} := j_\varepsilon/(1-\alpha/2)$. For large j (i.e., $j \ge j_{1\varepsilon}$), we use (A.26) after noting that $j_* \le j_\varepsilon$,

$$\Omega^2(2\varepsilon t_j^F;\Theta^{(j)}) \le c\varepsilon^{2r}j^r 2^{-2\eta|j-j_\varepsilon|} = o(\varepsilon^{2r}).$$

Because the bounds for both $\Theta^{(j)}$ and $\bar\Theta^{(j)}$ decay geometrically with j, it follows that the sum $S_{12\varepsilon} = o(\varepsilon^{2r})$ also.

We now turn to $S_{21\varepsilon}$. So that we may apply Theorem 4(b) to obtain the bound (A.20), it is necessary that j_0 be chosen so that $\Theta^{(j)} \subset \frac{1}{3}\bar\Theta^{(j)}$ for all $j \ge j_0$. Because $B_{p,n}(r_1) \subset B_{2,n}(r_2)$ iff $n^{1-2/p\vee 2}r_1^2 \le r_2^2$, this requires that $2^{j(1-2/p\vee 2)}C^2 2^{-2sj} \le \frac{1}{3}\varepsilon^2 2^j\gamma_j$ for $j \ge j_0$, which in turn amounts to the requirement that

$$(2s + 2/2\vee p)j_0 + \log_2\gamma_{j0}/3 \ge (2\sigma+1)j_\varepsilon. \tag{A.29}$$

Using (A.24), (A.26), and $j_* \le j_\varepsilon$,

$$S_{21\varepsilon} \le c\sum_{j>j_0}\Omega^2(2\varepsilon t_j^F;\Theta^{(j)}) \tag{A.30}$$

$$\le c\varepsilon^{2r}C^{2-2r}\sum_{j>j_0} j^r 2^{-2\eta|j-j_\varepsilon|} = o(\varepsilon^{2r}), \tag{A.31}$$

as long as $j_0 - j_\varepsilon \to \infty$.

Suppose that γ_j equals either $3.2^{-\gamma j}$ (for some fixed $\gamma \in (0, \frac{1}{2})$) or $j^{3/2}2^{-j/2}$. In either case, set $j_0 = aj_\varepsilon$. Condition (A.23) is satisfied if $a < 2$. Condition (A.29) holds if $a \geq \underline{a} = (2\sigma + 1)/(2s - \gamma + 2/2 \vee p)$. Simple algebra shows that it is possible to choose a to satisfy the two conditions simultaneously as long as $s > |1/p - \frac{1}{2}| + \gamma - \frac{1}{2}$.

Proof of Theorem 5

We first recall that the risk of the positive-part James–Stein estimator is no larger than that of the original James–Stein estimator, $\bar{\boldsymbol{\mu}}^{\mathrm{JS}}$, in which the restriction that the shrinkage coefficient be positive is dropped. Then using Stein's (1981) unbiased estimate of risk (or, alternatively, Lehmann 1983, p. 300) and Jensen's inequality, we have for $d > 2$,

$$\begin{aligned} E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\|_2^2 \leq E_{\boldsymbol{\mu}}\|\bar{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\|_2^2 &= d - (d-2)^2 E_{\boldsymbol{\mu}}\|\mathbf{x}\|_2^{-2} \\ &\leq d - (d-2)^2/(\|\boldsymbol{\mu}\|_2^2 + d). \end{aligned}$$

By direct calculation,

$$E_{\boldsymbol{\mu}}\|\bar{\boldsymbol{\mu}}^{\mathrm{IS}} - \boldsymbol{\mu}\|_2^2 = \|\boldsymbol{\mu}\|_2^2/(\|\boldsymbol{\mu}\|_2^2 + d). \tag{A.32}$$

The difference of the two expressions is thus bounded by $(2d - 4)/(\|\boldsymbol{\mu}\|_2^2 + d) \leq 2$.

