Open access • Journal Article • DOI:10.1093/BIOMET/81.3.425

Ideal spatial adaptation by wavelet shrinkage — Source link

David L. Donoho, Jain M. Johnstone

Institutions: Stanford University

Published on: 01 Sep 1994 - Biometrika (Oxford University Press)

Topics: Piecewise, Oracle, Spline (mathematics), Wavelet and Estimator

Ideal Spatial Adaptation by Wavelet Shrinkage

David L. Donoho
Iain M. Johnstone
Department of Statistics, Stanford University, Stanford, CA, 94305-4065, U.S.A.

June 1992
Revised April 1993

Abstract

With *ideal spatial adaptation*, an oracle furnishes information about how best to adapt a spatially variable estimator, whether piecewise constant, piecewise polynomial, variable knot spline, or variable bandwidth kernel, to the unknown function. Estimation with the aid of an oracle offers dramatic advantages over traditional linear estimation by nonadaptive kernels; however, it is *a priori* unclear whether such performance can be obtained by a procedure relying on the data alone. We describe a new principle for spatially-adaptive estimation: *selective wavelet reconstruction*. We show that variable-knot spline fits and piecewise-polynomial fits, when equipped with an oracle to select the knots, are not dramatically more powerful than selective wavelet reconstruction with an oracle. We develop a practical spatially adaptive method, *RiskShrink*, which works by shrinkage of empirical wavelet coefficients. *RiskShrink* mimics the performance of an oracle for selective wavelet reconstruction as well as it is possible to do so. A new inequality in multivariate normal decision theory which we call the *oracle inequality* shows that attained performance differs from ideal performance by at most a factor $\sim 2 \log n$, where n is the sample size. Moreover no estimator can give a better guarantee than this. Within the class of spatially adaptive procedures, *RiskShrink* is essentially optimal. Relying only on the data, it comes within a factor $\log^2 n$ of the performance of piecewise polynomial and variable-knot spline methods equipped with an oracle. In contrast, it is unknown how or if piecewise polynomial methods could be made to function this well when denied access to an oracle and forced to rely on data alone.

Keywords: Minimax estimation subject to doing well at a point; Orthogonal Wavelet Bases of Compact Support; Piecewise-Polynomial fitting; Variable-Knot Spline.

1 Introduction

Suppose we are given data

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

$t_i = i/n$, where e_i are independently distributed as $N(0, \sigma^2)$, and $f(\cdot)$ is an unknown function which we would like to recover. We measure performance of an estimate $\hat{f}(\cdot)$ in terms of quadratic loss at the sample points. In detail, let $f = (f(t_i))_{i=1}^n$ and $\hat{f} = (\hat{f}(t_i))_{i=1}^n$ denote the vectors of true and estimated sample values, respectively. Let $\|v\|_{2,n}^2 = \sum_{i=1}^n v_i^2$ denote the usual squared ℓ_n^2 norm; we measure performance by the risk

$$R(\hat{f}, f) = n^{-1} E\|\hat{f} - f\|_{2,n}^2,$$

which we would like to make as small as possible. Although the notation f suggests a function of a real variable t, in this paper we work only with the equally spaced sample points t_i.

1.1 Spatially Adaptive Methods

We are particularly interested in a variety of spatially adaptive methods which have been proposed in the statistical literature, such as *CART* (Breiman, Friedman, Olshen and Stone, 1983), *Turbo* (Friedman and Silverman, 1989), *MARS* (Friedman, 1991), and variable-bandwidth kernel methods (Müller and Stadtmuller, 1987).

Such methods have presumably been introduced because they were expected to do a better job in recovery of the functions actually occurring with real data than do traditional methods based on a fixed spatial scale, such as Fourier series methods, fixed-bandwidth kernel methods, and linear spline smoothers. Informal conversations with Leo Breiman and Jerome Friedman have confirmed this assumption.

We now describe a simple framework which encompasses the most important spatially adaptive methods, and allows us to develop our main theme efficiently. We consider estimates $\hat{f}$ defined as

$$\hat{f}(\cdot) = T(y, d(y))(\cdot) \tag{2}$$

where $T(y, \delta)$ is a *reconstruction formula* with "spatial smoothing" parameter δ, and $d(y)$ is a data-adaptive choice of the spatial smoothing parameter δ. A clearer picture of what we intend emerges from five examples.

[1]. Piecewise Constant Reconstruction $T_{PC}(y, \delta)$. Here δ is a finite list of, say, L real numbers defining a partition $(I_1, \ldots, I_L)$ of $[0, 1]$ via $I_1 = [0, \delta_1), I_2 = [\delta_1, \delta_1 + \delta_2), \ldots, I_L = [\delta_1 + \cdots + \delta_{L-1}, \delta_1 + \cdots + \delta_L]$,so that $\sum_1^L \delta_i = 1$. Note that L is a variable. The reconstruction formula is

$$T_{PC}(y, \delta)(t) = \sum_{\ell=1}^{L} \mathrm{Ave}(y_i : t_i \in I_\ell) 1_{I_\ell}(t);$$

piecewise constant reconstruction using the mean of the data within each piece to estimate the pieces.

[2]. Piecewise Polynomials $T_{PP(D)}(y, \delta)$. Here the interpretation of δ is the same as in [1], only the reconstruction uses polynomials of degree D.

$$T_{PP(D)}(y, \delta)(t) = \sum_{\ell=1}^{L} \hat{p}_\ell(t) 1_{I_\ell}(t),$$

where $\hat{p}_\ell(t) = \sum_{k=0}^{D} a_k t^k$ is determined by applying the least squares principle to the data arising for interval I_ℓ

$$\sum_{t_i \in I_\ell} (\hat{p}_\ell(t_i) - y_i)^2 = \min!$$

[3]. Variable-Knot Splines $T_{spl,D}(y, \delta)$. Here δ defines a partition as above, and on each interval of the partition the reconstruction formula is a polynomial of degree D, but now the reconstruction must be continuous and have continuous derivatives through order $D-1$. In detail, let τ_ℓ be the left endpoint of I_ℓ, $\ell = 1, \ldots, L$. The reconstruction is chosen from among those piecewise polynomials $s(t)$ satisfying

$$\left(\frac{d}{dt}^k s\right)(\tau_\ell -) = \left(\frac{d}{dt}^k s\right)(\tau_\ell +)$$

for $k = 0, \ldots, D-1$, $\ell = 2, \ldots, L$; subject to this constraint, one solves

$$\sum_{i=1}^{n} (s(t_i) - y_i)^2 = \min!$$

[4]. Variable Bandwidth Kernel Methods $T_{VK,2}(y, \delta)$. Now δ is a *function* on $[0,1]$; $\delta(t)$ represents the "bandwidth of the kernel at t"; the smoothing kernel K is a C^2 function of compact support which is also a probability density, and if $\hat{f} = T_{VK,2}(y, \delta)$ then

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} y_i K\left(\frac{t - t_i}{\delta(t)}\right) \Big/ \delta(t). \qquad (3)$$

More refined versions of this formula would adjust K for boundary effects near $t = 0$ and $t = 1$.

[5]. Variable-Bandwidth High-Order Kernels $T_{VK,D}(y, \delta)$, $D > 2$. Here δ is again the local bandwidth, and the reconstruction formula is as in (3), only $K(\cdot)$ is a C^D function integrating to 1, with vanishing intermediate moments

$$\int t^j K(t)\, dt = 0, \qquad j = 1, \ldots, D-1.$$

As $D > 2$, $K(\cdot)$ cannot be nonnegative.

These reconstruction techniques, when equipped with appropriate selectors of the spatial smoothing parameter δ, duplicate essential features of certain well-known methods.

[1] The piecewise constant reconstruction formula T_{PC}, equipped with choice of partition δ by recursive partitioning and cross-validatory choice of "pruning constant" as described by Breiman, Friedman, Olshen and Stone (1983) results in the method *CART* applied to 1-dimensional data.

[2] The spline reconstruction formula $T_{\mathrm{spl},D}$, equipped with a backwards deletion scheme models the methods of Friedman and Silverman (1989) and Friedman (1991) applied to 1-dimensional data.

[3] The kernel method $T_{K,2}$ equipped with the variable bandwidth selector described in Brockmann, Gasser and Herrmann (1992) results in the "Heidelberg" variable bandwidth smoothing method. Compare also Terrell and Scott (1992).

These schemes are computationally feasible and intuitively appealing. However, very little is known about the theoretical performance of these adaptive schemes, at the level of uniformity in f and N that we would like.

1.2 Ideal Adaptation with Oracles

To avoid messy questions, we abandon the study of specific δ-selectors and instead study *ideal* adaptation.

For us, ideal adaptation is the performance which can be achieved from smoothing with the aid of an *oracle*. Such an oracle will not tell us f, but will tell us, for our method $T(y,\delta)$, the "best" choice of δ for the true underlying f. The oracle's response is conceptually a selection $\Delta(f)$ which satisfies

$$R(T(y,\Delta(f)),f) = \mathcal{R}_{n,\sigma}(T,f)$$

where $\mathcal{R}_{n,\sigma}$ denotes the *ideal risk*

$$\mathcal{R}_{n,\sigma}(T,f) = \inf_{\delta} R(T(y,\delta),f).$$

As $\mathcal{R}$ measures performance with a selection $\Delta(f)$ based on full knowledge of f rather than a data-dependent selection $d(y)$, it represents an ideal we cannot expect to attain. Nevertheless it is the target we shall consider.

Ideal adaptation offers, in principle, considerable advantages over traditional nonadaptive linear smoothers. Consider the case of a function f which is a piecewise polynomial of degree D, with a finite number of pieces $I_1,\ldots,I_L$, say:

$$f = \sum_{\ell=1}^{L} p_\ell(t) 1_{I_\ell}(t). \tag{4}$$

Assume that f has discontinuities at some of the break-points $\tau_2,\ldots,\tau_L$.

The risk of ideally adaptive piecewise polynomial fits is essentially $\sigma^2 L(D+1)/n$. Indeed, an oracle could supply the information that one should use $I_1,\ldots,I_L$ rather than some other partition. Traditional least-squares theory says that, for data from the traditional linear model $Y = X\beta + E$, with noise E_i independently distributed as $N(0,\sigma^2)$, the traditional least-squares estimator $\hat\beta$ satisfies

$$E\|X\beta - X\hat\beta\|_2^2 = (\text{number of parameters in } \beta)(\text{variance of noise})$$

Applying this to our setting, fitting a function of the form (4) requires fitting (# pieces)(degree+1) parameters, so for the risk $R(\hat f,f) = n^{-1}E\|\hat f - f\|_{2,n}^2$ we get $L(D+1)\sigma^2/n$ as advertised.

