Adapting to Unknown Smoothness via Wavelet Shrinkage

David L. Donoho
Iain M. Johnstone
Department of Statistics
Stanford University

July 20, 1994

Abstract

We attempt to recover a function of unknown smoothness from noisy, sampled data. We introduce a procedure, *SureShrink*, which 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, the reconstruction (essentially) does also; if the unknown function has a smooth piece, 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. Minimax Decision theory. Stein's Unbiased Estimate of Risk. James-Stein Estimator. Besov, Hölder, Sobolev, Triebel Spaces. Nonlinear Estimation. Thresholding of Wavelet Coefficients. White Noise Model. Nonparametric regression. Orthonormal Bases of Compactly Supported Wavelets.

Acknowledgements. The first author was supported at U.C. Berkeley by NSF DMS 88-10192, by NASA Contract NCA2-488, and by a grant from ATT Foundation. The second author was supported in part by NSF grants DMS 84-51750, 86-00235, and NIH PHS grant GM21215-12, and by a grant from ATT Foundation.

Presented as "Wavelets + Decision Theory = Optimal Smoothing" at "Wavelets and Applications" Workshop, Luminy, France, March 10, 1991, and at Workshop on "Trends in the Analysis of Curve Data" University of Heidelberg, March 22, 1991.

Contents

1 Introduction

Suppose 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, i.e. to find an estimate $\hat{\mathbf{f}}$ depending on $y_1, \ldots, y_N$ with small *risk* $R(\hat{\mathbf{f}}, \mathbf{f}) = N^{-1} \cdot E||\hat{\mathbf{f}} - \mathbf{f}||_2^2 = E\,\mathrm{Ave}_i(\hat{f}(t_i) - f(t_i))^2$.

In order 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 which are of considerable interest: Stone (1982), Nussbaum (1985), Nemirovskii, Polyak, and Tsybakov (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 m-th derivative):

$$W_2^m(C) = \{f : ||\frac{d^m}{dt^m}f||_2 \leq C\}.$$

Work of Efroimovich and Pinsker (1984) and Nussbaum and Golubev (1990), for example, shows how to construct estimates which 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

$$W_p^m(C) = \{f : ||\frac{d^m}{dt^m}f||_p \leq C\}.$$

If $p < 2$, linear methods cannot attain the optimal rate of convergence over such a class when m and C are known (Nemirovskii, 1985), (Donoho and Johnstone, 1992a). 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. It is based on new results in multivariate normal decision theory which are interesting in their own right.

SureShrink has the following ingredients:

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 $\eta_t(y) = \mathrm{sgn}(y)(|y| - t)_+$ denote the *soft threshold* which sets to zero data y below t in absolute value, and which pulls other data towards 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^*.

3. *Stein's Unbiased Estimate of Risk for Threshold Choice.* The level-dependent thresholds are arrived at 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 $y_{j,k} = w_{j,k} + \epsilon z_{j,k}$, $k = 0, \ldots, 2^j - 1$ 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 slight modification of this recipe is employed in case the data vector has a very small ℓ_2 norm, in which case the the Unbiased risk estimate is very noisy and a fixed threshold is employed).

We briefly describe some examples of the method in action. Figure 1 depicts four specific functions f which we will wavelet-analyze repeatedly in this paper.

(1.a) *Blocks.* A piecewise constant function, with jumps at $\{.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81\}$.

(1.b) *Bumps.* A sum of bumps $\sum_{i=1}^{1} 1h_ib((t-t_i)/s_i)$ with locations t_i at the same places as jumps in *Blocks*; the heights h_i and widths s_i vary; the individual bumps are of the form $b(t) = 1/(1+t^4)$.

(1.c) *HeaviSine.* A sinusoid of period 1 with two jumps, at $t_1 = .3$ and $t_2 = .72$.

(1.d) *Doppler.* The variable frequency signal $f(t) = \sqrt{t(1-t)}sin(2\pi \cdot \frac{1.05}{t+.05})$.

Precise formulas appear in Table 1. These examples have been chosen to represent various *spatially inhomogeneous* phenomena. We regard *Blocks* as a caricature of the acoustic 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* in this case. The results are qualitatively appealing; the reconstructions jump where the true object jumps; the reconstructions 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 which shows 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 whatever is 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 high order wavelets (i.e. 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 directly proportional to the smoothness of the wavelet employed.]

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, we give in Figure 4 an example of what can be accomplished by a representative adaptive linear method. The method applies the James-Stein shrinker (which may be interpreted as an adaptive linear shrinker, see Section 4.1 below) to Dyadic Fourier Corona, or "Littlewood-Paley Blocks". The method is related, as we describe in section 4.2 below, to the proposal of Efroimovich and Pinsker

(1984). The method 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 emphasise that the method *SureShrink* described in this paper differs from variants *RiskShrink* and *VisuShrink* discussed in DJ(1992c) and Donoho, Johnstone, Kerkyacharian and Picard (1993) only in the choice of threshold. Through use of a data based choice of threshold, *SureShrink* is more explicitly adaptive to unknown smoothness and has better large sample mean square error 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 $N = 2^n$. 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 construction relies heavily on concepts in Daubechies (1992), Meyer (1990), (1991) and Cohen et. al. (1993). 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 employ 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 are made in Section 4.6 of DJ(1992c).

A crucial detail: the transform is implemented not by matrix multiplication, but by a sequence of special finite-length filtering steps which result in an order $O(N)$ transform. The choice of wavelet transform is essentially a choice of filter. See Strang (1989) and Daubechies (1992).

