Exact risk analysis of wavelet regression

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H. and Patil, P.

January 2, 1998

Abstract

Wavelets have motivated development of a host of new ideas in nonparametric regression smoothing. Here we apply the tool of exact risk analysis, to understand the small sample behavior of wavelet estimators, and thus to check directly the conclusions suggested by asymptotics. Comparisons between some wavelet bases, and also between hard and soft thresholding are given from several viewpoints. Our results provide insight as to why the viewpoints and conclusions of Donoho and Johnstone differ from those of Hall and Patil.

1. Introduction

In a series of papers, Donoho and Johnstone (1992 [9],1994a [10], 1995 [13]) and Donoho, Johnstone, Kerkyacharian and Picard (1995) [14] developed nonlinear wavelet shrinkage technology in nonparametric regression. For other work relating wavelets and nonparametric estimation, see Doukhan (1988) [15], Kerkyacharian and Picard, (1992) [21], Antoniadis (1994) [1] and Antoniadis, Gregoire and McKeague (1994) [2]. These papers have both introduced a new class of estimators, and have also provided new viewpoints for understanding other nonparametric regression smoothers as well. In particular, these papers study curve estimation from a minimax viewpoint, using some important function classes not previously considered in statistics, which model the notion of "different amounts of smoothness in different locations" more effectively than the usual classes. Hall and Patil (1995a) [17], studied wavelet based methods from the different viewpoint of a fixed target function, as opposed to the minimax approach of Donoho and Johnstone.

To date, with the exception of Donoho and Johnstone (1994a) [10], the study of these methods has been mostly asymptotic in character. As with any asymptotic result, there remain doubts as to how well the asymptotics describe small sample behavior. How large the sample should be before the asymptotic lessons apply is an important question. Here we address these issues using the tool of exact risk analysis, which was developed in Gasser and Müller (1984) [16] and Marron and Wand (1992) [25], and first applied to wavelet estimators by Antoniadis, Gregoire and McKeague (1994) [2]. This type of analysis is more efficient than simulation. This efficiency allows deeper study of the issues at hand.

Important lessons learned about wavelets in nonparametric regression are that "hard thresholding" (defined in Section 1.4), with a simple threshold value, as studied in Donoho and Johnstone (1994a) [10], performs as asymptotically predicted for reasonable sample sizes. In particular their performance is not far from the optimum, across a wide variety of contexts. However "soft thresholding" (also defined in Section 1.4), using the asymptotic minimax optimal threshold derived in Donoho and Johnstone (1994a) [10] did not perform so well in small samples, and seems to require large samples before the asymptotic lessons apply. Insight is provided as to why the soft threshold based method requires larger samples for the asymptotic effects to be dominant.

The theoretical literature contains some differing viewpoints and conclusions. In particular the asymptotic results and ideas in a series of papers by Donoho and Johnstone and co-workers are rather different from those in papers by Hall and Patil and coauthors. In Section 2.5 it is seen that both viewpoints are relevant in small samples, and insight is given as to what aspects each viewpoint is studying.

Wavelet estimators are a new subset of an old class of nonparametric regression estimators: orthogonal series methods. The basics of this are reviewed in Section 1.1. An important contribution of wavelet ideas to the classical theory is a useful new set of bases, which allow characterizations in terms of both "time" and "frequency". An overview of these is given in Section 1.2.

In Section 1.3 we introduce a set of test regression functions, that we used in this research. These include four functions of Donoho and Johnstone that have become quite well known, but others are added with the intent of studying additional issues. Important features of these functions are discussed. Insight into how the various estimators perform in the exact risk calculations comes from studying their wavelet spectra, which are also presented. In particular, it is seen that when the wavelet basis has sufficiently many vanishing moments, the nature and order of a discontinuity can be obtained from the slope of linear decay of the coefficients that intersect the singularity.

The exact risk calculations are developed in Section 2. Substantial insight comes from applying these ideas to a single coefficient (i.e. just to estimation of the mean of a single Gaussian random variable), which is done in Section 2.1. Exact risk from a variety of other viewpoints is studied in the remaining subsections.

In Section 2.2 exact risk is studied as a function of the threshold scale. This allows simple comparison of bases, which shows that no basis is uniformly best, although smooth wavelet bases are effective in all around sense. For each basis, its risk performance is well predicted by a "row-wise power remaining". We also do a comparison of threshold types, and show that the "hard denoising threshold" is usually superior to the "soft minimax optimal threshold" (these are defined in Section 1.4), despite their similar asymptotic performance.

In Sections 2.3 and 2.4, we use the optimal value of the threshold scale, and study exact risk as a function of sample size, and also of the noise level (i.e. the variance of the residuals). Slopes of these curves give "finite sample rates of convergence", and show how better signal compression gives a faster rate of convergence. It is seen that is some cases it can take surprisingly long for "the asymptotics to take effect". This viewpoint also provides confirmation that wavelet bases have good all around performance, although none is uniformly best, and that the "hard denoising threshold" is generally better than the "soft minimax optimal threshold".

In Section 2.5, we study risk as a function of both the threshold scale, and the threshold level. We use this to address the issue raised by Hall and Patil (1995c) [19], of "where is the smoothing parameter in wavelet methods?". We see that indeed these quantities work like smoothing parameters, in the sense that various trade offs of variance and bias can be created by appropriate adjustment of these parameters. However, for hard thresholding, with the denoising threshold, there are simple choices that are surprisingly effective in general. While they are not always close to the optimal values, the Risk at those values is never too far from the minimizing Risk. This is a visual demonstration of the ideas in Donoho and Johnstone (1994a) [10] and Donoho and Johnstone (1992) [9].

The above points are illustrated here using only a few carefully chosen examples, but we have studied many more. To assist readers who would like to look at more, we have constructed a MATLAB browser, which provides easy menu driven access to construction of linked versions of these pictures. This is available on the World Wide Web at the URL: `http://playfair.stanford.edu/~wavelab`, or by anonymous ftp at: `ftp://playfair.stanford.edu/pub/wavelab`.

The exact risk development in this paper is for nonparametric regression, but somewhat related calculations might also be done in wavelet density estimation, using the formulas in Hall and Patil (1995b) [18]. In that context, the continuous case can be treated as well as the discrete case considered here.

1.1. Setting and orthogonal series estimation

In this paper we study nonparametric regression, in the case of a fixed, equally spaced design, with homoscedastic Gaussian errors. The data are of the form:

$$Y_i = m(x_i) + \varepsilon_i, \; i = 1, ..., n$$

where the x_i are equally spaced on $[0, 1]$, where n is a power of 2, where the ε_i are independent and identically distributed $N(0, \sigma^2)$ and where the curve $m(x)$ is "smooth" in some sense. The goal is to use the data $\mathbf{Y} = (Y_1, \cdots, Y_n)^t$ to estimate the curve $m(x)$.

The classical orthogonal series estimation is motivated as follows. Given a basis $\{\psi_1, ..., \psi_n\}$ of $\Re^n$, that is orthonormal with respect to counting measure on $\{x_1, ..., x_n\}$ (e.g. the classical discrete Fourier basis), the vector $\mathbf{m} = (m(x_1), \cdots, m(x_n))^t$ has the "spectral representation"

$$m(x_i) = \sum_{i'=1}^{n} \theta_{i'} \psi_{i',i}, \; i = 1, ..., n$$

where the coefficients are given by

$$\theta_{i'} = \langle \mathbf{m}, \psi_{i'} \rangle = \mathbf{m}^t \psi_{i'}, \tag{1.1}$$

and where $\psi_{i',i}$ is the i-th entry of the vector $\psi_{i'}$. The vector $\theta = (\theta_1, ..., \theta_n)^t$ is called the transform of $\mathbf{m}$ and is just a rotation of $\mathbf{m}$ in $\Re^n$. A useful way to view the problem of using $\mathbf{Y}$ to estimate $\mathbf{m}$ is to think of estimating the $\theta_{i'}$ by the empirical coefficients

$$\widetilde{\theta}_{i'} = \langle \mathbf{Y}, \psi_{i'} \rangle = \mathbf{Y}^t \psi_{i'}. \tag{1.2}$$

The vector $\widetilde{\theta} = (\widetilde{\theta}_1, ..., \widetilde{\theta}_n)^t$ is the transform of $\mathbf{Y}$. For a good choice of basis, most of the "power of $\mathbf{m}$", which is conveniently quantified as

$$P_{\mathbf{m}} = \sum_{i=1}^{n} m(x_i)^2 = \sum_{i=1}^{n} \theta_i{}^2,$$

(where the last equality follows by the Parseval identity), will be "contained in a few" of the θ_i. For example, when the underlying curve $m(x)$ is smooth and periodic, most of the power of the Fourier representation will be concentrated in the lower frequency terms. In this case, a reasonable reconstruction of the "signal" $\mathbf{m}$ can be obtained from the data $\mathbf{Y}$ by inverting the transformation, but using only the important coefficients:

$$\widehat{m}(x_i) = \sum_{i' \in S} \widetilde{\theta}_{i'} \psi_{i',i},$$

where S is some set of "high power coefficients". If the set S is small, but at the same time $\sum_{i \in S} \theta_i^2$ is a large part of $P_{\mathbf{m}}$, then this estimator will be quite effective. This is because in that case most of the power of the noise will be in the other coefficients, and hence eliminated. For example this happens when using the Fourier basis with a smooth, periodic function m, because the resulting $P_{\mathbf{m}}$ is concentrated in the low frequency coefficients.