On the other hand, the risk of a spatially-non-adaptive procedure is far worse. Consider kernel smoothing. Because f has discontinuities, no kernel smoother with fixed non-spatially varying bandwidth attains a risk $R(\hat f,f)$ tending to zero faster than $Cn^{-1/2}$, $C = C(f,\text{kernel})$. The same result holds for estimates in orthogonal series of polynomials or sinusoids, for smoothing splines with knots at the sample points and for least squares smoothing splines with knots equispaced.

Most strikingly, even for piecewise polynomial fits with equal-width pieces, we have that $R(\hat f,f)$ is of size $\asymp n^{-1/2}$ unless the breakpoints of f form a subset of the breakpoints of $\hat f$. But this can happen only for very special n, so in any event

$$\limsup_{N\to\infty} R(\hat f,f)n^{1/2} \geq C > 0.$$

In short, oracles offer an improvement—ideally—from risk of order $n^{-1/2}$ to order n^{-1}. No better performance than this can be expected, since n^{-1} is the usual "parametric rate" for estimating finite-dimensional parameters.

Can we approach this ideal performance with estimators using the data alone?

1.3 Selective Wavelet Reconstruction as a Spatially Adaptive Method

A new principle for spatially adaptive estimation can be based on recently developed "wavelets" ideas. Introductions, historical accounts and references to much recent work may be found in the books by Daubechies (1992), Meyer (1990), Chui (1992) and Frazier, Jawerth and Weiss (1991). Orthonormal bases of compactly supported wavelets provide a powerful complement to traditional Fourier methods: they permit an analysis of a signal or image into localised oscillating components. In a statistical regression context, this spatially varying decomposition can be used to build algorithms that adapt their effective "window width" to the amount of local oscillation in the data. Since the decomposition is in terms of an orthogonal basis, analytic study in closed form is possible.

For the purposes of this paper, we discuss a *finite, discrete, wavelet transform.* This transform, along with a careful treatment of boundary correction, has been described by Cohen, Daubechies, Jawerth, and Vial (1993), with related work in Meyer (1991) and Malgouyres (1991). To focus attention on our main themes, we employ a simpler *periodised* version of the finite discrete wavelet transform in the main exposition. This version yields an *exactly* orthogonal transformation between data and wavelet coefficient domains. Brief comments on the minor changes needed for the boundary corrected version are made in Section 4.6.

Suppose we have data $y = (y_i)_{i=1}^n$, with $n = 2^{J+1}$. For various combinations of parameters M (number of vanishing moments), S (support width), and j_0 (Low-resolution cutoff), one may construct an n-by-n orthogonal matrix $\mathcal{W}$—the finite wavelet transform matrix. Actually there are many such matrices, depending on special filters: in addition to the original Daubechies wavelets there are the Coiflets and Symmlets of Daubechies (1993). For the figures in this paper we use the Symmlet with parameter $N = 8$. This has $M = 7$ vanishing moments and support length $S = 15$.

This matrix yields a vector w of the *wavelet coefficients* of y via—

$$w = \mathcal{W}y;$$

and because the matrix is orthogonal we have the inversion formula $y = \mathcal{W}^T w$.

The vector w has $n = 2^{J+1}$ elements. It is convenient to index dyadically $n-1 = 2^{J+1}-1$ of the elements following the scheme

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

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

$$y_i = \sum_{j,k} w_{j,k} W_{jk}(i),$$

expressing y as a sum of basis elements W_{jk} with coefficients $w_{j,k}$. We call the W_{jk} *wavelets.*

The vector W_{jk}, plotted as a function of i, looks like a localized wiggle, hence the name "wavelet". For j and k bounded away from extreme cases by the conditions

$$j_0 \leq j < J - j_1, \qquad S < k < 2^j - S,$$

we have the approximation

$$n^{1/2} W_{jk}(i) \approx 2^{j/2} \psi(2^j t - k) \qquad t = i/n$$

where ψ is a fixed "wavelet" in the sense of the usual wavelet transform on $\mathbb{R}$ (Meyer, 1990), Daubechies (1988). This approximation improves with increasing n and increasing j_1. Here ψ is an oscillating function of compact support, usually called the mother wavelet. We therefore speak of W_{jk} as being localized to spatial positions near $t = k2^{-j}$ and frequencies near 2^j.

The wavelet ψ can have a smooth visual appearance, if the parameters M and S are chosen sufficiently large, and favorable choices of so-called quadrature mirror filters are made in the construction of the matrix $\mathcal{W}$. Daubechies (1988) described a particular construction with $S = 2M + 1$ for which the smoothness (number of derivatives) of ψ is proportional to M.

For our purposes, the only details we need are

[W1] W_{jk} has vanishing moments through order M, as long as $j \geq j_0$:

$$\sum_{i=0}^{n-1} i^\ell W_{jk}(i) = 0 \qquad \ell = 0, \ldots, M, \quad j \geq j_0, \quad k = 0, \ldots, 2^j - 1.$$

[W2] W_{jk} is supported in $[2^{J-j}(k - S), 2^{J-j}(k + S)]$, provided $j \geq j_0$.

Because of the spatial localization of wavelet bases, the wavelet coefficients allow one to easily answer the question "is there a significant change in the function near t?" by looking at the wavelet coefficients at levels $j = j_0, \ldots, J$ at spatial indices k with $k2^{-j} \approx t$. If these coefficients are large, the answer is "yes."

Figures 1 displays four functions – *Bumps, Blocks, HeaviSine* and *Doppler* – which have been chosen because they caricature spatially variable functions arising in imaging, spectroscopy and other scientific signal processing. For all figures in this article, $n = 2048$.

Figure 2 depicts the wavelet transforms of the four functions. The large coefficients occur exclusively near the areas of major spatial activity. This property suggests that a spatially adaptive algorithm could be based on the principle of *selective wavelet reconstruction.* Given a finite list δ of (j,k) pairs, define $T_{SW}(y, \delta)$ by

$$T_{SW}(y, \delta) = \hat{f} = \sum_{(j,k) \in \delta} w_{j,k} W_{jk}. \tag{5}$$

This provides reconstructions by selecting only a subset of the empirical wavelet coefficients.

Our motivation in proposing this principle is twofold. First, for a spatially inhomogeneous function, "most of the action" is concentrated in a small subset of (j,k)-space. Second, under the noise model underlying (1), noise contaminates all wavelet coefficients equally. Indeed, the noise vector $e = (e_i)$ is assumed to be a white noise; so its orthogonal transform $z = \mathcal{W}e$ is also a white noise. Consequently, the empirical wavelet coefficient

$$w_{j,k} = \theta_{j,k} + z_{j,k}$$

where $\theta = \mathcal{W}f$ is the wavelet transform of the noiseless data $f = (f(t_i))_{i=0}^{n-1}$.

Every empirical wavelet coefficient therefore contributes noise of variance σ^2, but only a very few wavelet coefficients contribute signal. This is the heuristic of our method.

Ideal spatial adaptation can be defined for selective wavelet reconstruction in the obvious way. For the risk measure (1) the ideal risk is

$$\mathcal{R}_{n,\sigma}(SW, f) = \inf_{\delta} R_{n,\sigma}(T_{SW}(y,\delta), f)$$

with optimal spatial parameter $\Delta(f)$ a list of (j,k) indices attaining

$$R_{n,\sigma}(T_{SW}(y,\Delta(f)), f) = \mathcal{R}_{n,\sigma}(SW, f).$$

Figures 3-6 depict the results of ideal wavelet adaptation for the 4 functions displayed in Figure 2. Figure 3 shows noisy versions of the four functions of interest; the signal-to-noise ratio $\|\text{signal}\|_{2,n}/\|\text{noise}\|_{2,n}$is 7. Figure 4 shows the noisy data in the wavelet domain. Figure 5 shows the reconstruction by selective wavelet reconstruction using an oracle; Figure 6 shows the situation in the wavelet domain. Because the oracle helps us to select the important wavelet coefficients, the reconstructions are of high quality.

The theoretical benefits of ideal wavelet selection can again be seen in the case (4) where f is a piecewise polynomial of degree D. Suppose we use a wavelet basis with parameter $M \geq D$. Then properties [W1] and [W2] imply that the wavelet coefficients $\theta_{j,k}$ of f all vanish except for

(i) coefficients at the coarse levels $0 \leq j < j_0$

(ii) coefficients at $j_0 \leq j \leq J$ whose associated interval $[2^{-j}(k-S), 2^{-j}(k+S)]$ contains a breakpoint of f.

There are a fixed number 2^{j_0} of coefficients satisfying (i), and, in each resolution level j, $(\theta_{j,k}, k = 0, \ldots, 2^j - 1)$ at most (# breakpoints)$(2S+1)$ satisfying (ii). Consequently, with L denoting again the number of pieces in (4), we have

$$\#\{(j,k) : \theta_{j,k} \neq 0\} \leq 2^{j_0} + (J + 1 - j_0)(2S+1)L.$$

Let $\delta^* = \{(j,k) : \theta_{j,k} \neq 0\}$. Then, because of the orthogonality of the (W_{jk}), $\sum_{(j,k)\in\delta^*} w_{j,k} W_{jk}$ is the least-squares estimate of f and

$$\begin{aligned} R(T(y,\delta^*), f) &= n^{-1}\#(\delta^*)\sigma^2 \\ &\leq (C_1 + C_2 J)L\sigma^2/n \qquad \text{for all } n = 2^{J+1} \end{aligned} \tag{6}$$

with certain constants C_1, C_2, depending linearly on S, but not on f. Hence

$$\mathcal{R}_{n,\sigma}(SW, f) = O\left(\frac{\sigma^2 \log n}{n}\right). \tag{7}$$

for every piecewise polynomial of degree $D \leq M$. This is nearly as good as the bound $\sigma^2 L(D+1)n^{-1}$ of ideal piecewise polynomial adaptation, and considerably better than the rate $n^{-1/2}$ of usual nonadaptive linear methods.

1.4 Near-Ideal Spatial Adaptation by Wavelets

Of course, calculations of ideal risk which point to the benefit of ideal spatial adaptation prompt the question: How nearly can one approach ideal performance when no oracle is available and we must rely on data only, and no side information about f?

The benefit of the wavelet framework is that we can answer such questions precisely. In Section 2 of this paper we develop new inequalities in multivariate decision theory which furnish an estimate $\hat{f}^*$ which, when presented with data y and knowledge of the noise level σ^2, obeys

$$R_{n,\sigma}(\hat{f}^*, f) \leq (2\log n + 1)\{\mathcal{R}_{n,\sigma}(SW, f) + \frac{\sigma^2}{n}\} \tag{8}$$

for every f, every $n = 2^{J+1}$, and every σ.