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

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

the remaining element we label $w_{-1,0}$. 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.

In the case $M > 0$, the building blocks of the transform are smoother than square waves. In that case, the vector $\mathbf{W}_{j,k}$, plotted as a function of i, has a continuous, wiggly, localized appearance which motivates the label "wavelet". For j and k bounded away from extreme cases by the condition

$$L < j << n, \qquad 0 << k << 2^j, \tag{2}$$

we have the approximation

$$\sqrt{N} \cdot \mathbf{W}_{j,k}(i) \approx 2^{j/2}\psi(2^j t) \qquad t = i/N - k2^{-j}, \tag{3}$$

where ψ is the mother wavelet arising in a wavelet transform on $I\!R$, as described in Daubechies (1988,1992). This approximation improves with increasing N. ψ is an oscillating function of compact support. We therefore 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 (number of derivatives) of ψ; the smoothness is proportional to M.

The vectors $\mathbf{W}_{j,k}$ outside the range of (2) 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 $j \geq L$ which have k near the boundaries 0 and 2^j. These cases fail to satisfy (3). If the transform is periodized, this is because $\mathbf{W}_{j,k}$ is actually approximated by dilation of 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 5 displays $\mathbf{W}_{j,k}$ for $j = 6$, $k = 32$ (and $N = 2048$), in four specific cases: (1.a) Haar Wavelet $L = 0, M = 0$; (1.b) Daubechies D4 Wavelet $L = 2, M = 2$; (1.c) Coiflet C3 $M = 9$; (1.d) Daubechies "Nearly Linear Phase" S8 Wavelet $M = 9$. The smoother wavelets have broader support.

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

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

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

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

denote the partial reconstruction from terms at resolution level j, or scale 2^{-j}. Then $\mathbf{x}$ can be recovered from these components via $\mathbf{x} = V_L\mathbf{x} + \sum_{L \leq j < n} W_j\mathbf{x}$, 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 $j = L, L+1, ..., n-1$. In Figure 6, we do this for our 4 functions and the $S8$ wavelet. In Figure 7, for contrast we look just at the *Blocks* and *HeaviSine* functions to see how the Haar Transform behaves (Figs. 7.a and 7.b); and how the Daubechies D4 Transform behaves (Figs. 7.c and 7.d).

A less usual way to display wavelet transforms is to look at the wavelet coefficients directly. We do this in Figure 8. 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 coefficents 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 about 1900 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 inhomogeneities 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 9 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 not very pronounced for object *HeaviSine* – much less so than for the $S8$-based transform.

2.2 Thresholding of Noisy Wavelet Coefficients

The orthogonality of the 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

$$y_{j,k} = w_{j,k} + z_{j,k} \tag{4}$$

where $z_{j,k}$ is an i.i.d. $N(0, \sigma^2)$ 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 data compression remarks above were meant to create in the reader the mental picture 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 which "kills" small $y_{j,k}$ and "keeps" large $y_{j,k}$. The particular soft thresholding scheme we introduced above is an instance of this.

Figure 3 has already shown the results such a scheme can provide, in the case of the S8 Wavelet Transform. To illustrate how this works in the wavelet domain, we display in Figure (10.c) the Haar transform of a noisy version of *Blocks*. We also display a thresholded version of this transform (10.d), as well as the raw data (10.a) and the reconstruction (10.b).

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

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

$$E_\mu \|\hat{\mu}(\mathbf{x}) - \mu\|^2 = d + E_\mu\{\|\mathbf{g}(\mathbf{x})\|^2 + 2\nabla \cdot \mathbf{g}(\mathbf{x})\}, \tag{5}$$

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

Now consider the soft threshold estimator $\hat{\mu_i}^{(t)} = \eta_t(x_i)$, and apply Stein's result. $\hat{\mu}^{(t)}$ is weakly differentiable in Stein's sense, and so we get from (5) that the quantity

$$SURE(t; \mathbf{x}) = d - 2 \cdot \#\{i : |x_i| \le t\} + \sum_{i=1}^d (|x_i| \wedge t)^2. \tag{6}$$

is an unbiased estimate of risk: $E_\mu \|\hat{\mu}^{(t)}(\mathbf{x}) - \mu\|^2 = E_\mu SURE(t; \mathbf{x})$.

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

$$t^S = \text{arg min}_{t \ge 0} \, SURE(t; \mathbf{x}). \tag{7}$$

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.

Computational evidence that t^S is a reasonable threshold selector is given in Figure 11. A vector μ of dimension $d = 128$ consists of 16 consecutive 4's, followed by all zeros. White Gaussian noise of variance 1 was added (11.c). The profile of SURE(t) is displayed in (11.a); it resembles quite closely the actual loss (11.b), which we of course know in this (artificial) example. The SURE principle was used to select a threshold which is applied to the data resulting in estimate an estimate of the mean vector (11.d). This estimate is sparse and much less noisy than the raw data (11.c). Note also the shrinkage of the non-zero part of the signal.

The optimization problem (7) 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 which lie between two values of $|x_i|$, $SURE(t)$ is strictly increasing. 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 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* employs a Hybrid scheme.

Figure (12.a) depicts results of a small-scale simulation study. A vector μ of dimension $d = 1024$ contained $\lfloor \epsilon \cdot d \rfloor$ 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 $\epsilon = \{.005, .01, .02(.02).20, .25\}$ were studied. 25 replications were tried at each parameter combination, and the root mean squared errors were displayed in the Figure. Evidently, the root MSE does not tend to zero linearly as the sparsity tends to 0. For the theoretical results of section 3, such behavior would be unacceptable.