Adaptive choice of the set S is an important aspect of the "thresholded estimators" considered below. Also considered below are estimators which modify $\widehat{m}$ through "shrinkage" of the $\widetilde{\theta}_i$, to reduce variability.

The assumption of i. i. d. Gaussian errors could be viewed as a very strong one in a nonparametric setting. However, the main ideas studied in this paper are approximately true even in non-Gaussian, and also some dependent cases, through an asymptotic risk equivalence to the Gaussian case, in a wide variety of settings, as noted in Neumann (1995) [28]. Details are given for regression by Neumann and Spokoiny (1995) [29], and for spectral density estimation in Neumann (1994) [27].

More serious departures from our assumptions include heteroscedasticity and a nonequally spaced design. While wavelet methods can be adapted for these cases, study of these using exact risk tools is beyond the scope of this paper.

1.2. Wavelet bases

The Fourier basis can effectively compress signals (i.e. pack most of the power into a few coefficients) when they are very smooth everywhere and also periodic. But in a wide variety of other cases other bases are preferable. Good wavelet bases are useful both for smooth targets, and also for those which are "somewhat unsmooth in some locations".

As with Fourier theory, wavelets have closely parallel discrete and continuous theories, which are connected by Riemann summation. Here, and in the following, vectors are often usefully viewed as being formed from evaluating a function (of $x \in \Re$) at $x_1, ..., x_n$. Often the distinction between is blurred, but everything in this paper should be viewed as discrete.

The standard discrete orthogonal transform provides coefficients θ_i, often reindexed as $\theta_{j,k}$, which give both a "frequency" and a "location" decomposition of m. The intuitive idea of "local frequency" is modeled with a collection of basis functions which work like "single period wave pieces". Different frequencies are modeled through various scalings of these wave pieces. Hence the common intuitive notion of "frequency" is equivalent to "scale" of the basis functions. Following standard terminology in signal processing, we will use the latter term. A convenient index for scale is $j = 0, ..., \log_2(n/2)$ (recall n is assumed to be a power of 2). At scale j, there are 2^j basis vectors that are shifts of each other indexed by $k = 0, ...2^j - 1$. See Strang (1989) [30] for intuitive introduction to the specifics of wavelet bases. Daubechies (1992) [7] gives methods of construction of discrete bases of the form:

$$\varphi_{0,0},\ \psi_{j,k},\ j = 0, ..., \log_2(n/2),\ k = 0, ...2^j - 1$$

which are orthonormal. The simplest of these is the Haar basis which consists of step functions. The smoother wavelet bases which better compress smooth signals are more complicated.

In this paper, we will concentrate on the Haar wavelet basis, and on the Symmlet 8, as described on page 198 of Daubechies (1992) [7]. In that book, this basis is called "least asymmetric" (because of how it was derived), but here we use the shorter name of "Symmlet". In some of the work described in this paper, we also considered the Coiflet 3 basis, as described on page 258 of Daubechies (1992) [7], but the results were sufficiently similar to those for the Symmlet 8 that we do not include them here. For such a basis the coefficients defined at (1.1) have the form

$$f_{0,0} = \langle \mathbf{m}, \varphi_{0,0} \rangle,\ \theta_{j,k} = \langle \mathbf{m}, \psi_{j,k} \rangle,\ j = 0, ..., \log_2(n/2),\ k = 0, ...2^j - 1,$$

and the empirical coefficients from (1.2) become

$$\tilde{f}_{0,0} = \langle \mathbf{Y}, \varphi_{0,0} \rangle,\ \tilde{\theta}_{j,k} = \langle \mathbf{Y}, \psi_{j,k} \rangle,\ j = 0, ..., \log_2(n/2),\ k = 0, ...2^j - 1.$$

Just as the Fast Fourier Transform allows very fast calculation of the Fourier coefficients , the wavelet coefficients can be calculated much more efficiently than by the n inner products (i.e. an $O(n^2)$ operation matrix multiplication) suggested by the formula (1.1). Fast wavelet algorithms were developed by Mallat (1989a [23], 1989b [24]), based on ideas for sub-band coding in the electrical engineering literature. Some of the history of these ideas is summarized in Daubechies (1988 [6], 1992 [7]). These ideas result in an $O(n)$ algorithm which is faster and in our opinion simpler than the FFT.

Wavelet bases that are smoother than the Haar basis have difficulty in handling data near the edges of $[0, 1]$. One approach to this problem is based on "boundary filters", see Cohen, Daubechies and Vial (1993) [3] and Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993) [4]. In this paper the boundary problem is addressed (both for the Fourier and Wavelet bases) by assuming a "circular design", i.e. by "wrapping the data", and by using periodic target functions.

1.3. Test functions and spectra

How well the power of a signal may be compressed into relatively few coefficients is crucial to the choice of basis, as nicely shown in Donoho (1993) [8]. In this Section we develop a set of examples which are expected to work both quite well and also quite poorly for a variety of bases.

A set of test functions, $m(x)$, is shown in Figure 1. Explicit parameterizations of these curves are given in the appendix.

[put Figure 1 about here]

FIGURE 1: *10 regression functions, used in examples in this paper. Lower right figures give visual impression of "low noise" and "high noise" Gaussian errors, also used later.*

Here is a summary of the motivation behind each, and its special features and properties:

1. Step. This function should be very hard to compress with the Fourier basis, because of its jumps, but relatively easy for the smooth wavelet bases, and very easy for the Haar basis. The first jump is at $x = \frac{1}{3}$, which means that the Haar basis will spread power to a single coefficient at every scale. The second jump is at $x = \frac{3}{4}$, so its power will be entirely contained in a single coefficient, for the Haar basis.

2. Wave. This is a sum of two periodic sinusoids. Hence the Fourier basis can compress the power of this signal into just two coefficients. On the other hand, the Haar basis has a good deal of difficulty with compression, because the power of the signal is spread widely across the coefficients. The smoother wavelets are in between, since this signal is smooth, so much of the power appears at coarser scales, although the time localization property of the wavelets is not useful here.

3. Blip. This is essentially the sum of a linear function with a Gaussian density, and has been often used as a target function in nonparametric regression. To make the jump induced by the assumed periodicity visually clear, the function has been periodically rotated so the jump is at $x = .8$. This function does not compress well in the Fourier basis, because of the jump. It does not compress well in the Haar basis because much of the function is smooth. However, the smooth wavelet bases give good compression of this signal.

4. Blocks. This step function has many more jumps than the Step above, and has been used in several Donoho and Johnstone papers, for example Donoho and Johnstone (1994a) [10]. Compression is worst for the Fourier basis, and the many jumps create difficulties for the smoother wavelets as well.

5. Bumps. This also comes from Donoho and Johnstone, and is very challenging for any basis to compress. The smooth wavelets do the best job, but only for very large n.

6. HeaviSine. Another Donoho and Johnstone example. This looks promising for the Fourier basis, except for the two jumps. The Haar basis is not good here because it cannot effectively compress the smooth parts of signal. The smooth wavelet bases compress this signal quite well.

7. Doppler. The final Donoho and Johnstone example. The time varying frequency makes this very hard for the Fourier basis, with power spread all across the spectrum. It is more suitable for the wavelets, with their space time localization. The smooth wavelets have an advantage over the Haar, because the function is smooth.

8. Angles. This function is piecewise linear, and continuous, but has big jumps in its first derivatives. It is ideal for the Daubechies 4 wavelet, which passes lines through its low pass filter (but no higher degree polynomials). The Haar basis should be somewhat worse than the smoother wavelet, because more of the power of the signal will be spread to finer scales. The Fourier basis is not expected to compress this well, although the performance should not be so bad as most of these examples, since this one is continuous,

9. Parabolas. This function is piecewise parabolic. The function and its first derivative are continuous, but there are big jumps in its second derivative. It is ideal for the Daubechies 6, and Symmlet 6 bases. It is too smooth to compress well in the Haar basis. Compression should be reasonable for the Fourier basis, because both the function and its first derivative is continuous.

10. Time Shifted Sine. This is a time shifted sine wave. It is intended to be a very smooth function, but rather far from a linear combination of sine waves. We view this as representing the type of curve that "traditional smoothers" would consider estimating. The Haar basis should do a poor job of compression, but the smoother bases should do well.

A visual idea of two noise levels that are used later in the paper is also given in Figure 1. "Low Noise", $\sigma = 0.02$, is usually somewhat lower than the noise in examples by Donoho and Johnstone. "High Noise", $\sigma = 0.1$, is intended to be a "usual amount" in a conventional nonparametric regression setting.