Thus, in complete generality, it is possible to come within a $2\log n$ factor of the performance of ideal wavelet adaptation. In small samples n, the factor $(2\log n + 1)$ can be replaced by a constant which is much smaller: e.g., 5 will do if $n \leq 256$; 10 will do if $n \leq 16384$. On the other hand, no radically better performance is possible: to get an inequality valid for all f, all σ, and all n, we cannot even change the constant 2 to $2 - \epsilon$ and still have (8) hold, neither by $\hat{f}^*$ nor by any other measurable estimator sequence.

To illustrate the implications, Figures 7 and 8 show the situation for the four basic examples, with an estimator $\tilde{f}_n^*$ which has been implemented on the computer, as described in section 2.3 below. The result, while slightly noisier than the ideal estimate, is still of good quality – and requires no oracle.

The theoretical properties are also interesting. Our method has the property that for every piecewise polynomial (4) of degree $D \leq M$ with $\leq L$ pieces,

$$R_{n,\sigma}(\hat{f}^*, f) \leq (C_1 + C_2 \log n)(2\log n + 1)L\sigma^2/n,$$

where C_1 and C_2 are as in (6); this result is merely a combination of (7) and (8). Hence in this special case we have an actual estimator coming within $C\log^2 n$ of ideal piecewise polynomial fits.

1.5 Universality of Wavelets as a Spatially Adaptive Procedure

This last calculation is not essentially limited to piecewise polynomials; something like it holds for *all* f. In section 3 we show that, for constants C_i not depending on f, n, or σ,

$$\mathcal{R}_{n,\sigma}(SW, f) \leq (C_1 + C_2 J)\mathcal{R}_{n,\sigma}(PP(D), f)$$

for every f, every $n = 2^{J+1}$ and every $\sigma > 0$.

We interpret this result as saying that *selective wavelet reconstruction is essentially as powerful as variable-partition piecewise constant fits, variable-knot least-squares splines, or piecewise polynomial fits.* Suppose that the function f is such that, furnished with an oracle, piecewise polynomials, piecewise constants, or variable-knot splines would improve the rate of convergence over traditional fixed-bandwidth kernel methods, say from rate of convergence n^{-r_1} (with fixed-bandwidth) to n^{-r_2}, $r_2 > r_1$. Then, furnished with an oracle, selective wavelet adaptation offers an improvement to $\log^2 n n^{-r_2}$; this is essentially the same benefit at the level of rates.

We know of no proof that existing procedures for fitting piecewise polynomials and variable-knot splines, such as those current in the statistical literature, can attain anything like the performance of ideal methods.

In contrast, for selective wavelet reconstruction, it is easy to offer performance comparable to that with an oracle, using the estimator $\hat{f}^*$. And wavelet selection with an oracle offers the advantages of other spatially-variable methods.

The main assertion of this paper is therefore that, from this (theoretical) perspective, it is cleaner and more elegant to abandon the ideal of fitting piecewise polynomials with optimal partitions, and turn instead to *RiskShrink*, about which we have theoretical results, and an order $O(n)$ algorithm.

1.6 Contents

Section 2 discusses the problem of mimicking ideal wavelet selection; Section 3 shows why wavelet selection offers the same advantages as piecewise polynomial fits; Section 4 discusses variations and relations to other work. Section 5 contains certain proofs. Related manuscripts by the authors, currently under publication review and available as LaTeX files by anonymous ftp from playfair.stanford.edu, are cited in the text by [filename.tex].

2 Decision Theory and Spatial Adaptation

In this section we solve a new problem in multivariate normal decision theory and apply it to function estimation.

2.1 Oracles for Diagonal Linear Projection

Consider the following problem from multivariate normal decision theory. We are given observations $w = (w_i)_{i=1}^n$ according to

$$w_i = \theta_i + \epsilon z_i \qquad i = 1, \dots, n \tag{9}$$

where z_i are independent and identically distributed as $N(0,1)$, $\epsilon > 0$ is the (known) noise level, and $\theta = (\theta_i)$ is the object of interest. We wish to estimate with ℓ_2-loss and so define the risk measure

$$R(\hat{\theta}, \theta) = E\|\hat{\theta} - \theta\|_{2,n}^2. \tag{10}$$

We consider a family of diagonal linear projections:

$$T_{DP}(w, \delta) = (\delta_i w_i)_{i=1}^n, \qquad \delta_i \in \{0, 1\}.$$

Such estimators "keep" or "kill" each coordinate. Suppose we had available an oracle which would supply for us the coefficients $\Delta_{DP}(\theta)$ optimal for use in the diagonal projection scheme. These ideal coefficients are $\delta_i = 1_{\{|\theta_i| > \epsilon\}}$ meaning that ideal diagonal projection consists in estimating only those θ_i larger than the noise level. Supplied with such coefficients, we would attain the ideal risk

$$\mathcal{R}_\epsilon(DP, \theta) = \sum_{i=1}^n \rho_T(|\theta_i|, \epsilon)$$

with $\rho_T(\tau, \sigma) = \min(\tau^2, \sigma^2)$.

In general the ideal risk $\mathcal{R}_\epsilon(DP, \theta)$ cannot be attained for all θ by any estimator, linear or nonlinear. However surprisingly simple estimates do come remarkably close.

Motivated by the idea that only very few wavelet coefficients contribute signal, we consider *threshold* rules, that retain only observed data that exceeds a multiple of the noise level. Define 'hard' and 'soft' threshold non-linearities by

$$\eta_H(w,\lambda) = wI\{|w|>\lambda\} \tag{11}$$
$$\eta_S(w,\lambda) = \mathrm{sgn}(w)(|w|-\lambda)_+. \tag{12}$$

The hard threshold rule is reminiscent of subset selection rules used in model selection and we return to it later. For now, we focus on soft thresholding.

Theorem 1 *Assume model (9)–(10). The estimator*

$$\hat\theta_i^u = \eta_S(w_i, \epsilon(2\log n)^{1/2}) \qquad i=1,\ldots,n$$

satisfies

$$E\|\hat\theta^u - \theta\|_{2,n}^2 \le (2\log n+1)\{\epsilon^2 + \sum_{i=1}^n \min(\theta_i^2,\epsilon^2)\} \qquad \textit{for all } \theta\in I\!R^n. \tag{13}$$

In "Oracular" notation, we have

$$R(\hat\theta^*,\theta) \le (2\log n+1)(\epsilon^2 + \mathcal{R}_\epsilon(DP,\theta)). \qquad \theta\in I\!R^n$$

Now ϵ^2 denotes the mean-squared loss for estimating one parameter unbiasedly, so the inequality says that we can mimick the performance of an oracle plus one extra parameter to within a factor of essentially $2\log n$.

A short proof appears in the Appendix. However it is natural and more revealing to look for 'optimal' thresholds λ_n^* which yield the smallest possible constant Λ_n^* in place of $2\log n+1$ among soft threshold estimators. We give the result here and outline the approach in Section 2.4.

Theorem 2 *Assume model (9)–(10). The minimax threshold λ_n^* defined at (20) and solving (22) below yields an estimator*

$$\hat\theta_i^* = \eta_S(w_i, \lambda_n^*\epsilon) \qquad i=1,\ldots,n \tag{14}$$

which satisfies

$$E\|\hat\theta^* - \theta\|_{2,n}^2 \le \Lambda_n^*\{\epsilon^2 + \sum_{i=1}^n \min(\theta_i^2,\epsilon^2)\} \qquad \textit{for all } \theta\in I\!R^n. \tag{15}$$

The coefficient Λ_n^*, defined at (19), satisfies $\Lambda_n^* \le 2\log n+1$, and the threshold $\lambda_n^* \le (2\log n)^{1/2}$. Asymptotically

$$\Lambda_n^* \sim 2\log n, \qquad \lambda_n^* \sim (2\log n)^{1/2}, \qquad n\to\infty.$$

Table 1 shows that this constant Λ_n^* is much smaller than $2\log n+1$ when n is on the order of a few hundred. For $n=256$, we get $\Lambda_n^* \approx 4.44$. For large n, however, the $\sim 2\log n$ upper bound is sharp. This holds even if we extend from soft co-ordinatewise thresholds to allow completely arbitrary estimator sequences into contention.

Theorem 3

$$\inf_{\hat\theta} \sup_{\theta \in \mathbb{R}^n} \frac{E\|\hat\theta - \theta\|_{2,n}^2}{\epsilon^2 + \sum_1^n \min(\theta_i^2, \epsilon^2)} \sim 2\log n \qquad \text{as } n \to \infty. \tag{16}$$

Hence an inequality of the form (13) or (15) cannot be valid for any estimator sequence with $(2 - \epsilon + o(1))\log n$ in place of Λ_n^*. In this sense, an oracle for diagonal projection cannot be mimicked essentially more faithfully than by $\hat\theta^*$.

The use of soft thresholding rules (12) was suggested to us in prior work on multivariate normal decision theory by Bickel (1983) and ourselves [mrlp.tex]. However it is worth mentioning that a more traditional hard threshold estimator (11) exhibits the same asymptotic performance.

Theorem 4 *With (ℓ_n) a thresholding sequence sufficiently close to $(2\log n)^{1/2}$, the hard threshold estimator*

$$\hat\theta_i^+ = w_i 1_{\{|w_i| > \ell_n \epsilon\}}$$

satisfies, for an $L_n \sim 2\log n$, the inequality

$$R(\hat\theta^+, \theta) \le L_n\{\epsilon^2 + \sum_{i=1}^n \min(\theta_i^2, \epsilon^2)\} \qquad \text{for all } \theta \in \mathbb{R}^n.$$

Here, sufficiently close to $(2\log n)^{1/2}$ means $(1-\gamma)\log\log n \le \ell_n^2 - 2\log n \le o(\log n)$ for some $\gamma > 0$.