In contrast, Figure (12.b) 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 $\epsilon >> 0$, but much smaller for ϵ near zero. The rationale for the choice $\sqrt{2 \log(d)}$ is developed by the authors thoroughly in [DJ92b].

Figure (12.c) displays the results of applying a hybrid method which 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.

In detail, the Hybrid method works as follows: Let $\eta_d = \log_2(d)^{3/2}$ and define $s_d^2 = d^{-1} \sum_i (x_i^2 - 1)$. 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 with respect to the respective subsets of indices, only with an additional restriction on the search range:

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

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

$$\hat{\mu}^*(\mathbf{x})_i = \begin{cases} \eta_{t_d^F}(x_i) & s_d^2 \le \eta_d/\sqrt{d} \\ \eta_{t_I^S}(x_i) & i \in I' \text{ and } s_d^2 > \eta_d/\sqrt{d} \\ \eta_{t_{I'}^S}(x_i) & i \in I \text{ and } s_d^2 > \eta_d/\sqrt{d} \end{cases} \quad (8)$$

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 non-negligible, we set the threshold to $\sqrt{2\log(d)}$.

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

$$\hat{\mu}^+(\mathbf{x})_i = \begin{cases} \eta_{t_d^F}(x_i) & s_d^2 \le \eta_d/\sqrt{d} \\ \eta_{t^S}(x_i) & s_d^2 > \eta_d/\sqrt{d} \end{cases}$$

offers the same performance benefits in simulations. See Figure (12.d).

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* $N = 2^n$ *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*

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

the estimator $\hat{\mathbf{f}}^*$ *derives from this via inverse discrete Wavelet transform.*

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 $N = 4096$.

3 Main Result

In this section we investigate the adaptivity of *SureShrink* to unknown degree of smoothness. To state our result, we must define Besov spaces. We follow De Vore and Popov (1988). Let $\Delta_h^{(r)} f$ denote the r-th difference $\sum_{k=0}^r \binom{r}{k}(-1)^k f(t+kh)$. The r-th 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 defined 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* $B^\sigma_{p,q}(C)$ is then the class of functions $f : [0,1] \to \mathbb{R}$ satisfying $f \in L^p[0,1]$ and $|f|_{B^\sigma_{p,q}} \le C$. Standard references on Besov spaces are Peetre (1976) and Triebel (1983).

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^m_2 the L^2 Sobolev space as in the introduction, $f \in W^m_2$ iff $|f|_{B^m_{2,2}} < \infty$.

The Besov scale essentially includes other less traditional spaces as well. For example, 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^m_p contain $B^m_{p,1}$ and are contained in $B^m_{p\infty}$.

Theorem 1 *Let the discrete wavelet analysis correspond to a wavelet ψ having r null moments and r continuous derivatives, $r > \max(1, \sigma)$. Let the minimax risk be denoted by*

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

Then, **SureShrink** *is simultaneously 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 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 which 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 the introduction that no linear estimator achieves the optimal rate over 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 $\le C$. Let now $\hat{f}^*$ denote the applicationof SureShrink in the Haar basis. 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 10 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 [DJ92a]. 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. Due to (3), the empirical wavelet coefficient $y_{j,k} = w_{j,k} + z_{j,k}$, 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 $\theta_{j,k} = \int f(t)\psi_{j,k}(t)dt$, then, it is tempting to act as though our observations were actually

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

or, what is the same thing,

$$\theta_{j,k} + \epsilon z_{j,k};$$

where $\epsilon = \frac{\sigma}{\sqrt{N}}$ and now $z_{j,k}$ is a standard i.i.d. $N(0,1)$ sequence. Moreover, due to the Parseval relation $\|\hat{\mathbf{f}} - \mathbf{f}\|_2 = \|\hat{\mathbf{w}} - \mathbf{w}\|_2$ and the above 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

$$y_{j,k} = \theta_{j,k} + z_{j,k} \qquad j \geq 0, k = 0, \ldots, 2^j - 1, \tag{9}$$

where $z_{j,k}$ are i.i.d. $N(0, \epsilon^2)$ 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 $\Theta(s, p, q, C)$ 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 $\hat{\theta}$ which is simultaneously nearly minimax over a range of $\Theta(s, p, q, C)$.

Suppose 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 which 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. The approximation arguments necessary to establish this correspondence are discussed in [DJ92a] and for reasons of space we omit them. See also Brown and Low (1992).

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 which [DJ92a] call *Besov Bodies*. These are sets $||\theta||_{\mathbf{b}^s_{p,q}} \leq C$, where

$$||\theta||_{\mathbf{b}^s_{p,q}} = \left(\sum_{j \geq 0} \left(2^{js} \left(\sum_{0 \leq k < 2^j} |\theta_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q}. \tag{10}$$