Transforms for some of the curves in Figure 1 are shown in Figure 2, for $n = 256$. The rows correspond respectively to the target curves # 1,8,10. All these plots display magnitudes of the transformed coefficients, $|\theta_i|$, on the scale of $\log_{10}$, since they have a large dynamic range. The first column shows a visual display of the magnitudes of the $\log_{10}|\theta_{j,k}|$ shown as gray levels with black meaning essentially 0 and lighter representing larger magnitude, for the Haar basis. Locations on the image reflect locations in "time-frequency" space, with scales indexed by $j = 0, ..., 7$ shown vertically, and x locations on the same scale as in Figure 1, shown horizontally. For each scale j, there are 2^j gray bars of equal length, which represent the influence of each $\theta_{j,k}$ over the support of the corresponding basis function. The second column is a similar gray level representation of the $\log_{10}|\theta_{j,k}|$ for the Symmlet 8 basis. The third column shows the $\log_{10}|\theta_i|$ for the Fourier basis, where the ordering is based on frequency. The fourth column allows comparison of the performance of the wavelet bases with the Fourier basis in compressing these signals. It shows the same set of magnitudes as shown in the first three columns, but now they are in decreasing order.

[put Figure 2 about here]

FIGURE 2: *Logs of absolute values of transformed coefficients of target functions, with respect to different bases. First and second columns represent magnitudes of* $\log_{10}|\theta_i|$ *as gray level images, for the Haar and Symmlet 8 bases. Third column shows magnitudes as heights for the Fourier basis, in order of increasing frequency. Last column shows same magnitudes, but sorted in decreasing order, to allow comparison of compression across bases.*

A key point visible in these plots: so long as the wavelet basis has sufficiently many vanishing moments, the nature and order of a discontinuity can be read off from the slope of linear decay (on the log scale) of the wavelet coefficients that intersect that singularity. This is described for the continuous wavelet transform in Daubechies (1992, pp45, 48) [7].

For the #1 Step regression function shown in the top row, the good compression provided by the Haar is visually clear. The gray level display shows that the jump at $x = 1/3$ gives a single large coefficient at each scale (recall black means the coefficient has essentially zero magnitude), while the jump at $x = 3/4$ appears only for $j = 1$, but for no finer scales, because it is represented completely by the basis function $\psi_{1,1}$. The Symmlet 8 basis has some coefficients at all scales that are affected by both of the jumps (i.e. are light at fine scales), but the number is not terribly large, so the compression is good for this basis too (more dark locations means better signal compression). However, the Fourier basis is severely affected by the jumps, and spreads the power of this signal all across the spectrum.

For the #8 Angles regression function shown in the middle row, the Haar basis has some coefficients that are 0 for locations $x \in (0.2, 0.5)$, where the target curve is flat, but is nonzero in most locations, because the differencing operation captures some of the power of the sloped parts. The sorted spectrum on the right has steps, because many coefficients are the same as each other, coming from regions where the regression curve has constant slope. The Symmlet 8 basis compresses this signal better than the Haar basis. At finer scales, it gives 0 coefficients everywhere except near the kinks. The Fourier basis now gives somewhat better compression (than it did for the #1 Step target), because this regression function is continuous.

For the #10 Time Shifted Sine, the Haar basis again gives very poor compression. The Symmlet 8 and the Fourier both give very good compression of this smooth signal. Versions of these plots have been made for the other target functions, but these are not shown here to save space, because most of the main ideas are contained in the ones selected for Figure 2, and because they can be constructed using our MATLAB browser.

An important point is that there is no single basis which is effective for the good compression (and hence estimation) of all plausible signals. This has lead to recent research in fast algorithms for "best basis selection", see e.g. Coifman, et. al. (1992) [5]. Some theoretical results on denoising in the context of best basis algorithms are contained in Donoho and Johnstone (1994c) [12].

Figure 3 gives an indication of what compressibility of a target function by a basis implies in terms of possible performance by an orthogonal series estimator. A benchmark for this comes from having an "oracle" which indicates the order of the $|\theta_i|$. With this extra information, a sensible estimator (the unordered version is called the "projection estimator" in Section 1.4) would involve the first few coefficients. More precisely, let $\left|\theta_{(1)}\right|, ..., \left|\theta_{(n)}\right|$ be a decreasing ordering of $|\theta_1|, ..., |\theta_n|$. Given a cutoff value i_0 define the "oracle projection estimator"

$$\widehat{m}_{O,i_0}(x_i) = \sum_{i' \leq i_0} \widetilde{\theta}_{(i')} \psi_{(i'),i},$$

Performance of this estimator can be assessed by its expected averaged squared error, called "risk" in Section 2,

$$R_O(i_0) = En^{-1} \sum_{i=1}^{n} [\widehat{m}_{O,i_0}(x_i) - m(x_i)]^2.$$

Insight into the risk comes from writing it as the sum of the averaged variance

$$AV_O(i_0) = n^{-1} \sum_{i=1}^{n} var(\widehat{m}_{O,i_0}(x_i)) = n^{-1} \sum_{i' \leq i_0} \sigma^2 = i_0 \sigma^2 / n,$$

and the averaged squared bias,

$$ASB_O(i_0) = n^{-1} \sum_{i=1}^{n} [E\widehat{m}_{O,i_0}(x_i) - m(x_i)]^2 = n^{-1} \sum_{i' > i_0} \theta_{(i')}^2.$$

The curves $R_O(i_0)$, $AV_O(i_0)$, and $ASB_O(i_0)$ are shown in Figure 3, for the #1 Step regression function. The $AV_O(i_0)$ curve is the same for all bases, as this depends only on σ^2.

[put Figure 3 about here]

FIGURE 3: *Mean Squared Errors, with variance and squared bias decompositions, for the oracle based projection estimator, with respect to different bases, for different sample sizes. The number of terms used by the estimator is i_0. with $\log_2(i_0)$ shown on the horizontal axis. Noise levels are "High", $\sigma = 0.02$ and "Low" $\sigma = 0.1$, which give the two values of $AV_O(i_0)$ (since these are the same regardless of basis). Thin line types for the averaged squared bias $ASB_O(i_0)$, and thick line types for the risk $R_O(i_0) = AV_O(i_0) + ASB_O(i_0)$.*

Performance in Figure 3 of the different bases with respect to $R_O(i_0)$ is driven by the compressibility indicated in the right hand column of Figure 2. A useful quantification of compression comes from the functional

$$\Omega_\sigma = \sum_i \min(\sigma^2, \theta_i^2).$$

This has a close relationship to R_O as shown in Figure 3, since

$$\Omega_\sigma = i(\sigma)\frac{\sigma^2}{n} + \sum_{i>i(\sigma)} \theta_{(i)}^2 = AV_O\left(i(\sigma)\right) + ASB_O\left(i(\sigma)\right) = R_O\left(i(\sigma)\right),$$

where $i(\sigma) = \#\{i : \theta_i^2 > \sigma^2\}$. The Haar basis gives the best compression of the signal, which results in the smallest $ASB_O(i_0)$ and $R_O(i_0)$. The Fourier basis is the worst, and the Symmlet 8 basis is in between. In general, when bias is worse, the optimal choice of i_0 is larger, and the minimum $R_O(i_0)$ is increased. As n increases, $AV_O(i_0)$ decreases, with a resulting improvement in the minimum $R_O(i_0)$.

This estimator $\widehat{m}_{O,i_0}$ makes use of the ordering of the magnitudes of the coefficients θ_i, which is unavailable in practice. Thresholding methods, discussed in Section 1.4, attempt to give similar performance, without making use of this ordering.

A useful diagnostic tool for understanding thresholded estimates is given in Figure 4. For each signal, and each basis, again let $\left|\theta_{(1)}\right|, ..., \left|\theta_{(n)}\right|$ be decreasing ordering of $|\theta_1|, ..., |\theta_n|$. For a proportion p, the "power remaining in the last np of the coefficients" is

$$PR(p) = \sum_{i>np} \theta_{(i)}^2.$$

See Donoho (1993) [8] for discussion of a closely related quantity. This is a relabelling of the curve ASB_O shown in Figure 3, the averaged squared bias of the "oracle projection estimator". It is also a single target function version of the minimax notion of "tail-n-widths" of sets of coefficients arising from certain smoothness classes, see definition 4 in Section 4.3.3 of Donoho, Johnstone, Kerkyacharian and Picard (1995) [14]. It is the part of the risk due to the last np of the ordered coefficients, if they are simply all set to 0.

Two such curves are shown in the lower right of Figure 4. The curves go down quite rapidly in p when there is good compression of the signal, and go down slowly when the power of the signal is spread among many coefficients. There is a strong qualitative relation of these curves to those in the right hand column of Figure 2, which provides a somewhat different measure of how well each basis can compress the signals. In particular the comparisons as to how well each basis compresses each signal are the same from either viewpoint. The curves in Figure 4 look visually smoother mostly because they are cumulative in nature. We found them somewhat more useful for understanding the behavior of different estimators seen in later sections of this paper, probably because they are more intimately connected to the bias of the estimation process.

To understand how the relative magnitudes of the coefficients relate to the size of the two noise levels, shown in the lower right hand parts of Figure 1, we also include "power remaining in the noise" curves in the lower right hand part of Figure 4. These are similar to $PR(p)$, except that each θ_i^2 is replaced by a realization of the noisy version, $\widetilde{\theta}_i^2$, where the regression curve is taken to be 0, so this is a "pure noise process". These give some insight into which parts of the power remaining curves are "negligibly small".