2.2 Adaptive Wavelet Shrinkage

We now apply the preceding results to function estimation. Let $n = 2^{J+1}$, and let $\mathcal{W}$ denote the wavelet transform mentioned in section 1.3. $\mathcal{W}$ is an orthogonal transformation of $\mathbb{R}^n$ into $\mathbb{R}^n$. In particular, if $f = (f_i)$ and $\hat f = (\hat f_i)$ are two n-vectors and $(\theta_{j,k})$ and $(\hat\theta_{j,k})$ their $\mathcal{W}$ transforms, we have the Parseval relation

$$\|f - \hat f\|_{2,n} = \|\theta - \hat\theta\|_{2,n} \tag{17}$$

Now let (y_i) be data as in model (1) and let $w = \mathcal{W}y$ be the discrete wavelet transform. Then with $\epsilon = \sigma$

$$w_{j,k} = \theta_{j,k} + \epsilon z_{j,k} \qquad j = 0, \ldots, J; \quad k = 0, \ldots, 2^j - 1.$$

As in the introduction, we define selective wavelet reconstruction via $T_{SW}(y, \delta)$, c.f. (5), and observe that

$$T_{SW} = \mathcal{W}^T \circ T_{DP} \circ \mathcal{W} \tag{18}$$

in the sense that (5) is realized by wavelet transform, followed by diagonal linear projection or shrinkage, followed by inverse wavelet transform. Because of the Parseval relation (17), we have

$$E\|T_{SW}(y,\delta) - f\|_{2,n}^2 = E\|T_{DP}(w,\delta) - \theta\|_{2,n}^2.$$

Also if $\hat\theta^*$ denotes the nonlinear estimator (14) and

$$\hat f^* \equiv \mathcal{W}^T \circ \hat\theta^* \circ \mathcal{W}$$

then again by Parseval $E\|\hat f^* - f\|_{2,n}^2 = E\|\hat\theta^* - \theta\|_{2,n}^2$, and we conclude immediately:

Corollary 1 *For all f and all $n = 2^{J+1}$,*

$$R(\hat{f}^*, f) \le \Lambda_n^* \{ \frac{\sigma^2}{n} + \mathcal{R}_{n,\sigma}(SW, f) \}.$$

Moreover, no estimator can satisfy a better inequality than this for all f and all n, in the sense that for no measurable estimator can such an inequality hold, for all n and f, with Λ_n^* replaced by $(2 - \epsilon + o(1)) \log n$. The same type of inequality holds for an estimator $\hat{f}^+ = \mathcal{W}^T \circ \hat{\theta}^+ \circ \mathcal{W}$ derived from hard thresholding, with L_n in place of Λ_n^*.

Hence, we have achieved, by very simple means, essentially the best spatial adaptation possible via wavelets.

2.3 Implementation

We have developed a computer software package which runs in the numerical computing environment *Matlab*. In addition, an implementation by G.P. Nason in the *S* language is available by anonymous ftp from *Statlib* at `lib.stat.cmu.edu` ; other implementations are also in development. They implement the following modification of $\hat{f}^*$.

Definition 1 *Let $\tilde{\theta}^*$ denote the estimator in the wavelet domain obtained by*

$$\tilde{\theta}^*_{j,k} = \begin{cases} w_{j,k} & j < j_0 \\ \eta_S(w_{j,k}, \lambda_n^* \sigma) & j_0 \le j \le J \end{cases} .$$

RiskShrink *is the estimator*

$$\tilde{f}^*_n \equiv \mathcal{W}^T \circ \tilde{\theta}^* \circ \mathcal{W}$$

The name *RiskShrink* for the estimator emphasises that shrinkage of wavelet coefficients is performed by soft thresholding, and that a mean squared error , or "risk" approach has been taken to specify the threshold. Alternative choices of threshold lead to the estimators *VisuShrink* introduced in Section 4.2 below, and *SureShrink* discussed in our report [ausws.tex].

The rationale behind this rule is as follows. The wavelets $W_{j,k}$ at levels $j < j_0$ do not have vanishing means, and so the corresponding coefficients $\theta_{j,k}$ should not generally cluster around zero. Hence, those coefficients (a fixed number, independent of n) should not be shrunken towards zero. Let $\widetilde{SW}$ denote the selective wavelet reconstruction where the levels below j_0 are never shrunk. We have, evidently, the risk bound

$$R(\tilde{f}^*, f) \le \Lambda_n^* \{ \frac{\sigma^2}{n} + \mathcal{R}_{n,\sigma}(\widetilde{SW}, f) \}$$

and of course $\mathcal{R}_{n,\sigma}(\widetilde{SW}, f) \le 2^{j_0}\sigma^2/n + \mathcal{R}_{n,\sigma}(SW, f)$, so RiskShrink is never dramatically worse than $\hat{f}^*$; it is typically much better on functions having non-zero average values.

Figure 7 shows the reconstructions of the four test functions; Figure 8 shows the situation in the wavelet domains. Evidently the methods do a good job of adapting to the spatial variability of functions.

The reader will note that occasionally these reconstructions exhibit fine scale noise artifacts. This is to some extent inevitable: no hypothesis of smoothness of the underlying function is being made.

2.4 Proof Outline for Theorem 2

Suppose we have a single observation $Y \sim N(\mu, 1)$. Define the function $\rho_{ST}(\lambda, \mu) = E(\eta(Y,\lambda) - \mu)^2$. See e.g. Bickel (1983). Qualitatively, $\mu \to \rho_{ST}(\lambda, \mu)$ increases from 0 to a maximum of $1 + \lambda^2$ at $\mu = \infty$. Some explicit formulas and properties are given in the Appendix.

The main idea is to define the minimax quantities

$$\begin{aligned} \Lambda_n^* &\equiv \inf_\lambda \sup_\mu \frac{\rho_{ST}(\lambda, \mu)}{n^{-1} + \min(\mu^2, 1)} && (19) \\ \lambda_n^* &\equiv \text{the largest } \lambda \text{ attaining } \Lambda_n^* \text{ above.} && (20) \end{aligned}$$

The key inequality (13) follows immediately: first assume $\epsilon = 1$. Set $\hat{\theta}_i^* = \eta_S(w_i, \lambda_n^*)$.

$$\begin{aligned} E\|\hat{\theta}^* - \theta\|_2^2 = \sum_{i=1}^n \rho_{ST}(\lambda_n^*, \theta_i) &\leq \sum_{i=1}^n \Lambda_n^*\{n^{-1} + \min(\theta_i^2, 1)\} \\ &= \Lambda_n^*\{1 + \sum_{i=1}^n \min(\theta_i^2, 1)\} \end{aligned}$$

If $\epsilon \neq 1$, then for $\hat{\theta}_i^* = \eta_S(w_i, \lambda_n^* \epsilon)$ we get by rescaling that

$$E\|\hat{\theta}^* - \theta\|_2^2 = \sum \rho_{ST}(\lambda_n^*, \theta_i/\epsilon)\epsilon^2$$

and the inequality (15) follows.

Consequently, Theorem 2 follows from asymptotics for Λ_n^* and λ_n^*. To obtain these, consider the analogous quantities where the supremum over the interval $[0, \infty)$ is replaced by the supremum over the endpoints $\{0, \infty\}$.

$$\Lambda_n^0 \equiv \inf_\lambda \sup_{\mu \in \{0,\infty\}} \frac{\rho_{ST}(\lambda, \mu)}{n^{-1} + \min(\mu^2, 1)}, \tag{21}$$

and λ_n^0 is the largest λ attaining Λ_n^0. In the Appendix we show that $\Lambda_n^* = \Lambda_n^0$ and $\lambda_n^* = \lambda_n^0$.

We remark that $\rho(\lambda, \infty)$ is strictly increasing in λ and $\rho(\lambda, 0)$ is strictly decreasing in λ, so that at the solution of (21),

$$(n+1)\rho_{ST}(\lambda, 0) = \rho_{ST}(\lambda, \infty). \tag{22}$$

Hence this last equation defines λ_n^0 uniquely, and, as is shown in the appendix, leads to

$$\begin{aligned} \lambda_n^0 &\leq (2\log n)^{1/2}, && n \geq 2 \\ (\lambda_n^0)^2 &= 2\log(n+1) - 4\log\log(n+1) - \log 2\pi + o(1), && n \to \infty. && (23) \end{aligned}$$

To complete this outline, we note that the balance condition (22) together with $\rho_{ST}(\lambda_n^0, \infty) = 1 + (\lambda_n^0)^2$ gives

$$\Lambda_n^0 = \frac{(\lambda_n^0)^2 + 1}{1 + n^{-1}} \sim 2\log n, \qquad n \to \infty.$$

3 Piecewise Polynomials are not more powerful than Wavelets

We now show that wavelet selection using an oracle can closely mimick piecewise polynomial fitting using an oracle.

Theorem 5 *Let $D \leq M$ and $n = 2^{J+1}$. With constants C_i depending on the wavelet transform alone,*

$$\mathcal{R}_{n,\sigma}(SW, f) \leq (C_1 + C_2 J)\mathcal{R}_{n,\sigma}(PP(D), f) \tag{24}$$

for all f, for all $\sigma > 0$.

Hence for every function, wavelets supplied with an oracle have an ideal risk that differs by at most a logarithmic factor from the ideal risk of the piecewise polynomial estimate. Since variable-knot splines of order D are piecewise polynomials of order D, we also have

$$\mathcal{R}_{n,\sigma}(SW, f) \leq (C_1 + C_2 J)\mathcal{R}_{n,\sigma}(Spl(D), f). \tag{25}$$

Note that the constants are not necessarily the same at each appearance : see the proof below. Since piecewise-constant fits are piecewise polynomials of degree $D = 0$, we also have

$$\mathcal{R}_{n,\sigma}(SW, f) \leq (C_1 + C_2 J)\mathcal{R}_{n,\sigma}(PC, f).$$

Hence, if one is willing to neglect factors of $\log n$ then selective wavelet reconstruction, with an oracle, is as good as these other methods, with their oracles.

We note that one should not expect to get better than a $\log n$ worst-case ratio, essentially for the reasons given in section 1.2. If f is a piecewise polynomial, so that it is perfectly suited for piecewise polynomial fits, then wavelets should not be expected to be also perfectly suited: wavelets are not polynomials. On the other hand, if f were precisely a finite wavelet sum, then one could not expect for piecewise polynomials to be perfectly suited to reconstructing f; some differences between different spatially adaptive schemes are inevitable.

The theorem only compares ideal risks. Of course, the ideal risk for wavelet selection is nearly attainable. We know of no parallel result for the ideal risk of piecewise polynomials. In any event, we get as a corollary that the estimator $\hat{f}^*$ satisfies

$$R(\hat{f}^*, f) \leq (C_1 + C_2 \log_2 n)(2 \log n + 1)\mathcal{R}_{n,\sigma}(PP(D), f)$$

so that $\hat{f}^*$ comes within a factor $\log^2 n$ of ideal piecewise polynomial fits. Thus, there is a way to mimick an oracle for piecewise polynomials: to abandon piecewise-polynomial fits and to use wavelet shrinkage.