Consider the problem of estimating θ when it is observed in a Gaussian white noise, and is known *a priori* to lie in a certain convex set $\Theta^s_{p,q}(C) \equiv \{\theta : ||\theta||_{\mathbf{b}^s_{p,q}} \leq C\}$. 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*

$$R^*(\epsilon; \Theta^s_{p,q}) = \inf_{\hat{\theta}} \sup_{\Theta^s_{p,q}} E||\hat{\theta} - \theta||_2^2 \tag{11}$$

and the minimax risk among threshold estimates is

$$R^*_T(\epsilon; \Theta^s_{p,q}) = \inf_{(t_j)} \sup_{\Theta^s_{p,q}} E||\hat{\theta}_{(t_j)} - \theta||_2^2 \tag{12}$$

where $\hat{\theta}_{(t_j)}$ stands for the estimator $(\eta_{t_j}(y_{j,k}))_{j,k}$. [DJ92a] shows that $R^*_T \leq \Lambda(p) \cdot R^* \cdot (1 + o(1))$ with, e.g. $\Lambda(1) \approx 1.6$. Hence threshold estimators are nearly minimax. Furthermore, [DJ92a] show that the minimax risk and minimax threshold risk over sets $\Theta^s_{p,q}(C)$ is equivalent, to within constants, so that over sets $\Theta(\sigma, p, q, C)$, provided σ is large enough, and we make the calibration $s = \sigma + 1/2 - 1/p$.

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

$$\begin{aligned} \hat{\theta}^*_{j,k}(\mathbf{y}) &= y_{j,k}, \qquad j < L, & (13) \\ \hat{\theta}^*_{j,k}(\mathbf{y}) &= \epsilon \cdot \hat{\mu}^*(\mathbf{x}_j) \qquad j \geq L. & (14) \end{aligned}$$

This is a particular adaptive threshold estimator.

Theorem 3 *Let $s > 1/p - 1/2$. Then*

$$\sup_{\Theta^s_{p,q}(C)} E_\theta ||\hat{\theta}^* - \theta||_2^2 \leq R^*_T(\epsilon; \Theta^s_{p,q}(C))(1 + o(1)) \qquad \epsilon \to 0.$$

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

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

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 estimator $\hat{\mu}^*$ almost performs as if we did know this ideal threshold.

Theorem 4 *(a) Uniformly in* $\mu \in I\!R^d$

$$d^{-1} E_\mu ||\hat{\mu}^* - \mu||_2^2 \leq \tilde{R}(\mu) + c(\log(d))^{5/2} d^{-1/2}.$$

(b) For any given $\gamma > 0$*, uniformly in* $d^{-1}\sum_i \mu_i^2 \leq \frac{1}{3}\eta_d d^{-1/2}$ *we have*

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

4 Comparison with Adaptive Linear Estimates

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

4.1 Adaptive Linear Estimation via James-Stein

In the multivariate normal setting of Section 2.3, the James-Stein (positive part) estimate is

$$\hat{\mu}_i^{JS} = c^{JS}(\mathbf{x}) \cdot x_i, i = 1, \ldots, d$$

where the shrinkage coefficient $c^{JS}(\mathbf{x}) = (\|\mathbf{x}\|_2^2 - (d-2))_+ / \|\mathbf{x}\|_2^2$. Among all linear estimators $\hat{\mu} = c \cdot \mathbf{x}$, the one with smallest risk at μ uses the coefficient

$$\tilde{c}(\mu) = \|\mu\|_2^2 / (\|\mu\|_2^2 + d).$$

Since μ is unknown (it is after all the quantity we are trying to estimate), this linear shrinker represents an unattainable ideal. From $E\|\mathbf{x}\|_2^2 = \|\mu\|_2^2 + d$ we see the James-Stein shrinkage coefficient $c^{JS}(\mathbf{x})$ is essentially an estimate of the ideal shrinkage coefficient $\tilde{c}$.

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

Theorem 5 *Consider the ideal estimator (not a statistic!)* $\tilde{\mu}^{IS}(\mathbf{x}) = \tilde{c}(\mu)\mathbf{x}$*. For all* $d > 2$*, and for all* $\mu \in R^d$

$$E_\mu ||\hat{\mu}^{JS} - \mu||_2^2 \leq 2 + E_\mu ||\tilde{\mu}^{IS} - \mu||_2^2$$

We pay a price of at most 2 for using the James-Stein shrinker rather than the ideal shrinker. In high dimensions d, this price is negligible.

Apply James-Stein in the wavelet domain:

$$(\hat{w}_{j,k})_{0\leq k<2^j} = \hat{\mu}^{JS}(\mathbf{x}_j).$$

Inverting the wavelet transform gives an estimate $\hat{f}^{WJS}$ which we call *WaveJS*.

A number of nice adaptivity properties of *WaveJS* follow immediately from Theorem 4. Consider the ideal linear shrinkage estimator (again not a statistic)

$$(\tilde{w}_{j,k})_{0\leq k<2^j} = \tilde{\mu}^{WIS}(\mathbf{x}_j),$$

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

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

It is not hard to see that for every space in the Besov scale covered by Theorem 1, the ideal estimator achieves within a constant factor of the minimax risk for *linear* estimators. Moreover the minimax risk measured as above behaves like N^{-r} for a certain $r \in [0,1]$. It is not however, a statistic; the James-Stein estimate is a statistic; and because $4\log_2(N)/N = o(N^{-r})$, it follows that $\hat{\mathbf{f}}^{WJS}$ 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.

However, theory aside, such an estimate is not very good in practice. Figure 13 gives an example on the same cases as figures 1-3. The *WaveJS* reconstruction is much noisier than *SureShrink*. This could be seen in the display of wavelet coefficients; if in one resolution level there are significant coefficients which 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, since it operates on coordinates at each level by the same multiplicative factor.