Even more useful than power remaining, for understanding the behavior of estimators, was a "row-wise" version of the power remaining plots, which is shown in the rest of Figure 4. This is crucial in understanding the exact risk calculations, because much of what is seen there depends on whether or not the finer scale terms have some significantly large coefficients. For each scale, indexed by $j = 0, ..., 7$, the coefficients $\theta_{j,0}^2, ..., \theta_{j,2^j-1}^2$ are ordered as $\theta_{j,(0)}^2, ..., \theta_{j,(2^j-1)}^2$. For a proportion p, the "power remaining in the last p of the coefficients for scale j" is

$$RPR(p) = \sum_{i>np} \theta_{j,(i)}^2.$$

These curves are then arranged in order of finer scale. The finest scale part, $j = 7$ appears in the plot above the interval $[7, 8)$, the next finest scale, $j = 6$, is above the interval $[6, 7]$, etc.

[put Figure 4 about here]

FIGURE 4: *Bottom row, two right hand plots are power remaining, $PR(p)$ in the last np coefficients, for each target curve, for $n = 256$, overlaid with pure noise power remaining. Other plots are row-wise power remaining, $RPR(p)$. Part above the interval $[j, j+1]$ is $RPR(p)$ for the given j, i.e. for that "row" of wavelet coefficients having all the same scale. These allow comparison of how well the bases compress signals, in this row-wise sense, which will give insights in later sections.*

An important lesson of Figure 4 is that the Symmlet 8 basis very often gives either the best compression of the signal, or is not far from the basis which is best. Both the Haar and the Fourier bases have cases where they are the best, but all too many spectacular failures. This is a graphical demonstration of why smooth wavelet bases, such as the Symmlet 8, have generated considerable interest. In particular they provide a reasonable balance between two goals. First, they are localized in time (in contrast to the Fourier basis), which allows greater spatial adaptivity. Second, when m is sufficiently smooth, power at each location is packed into relatively few coefficients (the Haar basis is poor at this). See for example Section III.11 of Volume I of Meyer (1990) [26], where an explicit comparison of Fourier and wavelet bases is carried out.

In studying the Symmlet 8 basis, a key idea for understanding the exact risks below is that the regression functions can be divided into two groups. The first group has some significantly large coefficients at each scale, and includes #1, #3, #4, #5, #6, #7. These are functions which are essentially smooth in a spatially inhomogeneous sense (modeled as regions in certain types of Besov-Triebel spaces), in particular allowing some nonsmooth features, but only at a few locations. The second group has no significant coefficients for scales finer than a certain level, including #2, #8, #9, #10. The second group are those that are essentially smooth in a more commonly considered spatially homogeneous sense (often modeled as Sobolev classes). Here the amount of smoothness is roughly the same at all locations.

1.4. Thresholded wavelet estimators

Because of the generally good compression properties of the wavelet bases, the double index structure will be frequently used in the rest of this paper. Hence wavelet transforms of $\mathbf{m}$ and $\mathbf{Y}$ will be

$$\theta = \begin{pmatrix} f_{0,0} \\ \theta_{0,0} \\ \theta_{1,0} \\ \theta_{1,1} \\ \theta_{2,0} \\ \vdots \end{pmatrix} \text{ and } \widetilde{\theta} = \begin{pmatrix} \widetilde{f}_{0,0} \\ \widetilde{\theta}_{0,0} \\ \widetilde{\theta}_{1,0} \\ \widetilde{\theta}_{1,1} \\ \widetilde{\theta}_{2,0} \\ \vdots \end{pmatrix}$$

where obvious analogs of (1.1) and (1.2) are used, and basis vectors are appropriately relabelled. This structure will be used even for the Fourier basis, where the terms are rather arbitrarily grouped so that smaller j corresponds to lower frequency. This is somewhat restrictive for the Fourier basis, because only $\log_2(n)$ different projections will thus be considered in the comparisons to come (instead of the usual full range of frequencies).

A simple orthogonal series estimator, called here the "projection estimator", is based on the hope that most of the power of the signal is concentrated in the lowest frequency components, while the noise is spread evenly across coefficients, and thus most of it will be in the high frequency coefficients. A reasonable estimator is then simply based on the coarse scale (i.e. low frequency) empirical wavelet coefficients $\widetilde{\theta}_{j,k}$, defined at (1.2). This is equivalent to setting the fine scale coefficients to 0 and inverting the transform. Suppose scales of level j_0 and finer are to be truncated. Then we get the estimator

$$\widehat{m}_P(x_i) = \widehat{m}_P(x_i, j_0) = \widetilde{f}_{0,0}\varphi_{0,0} + \sum_{j=0}^{j_0-1} \sum_{k=0}^{2^j-1} \widetilde{\theta}_{j,k}\psi_{j,k,i}, \tag{1.3}$$

where $\psi_{j,k,i}$ denotes the i-th entry of the j,k-th basis vector, $\psi_{j,k}$. An alternate way of writing this estimator, which is convenient for generalization to "thresholded estimators" is based on the concept of "estimated coefficients"

$$\widehat{\theta}_{j,k,T} = \begin{cases} \widetilde{\theta}_{j,k} & j < j_0 \\ \eta_T(\widetilde{\theta}_{j,k}, \lambda\sigma) & j \geq j_0 \end{cases},$$

which uses some "thresholding function", $\eta_T(\widetilde{\theta}_{j,k}, \lambda\sigma)$, and results in the general form

$$\widehat{m}_T(x_i) = \widehat{m}_T(x_i, j_0) = \widetilde{f}_{0,0}\varphi_{0,0} + \sum_{j=0}^{\log_2(n/2)} \sum_{k=0}^{2^j-1} \widehat{\theta}_{j,k,T}\psi_{j,k,i}. \tag{1.4}$$

The case $T = P$ comes from $\eta_P(\widetilde{\theta}_{j,k}, \lambda\sigma) = 0$.

The hard thresholding idea can be viewed as an extension of (1.3) where a few finer scale terms are very selectively added in. Terms are included if their magnitudes are larger than some "threshold value" λ times σ (or an estimated value in practice). This idea is summarized mathematically by a threshold function of the form

$$\eta_H(\theta, \lambda) = \theta 1_{\{|\theta| > \lambda\}}.$$

The resulting estimator is (1.4) with $T = H$.

Soft thresholding is a related idea, in particular the same type of choice is made as to which terms to include. The difference between soft and hard thresholding is that in the soft case, the finer scale terms that are included are also shrunk. This shrinking has the effect of reducing variance, but at some cost in increased bias (this will be made more clear below). More precisely define the threshold function

$$\eta_S(\theta, \lambda) = sgn(\theta)(|\theta| - \lambda)_+.$$

The resulting estimator is (1.4) with $T = S$.

An attractively simple choice of threshold value λ, is based on the idea of "denoising", which attempts to eliminate terms which are pure noise. For homoscedastic noise, σ can be very well estimated, say by a robust scale estimate applied to the finest scale wavelet coefficients, so we will assume that σ is known. Assuming the errors are Gaussian allows use of the interesting property that for $X_1, ..., X_n$ i.i.d. $N(0, \sigma^2)$,

$$P\left(\max_{i=1,...,n} |X_i| \leq \sigma\sqrt{2\log(n)}\right) \to 1,$$

e.g. see Theorem 1.5.3 of Leadbetter, Lindgren and Rootzén (1983) [22], where log denotes the natural logarithm. Hence the threshold $\lambda_D = \sqrt{2\log(n)}$ will zero out every term that has all noise, and no signal. This threshold choice gives some interesting asymptotic near-minimax properties, as shown in Donoho et. al. (1995) [14].

A smaller threshold choice, which adjusts for some of the bias problems of soft thresholding, has been proposed by Donoho and Johnstone (1994a) [10], using finite sample minimax considerations. We call this choice (which depends on n) λ_{MO}, for Minimax Optimal.

2. Exact risk for thresholded estimators

A commonly considered measure of the performance of an estimator, $\widehat{m}(x)$ of the curve $m(x)$, (i.e. of the vector $\widehat{\mathbf{m}}$ as an estimate of $\mathbf{m}$) is the expected averaged squared error risk function

$$R(\widehat{m}) = R(\widehat{\mathbf{m}}) = n^{-1}E(\widehat{\mathbf{m}} - \mathbf{m})^t(\widehat{\mathbf{m}} - \mathbf{m}) = En^{-1}\sum_{i=1}^{n}[\widehat{m}(x_i) - m(x_i)]^2 .$$

For estimators of the form introduced in Section 1.4 (now replacing the j, k indexing by an equivalent $i = 1, ..., n$), a useful representation of the risk, by Parseval's identity is:

$$R(\widehat{m}) = n^{-1}\sum_{i=1}^{n} E\left[\widehat{\theta}_i - \theta_i\right]^2 .$$

Note that under the assumption of independent normal errors, $\varepsilon \sim N(0, \sigma^2\mathbf{I})$, where $\varepsilon = (\varepsilon_1 \cdots \varepsilon_n)^t$, the orthonormality of the transformation gives $\widetilde{\theta} \sim \mathbf{N}(\theta, \sigma^2\mathbf{I})$, where $\theta = (\theta_1, ..., \theta_n)^t$ and $\widetilde{\theta} = (\widetilde{\theta}_1, ..., \widetilde{\theta}_n)^t$. Hence it is simple to calculate the risk of the projection estimator $\widehat{m}_P$. When $\widehat{m}$ is a thresholded orthogonal series estimator, calculation of $R(\widehat{m})$ is complicated by the fact that $\widehat{m}$ is not a linear estimator. However, closed forms for the $E\left[\widehat{\theta}_i - \theta_i\right]^2$ have been derived under the normal error assumption, by Donoho and Johnstone (1994b) [11].