Proof of Theorem 6

We use a notation and decomposition similar to that of the proof of Theorem 3. Thus

$$\begin{aligned} E_\theta\|\hat{\theta}^{\mathrm{WJS}} - \theta\|^2 &= 2^L\varepsilon^2 + \varepsilon^2 \sum_{j=L}^{J} E\|\hat{\mu}^{\mathrm{JS}}(x_j) - \mu_j\|^2 + \sum_{j>J}\|\theta_j\|^2 \\ &\leq O(\varepsilon^2) + T_{1\varepsilon} + T_{2\varepsilon}. \end{aligned}$$

The maximum over $\Theta_{p,q}^s$ of the tail term $T_{2\varepsilon}$ is given via the tail n-width,

$$\Delta(\varepsilon, \|\cdot\|; \Theta) = \sup\{\|\theta\| : \theta \in \Theta, \theta_{jk} = 0 \text{ if } j \leq J(\varepsilon)\}$$

(recall that $J(\varepsilon) = \log_2 \varepsilon^{-2}$). From theorem 5 of Donoho et al. (1994), we have

$$T_{2\varepsilon} = \Delta^2(\varepsilon, \|\cdot\|_2; \Theta_{p,q}^s) \leq c(\varepsilon^2)^{(\sigma - (1/p - 1/2)_+)}.$$

Applying (21) to $T_{1\varepsilon}$ yields

$$T_{1\varepsilon} \leq 2J\varepsilon^2 + \varepsilon^2 \sum_{j=L}^{J} E\|\hat{\boldsymbol{\mu}}^{\mathrm{IS}}(\mathbf{x}_j) - \boldsymbol{\mu}_j\|^2. \tag{A.33}$$

Write $\hat{\theta}_{\mathrm{DL}}$ for a diagonal linear estimate of the form $\hat{\theta}_j(y) = c_j y_j$. By the definition of the ideal estimator, the second term on the right of (A.33) is bounded for $\theta \in \Theta_{p,q}^s$ by

$$\inf_{\hat{\theta}_{\mathrm{DL}}} E\|\hat{\theta}_{\mathrm{DL}} - \theta\|^2 \leq R_L^*(\varepsilon, \Theta_{p,q}^s).$$

From other work (Donoho and Johnstone 1995, sec. 6), it is known that $R_L^*(\varepsilon, \Theta_{p,q}^s) \sim \varepsilon^{2r'}$, where $r' = \sigma/(\sigma + 1/2)$ if $p \geq 2$, and $s/(s + 1/2)$ if $p \leq 2$. Thus in particular $r' < 1$, and so all other remainder terms are $o(\varepsilon^{2r'})$, and the proof is complete.

[Received June 1993. Revised January 1995.]

REFERENCES

Brockmann, M., Gasser, T., and Herrmann, E. (1993), "Locally Adaptive Bandwidth Choice for Kernel Regression Estimators," *Journal of the American Statistical Association,* 88, 1302–1309.

Brown, L. D., and Low, M. G. (1995), "Asymptotic Equivalence of Nonparametric Regression and White Noise," submitted to *The Annals of Statistics.*

Bruce, A. G., and Gao, H.-Y. (1995), S + WAVELETS Toolkit Statistics, a division of MathSoft Inc., Seattle WA.

Cleveland, W. S. (1979), "The Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association,* 74, 829–836.

Cohen, A., Daubechies, I., Jawerth, B., and Vial, P. (1993), "Multiresolution Analysis, Wavelets, and Fast Algorithms on an Interval," *Comptes Rendus Academie des Sciences, Paris (A),* 316, 417–421.

Daubechies, I. (1988), "Orthonormal Bases of Compactly Supported Wavelets," *Communications in Pure and Applied Mathematics,* 41, 909–996.

——— (1992), *Ten Lectures on Wavelets,* CBMS-NSF Series in Applied Mathematics, No. 61, Philadelphia: SIAM.

DeVore, R. A., and Popov, V. A. (1988), "Interpolation of Besov Spaces," *Transactions of the American Mathematical Society,* 305, 397–414.