Proof of Theorem 5. Let $\Delta(f)$ be the partition supplied by an oracle for piecewise polynomial fits. Suppose that this optimal partition contains L elements. Let s be the least-squares fit, using this partition, to noiseless data. We have the $Bias^2 + Variance$ decomposition of ideal risk

$$R(T_{PP(D)}(y, \Delta(f)), f) = n^{-1}\|f - s\|_{2,n}^2 + (D+1)L\sigma^2/n. \tag{26}$$

Now let $\theta = \mathcal{W}s$ be the wavelet transform of s. Then, as s is a piecewise polynomial, the argument leading to (6) tells us that most of the wavelet coefficients of s vanish. Let $\delta^* = \{(j,k) : \theta_{j,k} \neq 0\}$. Then

$$\#(\delta^*) \leq (C_1 + C_2 J)L.$$

Consider the use of δ^* as spatial parameter in selective wavelet reconstruction.

$$R(T_{SW}(y,\delta^*),f) = n^{-1}||f-s||_{2,n}^2 + \#(\delta^*)\sigma^2/n. \tag{27}$$

Comparing this with (26), we have

$$R(T_{SW}(y,\delta^*),f) \leq \{1+(C_1+C_2J)/(D+1)\}R(T_{PP(D)}(y,\Delta),f);$$

the theorem now follows from the assumption

$$\mathcal{R}_{n,\sigma}(PP(D),f) = R(T_{PP(D)}(y,\Delta(f)),f)$$

and the definition

$$\mathcal{R}_{n,\sigma}(SW,f) \leq R(T_{SW}(y,\delta^*),f).$$

Finally, to verify (25) observe that the optimal variable knot spline $\tilde{s}$ of order D for noiseless data is certainly a piecewise polynomial, so $||f-s||^2 \leq ||f-\tilde{s}||^2$. It depends on at least L unknown parameters and so for noisy data has variance term at least $1/(D+1)$ times that of (26). Therefore,

$$\mathcal{R}_{n,\sigma}(PP(D),f) \leq (D+1)\mathcal{R}_{n,\sigma}(Spl(D),f)$$

which, together with (24) establishes (25).

4 Discussion

4.1 Variations on Choice of Oracle

An alternate family of estimators for the multivariate normal estimation problem (9) is given by diagonal linear shrinkers:

$$T_{DS}(w,\delta) = (\delta_i w_i)_{i=1}^n, \qquad \delta_i \in [0,1].$$

Such estimators shrink each coordinate towards 0, different coordinates being (possibly) treated differently. An oracle $\Delta_{DS}(\theta)$ for this family of estimators provides the ideal coefficients $(\delta_i) = (\theta_i^2/(\theta_i^2+\epsilon^2))_{i=1}^n$ and would yield an ideal risk

$$\mathcal{R}_\epsilon(DS,\theta) = \sum_{i=1}^n \frac{\theta_i^2\epsilon^2}{\theta_i^2+\epsilon^2} = \sum_{i=1}^n \rho_L(\theta_i,\epsilon), \text{ say.}$$

There is an oracle inequality for diagonal shrinkage also.

Theorem 6 *(i) The soft thresholding estimator $\hat{\theta}^*$ with threshold λ_n^* satisfies*

$$R(\hat{\theta}^*,\theta) \leq \tilde{\Lambda}_n\{\epsilon^2 + \sum_{i=1}^n \frac{\theta_i^2\epsilon^2}{\theta_i^2+\epsilon^2}\} \tag{28}$$

for all $\theta \in \mathbb{R}^n$, with $\tilde{\Lambda}_n \sim 2\log n$.

(ii) More generally, the asymptotic inequality (28) continues to hold for soft threshold sequences (λ_n) and hard threshold estimators with threshold sequences (ℓ_n) satisfying respectively

$$5\log\log n \leq \lambda_n^2 - 2\log n \leq o(\log n) \tag{29}$$
$$(1-\epsilon)\log\log n \leq \ell_n^2 - 2\log n \leq o(\log n). \tag{30}$$

(iii) Theorem 3 continues to hold, a fortiori, if the denominator $\epsilon^2 + \sum_{i=1}^n \min(\theta_i^2, \epsilon^2)$ *is replaced by* $\epsilon^2 + \sum_{i=1}^n \theta_i^2 \epsilon^2/(\theta_i^2 + \epsilon^2)$. *So oracles for diagonal shrinkage can be mimicked to within a factor* $\sim 2\log n$ *and not more closely.*

In the Appendix is a proof of Theorem 6 that covers both soft and hard threshold estimators and both DP and DS oracles. Thus the proof also establishes Theorem 4 and an asymptotic version of Theorem 2 for thresholds in the range specified in (29).

These results are carried over to adaptive wavelet shrinkage just as in Section 2.2 by defining wavelet shrinkage in this case by the analog of (18)

$$T_{WS} = \mathcal{W}^T \circ T_{DS} \circ \mathcal{W}.$$

Corollary 1 extends immediately to this case.

4.2 Variations on Choice of Threshold

Optimal Thresholds. In Theorem 1 we have studied λ_n^*, the minimax threshold for the *soft threshold* nonlinearity, with comparison to a *projection oracle.* A total of 4 minimax quantities may be defined, by considering various combinations of threshold type *(soft, hard)* and oracle type *(projection,shrinkage).*

We have computer programs for calculating λ_n^* which have been used to tabulate $\lambda_{2^j}^*$ for $j = 6, 7, \ldots, 16$ (cf. Table 1). These have also been embedded as look-up tables in the *RiskShrink* software mentioned earlier.

Implementation of any of the other optimal thresholds would require a computational effort to tabulate the thresholds for various values of n. However, this computational effort would be far greater in the other three cases than in the case we have studied here, essentially because there is no analog of the simplification that occurs through replacing (19) with (21).

Remark: A drawback of using optimal thresholds is that the threshold which is precisely optimal for one of the four combinations may not be even asymptotically optimal for another of the four combinations. Comparing (23) with (30) shows that λ_n^* used with hard thresholding can only mimick the oracle to within a factor $a \log n$, for some $a > 2$.

Universal Thresholds. As an alternative to the use of minimax thresholds, one could simply employ the universal sequence $\lambda_n^u = (2\log n)^{1/2}$. The sequence is easy to remember; implementation in software requires no costly development of look-up tables; and it is asymptotically optimal for each of the four combinations of threshold nonlinearity and oracle discussed above. In fact, finite-n risk bounds may be developed for this threshold by examining closely the proofs of Theorems 4 and 6.

Theorem 7

$$\begin{aligned}
\rho_{ST}(\lambda_n^u, \mu) &\leq (2\log n + 1)\{n^{-1} + \rho_T(\mu, 1)\}, && n = 1, 2, \ldots \\
\rho_{ST}(\lambda_n^u, \mu) &\leq (2\log n + 2.4)\{n^{-1} + \rho_L(\mu, 1)\}, && n = 4, 5, \ldots \\
\rho_{HT}(\lambda_n^u, \mu) &\leq (2\log n + 2.4)\{n^{-1} + \rho_T(\mu, 1)\}, && n = 4, 5, \ldots \\
\rho_{HT}(\lambda_n^u, \mu) &\leq (2\log n + 2.4)\{n^{-1} + \rho_L(\mu, 1)\}, && n = 4, 5, \ldots
\end{aligned}$$

The drawback of this simple threshold formula is that in samples on the order of dozens or hundreds, the mean square error performance of minimax thresholds is noticeably better.

VisuShrink. On the other hand (λ_n^u) has an important *visual* advantage: *the almost "noise-free" character of reconstructions.* This can be explained as follows. The wavelet transform of many noiseless objects, such as those portrayed in figure 1, is very sparse, and filled with essentially zero coefficients. After contamination with noise, these coefficients are all nonzero. If a sample that in the noiseless case ought to be zero is in the noisy case nonzero, and that character is preserved in the reconstruction, the reconstruction will have an annoying visual appearance – it will contain small blips against an otherwise clean background.

The threshold $(2 \log n)^{1/2}$ avoids this problem because of the fact that when (z_i) is a white noise sequence i.i.d. $N(0,1)$, then

$$\mathrm{pr}\{\max_i |z_i| > (2\log n)^{1/2}\} \to 0, \qquad n \to \infty. \tag{31}$$

So, with high probability, *every* sample in the wavelet transform in which the underlying signal is exactly zero will be estimated as zero.

Figure 9 displays the results of using this threshold on the noisy data of Figures 3 and 4. The almost "noise free" character of the plots is striking.

Definition 2 *Let $\tilde{\theta}^v$ denote the estimator in the wavelet domain obtained by*

$$\tilde{\theta}^v = \begin{cases} w_{j,k} & j < j_0 \\ \eta_S(w_{j,k}, \sigma(2\log n)^{1/2}) & j_0 \le j \le J \end{cases} .$$

VisuShrink *is the estimator*

$$\tilde{f}_n^v \equiv \mathcal{W}^T \circ \tilde{\theta}^v \circ \mathcal{W}.$$

Not only is the method better in visual quality than *RiskShrink*, the asymptotic risk bounds are no worse:

$$R(\tilde{f}_n^v, f) \le (2\log n + 1)\{\frac{\sigma^2}{n} + \mathcal{R}_{n,\sigma}(\widetilde{SW}, f)\}.$$

This estimator is discussed further in our report [asymp.tex].

Estimating the Noise Level. Our software estimates the noise level σ as the median absolute deviation of the wavelet coefficients at the finest level J, divided by .6745. In our experience, the empirical wavelet coefficients at the finest scale are, with a small fraction of exceptions, essentially pure noise. Naturally, this is not perfect; we get an estimate that suffers an upward bias due to the presence of some signal at that level. By using the median absolute deviation, this bias is effectively controlled. Incidentally, upward bias is not disastrous; if our estimate is biased upwards by, say 50%, then the same type of risk bounds hold, but with with a $3 \log n$ in place of $2 \log n$.

4.3 Adaptation in Other Bases

A considerable amount of Soviet literature in the 1980's – for example, Efroimovich and Pinsker (1984) et seq. – concerns what in our terms could be called mimicking an oracle in the Fourier basis. Our work is an improvement in two respects:

1. For the type of objects considered here, a Wavelet Oracle is more powerful than a Fourier Oracle. Indeed, a Fourier oracle can never give a rate of convergence faster than $n^{-1/2}$ on any discontinuous object, while the Wavelet oracle can achieve rates as fast as $\log n/n$ on certain discontinuous objects. Figure 10 displays the results of using a Fourier-domain oracle with our four basic functions; this should be compared with Figure 5. Evidently, the Wavelet oracle is visually better in every case. It is also better in mean square.

2. The Efroimovich-Pinsker work did not have access to the oracle inequality and used a different approach, not based on thresholding but instead on grouping in blocks and adaptive linear damping within blocks. Such an approach cannot obey the same risk bounds as the oracle inequality, and can easily be off of ideal risk by larger than logarithmic factors. Indeed, from a "minimax over L^2-Sobolev balls" point of view, for which the Efroimovich-Pinsker work was designed, the adaptive linear damping is essentially optimal; compare comments in our report [ausws.tex, Section 4]. Actual reconstructions by *RiskShrink* and by the Efroimovich-Pinsker method on the data of Figure 3 show that *RiskShrink* is much better for spatial adaptation; see Figure 4 of [ausws.tex].

4.4 Numerical measures of fit

Table 2 contains the average (over location) squared error of the various estimates from our four test functions for the noise realisation and the reconstructions shown in Figures 2 - 10. It is apparent that the ideal wavelets reconstruction dominates ideal Fourier and that the genuine estimate using soft threshold at λ_n^* comes well within the factor 6.824 of the ideal error predicted for $n = 2048$ by Table 1. Although the $(2 \log n)^{1/2}$ threshold is visually preferable in most cases, it has uniformly worse squared error than λ_n^*, which reflects the well-known divergence between the usual numerical and visual assessments of quality of fit.