For a threshold value λ, not based on the data, calculation of each $E\left[\widehat{\theta}_i - \theta_i\right]^2$ is equivalent to the risk of estimating a single normal mean θ, based on a single observation $X \sim N(\theta, \sigma^2)$. Hard and soft thresholded estimates $\widehat{\theta}_H = \eta_H(X, \lambda\sigma)$ and $\widehat{\theta}_S = \eta_S(X, \lambda\sigma)$ may be employed in this context as well. The risks are:

$$E\left[\widehat{\theta}_H - \theta\right]^2 = \sigma^2 r_H(\theta/\sigma, \lambda), \tag{2.1}$$

and

$$E\left[\widehat{\theta}_S - \theta\right]^2 = \sigma^2 r_S(\theta/\sigma, \lambda) \tag{2.2}$$

where

$$r_H(\mu,\lambda) = \mu^2\left[\Phi(\lambda-\mu) - \Phi(-\lambda-\mu)\right] + \widetilde{\Phi}(\lambda-\mu) + \widetilde{\Phi}(\lambda+\mu) \\ +(\lambda-\mu)\phi(\lambda-\mu) + (\lambda+\mu)\phi(\lambda+\mu) \tag{2.3}$$

and

$$r_S(\mu,\lambda) = 1 + \lambda^2 + (\mu^2-\lambda^2-1)\left[\Phi(\lambda-\mu) - \Phi(-\lambda-\mu)\right] \\ -(\lambda-\mu)\phi(\lambda+\mu) - (\lambda+\mu)\phi(\lambda-\mu) \tag{2.4}$$

for Φ the standard normal cumulative distribution function, and $\widetilde{\Phi}$ it's complement, given by $\widetilde{\Phi}(x) = 1 - \Phi(x)$. These formulas are summed to give the exact risks used in the rest of the paper.

2.1. The single coefficient case

The formulas (2.3) and (2.4) provide substantial insight into the relative behaviors of hard and soft thresholding. Hence we study these functions first. Plots are shown for both hard and soft thresholding in Figure 5. They show single coefficient risks in the case $\sigma^2 = 1$. For general σ^2, the picture is the same, except scaled vertically by σ^2, and the θ axis should be replaced by $\mu = \theta/\sigma$.

[put Figure 5 about here]

FIGURE 5: *Single coefficient exact risks. Hard thresholding in the upper left, soft thresholding in the upper right, the difference in the lower left. The top shows that hard thresholding has smaller risk in most regions. A different display is used in the bottom, to show that the soft threshold is slightly smaller for small θ and moderate λ, which is where many terms should be when there is good signal compression.*

First consider the hard thresholding display in Figure 5, think of fixed λ, and study the risk as a function of θ. For $\theta < 1$, i.e. when the power of the signal is less than the noise, the risk is less than 1 ($= \sigma^2$), which is the risk from simply using the raw data as the estimate, i.e. $\widehat{\theta} = \widetilde{\theta}$. This value is smaller for larger λ. When $\lambda > 1$ (usually true in this context), for θ roughly between 1 and 2λ, the risk is quite a bit larger, with a peak at θ near λ. This peak is taller for larger λ. Hard thresholding is worse than estimating with the raw data when θ is in this range, because there is a large probability (highest near $\theta \approx \lambda$) of zeroing out a relatively large coefficient (the height of the curve is essentially the "squared bias", which is the product of θ^2 and the probability of its being mistakenly zeroed). The risk decreases for $\theta >> \lambda$, because then there is only a small probability that such a coefficient will be zeroed. For very large θ, the risk is essentially 1, since $\widehat{\theta} \approx \widetilde{\theta}$, so $R(\widehat{\theta}) \approx var(\widetilde{\theta}) = \sigma^2 = 1$, i.e. the performance is essentially the same as using $\widehat{\theta} = \widetilde{\theta}$. Hard thresholding will be effective when few of the coefficients are close to the threshold $\lambda\sigma$. Thresholding only those terms with scale finer than j_0 can help avoid having too many terms in this "boundary" area.

Next consider the soft threshold risk in Figure 5, as a function of θ, for each fixed λ. Soft thresholding has a higher maximum risk over this range than hard thresholding. Again for $\theta \leq 1$, i.e. the signal is smaller in magnitude than the noise, the risk is less than 1, and there is an advantage to thresholding. As for hard thresholding though, when the signal is bigger than the noise, soft thresholding is worse than using the raw data as the estimator. Unlike hard thresholding, this effect does not diminish for larger values of θ, because the shrinkage in the soft threshold induces some bias. Instead of going down to 1, the risk instead increases monotonically to $1 + \lambda^2$. This has the potential to create severe bias problems, because some of the coarser scale coefficients are expected to be quite large. Hence for soft thresholding it is important (and more important than for hard thresholding) to threshold only those terms with scale finer than a reasonably large value of j_0. Furthermore the value of j_0 that is used will have a stronger effect than for hard thresholding. This effect is seen in Figure 7 below.

Comparison between the risks is given in the bottom part of Figure 5. This is an overlay of functions of θ, for several λ's, since it is hard to tell which method is better from a surface plot as in the other parts. Note that soft thresholding is slightly better than hard for smaller θ. This is because the shrinkage effect of soft thresholding reduces variability, and bias is not important. But for larger θ, bias effects become dominant, and hard thresholding is superior. Note that the latter effect is much larger in magnitude, especially for larger values of λ.

The risk $R(\widehat{m})$ is an aggregate of such single coefficient risks. Soft thresholding will be superior to hard when a large enough share of the large coefficients are at coarser scales than j_0. But it does not take too many large coefficients at finer scales to put hard thresholding ahead of soft. This will happen in a number of the following examples. However it is important to keep in mind that we look only from the particular view of squared error risk. Soft thresholding has attractions of other types. One of these is that it is theoretically more tractable, so it is sensible to first analyze new settings in these terms, to give insights into handling the theory for the technically demanding case of hard thresholding. Another attraction is the intuitive connection to the deterministic theory of optimal recovery, as discussed in section 4 of Donoho, et. al. (1995) [14], which gives a different and also important type of benefit: a high probability of no sampling artifacts appearing (which comes at a cost of increased bias).

2.2. As a function of initial threshold scale

In this section we study how the risk $R(\widehat{m})$ depends on the scale j_0 beyond which terms are thresholded (or zeroed in the case of the projection estimator).

The first comparison is of the bases, as shown in Figure 6, where comparisons are done for hard thresholding, and the threshold value λ_D, since this is among the best as seen below. The examples in Figure 6 are based on $n = 1024$ observations. The rows of Figure 6 correspond to different signals, and the two columns show the log of the risk, for the low and high noise levels. The curves are all much higher in the second column because it is harder to estimate in the presence of more noise. The upper horizontal line in each plot corresponds to the "Raw Data Estimate", $\widehat{\mathbf{m}} = \mathbf{Y}$. At $j_0 = 10$, each estimator is the same as this estimator, because then no terms are thresholded so $\widehat{\mathbf{m}} = \widetilde{\theta} = \mathbf{Y}$. The other horizontal lines, that are thin and have the same line types as the curve for each basis, indicate the "ideal projection Risk", $\min_{i_0} R_O(i_0)$ (i.e. the minimum of the curve of the type shown in Figure 3) which is the risk of a projection estimator, when an oracle indicates exactly which terms should be included. Note that this ideal risk depends critically on the basis, since it is essentially a measure of how well the signal compresses in that basis.

[put Figure 6 about here]

FIGURE 6: *Exact Risk as a function of initial threshold scale j_0, shown as thick curves, allowing comparison of bases. For hard thresholding with denoising threshold λ_D. Target functions selected here are #1 Step, #2 Wave and #6 HeaviSine. Both noise levels are shown. Bases are Haar, Symmlet 8 and Fourier, with indicated line types. "Ideal Risk", i.e.* $\min_{i_0} R_O(i_0)$*, is shown for each basis as the thin horizontal line with the same line type.*

The top row provides a comparison of the performance of bases for the #1 Step regression function. Recall from Figure 4 that the Haar basis compresses this signal very well (i.e. a great deal of the power of the signal is contained in relatively few coefficients), so its ideal risk is quite small. This results in superior risk for this basis, compared to the others. For the Haar basis, the risk is also better for smaller j_0, since this allows more terms to be zeroed by the thresholding process (not zeroing them introduces additional variance). The Fourier basis gives very poor performance (for low noise and $j_0 < 8$, it is even worse than the estimator $\widehat{\mathbf{m}} = \mathbf{Y}$, and its ideal risk is also quite poor). The reason for this is seen in Figure 4: the power of the signal is spread among very many coefficients. A large proportion of the coefficients are in the peak area of the hard thresholding part of Figure 5. The effect is less for the high noise, because many coefficients are then relatively smaller than the threshold $\lambda\sigma$. The performance of the Symmlet 8 basis is between that of the others, which fits with the fact that its compression of this signal is in between as well.