4.2 Linear Adaptation using Fourier Coronae

Suppose we identify 0 with 1, so that $[0,1]$ has a circular interpretation. Work by Efroimovich and Pinsker (1984), and other recent Soviet literature, would consider the use of adaptive linear estimators based on empirical Fourier Coefficients $(\hat{v}_\ell)$. One divides the frequency domain into coronae $\ell_i \leq \ell < \ell_{i+1}$, and within each corona, one uses a linear shrinker

$$\hat{f}_\ell = c_i \cdot \hat{v}_\ell \qquad \ell_i \leq \ell < \ell_{i+1}$$

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

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

We propose an adaptive linear scheme which differs from the Efroimovich-Pinsker choice in two ways. First, we propose to use Dyadic coronae $\ell_i = 2^{i+L}$. Such Dyadic Fourier Coronae occur frequently in Littlewood-Paley theory: Frazier, Jawerth, and Weiss (1991), Peetre (1976), Triebel (1983). Second, within each corona, we shrink via the James-Stein estimate $c_i = c^{JS}(\mathbf{v}_i)$, which has nicer theoretical properties than the Efroimovich-Pinsker choice. The estimator 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 $\frac{4\log_2(N)}{N}$ 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 Coronae, has performance not worse than a constant factor times the performance of so-called Pinkser weights, and hence we conclude that, except for constants, *LPJS* replicates the adaptive-rate advantages of the Efroimovich-Pinsker choice of coronae. *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.

However, we have already seen the *LPJS* reconstructions in Figure 4. 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, Jawerth, and Weiss, 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 8 different methods to noisy versions of the data: *SureShrink* in the Haar, Db4, C3, and S8 Wavelet bases, *WaveJS* in the S8 Wavelet Basis, *LPJS*, and finally the procedure *RiskShrink* [DJ 1992c] using the C3 and S8 wavelet bases (denoted "ThrC3" and "ThrS8"). *RiskShrink* uses a fixed threshold chosen to yield minimax performance for mean square error against an 'oracle'. These threshold values are tabulated in [DJ 1992c]. Dyadic sample sizes $N = 2^n$, from $N = 128$ to $N = 16,384$ were studied.

Sample results are given in Table 2, which reports the root-loss $N^{-1/2}\|\hat{\mathbf{f}} - \mathbf{f}\|_2$ (not its square). We decided not to report risk (i.e. loss averaged over an ensemble of realizations), because replications told nearly the same story.

In all examples, there is little quantitative difference bewteen the methods at small N. There is a visual difference, however. For large N, *SureShrink* with the C3 and S8 wavelets consistently outperforms the linear adaptive shrinkers, obtaining equivalent precisions with half or less than half the sample size. The most extreme case is object *Blocks*, where the performance of shrinkage in the Haar basis at sample size 1024 is comparable to the performance of *LPJS* at sample size 8192. The results for *SureShrink* and *RiskShrink* are remarkably similar here.

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 tradeoff between noise and bias. The ℓ^2 norm balances these in equilibrium, which insures that some noise artifacts will be visible.

Enhanced visual quality can be obtained by keeping the noise term in the tradeoff 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 $\sqrt{2\log(N)}$ rule (called *VisuShrink* and the connection with optimum "Visual Quality" may be found in [DJ92b].

5.3 Hard Thresholding

Many persons have asked us if it would be possible to use "Hard Thresholding"

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

in place of soft thresholding η_t. Indeed, Hard thresholding seems more natural to non-statisticians. We prefer soft thresholding because of various statistical advantages (continuity of the rule; simplicity of the SURE formula). However, in principle, the results above could have equally well been derived for Hard Thresholding. A more complicated SURE formula would be required to implement the idea on data. The proofs would also be more complicated. The resulting estimator might be called *WaveChop*.

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: $\hat{\sigma} = Median(|y_{n-1,k}| : 0 \leq k < 2^{n-1})$. We believe 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

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 there. Nor have we discussed earlier applications of SURE with linear estimates; compare Li (1986) and references there. Finally we have not discussed applications of wavelet thresholding in Density Estimation – Johnstone, Kerkyacharian, Picard (1992). All of these are interesting topics which we have omitted for reasons of space.

6 Appendix: Proofs of Theorems 3, 4 and 5

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-square variates:

A) Let $Y_1, \cdots, Y_n$ 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 $t > 0$,

$$P\{|\bar{Y}_n - \mu| \ge t\} \le 2\exp\{-2n^2t^2/\sum_1^n (b_i - a_i)^2\}. \tag{15}$$

B) Let $X_1, \cdots, X_m$ be sampled *without replacement* from $\{c_1, \cdots, c_n\}$. 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 $t > 0$,

$$P\{|\bar{X}_m - \mu| \ge t\} \le 2\exp\{-2nt^2/(b-a)^2\}. \tag{16}$$

C1) Let $Z_1, \ldots, Z_n$ be i.i.d $N(0,1)$. Then by elementary arguments,

$$P\{|\Sigma\alpha_j(z_j^2 - 1)| > t\} \le 2e^{2s^2\Sigma\alpha_j^2 - |s|t} \quad \text{for} \quad |s| \le 1/(4\max(|\alpha_j|)).$$

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

$$P\{|n^{-1}\sum(z_j^2 - 1)| > t\} \le 2e^{-nt(t\wedge 1)/8}. \tag{17}$$

Preparatory Propositions. We use (A) to bound the deviation of the unbiased risk estimate 6 from its expectation. To recapitulate the setting of Section 2.3, suppose $x_i \sim N(\mu_i, 1)$, $i = 1, \cdots, d$ are independent. Let F_d denote the empirical distribution function of $\{\mu_i\}$. As above, let $r(t, \mu_i) = E[\eta_t(x_i) - \mu_i]^2$ denote the mean squared error of the soft threshold estimate of a single co-ordinate, and define