Table 3 shows the results of a very small simulation comparison of the same four techniques as sample size is varied dyadically from $n = 256$ through 8192, and using 10 replications in each case. The same features noted in Table 2 extend to the other sample sizes. In addition, one notes that, as expected, the average squared errors decline more rapidly with sample size for the smoother signals *HeaviSine* and *Doppler* than for the rougher *Blocks* and *Bumps* .

4.5 Other Adaptive Properties

The estimator proposed here has a number of optimality properties in minimax decision theory. In recent work, we consider the problem of estimating f at a single point $f(t_0)$ is discussed, where we believe that f is in some Hölder class, but we are not sure of the exponent nor the constant of the class. *RiskShrink* is adaptive in the sense that it achieves, within a logarithmic factor, the best risk bounds that could be had if the class were known; and the logarithmic factor is necessary when the class is unknown, by work of Brown and Low (1993) and Lepskii (1990). Other near-minimax properties are described in detail in our report [asymp.tex].

4.6 Boundary correction

As described in the Introduction, Cohen, Daubechies, Jawerth and Vial (1993), have introduced separate 'boundary filters' to correct the non-orthogonality on $[0,1]$ of the restriction to $[0,1]$ of basis functions that intersect $[0,1]^c$. To preserve the important property [W1] of orthogonality to polynomials of degree $\leq M$, a further 'preconditioning' transformation P of the data $\mathbf{y}$ is necessary. Thus, the transform may be represented as $\mathcal{W} = U \circ P$, where U is the orthogonal transformation built from the quadrature mirror filters and their boundary versions via the cascade algorithm. The preconditioning transformation affects only the $N = M+1$ left-most and the N right-most elements of $\mathbf{y}$: it has block diagonal structure $P = diag(P_L \mid I \mid P_R)$. The key point is that the size and content of the boundary blocks P_L and P_R do not depend on $n = 2^{J+1}$. Thus the Parseval relation (17) is modified to

$$\gamma_1 ||\theta||_{2,n}^2 \leq ||f||_{2,n}^2 \leq \gamma_2 ||\theta||_{2,n}^2,$$

where the constants γ_i correspond to the smallest and largest singular values of P_L and P_R, and hence do not depend on $n = 2^{J+1}$. Thus all the ideal risk inequalities in the paper remain valid, with only an additional dependence for the constants on γ_1 and γ_2. In particular, the conclusions concerning logarithmic mimicking of oracles are unchanged.

4.7 Relation to Model Selection

RiskShrink may be viewed by statisticians as an automatic model selection method, which picks a subset of the wavelet vectors and fits a "model", consisting only of wavelets in that subset, to the data by ordinary least-squares. Our results show that the method gives almost the same performance in mean-square error as one could attain if one knew in advance which model provided the minimum mean-square error.

Our results apply equally well in orthogonal regression. Suppose we have $Y = X\beta + E$, with noise E_i independent and identically distributed as $N(0, \sigma^2)$, and X an n by p matrix. Suppose that the predictor variables are orthogonal: $X^T X = I_p$. Theorem 1 shows that the estimator $\tilde{\beta}^* = \theta^* \circ X^T Y$ achieves a risk not worse than $p^{-1} + \mathcal{R}_{p,\sigma}(DP, \beta)$ by more than a factor $2\log p + 1$. This point of view has amusing consequences. For example, the hard thresholding estimator $\tilde{\beta}^+ = \theta^+ \circ X^T Y$ amounts to "backwards-deletion" variable selection; one retains in the final model only variables which had Z-scores larger than λ in the original least-squares fit of the full model. In small dimensions p, this actually corresponds to current practice; the "5% significance" rule $\lambda \approx 2$ is near-minimax, in the sense of Theorem 2, for $p \approx 200$.

For lack of space, we do not pursue the model-selection connection here at length, except for two comments.

1. George and Foster (1990) have proved two results about model selection which it is interesting to compare with our Theorem 4. In our language, they show that one can mimick the "nonzeroness" oracle $\rho_Z(\theta, \epsilon) = \epsilon^2 1_{\{\theta \neq 0\}}$ to within $L_n = 1 + 2\log(n+1)$ by hard thresholding with $\lambda_n = (2\log(n+1))^{1/2}$. They also show that for what we call the hard thresholding nonlinearity, no other choice of threshold can give a worst-case performance ratio, which they call a "Variance Inflation Factor", asymptotically smaller than $\sim 2\log n$ as $n \to \infty$. Compare also Bickel(1983). Our results here differ because we attempt to mimick more powerful oracles, which attain optimal mean-squared errors. The increase in power of our oracles is expressed by

$\rho_Z(\mu,1)/\rho_L(\mu,1) \to \infty$ as $\mu \to 0$. Intuitively, our oracles achieve significant risk savings over the nonzeroness oracle for the case when the true parameter vector has many coordinates which are nearly, but not precisely zero. We thank Dean Foster and Ed George for calling our attention to this interesting work, which also describes connections with "classical" model selection, such as Gideon Schwarz' BIC criterion.

2. Alan Miller (1984, 1990) has described a model selection procedure whereby an equal number of "pure noise variables", namely column vectors independent of Y, are appended to the X matrix. One stops adding terms into the model at the point where the next term to be added would be one of the artificial, pure noise variables. This simulation method sets, implicitly, a threshold at the maximum of a collection of n Gaussian random variables. In the orthogonal regression case, this maximum behaves like $(2\log n)^{1/2}$, i.e. (λ_n^u) (compare (31)). Hence Miller's method is probably not far from minimaxity with respect to an MSE-oracle.

ACKNOWLEDGEMENTS

This paper was completed while Donoho was on leave from the University of California, Berkeley, where this work was supported by grants from NSF and NASA. Johnstone was supported in part by grants from NSF and NIH. Helpful comments of a referee are gratefully acknowledged. We are also most grateful to Carl Taswell, who carried out the simulations reported in Table 3.

5 Appendix: Proofs

5.1 Proof of Theorem 1

It is enough to verify the univariate case, for the multivariate case follows by summation. So, let $X \sim N(\mu,1)$, and $\eta_t(x) = \mathrm{sgn}(x)(|x|-t)_+$. In fact we show, for all $\delta \le 1/2$ and with $t = (2\log\delta^{-1})^{1/2}$ that

$$E(\eta_t(X)-\mu)^2 \le (2\log\delta^{-1}+1)(\delta+\mu^2\wedge 1).$$

Regard the right side above as the minimum of two functions and note first that

$$\begin{aligned} E(\eta_t(X)-\mu)^2 &= 1-2\mathrm{pr}_\mu(|X|<t)+E_\mu\, X^2\wedge t^2 \le 1+t^2 \\ &\le (2\log\delta^{-1}+1)(\delta+1), \end{aligned} \tag{32}$$

where we used $X^2\wedge t^2 \le t^2$. Using instead $X^2\wedge t^2 \le X^2$, we get from (32)

$$E(\eta_t-\mu)^2 \le 2\mathrm{pr}_\mu(|X|\ge t)+\mu^2. \tag{33}$$

The proof will be complete if we verify that

$$g(\mu) = 2\mathrm{pr}_\mu(|X|\ge t) \le \delta(2\log\delta^{-1}+1)+(2\log\delta^{-1})\mu^2.$$

Since g is symmetric about 0,

$$g(\mu) \le g(0)+(1/2)(\sup|g''|)\mu^2. \tag{34}$$

Finally, some calculus shows that $g(0) = 4\Phi(-t) \le \delta(2\log\delta^{-1}+1)$ and that $\sup|g''| \le 4\sup|x\phi(x)| \le 4\log\delta^{-1}$ for all $\delta \le 1/2$.

5.2 Mean squared error properties of univariate thresholding.

We begin a more systematic summary by recording

$$\begin{aligned} \rho_{ST}(\lambda,\mu) &= 1+\lambda^2+(\mu^2-\lambda^2-1)\{\Phi(\lambda-\mu)-\Phi(-\lambda-\mu)\} \\ &\quad -(\lambda-\mu)\phi(\lambda+\mu)-(\lambda+\mu)\phi(\lambda-\mu), \qquad (35) \\ \rho_{HT}(\lambda,\mu) &= \mu^2\{\Phi(\lambda-\mu)-\Phi(-\lambda-\mu)\}+\tilde{\Phi}(\lambda-\mu)+\tilde{\Phi}(\lambda+\mu) \\ &\quad +(\lambda-\mu)\phi(\lambda-\mu)+(\lambda+\mu)\phi(\lambda+\mu), \qquad (36) \end{aligned}$$

where ϕ, Φ are the standard Gaussian density and cumulative and $\tilde{\Phi}(x) = 1-\Phi(x)$.

Lemma 1 *For both $\rho = \rho_{ST}$ and $\rho = \rho_{HT}$*

$$\begin{aligned} \rho(\lambda,\mu) &\le \lambda^2+1, && \text{for all } \mu \in \boldsymbol{R}, \quad \lambda > c_1 && (37) \\ \rho(\lambda,\mu) &\le \mu^2+1, && \text{for all } \mu \in \boldsymbol{R} && (38) \\ \rho(\lambda,\mu) &\le \rho(\lambda,0)+c_2\mu^2 && 0<\mu<c_3 && (39) \end{aligned}$$

For soft thresholding, (c_1, c_2, c_3) may be taken as $(0, 1, \infty)$ and for hard thresholding as $(1, 1.2, \lambda)$. At $\mu = 0$, we have the inequalities

$$\begin{aligned} \rho_{ST}(\lambda,0) &\le 4\lambda^{-3}\phi(\lambda)(1+1.5\lambda^{-2}) && (40) \\ \rho_{HT}(\lambda,0) &\le 2\phi(\lambda)(\lambda+1) \qquad \lambda>1. && (41) \end{aligned}$$

Proof. For soft thresholding, (37) and (38) follow from (32) and (33) respectively. In fact $\mu \to \rho_{ST}(\lambda,\mu)$ is monotone increasing, as follows from

$$(\partial/\partial\mu)\ \rho_{ST}(\lambda,\mu) = 2\mu\{\Phi(\lambda-\mu)-\Phi(-\lambda-\mu)\}. \qquad (42)$$

From (42) it follows that $(\partial^2/\partial\mu^2)\rho_{ST}(\lambda,\mu) \le 2$ for $\mu \ge 0$. Using (34) for $g = \rho_{ST}$ establishes (39). The inequality (40) follows from (35) and the alternating series bound for Gaussian tails: $\tilde{\Phi}(\lambda) \le \phi(\lambda)(\lambda^{-1}-\lambda^{-3}+3\lambda^{-5})$.