The second row provides a similar comparison for the #2 Wave regression function. Now the situation with respect to the bases is reversed, both in Figure 4, and here. For both noise levels, the Symmlet 8 basis has the same risk as the Fourier basis for $j_0 \geq 6$, because the power of the signal for both bases is concentrated in scales coarser than j_0 (as shown in Figure 4), so each estimator is based on using a set of 2^{j_0} raw data coefficients that contain essentially the full power of the signal. The Symmlet 8 becomes much worse for smaller j_0 because the signal does have substantial power (widely spread among coefficients, since the time localization property of the wavelets does not help in this example) for scales with $j < 6$ (this corresponds to $p < 0.0625$ in Figure 4). The picture for high noise is similar to that for low noise, except the right side is "lifted up" by the higher variability in the data.

The third row of Figure 6 is roughly representative of the performance for the large majority of regression functions that we considered (see our MATLAB browser for the rest of these). The Symmlet 8 basis was almost always superior, although sometimes one of the other bases was competitive, especially for high noise, because the superior compressibility properties of the Symmlet 8 does involve some "start up cost" as seen in Figure 4, meaning that there are relatively many "large" coefficients, while the "superior compressibility" applies to the majority of smaller coefficients. This confirms the assertion that the smooth wavelet bases give reasonable all around performance, see for example Donoho (1993) [8]. Hence in much of the following, we will focus on the Symmlet 8 basis.

The next comparison is of the different thresholding types, i.e. a comparison of $\widehat{m}_P$, $\widehat{m}_H$ and $\widehat{m}_S$, shown in Figure 7. Again the two rows show two different target functions. The format is the same as in Figure 6. This comparison is explicitly shown for the Symmlet 8 basis, for the low noise level $\sigma = 0.02$, for the threshold value λ_D, and for $n = 1024$. See our MATLAB browser for many more such plots.

[put Figure 7 about here]

FIGURE 7: *Exact Risk as a function of initial threshold scale j_0, allowing comparison of threshold types. Selected target functions are #10 Time Shifted Sine and #4 Blocks. Only low noise shown. Basis is Symmlet 8.*

The main ideas for the plots for all the different regression examples were of two different types, represented by #10 and #4 in Figure 7. Where the finer scale coefficients were all essentially negligibly small, the estimators gave the same performance (and indeed are practically the same as each other) for large j_0. But for smaller j_0 there is some benefit to thresholding, because many of the terms are quite small, but some should be included. The effect was usually larger for hard thresholding than for soft, because soft thresholding pays a higher price for the larger coefficients, as shown in the soft threshold part of Figure 5. A typical setting of this type is shown in the top row, where the target is the #10 Time Shifted Sine. In this case, the power of the signal is negligible for $j \geq 6$, as seen in Figure 4.

The other type of performance appeared in settings where the signal had appreciable power in some coefficients at all levels, as for the #4 Blocks regression in Figure 4. Here the three estimators are different even for the larger j_0. Soft thresholding encounters difficulty even for the larger j_0 since now there too many terms in the high region of the soft threshold part of Figure 5. The projection estimator is scarcely better than the Raw Data estimate, since zeroing all the terms at any scale incurs too much bias.

In all cases, the Projection estimator was often a lot worse, and essentially never better, than the thresholded estimators. This confirms the wavelet nonparametric regression folklore, see e.g. Donoho and Johnstone (1992) [9], indicating this estimator should be ruled out in most situations because it is not spatially adaptive. Another viewpoint on this is provided by Hall and Patil (1995d) [20], who show it is essentially a kernel estimator with a special bivariate kernel. The hard type of thresholding was typically much better, and essentially never worse, than the soft thresholding, in the sense of squared risk considered here. Note however, that this is only for the case of the threshold value λ_D. Soft thresholding improves dramatically for λ_{MO}.

The next comparison is of the thresholding types and values. In particular, this now includes λ_{MO}. We omit the combination of hard thresholding, with λ_{MO}, because that value was not intended for hard thresholding, and because it was almost always much worse than hard thresholding with the value λ_D, and was never significantly better. This comparison was also done using the Symmlet 8 basis, and $n = 1024$.

[put Figure 8 about here]

FIGURE 8: *Exact Risk as a function of initial threshold scale j_0, allowing comparison of threshold values and threshold types. In particular this compares hard thresholding with λ_D, vs. soft thresholding with λ_{MO}. Selected target functions are #7 Doppler and #6 HeaviSine. Both noise levels are shown. Basis is Symmlet 8.*

The top row of Figure 8 show this comparison for the #7 Doppler regression, which was typical of the majority of cases studied. For the low noise level, hard thresholding with λ_D was somewhat better. But for the high noise level, soft thresholding, with λ_{MO}. Soft thresholding with λ_D was almost always worse, because of the bias problems suggested by Figure 5. One exception to this is shown in the bottom row of Figure 8, for #6 HeaviSine regression, with high noise. Here the soft threshold, with λ_D, is slightly better than the others. In this case, the scales $j \geq 3$ had very few large coefficients (which are worst for this estimator, as shown in Figure 5), but relatively many that were "small but nonzero", and hence in the low region of the soft threshold part of Figure 5. The magnitude of λ_{MO} is smaller than λ_D, which results in increased risk, as shown in Figure 5.

2.3. As a function of sample size

In this section we study exact risk, as a function of sample size n. We consider the range $n = 64(2^6), ..., 16,384(2^{14})$. The parameter j_0 is handled by using the value that minimizes the risk in each case, so each curve here is constructed using only the minimum of the curves shown in the previous section. The midpoint of these plots, $\log_2(n) = 10$, is $n = 1024$, which is used for most other examples in this paper.

For comparison of the bases, the main ideas were similar to those in the preceding section. Figure 9 shows hard thresholding, with λ_D, but very similar pictures were obtained for soft thresholding, with λ_{MO}. In particular the Symmlet 8 basis was generally superior, being either best or nearly best in the great majority of cases. Fairly representative behavior is seen in Figure 9 for the low noise parts of #4 Blocks and #3 Blip. As expected, the Haar basis was excellent, and the Fourier basis very poor for the #1 Step (not shown to save space) and #4 Blocks regressions. An exception to the expected is shown for the high noise case, with #4 Blocks regression, where it took surprisingly large $n = 2048$ (2^{11}), before the Haar had smaller risk. This is a case where it takes surprisingly long for the asymptotic effect to be dominant. For the #2 Wave example (again not shown to save space), the Fourier Basis was far superior, and the Haar basis was much the worst. The Symmlet 8 basis was in between in these cases, and generally superior otherwise. An exception was the # 3 Blip regression, with high noise, where the Haar basis was surprisingly competitive with the Symmlet 8. This is because while there are a large number of Haar coefficients visible in Figure 4, most of them are much lower than the high pure noise level shown in the lower right parts of Figure 4, and thus are practically negligible. The Fourier basis gave surprisingly good performance for the Angles regression, for the same reason. The Symmlet 8 basis will eventually be superior, but this requires samples larger than $n = 16,384$. Again the ideal projection risk for the given function and basis has also been shown on the plot. The reason that some of the ideal risks decrease faster than the thresholded risks (even for large n) is the following. From Donoho and Johnstone [10] (1994a), we may derive the inequality

$$\log_2(\text{Risk}) \leq \log_2(2\log n) + \log_2(\text{Ideal Risk})$$

which indicates a slowly increasing upper bound on the separation between the actual and ideal risk lines for a given signal-basis combination.

[put Figure 9 about here]

FIGURE 9: *Exact Risk as a function of sample size n, shown as the thick curves, allowing comparison of bases. The initial threshold scale j_0 is chosen to minimize the risk in each case. The thin curves of each line type are the ideal projection risks,* $\min_{i_0} R_O(i_0)$ *which give a theoretical lower bound for the actual risk. Selected target functions are #4 Blocks, #3 Blip and #8 Angles. Both noise levels are shown. Bases are Haar, Symmlet 8 and Fourier. Thresholding is hard, with λ_D.*

Figure 10 allows comparison (this time looking at risk as a function of n) of threshold types and values, as in Figure 8. As noted in the discussion of Figure 7 above, the examples were of two main types. Again the types are determined by whether or not there are a few large fine scale components.

The case where there are fine scale components is typified by the #3 Blip regression, where there were important differences. For all of these, the hard thresholding with λ_D was the best in low noise contexts. In high noise cases, soft thresholding with λ_{MO}, was often better for smaller n, but this effect tended to disappear asymptotically and then hard thresholding with λ_D was superior.