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

In particular

$$r(t, F_d) = d^{-1}\Sigma r(t, \mu_i) = d^{-1}E_\mu||\hat{\mu}^{(t)} - \mu||^2. \tag{18}$$

Stein's unbiased risk estimate 6,

$$\begin{aligned} U_d(t) &= d^{-1}\,\mathrm{SURE}\,(t, \mathbf{x}) && (19)\\ &= 1 - 2d^{-1}\sum_i I\{x_i^2 \le t^2\} + d^{-1}\sum_i x_i^2 \wedge t^2 && (20)\\ &= d^{-1}\sum_i 1 - 2I\{x_i^2 \le t^2\} + x_i^2 \wedge t^2, && (21)\end{aligned}$$

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 $\mu \in I\!R^d$,*

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

Proof: Combining (18) and (19) 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 (15) gives, for a fixed t, and (for now) arbitrary $r_d > 1$,

$$P\{|Z_d(t)| \ge r_d d^{-1/2}\} \le 2 \exp\{-r_d^2/2(t^2+2)^2\}. \tag{22}$$

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

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

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

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

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 (15),

$$P\{d^{-1}|N_d(t_j, t_j + \delta_d) - EN_d| \ge r_d d^{-1/2}/3\} \ge e^{-2r_d^2/9}. \tag{23}$$

Finally, using (22), (23) and the cardinality of $\{t_j\}$,

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

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

$$P(A_d) \leq \frac{3t_d}{\delta_d d^b}. \tag{24}$$

Let $||Z_d|| = \sup\{|Z_d(t)| : 0 \leq t \leq t_d\}$ and $r_d^\circ = (2\log d)^{1/2}(t_d^2+2)$. We may rephrase (24) as

$$P_\mu\{\sqrt{d}||Z_d||/r_d^\circ > s\} \leq 3t_d\delta_d^{-1}e^{-s^2\log d}.$$

which suffices for the L_1 convergence claimed. ▮

Proposition 2 *Uniformly in $\mu \in I\!R^d$,*

$$E_I||r(\cdot,F_I) - r(\cdot,F)||_\infty = 0\left(\frac{\log^{3/2} d}{d^{1/2}}\right).$$

Proof. This is similar to that of Proposition 1, but is simpler and uses Hoeffding's inequality (16). In the notation of (16), set $n = d$, $c_i = r(t,\mu_i) \leq 1 + t_d^2$, $m = d/2$, so that $\mu = r(t,F_d), \bar{X}_m = r(t,F_I)$ and

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

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

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

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

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

$$\begin{aligned} P\{||Z_d|| \geq 2r_d d^{-1/2}\} &\leq P\{\sup_{j:j\delta_d \leq t_d} |Z_d(j\delta_d)| \geq r_d d^{-1/2}\} \\ &\leq \frac{2t_d}{\delta_d}\frac{1}{d^{4b}}. \end{aligned}$$

As for Proposition 1, this yields the result. ▮

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

$$\sup_{\tau^2 \geq \eta_d d^{-1/2}} (1+\tau^2)P\{s_d^2 \leq \eta_d d^{-1/2}\} = o(d^{-1/2}). \tag{25}$$

Proof.: This is a simple statement about the tails of the non-central chi-squared distribution. Write $x_i = z_i + \mu_i$ where $z_i \sim N(0,1)$. The event in 25 may be rewritten as

$$A_d = \{d^{-1}\Sigma(z_i^2 - 1) + d^{-1}\Sigma 2\mu_i z_i \le -(\tau^2 - \eta_d d^{-1/2})\} \tag{26}$$
$$\subset \{d^{-1}\Sigma(z_i^2 - 1) \le -\tau^2/3\} \cup \{d^{-1}\Sigma 2\mu_i z_i \le -\tau^2/3\} = B_d \cup C_d$$

from the lower bound on τ^2 in (25).

By elementary inequalities,

$$P(C_d) = \tilde{\Phi}\left(\frac{\tau^2}{3}\frac{d^{1/2}}{2\tau}\right) \le c_1 e^{-c_2 d\tau^2}. \tag{27}$$

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

For B_d, apply the exponential inequality (17) to obtain

$$(1+\tau^2)P(B_d) \le 2(1+\tau^2)\exp\{-d\tau^2(\tau^2 \wedge 3)/72\} \le c_3 \exp\{-c_4\eta_d^2\} = o(d^{-1/2})$$

since $\eta_d \ge \log d$. ∎

Proof of Theorem 4(a). Decompose the risk of $\hat{\mu}^*$ according to the outcome of the pre-test event $A_d = \{s_d^2 \le \eta_d d^{-1/2}\}$, with the goal of showing that

$$R_{1d}(\mu) = d^{-1}E[||\mu^* - \mu||^2, A_d] \le c(\log d)^{5/2}d^{-1/2}, \text{ and} \tag{28}$$
$$R_{2d}(\mu) = d^{-1}E[||\mu^* - \mu||^2, A_d^c] \le \tilde{R}(\mu) + c(\log d)^{5/2}d^{-1/2}. \tag{29}$$

On event A_d, fixed thresholding is used:

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

If $\tau^2 = d^{-1}\sum_1^d \mu_i^2 \le \eta_d d^{-1/2}$, then the oracle inequality of [DJ 92] shows that

$$R_{1d} \le (1 + 2\log d)(d^{-1} + d^{-1}\Sigma\min(1,\mu_i^2)) \le c(\log d)^{5/2}d^{-1/2}.$$

Conversely, if $\tau^2 \ge \eta_d d^{-1/2}$, then we first note that on event A_d,

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

Using Lemma 1, it follows that

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

Under either condition on τ^2, (28) holds true.