Turning now to hard thresholding, formula (38) (and (37) for $\mu \le \lambda$) follows by taking expectations in

$$(Y1_{\{|Y|>\lambda\}}-\mu)^2 \le (Y-\mu)^2+\mu^2.$$

Now consider (37). In the range $\mu \in [\lambda, \infty)$, we have

$$\begin{aligned} E_\mu(Y1_{\{|Y|>\lambda\}}-\mu)^2 &\le E_\mu(Y-\mu)^2+\mu^2\mathrm{pr}_\mu(|Y|\le\lambda) \\ &= 1+(\lambda+v)^2\tilde{\Phi}(v), \qquad v=\mu-\lambda. \end{aligned}$$

For $\lambda \ge 1$, we obtain (37) from

$$\lambda^{-2}(\lambda+v)^2\tilde{\Phi}(v) \le (1+v)^2\tilde{\Phi}(v) \le 1, \qquad \text{for all } v>0.$$

To prove (39) it suffices, as for $\rho_{ST}(\lambda,)$, to bound $(\partial^2/\partial\mu^2)\rho(\lambda,\mu) \le 2$. Differentiating (36) twice, using the inequalities $\lambda(\lambda \pm 2\mu) \le (\lambda \pm \mu)^2$ and (for $0 \le \lambda \le \mu$), $4\mu\phi(\lambda+\mu) - 4\mu\phi(\lambda-\mu) \le 0$, and finally substituting $s = \lambda+\mu$ and $s = \lambda-\mu$, we obtain, for $0 \le \lambda \le \mu$,

$$\frac{\partial^2}{\partial\mu^2}\rho(\lambda,\mu) \le 2+2\sup_{s>0}[\phi(s)(s^3-2s)-2\Phi(-s)] \le 2.4.$$

Finally (41) follows from (36) and $\tilde{\Phi}(\lambda) \le \lambda^{-1}\phi(\lambda)$ for $\lambda > 1$.

5.3 Theorem 2 : Asymptotics of λ_n^0

For the first half, let Φ and ϕ denote the distribution and density of the normal $N(0,1)$. The quantity λ_n^0 is the root of $p_n(\lambda) = (n+1)\rho(\lambda,0) - \rho(\lambda,\infty)$. Note that p_n is a continuous function, with one zero on $[0,\infty)$. Furthermore, $p_n(0) = n$, and $p_n(+\infty) = -\infty$. Now

$$p_n(\lambda) = (1+\lambda^2)\{2(n+1)\Phi(-\lambda) - 1\} - 2\lambda\phi(\lambda)(n+1), \qquad \lambda \geq 0. \tag{43}$$

Note that if the term in brackets is negative, the whole expression is negative on $[\lambda,\infty)$. Using the standard inequality $\Phi(-\lambda) \leq \lambda^{-1}\phi(\lambda)$, one verifies that this happens for $\lambda = (2\log n)^{1/2}$, for $n \geq 3$.This implies that the zero λ_n^0 of p_n is less than $(2\log n)^{1/2}$. For $n = 2$, the claim has been verified by direct computation.

For the second half, define $\lambda_{\eta,N}$ for all sufficiently large n via

$$\lambda_{\eta,n}^2 = 2\log(n+1) - 4\log\log(n+1) - \log 2\pi + \eta.$$

By using the standard asymptotic result $\Phi(-\lambda) \sim \lambda^{-1}\phi(\lambda)$ as $\lambda \to +\infty$, it follows that $p_n(\lambda_{\eta,n})$ converges to $-\infty$ (resp. ∞) according as $\eta > 0$ (resp. $\eta < 0$). This implies (23).

5.4 Theorem 2 : Equivalence of $\Lambda_n^* = \Lambda_n^0, \lambda_n^* = \lambda_n^0$.

We must prove that with λ_n^0 defined as before,

$$L(\lambda_n^0,\mu) := \sup_\mu \frac{\rho_{ST}(\lambda_n^0,\mu)}{n^{-1} + \min(1,\mu^2)}$$

attains its maximum at either $\mu = 0$ or $\mu = \infty$.

We consider two ranges. For $\mu \in [1,\infty]$, the numerator $\rho_{ST}(\lambda_n^0,\mu)$ is monotone increasing in μ, and the denominator is constant.

For $\mu \in [0,1]$, we apply (39) to $\rho_{ST}(\lambda_n^0,\mu)$. An argument similar to that following (43) shows that $p(n^{-1/2}) \geq 0$ for $n \geq 3$ so that $\lambda_n^0 \geq n^{-1/2}$. By the equation preceding (22), we conclude that $n\rho(\lambda_n^0,0) = \{1+(\lambda_n^0)^2\}/(1+n^{-1}) \geq 1$. Combining this with (39),

$$L(\lambda_n^0,\mu) \leq \frac{\rho(\lambda_n^0,0)+\mu^2}{n^{-1}+\mu^2} \leq n\rho(\lambda_n^0,0),$$

so that L attains its maximum over $\mu \in [0,1]$ at 0, establishing the required equivalence.

5.5 Theorem 3

The main idea is to make θ a random variable, with prior distribution chosen so that a randomly selected subset of about $\log n$ coordinates are each of size roughly $(2\log n)^{1/2}$, and to derive information from the Bayes risk of such a prior.

Consider the θ-varying loss

$$\tilde{L}_n(\hat{\theta},\theta) = \frac{\sum_{i=1}^n(\hat{\theta}_i - \theta_i)^2}{1+\sum_i \theta_i^2 \wedge 1}$$

and the resulting risk

$$\tilde{R}_n(\delta,\theta) = E_\theta \tilde{L}_n(\delta(w),\theta).$$

Let π be a prior distribution on θ and let

$$\tilde{r}_n(\delta, \pi) = E_\pi \tilde{R}_n(\delta, \theta);$$

finally, let

$$\tilde{\rho}_n(\pi) = \inf_\delta \tilde{r}_n(\delta, \pi)$$

denote the Bayes risk of the prior π. Call the corresponding Bayes rule $\tilde{\delta}_\pi$.

The minimax theorem of statistical decision theory applies to the loss $\tilde{L}_n(\hat{\theta}, \theta)$, and so, if we let $\tilde{m}_n$ denote the left side of (16), we have

$$\tilde{m}_n = \sup_\pi \tilde{\rho}_n(\pi).$$

Consequently, Theorem 2 is proved if we can exhibit a sequence of priors π_n such that

$$\tilde{\rho}_n(\pi_n) \geq 2 \log n(1 + o(1)), \qquad n \to \infty. \tag{44}$$

Consider the three-point prior distribution

$$F_{\epsilon,\mu} = (1 - \epsilon)\nu_0 + \epsilon(\nu_\mu + \nu_{-\mu})/2,$$

where ν_x denotes Dirac mass at x. Fix $a \gg 0$. Define $\mu = \mu(\epsilon, a)$ for all sufficiently small $\epsilon > 0$ by

$$\phi(a + \mu) = \epsilon\phi(a)$$

with ϕ the standard N(0,1) normal density. Then

$$\mu \sim (2 \log \epsilon^{-1})^{1/2}, \qquad \epsilon \to 0.$$

Our reports [mrlp.tex, mews.tex, ausws.tex] have considered the use of this prior in the scalar problem of estimating $\xi \sim F_{\epsilon,\mu}$ from data $v = \xi + z$ with $z \sim N(0, 1)$ and usual squared-error loss $E\{\delta(v) - \xi\}^2$. They show that the Bayes risk

$$\rho_1(F_{\epsilon,\mu}) \sim \epsilon\mu^2\Phi(a), \qquad \epsilon \to 0. \tag{45}$$

To apply these results in our problem, we will select $\epsilon = \epsilon_n = \log n/n$, so that $\mu = \mu_n = \mu(\epsilon_n, a) \sim (2 \log n - 2 \log \log n)^{1/2}$.

Consider the prior π_n which is i.i.d. F_{ϵ_n,μ_n}. This prior has an easily calculated Bayes risk $\rho_n(\pi_n)$ for the vector problem $w_i = \theta_i + z_i$, $i = 1, \ldots, n$, when the usual ℓ_n^2 loss $L_n(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_{2,n}^2$ is used. Applying (45),

$$\rho_n(\pi_n) = n\rho_1(F_{\epsilon_n,\mu_n}) \sim n\epsilon_n\mu_n^2\Phi(a).$$

Our aim is to use this fact to get a lower bound for the Bayes risk $\tilde{\rho}_n(\pi_n)$.

Consider the random variable $N_n = \#\{i : \theta_i \neq 0\}$; it has a binomial distribution with parameters n, ϵ_n. Set $\eta_n = (\log n)^{2/3}$ and define the event $A_n = \{N_n \leq n\epsilon_n + \eta_n\}$. By Chebyshev's inequality, $a_n \equiv P(A_n^c) \leq n\epsilon/\eta^2 \to 0$.

Let $\tilde{\delta}_n$ denote the Bayes rule for π_n with respect to the loss $\tilde{L}_n$. Then (step (*) is justified below)

$$\begin{aligned}
\tilde{\rho}(\pi_n) &= E_{\pi_n} E_\theta \frac{L_n(\tilde{\delta}_n, \theta)}{1 + \sum \theta_i^2 \wedge 1} \\
&= E_{\pi_n} E_\theta \frac{L_n(\tilde{\delta}_n, \theta)}{1 + N_n} \\
&\geq \frac{1}{1 + n\epsilon_n + \eta_n} E_{\pi_n} E_\theta L_n(\tilde{\delta}_n, \theta) 1_{A_n} \\
&\geq \frac{1 + o(1)}{1 + n\epsilon_n + \eta_n} E_{\pi_n} E_\theta L_n(\tilde{\delta}_n, \theta) \qquad (*) \\
&\geq \frac{1 + o(1)}{1 + n\epsilon_n + \eta_n} \rho_n(\pi_n) \\
&\sim \frac{1}{1 + n\epsilon_n + \eta_n} n\epsilon_n \mu^2 \Phi(a) \\
&\sim 2\log n \Phi(a) \qquad n \to \infty;
\end{aligned}$$

as a can be chosen arbitrarily large, this proves (44).

Finally, to justify (*), we must verify that $E_{\pi_n} E_\theta[||\tilde{\delta} - \theta||^2, A_n^c] = o\{\rho_n(\pi_n)\} = o(\mu_n^2 \log n)$. We focus only on the trickier term $E[||\tilde{\delta}||^2, A_n^c]$, where we use simply E to denote the joint distribution of θ and x. Set $p(\theta) = 1 + N_n(\theta)$. Using by turns the conditional expectation representation for $\tilde{\delta}_{n,i}(x)$, the Cauchy-Schwartz and Jensen inequalities, we find

$$\begin{aligned}
||\tilde{\delta}_n||^2 &\leq E[p(\theta)|x] E[||\theta||^2/p(\theta)|x], \quad \text{and} \\
E\{||\tilde{\delta}_n||^2, A_n^c\} &\leq \{Ep^4(\theta)\ \mathrm{pr}^2(A_n^c)\ E||\theta||^8/p^4(\theta)\}^{1/4} \\
&\leq C\mu_n^2\ \mathrm{pr}^{1/2}(A_n^c)\ \log n = o(\mu_n^2 \log n),
\end{aligned}$$

since $||\theta||^8 = N_n \mu_n^8$ and $EN_n^p = O(\log^p n)$.