The #8 Angles regression is typical of the case of essentially no large coefficients finer than a certain scale. Here all of the estimators had very similar performance, because they are essentially the same. In these cases, soft thresholding, with λ_D was sometimes slightly better than the others, for the same reason as given above as for #6 in Figure 8.

[put Figure 10 about here]

FIGURE 10: *Exact Risk as a function of sample size n, allowing comparison of threshold values and threshold types. In particular this compares hard thresholding with λ_D, vs. soft thresholding with λ_{MO}. The initial threshold scale j_0 is chosen to minimize the risk in each case. Selected target functions are #3 Blip and #8 Angles. Both noise levels are shown. Basis is Symmlet 8.*

2.4. As a function of noise level

Another interesting way to study the risk of wavelet estimators is as a function of the noise level σ. A comparison of the bases shows mostly the same lessons as above, so no pictures are included to save space. An interesting feature was that for bases which do a poor job of signal compression, small enough σ entailed that the best estimator (i.e. best choice of j_0) was simply to return the raw data, i.e. use $\widehat{\mathbf{m}} = \mathbf{Y}$. This is sensible, because in such situations, the data are very informative. However, bases which provide good signal compression (extreme examples are the Haar basis for the #1 Step regression, and the Fourier basis for the #2 Wave regression) always give improved performance, even in the limit $\sigma \to 0$.

A comparison of the thresholding types and values, again leads to two main types of examples, depending on whether there are large coefficients at finer scales or not. Representative examples are in Figure 11 (this time looking at risk as a function of σ) . When there are large coefficients at finer scales, there is a difference between the different types of estimator, as shown for the #5 Bumps regression, and hard thresholding, with λ_D, is eventually better as $\sigma \to 0$. This is because as σ gets smaller, more and more terms are in the area of the lower part of Figure 5 where hard thresholding is superior. But the other approaches can be better in other regions. In particular, note that soft thresholding with λ_{MO} is better at $\sigma = 0.1$, i.e. $\log_{10}(\sigma) = -1$. When there are no important coefficients for fine scales, e. g. for the #10 Time Shifted Sine example, all methods are roughly the same, as noted above.

[put Figure 11 about here]

FIGURE 11: *Exact Risk as a function of noise level* σ, for $n = 1024$, *and the Symmlet 8 basis, allowing comparison of threshold values and threshold types. In particular this compares hard thresholding with* λ_D, *vs. soft thresholding with* λ_{MO}. *The initial threshold scale* j_0 *is chosen to minimize the risk in each case. Vertical lines show the "low noise level",* $\sigma = 0.02$ *and the "high noise level",* $\sigma = 0.1$ *used in other examples. Selected target functions are #5 Bumps and #10 Time Shifted Sine.*

An important lesson from these sections is that the hard thresholding generally has better squared error risk than soft thresholding. The overall strong performance of hard thresholding with λ_D was perhaps surprising, since no attempt at optimizing risk is made, to the level done by λ_{MO}. In this sense the comparison is not fair to the hard thresholding approach. It would be interesting to see how similar attempts to optimize hard thresholding would perform.

Again the driving force behind the overall good performance of this simple hard thresholding seems to be that the gains in the single coefficient risk, as indicated in the lower part of Figure 5, outweigh the advantages of the careful soft threshold choice.

2.5. As a function of threshold value

An interesting question from a classical nonparametric regression viewpoint, considered by Hall and Patil (1995c) [19], is "where is the smoothing parameter in wavelet methods?" They suggest that for a reasonable range of level dependent threshold values, both the initial threshold scale and the threshold value function as smoothing parameters. In this section we investigate the joint effect of j_0 and λ on the Risk, as shown in Figure 12. The curves shown in Section 2.2, are slices of surfaces such as these, taken along the dotted rectangles.

[put Figure 12 about here]

FIGURE 12: *Exact Risk as a function of initial threshold scale* j_0 *and threshold value* λ, for $n = 1024$, *the Symmlet 8 basis, and hard thresholding. The threshold values* λ_D *and* λ_{MO}. *are shown as dotted rectangles. The lowest point on the surface, hence the minimizing values of* j_0 *and* λ, *is indicated by the vertical dashed line. Selected target functions are #4 Blocks and #8 Angles. Both noise levels are shown.*

These surfaces had the same general shape for all of our examples. For j_0 close to 10, the surfaces are high, because there is too much variance. This is essentially the right hand side of each part of Figure

3, where AV dominates R. The variance comes from the estimator having too many unthresholded terms. Recall the extreme case $j_0 = 10$ is the raw data estimator $\widehat{\mathbf{m}} = \mathbf{Y}$. Similarly, the surfaces are high for small values of λ, again because of high variance from too many terms in the model, but this time they come from the low threshold (again the extreme case is the raw data estimator). When j_0 is small, and λ is large, the surface is high because the bias is too large. This is essentially the left side of each part of Figure 3, where ASB dominates R. The bias comes from not enough terms being used in the estimator.

In between these high regions are two valleys, which both start in the central area. One valley is in the direction of smaller j_0. This shows that for good choice of λ, there is usually not a large penalty for smaller j_0, i.e. for applying the thresholding operation to more terms. The other valley is in the direction of large λ, and suggests there is often not a large penalty for raising the threshold level. This second valley has some important differences depending on the particular regression function, in a manner similar to that noted above. When there are a significant number of large coefficients at fine scales (see Figure 4), as in the #4 Blocks case, this valley is relatively high, because bias is created by important terms being eliminated through the higher threshold. In such cases the optimal choice of λ was usually close to the value λ_D. It was usually smaller as suggested by the intuition behind λ_D, with the one exception being the high noise case for the #4 Blocks regression. When all the finer scale terms are negligible (again see Figure 4), as in the #8 Angles case, this valley tended to be rather lower. This is because, for reasonable j_0, there are no terms to create additional bias from larger λ. In such cases, the optimum (over the λ's considered) was often at the largest value, $\lambda = 2\sqrt{2\log(n)}$.

These plots show that in an absolute sense, j_0 and λ do work like smoothing parameters, in the sense that various trade offs of variance and bias can be created by appropriate adjustment of these parameters, which illustrates the points of Hall and Patil (1995c) [19]. However, the choices of "j_0 fairly small" and λ_D, give results that are unusually good (in the broader context of smoothing parameter selection) for such simple methods. While they are not always close to the optimal values, the Risk at those values is never too far from the minimizing Risk. This is a visual demonstration of the ideas in Donoho and Johnstone (1994a) [10] and Donoho and Johnstone (1992) [9]. The two cases we find here correspond to different smoothness classes in paper [9]. The case of no significant coefficients at finer scales, e.g. the #8 Angles in Figure 12, is an example of spatially homogeneous smoothness, which is roughly constant with respect to location. The case of some significant coefficients at isolated locations at finer scales, e.g. the #4 Blocks in Figure 12 is an example of spatially inhomogeneous smoothness, which is far less homogeneous with respect to location. The appealing feature of hard thresholding with the threshold λ_D is that it is not far from optimal with respect to both types of smoothness.

In contrast, Hall and Patil (1995c) [19] use only spatially homogeneous smoothness conditions. This corresponds to the #8 Angles part of Figure 12. Note that near the optimal value of λ, the parameter j_0 works more like a smoothing parameter, which corresponds roughly to their observations. This part of Figure 12 also shows Hall and Patil's point that for values of λ that are smaller than the optimal, the performance is not too far from the optimal, see Section 2.7 of Hall and Patil (1995a) [17].

The Soft thresholding analogs of Figure 12 are shown in Figure 13. These pictures are roughly similar in structure to those in Figure 12, with two valleys for the same reason. As expected from the lower part of Figure 5, the case of small j_0, large λ is worse for soft thresholding, but the small λ variance is better.

[put Figure 13 about here]

FIGURE 13: *Exact Risk as a function of initial threshold scale j_0 and threshold value λ,* for $n = 1024$, *the Symmlet 8 basis, and soft thresholding. The threshold value λ_D and λ_{MO} are shown as dotted rectangles. The lowest point on the surface, hence the minimizing values*

of j_0 and λ, is indicated by the vertical dashed line. Selected target functions are #2 Wave and #4 Blocks. Both noise levels are shown.

An important difference though is that nearly optimal choice of the parameters is not so simply available for soft thresholding. There is no longer a common value of j_0 and λ whose risk is always reasonably close to the optimum. The threshold λ_{MO} performs well in an average sense (and in a number of the examples considered here), but there is substantial variability in each direction. Two extreme cases are shown in Figure 13. Another unfortunate feature is that in a number of our examples, even at λ_{MO}, choice of j_0 is important too. However our plots suggest that the use of λ_{MO} *coupled with a good choice of* j_0 does reasonably well. An interesting open problem is data based choice of j_0, which could be perhaps based on some estimate of the Risk.