On event A_n^c, the adaptive, half-sample based thresholding applies. Let E_μ denote expectation over the distribution of (x_i) and E_I denote expectation over the random choice of half sample I. Then

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

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

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

Thus, to complete the proof of (29), it suffices to show that

$$R_{3d}(\mu) = E_I E_\mu r(\hat{t}_I, F_{I'}) - \tilde{R}(\mu) \leq c (\log d)^{5/2} d^{-1/2}. \tag{30}$$

There is a natural decomposition

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

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

$$r(t, F_I) - r(t, F_d) = \tfrac{1}{2}[r(t, F_I) - r(t, F_{I'})] \tag{31}$$

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

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

using Proposition 2.

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

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

Putting all together, we obtain (30). ▮

Proof of Theorem 4(b). When $||\mu||$ is small, the pretest of $s_d^2 \leq \eta_d d^{-1/2}$ will with high probability lead to use of the fixed threshold t_d^F. The $O(d^{-1/2} \log^{5/2} d)$ 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 μ^* as in (28) and (29), but now $P(A_d) \nearrow 1$ as $d \nearrow \infty$. On A_d, fixed thresholding is used and we exploit the inequalities

$$\begin{aligned} r(t, \mu) &\leq r(t, 0) + \mu^2 \\ r(t, 0) &\leq 4\phi(t) t^{-3} (1 + \tfrac{3}{2} t^2) \end{aligned}$$

proved in [DJ 1992c] to conclude that

$$\begin{aligned} dR_{1d} &\leq \sum_1^d r(t_d^F, \mu_i) \leq dr(t, 0) + ||\mu||^2 \\ &\leq (\log d)^{-3/2} + ||\mu||^2 \end{aligned}$$

for large d.

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

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

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

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

Arguing similarly to (26), we note that the hypothesised bound on μ implies that

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

The chisquare exponential inequality (17) and standard Gaussian inequalities give

$$\begin{aligned} P(B_d) &\leq \exp\{-\log^2 d/72\}, \\ P(C_d) &\leq \exp\{-\eta_d d^{1/2}/24\} \end{aligned}$$

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

Proof of Theorem 3. We make use of the definitions (13) and (14) to write

$$E_\theta ||\hat{\theta}^* - \theta||^2 = 2^L \epsilon^2 + \epsilon^2 \sum_{j > L} E||\hat{\mu}^*(x_j) - \mu_j||^2$$

where $\mu_j = (\mu_{jk}) = (\theta_{jk}/\epsilon)$ and $\theta_j = (\theta_{jk})$. For a $j_0 = j_0(\epsilon, \sigma, p, q) \nearrow \infty$ to be specified below, we use Theorem 4(a) for levels $j \leq j_0$ and Theorem 4(b) for $j > j_0$:

$$\begin{aligned} E_\theta ||\hat{\theta}^* - \theta||^2 &\leq O(\epsilon^2) + S_{1\epsilon} + S_{2\epsilon}, \\ S_{1\epsilon} &\leq \epsilon^2 \sum_{j \leq j_0} \left\{ \inf_{t_j} \sum_k r(t_j, \mu_{jk}) + cj^{5/2} 2^{j/2} \right\} \\ S_{2\epsilon} &\leq \epsilon^2 \sum_{j > j_0} \left\{ ||\mu_j||^2 + cj^{-\frac{3}{2}} \right\} = \sum_{j > j_0} ||\theta_j||^2 + o(\epsilon^2). \end{aligned}$$

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

$$\sup_\Theta S_{1\epsilon} \leq \sup_\Theta \inf_{(t_j)} E||\hat{\theta}_{(t_j)} - \theta||^2 + c\epsilon^2 j_0^a 2^{j_0/2}.$$

The first term on the right side is precisely $R_T^*(\epsilon, \Theta_{p,q}^s) \asymp \epsilon^{2r}$ where $r = 2\sigma/(2\sigma + 1)$, and so it remains to verify that j_0 can be chosen so that all other error terms are $o(\epsilon^{2r})$. The error term in $S_{1\epsilon}$ is negligible if $j_0 + 2a \log j_0 - (2/(2\sigma + 1)) \log \epsilon^{-2} \rightarrow -\infty$. Since $p \leq 2, s_j^2 = \sup\{||\theta_j||^2, \theta \in \Theta_{p,q}^s(C)\} = C^2 2^{-5j}$ and so the term $S_{2\epsilon} \sim 2^{-2sj_0}$ is negligible if $j_0 - (\sigma/s(2\sigma+1)) \log \epsilon^{-2} \rightarrow \infty$. These two requirements on j_0 are compatible if $s > p^{-1} - 2^{-1}$.