5.6 Theorems 4 and 6

We give a proof that covers both soft and hard thresholding, and both DP and DS oracles. In fact, since $\rho_L < \rho_T$ it is enough to consider $\rho = \rho_L$. Let

$$L(\lambda, \mu) = \frac{\rho(\lambda, \mu)}{n^{-1} + \mu^2/(\mu^2 + 1)},$$

where ρ is either ρ_{ST} or ρ_{HT}. We show that $L(\lambda, \mu) \leq (2\log n)(1 + \delta_n)$ uniformly in μ so long as

$$c \log\log n \leq \lambda^2 - 2\log n \leq \epsilon_n \log n.$$

Here $\delta_n \to 0$ and depends only on ϵ_n and c in a way that can be made explicit from the proof. For ρ_{ST}, we require that $c < 5$ and for ρ_{HT}, that $c < 1$.

For $\mu \in [(2\log n)^{1/2}, \infty]$, the numerator of L is bounded above by $1 + \lambda^2$ (from (37)) and the denominator is bounded below by $2\log n/(2\log n + 1)$.

For $\mu \in [1, (2\log n)^{1/2}]$, bound the numerator by (38) to obtain

$$L(\lambda, \mu) \leq \mu^{-2}(1 + \mu^2)^2 \leq (2\log n)\{1 + o(1)\}.$$

For $\mu \in [0,1]$, use (39):

$$
\begin{aligned}
L(\lambda, \mu) &\leq \frac{\rho(\lambda, 0)}{n^{-1}} + \frac{\rho(\lambda,\mu) - \rho(\lambda, 0)}{\mu^2/(1+\mu^2)} \\
&\leq n\rho(\lambda, 0) + 2c_2. \qquad (46)
\end{aligned}
$$

If $\lambda_n(c) = (2\log n - c\log\log n)^{1/2}$, then $n\phi(\lambda_n(c)) = \phi(0)(\log n)^{c/2}$. It follows from (40) and (41) that $n\rho(\lambda, 0)$ and hence $L(\lambda,\mu) = o(\log n)$ if $\lambda > \lambda_n(c)$ where $c < 5$ for soft thresholding and $c < 1$ for hard thresholding. The expansion (23) shows that this range includes λ_n^* and hence $\hat{\theta}^*$.

5.7 Theorem 7

When $\lambda^2 = (2\log n)^{1/2}$, the bounds over $[1, (2\log n)^{1/2}]$ and $[(2\log n)^{1/2}, \infty]$ in the previous section become simply $[1 + 2\log n]^2/2\log n \leq 2\log n + 2.4$ for $n \geq 4$. For $\mu \in [0,1]$, the bounds follow by direct evaluation from (46), (40) and (41). We note that these bounds can be improved slightly by considering the cases separately.

References

[1] BICKEL, P. J. (1983). Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi, and D. Siegmund, eds.), Academic Press, New York, 511–528.

[2] BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A.,& STONE, C.J. (1983). *CART: Classification and Regression Trees*. Wadsworth: CBelmont, CA.

[3] BROCKMANN, M., GASSER, T., & HERRMANN, E. (1992). Locally Adaptive Bandwidth Choice for Kernel Regression Estimators. To appear. *J. Amer. Statist. Assoc.*.

[4] BROWN, L.D. & LOW, M.G. (1993). Superefficiency and lack of adaptability in nonparametric functional estimation. To appear, *Annals of Statistics.*

[5] COHEN, A., DAUBECHIES, I., JAWERTH, B. & VIAL, P. (1993). Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris* (A). **316**., 417–421.

[6] CHUI, C.K. (1992)., *An Introduction to Wavelets.* Academic Press, Boston, MA.

[7] DAUBECHIES, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics* **41**, Nov. 1988, pp. 909-996.

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

[9] DAUBECHIES, I. (1993). Orthonormal Bases of Compactly Supported Wavelets II: Variations on a theme. *SIAM J. Math. Anal.*, **24**, 499–519.

[10] EFROIMOVICH, S. YU. & PINSKER, M.S. (1984). A learning algorithm for nonparametric filtering. *Automat. i Telemeh.* **11** 58-65 (in Russian).

[11] FRAZIER M., JAWERTH B., & WEISS G. (1991). *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, **79**. American Math. Soc.: Providence, RI.

[12] FRIEDMAN, J.H. & SILVERMAN, B.W. (1989). Flexible Parsimonious Smoothing and Additive Modeling. (with discussion). *Technometrics* **31,** 3–39.

[13] FRIEDMAN, J.H. (1991). Multiple Additive Regression Splines (with discussion). *Annals of Statistics*, **19**, 1–67.

[14] GEORGE, E. I. & Foster, D. P.(1990). The risk inflation of variable selection in regression. Technical Report, University of Chicago.

[15] LEPSKII, O.V. (1990). On one problem of adaptive estimation on white Gaussian noise. *Teor. Veoryatnost. i Primenen.* **35** 459-470 (in Russian). *Theory of Probability and Appl.* **35**, 454-466 (in English).

[16] MALGOUYRES, G. (1991). Ondelettes sur l'Intervalle: algorithmes rapides. Prépublications Mathematiques Orsay.

[17] MEYER, Y. (1990). *Ondelettes et Opérateurs: I. Ondelettes* Hermann et Cie, Paris.

[18] MEYER, Yves (1991). Ondelettes sur l'intervalle. *Revista Matemática Ibero-Americana* **7** (2), 115-133.

[19] MILLER, A.J. (1984). Selection of subsets of regression variables (with discussion). *J. R. Statist. Soc. A.*, **147,**389–425.

[20] MILLER, A.J. (1990). *Subset Selection in Regression.* Chapman and Hall. London, New York.

[21] MÜLLER, Hans-Georg & STADTMULLER, Ulrich. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, **15**(1), 182–201.

[22] TERRELL, G.R. & SCOTT, D.W. (1992). Variable kernel density estimation. *Annals of Statistics.*, **20**, 1236 – 1265.

Table 1

λ_n^* and Related Quantities				
n	λ_n^*	$(2\log n)^{1/2}$	Λ_n^*	$2\log n + 1$
64	1.474	2.884	3.124	8.3178
128	1.669	3.115	3.755	9.7040
256	1.860	3.330	4.442	11.090
512	2.048	3.532	5.182	12.477
1024	2.232	3.723	5.976	13.863
2048	2.414	3.905	6.824	15.249
4096	2.594	4.079	7.728	16.635
8192	2.773	4.245	8.691	18.022
16384	2.952	4.405	9.715	19.408
32768	3.131	4.560	10.80	20.794
65536	3.310	4.710	11.95	22.181

Table 2

average square errors $\|\|\hat{f} - f\|\|_{2,n}^2/n$ in the Figures				
Figure	Blocks	Bumps	HeaviSine	Doppler
Fig 1: $\|\|f\|\|_{2,n}^2/n$	81.211	57.665	52.893	50.348
Fig 3: with noise	1.047	0.937	1.008	0.9998
Fig 5: ideal wavelets	0.097	0.111	0.028	0.042
Fig 10: ideal Fourier	0.370	0.375	0.062	0.200
Fig 7: threshold λ_n^*	0.395	0.496	0.059	0.152
Fig 9: threshold $(2\log n)^{1/2}$	0.874	1.058	0.076	0.324

Table 3

average square errors $\|\|\hat{f} - f\|\|_{2,n}^2/n$ from 10 replications				
n	Ideal Fourier	Ideal Wavelets	Threshold λ_n^*	Threshold $(2 \log n)^{1/2}$
Blocks				
256	0.717	0.367	0.923	2.072
512	0.587	0.243	0.766	1.673
1024	0.496	0.168	0.586	1.268
2048	0.374	0.098	0.427	0.905
4096	0.288	0.062	0.295	0.621
8192	0.212	0.035	0.204	0.412
Bumps				
256	0.913	0.411	1.125	2.674
512	0.784	0.291	0.968	2.310
1024	0.578	0.177	0.694	1.592
2048	0.396	0.109	0.499	1.080
4096	0.233	0.062	0.318	0.683
8192	0.144	0.037	0.208	0.430
HeaviSine				
256	0.168	0.136	0.222	0.244
512	0.132	0.079	0.155	0.186
1024	0.091	0.040	0.089	0.122
2048	0.065	0.026	0.060	0.083
4096	0.048	0.016	0.045	0.066
8192	0.033	0.008	0.030	0.047
Doppler				
256	0.711	0.220	0.473	0.951
512	0.564	0.146	0.341	0.672
1024	0.356	0.078	0.249	0.470
2048	0.208	0.039	0.151	0.318
4096	0.127	0.023	0.098	0.203
8192	0.071	0.012	0.055	0.113

List of Figures

1. Four spatially variable functions. n=2048. Formulas below.
2. The four functions in the wavelet domain. Most Nearly Symmetric Daubechies Wavelet with $N = 8$ was used. Wavelet Coefficients $\theta_{j,k}$ are depicted for $j = 5, 6, \ldots, 10$. Coefficients in one level (j constant) are plotted as a series against position $t = 2^{-j}k$. The vast majority of the coefficients are zero or effectively zero.
3. Four functions with Gaussian white noise, $\sigma = 1$, rescaled to have signal-to-noise ratio, $SD(f)/\sigma = 7$.
4. The four noisy functions in the wavelet domain. Compare Fig. 2. Only a small number of coefficients stand out against a noise background.
5. Ideal Selective Wavelet Reconstruction, with $j_0 = 5$. Compare Figs. 1,3.
6. Ideal Reconstruction, Wavelet Domain. Compare Figs. 2,4. Most of the coefficients in Figure 4 have been set to zero. The others have been retained as is.
7. *RiskShrink* Reconstruction using soft thresholding and $\lambda = \lambda_n^*$. Mimicking an oracle while relying on the data alone.
8. *RiskShrink*, Wavelet Domain. Compare Figs 2,4,6.
9. *VisuShrink* reconstructions using soft thresholding and $\lambda = (2 \log n)^{1/2}$. Notice "noise-free" character; compare Figs 1,3,5,7.
10. Ideal Selective Fourier Reconstruction. Compare Fig. 5. Superiority of Wavelet Oracle is evident.

Formulas for Test Functions

Blocks.

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

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

Bumps.

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

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

HeaviSine.

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

Doppler.

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