Donoho, D. (1992), "Interpolating Wavelet Transforms," Technical Report 408, Stanford University, Dept. of Statistics.

Donoho, D. L., and Johnstone, I. M. (1994a), "Ideal Spatial Adaptation via Wavelet Shrinkage," *Biometrika,* 81, 425–455.

——— (1994b), "Minimax Risk Over l_p-Balls for l_q-Error," *Probability Theory and Related Fields,* 99, 277–303.

——— (in press), "Neo-Classical Minimax Problems, Thresholding, and Adaptation," *Bernoulli.*

——— (1995), "Minimax Estimation via Wavelet Shrinkage," submitted to *The Annals of Statistics.*

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1994), "Universal Near Minimaxity of Wavelet Shrinkage," technical report, Stanford University, Dept. of Statistics.

——— (1995), "Wavelet Shrinkage: Asymptopia?" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 57, 301–369.

Efroimovich, S., and Pinsker, M. (1984), "A Learning Algorithm for Nonparametric Filtering," *Automat. i Telemeh.,* 11, 58–65 (in Russian).

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting, Variable Bandwidth, and Spatial Adaptation," *Journal of the Royal Statistical Society,* Ser. B, 57, 371–394.

Frazier, M., Jawerth, B., and Weiss, G. (1991), *Littlewood–Paley Theory and the Study of Function Spaces,* NSF-CBMS Regional Conference Series in Mathematics, 79, Providence, RI: American Mathematical Society.

Friedman, J. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics,* 19, 1–67.

Golubev, G. K., and Nussbaum, M. (1990), "A Risk Bound in Sobolev Class Regression," *The Annals of Statistics,* 18, 758–778.

Johnstone, I. M., and Hall, P. G. (1992), "Empirical Functionals and Efficient Smoothing Parameter Selection" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 54, 475–530.

Johnstone, I., Kerkyacharian, G., and Picard, D. (1992), "Estimation d'une Densité de Probabilité par Méthode d'Ondelettes," *Comptes Rendus Academie des Sciences Paris (A),* 315, 211–216.

Lehmann, E. L. (1983), *Theory of Point Estimation,* New York: John Wiley.

Li, K. C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-Validation," *The Annals of Statistics,* 13, 1352–1377.

Mallatt, S. G. (1989a), "The Multifrequency Channel Decompositions of Images and Wavelet Models," *IEEE Transactions on Acoustic Signal Speech Processes,* 37, 2091–2110.

——— (1989b), "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11, 674–693.

Meyer, Y. (1990), *Ondelettes et Opérateurs; I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III:* (with R. Coifman) *Opérateurs multilinéaires,* Paris: Hermann. (English translation of Vol. I published by Cambridge University Press).

——— (1991), "Ondelettes sur l'Intervalle," *Revista Matematica Iberoamericana,* 7, 115–133.

Nemirovskii, A. (1985), "Nonparametric Estimation of Smooth Regression Function," *Izv. Akad. Nauk. SSR Teckhn. Kibernet,* 3, 50–60 (in Russian), *Journal of Computer and System Sciences,* 23, 1–11 (in English).

Nemirovskii, A., Polyak, B., and Tsybakov, A. (1985), "Rate of Convergence of Nonparametric Estimates of Maximum-Likelihood Type," *Problems of Information Transmission,* 21, 258–272.

Nussbaum, M. (1985), "Spline Smoothing and Asymptotic Efficiency in l_2," *The Annals of Statistics,* 13, 984–997.

Peetre, J. (1975), *New Thoughts on Besov Spaces,* Duke University, Durham, NC: Dept. of Mathematics.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics,* 9, 1135–1151.

Stone, C. (1982), "Optimal Global Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics,* 10, 1040–1053.

Strang, G. (1989), "Wavelets and Dilation Equations: A Brief Introduction," *SIAM Review,* 31, 614–627.

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

——— (1990), *Theory of Function Spaces II,* Basel: Birkhäuser Verlag.