Finally, we note that the use of Theorem 4(b) requires that $2^{-j}||\mu_j||^2 \le \frac{1}{3}\eta_{2^j} 2^{-j/2}$ for $j \ge j_0$, which is guaranteed if

$$s_j^2 = C^2 2^{-2sj} \le \epsilon^2 \tfrac{1}{3} j^{3/2} 2^{j/2} \qquad \forall j \le j_0.$$

This holds if $(2s+1/2)j_0 \ge \log \epsilon^{-2}$, which is again compatible with $j_0 << (2/(2\sigma+1)) \log \epsilon^{-2}$ so long as $s > p^{-1} - 2^{-1}$. ∎

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, $\bar{\mu}^{JS}$, 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 $d > 2$,

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

By direct calculation,

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

The difference of (32) and (32) is thus bounded by $(2d-4)/(||\mu||_2^2 + d) \le 2$. ∎

References

[1] Brown, L. D. and Low, M. G. (1990). Asymptotic equivalence of nonparametric regression and white noise. Manuscript.

[2] Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993) Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris* (A). **316**., 417–421.

[3] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comunications in Pure and Applied Mathematics*, **41**, 909–996.

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

[5] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Am. Math. Soc.*

[6] Donoho, D. L. and Johnstone, I. M (1992a). Minimax risk over ℓ_p-balls for l_q loss. Technical Report No. 401, Department of Statistics, Stanford University.

[7] Donoho, D. L. and Johnstone, I. M (1992b). Minimax Estimation via Wavelet Shrinkage. Technical Report No. 402, Department of Statistics, Stanford University.

[8] Donoho, D. L. and Johnstone, I. M (1992c). Ideal Spatial Adaptation via Wavelet Shrinkage. Technical Report, Department of Statistics, Stanford University. Tentatively accepted, *Biometrika* .

[9] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1993). Wavelet Shrinkage: Asymptopia? Technical Report No. 419, Department of Statistics, Stanford University.

[10] Efroimovich, S. Yu. and Pinsker, M.S. (1984) A learning algorithm for nonparametric filtering. *Automat. i Telemeh.* **11** 58-65 (in Russian).

[11] Frazier, M. and Jawerth, B. (1985). Decomposition of Besov spaces. *Indiana Univ. Math. J.*, 777–799.

[12] M. Frazier and B. Jawerth (1990) A discrete Transform and Decomposition of Distribution Spaces. *Journal of Functional Analysis* **93** 34-170.

[13] M. Frazier, B. Jawerth, and G. Weiss (1991) *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, **79**. American Math. Soc.: Providence, RI.

[14] Golubev, G.K. (1987) Adaptive asymptotically minimax estimates of smooth signals. *Problemy Peredatsii Informatsii* **23** 57-67.

[15] Golubev, G.K. and Nussbaum, M. (1990) A risk bound in Sobolev class regression. *Ann. Statist.* **18** (2), 758-778.

[16] Johnstone, I.M. and Hall, P.G. (1992) Empirical functionals and efficient smoothing parameter selection. *J. Roy. Stat. Soc.* B, **54**, to appear.

[17] Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1992) Estimation d'une densité de probabilité par méthode d'ondelettes. To appear *Comptes Rendus Acad. Sciences Paris* (A).

[18] Lehmann, E.L. (1983) *Theory of Point Estimation.* Wiley, New York.

[19] Lemarié, P.G. and Meyer, Y. (1986) Ondelettes et bases Hilbertiennes. *Revista Mathematica Ibero-Americana.* **2**, 1-18.

[20] Li, K.C. (1985) From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.* **13** 1352-1377.

[21] Mallat, S. (1989a). Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Mat. Soc.*, **315**, 69–87.

[22] Mallat, S. (1989b). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

[23] Mallat, S. (1989c). Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 2091–2110.

[24] Meyer, Y. (1990). *Ondelettes.* Paris: Hermann.

[25] Meyer, Y. (1991). Ondelettes sur l'Intervalle. *Revista Mathematica Ibero-Americana.*

[26] Nemirovskii, A.S. (1985) Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet.* **3**, 50-60 (in Russian). *J. Comput. Syst. Sci.* **23**, 6, 1-11, (1986) (in English).

[27] Nemirovskii, A.S., Polyak, B.T. and Tsybakov, A.B. (1985) Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems of Information Transmission* **21**, 258-272.

[28] Nussbaum, M. (1985). Spline smoothing and asymptotic efficiency in L_2. *Ann. Statist.*, **13**, 984–997.

[29] Pinsker, M.S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredatsii Informatsii* **16** 52-68 (in Russian); *Problems of Information Transmission* (1980) 120-133 (in English).

[30] Peetre, J. (1976). *New Thoughts on Besov Spaces.* Duke Univ. Math. Series. Number 1.

[31] Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**, 6, 1135-1151.

[32] Stone, C. (1982). Optimal global rates of convergence for nonparametric estimators. *Ann. Statist.* **10**, 1040-1053.

[33] Triebel, H. (1983) *Theory of Function Spaces.* Birkhäuser Verlag: Basel.

List of Figures

Table 1. Formulas for Test Functions

Blocks.

$$f(t) = \sum h_j K(t - t_j) \qquad K(t) = (1 + \mathrm{sgn}(t))/2.$$

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

Bumps.

$$f(t) = \sum h_j K((t - t_j)/w_j) \qquad K(t) = (1 + |t|^4)^{-1}.$$

$$\begin{array}{llllllllllll}
(t_j) & = t_{Blocks} & & & & & & & & & & \\
(h_j) & = (4, & 5, & 3, & 4, & 5, & 4.2, & 2.1, & 4.3, & 3.1, & 5.1, & 4.2) \\
(w_j) & = (.005, & .005, & .006, & .01, & .01, & .03, & .01, & .01, & .005, & .008, & .005)
\end{array}$$

HeaviSine.

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

Doppler.

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