The oracle inequalities of Donoho and Johnstone (1994a) [10] and the asymptotics of Donoho, Johnstone, Kerkyacharian and Picard (1995) [14] show that both hard and soft thresholding have near optimal rates of convergence over a very large range of functions classes. Our results suggest that for soft thresholding, careful choice of the level j_0 is needed for good mean squared error performance. On the other hand, hard thresholding with λ_D does surprisingly well for reasonable sample sizes, and is not adversely affected by a fixed and relatively small choice of j_0. This greater robustness of hard thresholding can be understood from the lower part of Figure 5, which suggests that carefully controlled circumstances are required for soft thresholding to be better than hard thresholding. While such circumstances do occur asymptotically, it is not surprising that this can often require very large sample sizes.

3. Appendix:

Here are the forms of the target functions introduced in Section 1.3.

1. Step:
$$m_1(x) = 0.2 + (0.6)\, 1_{(\frac{1}{3},0.75)}(x).$$
2. Wave:
$$m_2(x) = 0.5 + (0.2)\cos(4\pi x) + (0.1)\cos(24\pi x).$$
3. Blip:
$$m_3(x) = \left(0.32 + 0.6x + 0.3e^{-100(x-0.3)^2}\right) 1_{[0,0.8]} + \\ + \left(-0.28 + 0.6x + 0.3e^{-100(x-1.3)^2}\right) 1_{(0.8,1]}.$$
4. Blocks: this is Donoho and Johnstone's Blocks, vertically rescaled to $[0.2, 0.8]$, first define the sign function $sgn(x) = 1_{(0,\infty)}(x) - 1_{(-\infty,0)}(x)$, then define a shifted version $ssgn(x) = (1 + sgn(x))/2$, then
$$m_4(x) = \left(\tfrac{0.6}{9.2}\right)\{4ssgn(x-0.1) - 5ssgn(x-0.13)+\} \\ 3ssgn(x-0.15) - 4ssgn(x-0.23) + 5ssgn(x-0.25) \\ -4.2ssgn(x-0.4) + 2.1ssgn(x-0.44) + 4.3ssgn(x-0.65) \\ -3.1ssgn(x-0.76) + 2.1ssgn(x-0.78) - 4.2ssgn(x-0.81) + 2\} + 0.2.$$
5. Bumps: this is Donoho and Johnstone's Bumps, vertically rescaled to approximately $[0.2, 0.8]$, first define $K_w(x) = (1 + |\frac{x}{w}|)^{-4}$, then
$$m_5(x) = \left(\tfrac{0.6}{5.3437952}\right)\{4K_{0.005}(x-0.1) + 5K_{0.005}(x-0.13)+ \\ +3K_{0.006}(x-0.15) + 4K_{0.01}(x-0.23) + 5K_{0.01}(x-0.25)+ \\ +4.2K_{0.03}(x-0.4) + 2.1K_{0.01}(x-0.44) + 4.3K_{0.01}(x-0.65)+ \\ +3.1K_{0.005}(x-0.76) + 5.1K_{0.008}(x-0.78) + 4.2K_{0.005}(x-0.81)\} + 0.2.$$

6. HeaviSine: this is Donoho and Johnstone's HeaviSine, vertically rescaled to $[0.2, 0.8]$, again using $sgn(x) = 1_{(0,\infty)}(x) - 1_{(-\infty,0)}(x)$,

$$m_6(x) = \left(\frac{0.6}{9}\right)(4\sin(4\pi x) + 5 - sgn(x - 0.3) - sgn(x - 0.72)) + 0.2.$$

7. Doppler: this is Donoho and Johnstone's Doppler, vertically rescaled to $[0.2, 0.8]$

$$m_7(x) = 0.6\left(\sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x + 0.05}\right) + 0.5\right) + 0.2.$$

8. Angles:

$$\begin{gathered} m_8(x) = (2x + 0.5)\,1_{[0,0.15]}(x) + (-12(x - 0.15) + 0.8)\,1_{(0.15,0.2]}(x) + \\ + (0.2)\,1_{(0.2,0.5]}(x) + (6(x - .5) + 0.2)\,1_{(0.5,0.6]}(x) + \\ + (-10(x - .6) + 0.8)\,1_{(0.6,0.65]}(x) + (-0.5(x - .65) + 0.3)\,1_{(0.65,0.85]}(x) + \\ + (2(x - .85) + 0.2)\,1_{(0.85,1]}(x). \end{gathered}$$

9. Parabolas: first define the "quadratic ramp function" $r_2(x, c) = (x - c)^2 1_{(c,1]}(x)$, then

$$\begin{gathered} m_9(x) = 0.8 - 30r_2(x, 0.1) + 60r_2(x, 0.2) - 30r_2(x, 0.3) + \\ +500r_2(x, 0.35) - 1000r_2(x, 0.37) + 1000r_2(x, 0.41) - 500r_2(x, 0.43) + \\ +7.5r_2(x, 0.5) - 15r_2(x, 0.7) + 7.5r_2(x, 0.9). \end{gathered}$$

10. Time Shifted Sine: first define the transformation $g(x) = (1 - \cos(\pi x))/2$, then

$$m_{10}(x) = 0.3\sin\{3\pi[g(g(g(g(x)))) + x]\} + 0.5.$$

References

[1] Antoniadis, A. (1994) Smoothing noisy data with coiflets, *Statist. Sinica*, 4, 651-678.

[2] Antoniadis, A. Gregiore, G. and McKeague, I. W. (1994) Wavelet methods in curve estimation, *Journal of the American Statistical Association*, 89, 1340-1353.

[3] Cohen, A., Daubechies, I. and Vial, P. (1993) Wavelets on the interval and fast wavelet transform, *Appl. Comput. Harm. Analy.*, 1, 54-81.

[4] 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.

[5] Coifman, R. R., Meyer, Y., Quake, S. and Wickerhauser, M. V. (1992) Signal processing and compression with wave packets. *Proceedings of the International Conference on Wavelets, Marseille*, (Y. Meyer and S. Roques, eds.) Springer, Berlin.

[6] Daubechies, I. (1988) Orthonormal Bases of Compactly Supported Wavelets, *Comm. Pure and Applied Math.*, 41, 909–996

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

[8] Donoho, D. L. (1993) Unconditional bases are optimal bases for data compression and for statistical estimation, *Appl. Comput. Harm. Analy.*, 1, 100-115.

[9] Donoho, D. L. and Johnstone, I. M. (1992). Minimax estimation via wavelet shrinkage. Technical report No. 402, Department of Statistics, Stanford University.

[10] Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81, 425-455.

[11] Donoho, D. L. and Johnstone, I. M (1994b). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, 99, 277-303.

[12] Donoho, D. L. and Johnstone, I. M. (1994c). Ideal Denoising in an orthonormal basis chosen from a library of bases, *C. R. Acad. Sci. Paris, Ser. I*, 319, 000-000 (in press).

[13] Donoho, D. L. and Johnstone, I. M. (1995a). Adapting to unknown smoothness via Wavelet shrinkage, *J. Amer. Statist. Assoc.*, to appear.

[14] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Jour. Roy. Statist. Soc. Ser. B*, 57, 301–369.

[15] Doukhan, P. (1988). Formes de Toeplitz associées à une analyse multi-échelle, *Comptes rendus Acad. Sci. Paris*, 306, 663-668.

[16] Gasser, T. and Müller. H.-G. (1984) Estimating regression functions and their derivatives by the kernel method, *Scandinavian Journal of Statistics*, 11,171-185.

[17] Hall, P. and Patil, P. (1995a). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear methods, to appear in *Journal of the Royal Statistical Society, Series B.*

[18] Hall, P. and Patil, P. (1995b). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators, to appear in *Ann. Statist.* 23.

[19] Hall, P. and Patil, P. (1995c). Effect of threshold rules on performance of wavelet based curve estimators, to appear in *Statistica Sinica.*

[20] Hall, P. and Patil, P. (1995d). On wavelet methods for estimating smooth functions, to appear in *Bernoulli,* 1.

[21] Kerkyacharian, G. and Picard, D. (1992) Density estimation in Besov spaces. *Statist. Probab. Lett.*,13, 15-24.

[22] Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York.

[23] Mallat, S. G. (1989a) A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, *IEEE Trans. Patt. Anal. Mach. Intell.*, 11, 674–693

[24] Mallat, S. G. (1989b) Multifrequency Channel Decompositions of Images and Wavelet Models, *IEEE Trans. on Acoust. Signal Speech Process.*, 37, 2091–2110

[25] Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error, *Ann. Statist.*, 20, 712-736.

[26] Meyer, Y. (1990) Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III: (with R. Coifman), Opérateurs multilinéaires. Hermann, Paris. English translation of first volume is published by Cambridge University Press.

[27] Neumann, M. H. (1994). Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. Preprint No. 99, Institut für Angewandte Analysis und Stochastik, Berlin.

[28] Neumann, M. H. (1995). Comments on "Wavelet shrinkage: Asymptopia?". *Jour. Roy. Statist. Soc. Ser. B*, 57, 346–347.

[29] Neumann, M. H. and Spokoiny, V. G. (1995) On the efficiency of wavelet estimators under arbitrary error distributions, *Math. Methods of Statist.*, to appear.

[30] Strang, G. (1989) Wavelets and dilation equations: a brief introduction, *SIAM Review*, 31, 614-627.