Adaptive Bayesian Wavelet Shrinkage*

Hugh A. Chipman, Eric D. Kolaczyk, Robert E. McCulloch

*Hugh A. Chipman and Robert E. McCulloch are Assistant Professor and Professor, at the University of Chicago, Graduate School of Business, 1101 E. 58th Street, Chicago, IL 60637. Eric D. Kolaczyk is Assistant Professor at the University of Chicago, Department of Statistics, 5734 University Avenue, Chicago, IL 60637. Correspondence should be directed to Eric D. Kolaczyk at the address above, or by email at *kolaczyk@galton.uchicago.edu.*

Abstract

When fitting wavelet based models, shrinkage of the wavelet coefficients is an effective tool for de-noising the data. This paper outlines a Bayesian approach to shrinkage, obtained by placing priors on the wavelet coefficients. The prior for each coefficient consists of a mixture of two normal distributions with different standard deviations, a form simple enough to allow closed form calculation of the posterior distribution. This means that a closed form shrinkage function is available, and reconstructions for large data sets can be computed quickly. The simple and intuitive form of prior allows us to propose automatic choices of prior parameters. These parameters are chosen adaptively according to the resolution level of the coefficients, typically shrinking high resolution (frequency) coefficients more heavily. The closed form of the posterior distribution means that posterior variances of the coefficients may be used to assess uncertainty in the reconstruction. Several examples are used to illustrate the method, and comparisons are made with other shrinkage methods.

Key Words: Bayesian Estimation, Mixture Models, Uncertainty Bands.

1 Introduction

In many problems in which measurements are made at equally spaced points (usually over time) with noise, wavelets provide an effective model. They are popular tools for "de-noising" the data, and extracting a potentially complex signal from the data.

The essence of a wavelet based model is a one-to-one transformation of the observed data into a space of wavelet coefficients. The coefficient space is structured, roughly, according to the location and scale (frequency) of the functional information contained in each coefficient. In the wavelet space, coefficients may be shrunk or reduced in some other fashion. When the reduced coefficients are then transformed back to the data space, the reconstructed function typically has much of the noise removed. See Donoho, Johnstone, Kerkyacharian, and Picard

(1995), as well as DeVore and Lucier (1992).

Shrinkage in the wavelet coefficient space works best in problems where the set of coefficients is *sparse*. That is, most of the coefficients are small, and a few large coefficients explain most of the functional form of the observed data. By shrinking the empirical coefficients towards zero, the smaller coefficients (which contain primarily noise) may be reduced to negligible levels, de-noising the resultant signal.

One natural way to shrink the coefficients is via Bayesian methods. In the Bayesian approach, a prior distribution is placed on each wavelet coefficient. Our prior distribution is designed to capture the sparseness common to most wavelet applications. Some of the mass is concentrated on values close to zero. The rest of the mass is spread to accommodate the possibility of large coefficients. These heavy tailed priors give rise to shrinkage functions which vary the amount of shrinkage according to the magnitude of the coefficient. Smaller coefficients are essentially shrunk to zero, while larger coefficients, which contain more information, are shrunk less. Other Bayesian approaches include Vidakovic (1994) and Clyde, Parmigiani, and Vidakovic (1995).

We propose a simple Bayesian method for shrinking coefficients that gives a closed form result for the shrinkage function, provided that the noise level is known, or can be estimated. Since the majority of other accepted wavelet tools are in closed form, and execute in real time, this is an advantage. Previously suggested Bayesian methods are not in closed form. We present automatic procedures for fixing the prior parameters at each resolution level, resulting in level dependent shrinkage functions. The adaptive nature of the procedure gives rise to its name, 'Adaptive Bayesian Wavelet Shrinkage' (ABWS). Alternatively, the intuitive meaning of each of the prior parameters means that they may also be experimented with easily to adapt the degree of shrinkage and de-noising subjectively.

The reconstructed signal is based on the posterior average of the wavelet coefficients. The posterior variance of the coefficients may also be calculated in closed form, and used to provide uncertainty bands for the reconstruction. We suggest a computationally efficient algorithm for computing these bands.

The paper is organized as follows: In Section 2, the model and prior are outlined, and an example with a minimum of details is given, to illustrate the potential of this approach. Section 3 discusses the parameters of the model, and presents an automatic method for selecting their values. In Section 4, formulas for the posterior mean and variance are given. Section 5 gives a more detailed example of the performance our uncertainty bands. A simulation study is given in Section 6, with comparisons between the proposed method, and two existing methods. Conclusions and discussion of further work are given in Section 7.

2 The Model and an Example

2.1 The Model

The data are assumed to be of the form

$$y_i = f(i/n) + \sigma z_i, \qquad i = 0, 1, \ldots, n-1,$$

where the z_i are independent and identically distributed standard normal observations, i.e. $N(0,1)$. Typically n is an integer power of 2. We let $\{\psi_{j,k}\}_{j,k\in\mathbf{Z}}$ be an orthonormal wavelet basis of $L^2(\mathbb{R})$, where

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$$

is a dilation at scale j and translation by $k/2^j$ of the 'mother' wavelet function ψ. The corresponding wavelet coefficient for $f \in L^2(\mathbb{R})$ will be written as $\theta_{j,k} = \langle f, \psi_{j,k}\rangle$.

In the context of discrete data, there is an analogue to the above. For example, using Daubechies wavelets of compact support (Daubechies (1992)), there exists a corresponding periodic, discrete wavelet transform which takes our n-length vector of observations, $\mathbf{y}$, to a vector ω of equal length. This process may be represented as multiplication by an orthogonal matrix $\mathcal{W}$, yielding the relations

$$\begin{aligned}\omega &\equiv \mathcal{W}\mathbf{y} \\ &= \mathcal{W}\mathbf{f} + \sigma\mathcal{W}\mathbf{z} \\ &\equiv \theta + \sigma\mathbf{z}^* \ ,\end{aligned}$$

where $\theta \equiv \mathcal{W}\mathbf{f}$ is the n-length vector of discrete wavelet coefficients of $\mathbf{f}$, and $\mathbf{z}^*$ is an n-length vector of independent and identically distributed $N(0,1)$ observations. Note that in practice the periodic, discrete wavelet transform and its inverse may be computed in only $O(n)$ operations, using the pyramid filtering algorithm of Mallat (1989).

Our shrinkage functions are derived by incorporating a particular prior structure into the model, in the space of wavelet coefficients. In the prior, the coefficients are mutually independent, and each coefficient is a mixture of two normal distributions i.e.,

$$\theta_{j,k}|\gamma_{j,k} \sim \gamma_{j,k}\ \mathrm{N}(\,0\,,\,c_j^2\tau_j^2\,) + (1-\gamma_{j,k})\ \mathrm{N}(\,0\,,\,\tau_j^2\,)\ \ .$$

The mixture parameter $\gamma_{j,k}$ has its own prior distribution given by

$$P(\gamma_{j,k}=1) = 1 - P(\gamma_{j,k}=0) \equiv p_j\ \ .$$

The quantity σ is assumed known, and p_j, c_j, and τ_j are prior parameters to be chosen. Note that we use the same prior parameters for all coefficients at a given resolution level. The $N(0,\tau_j^2)$ component allows us to concentrate some of the mass near zero, while the $N(0,c_j^2\tau_j^2)$ component spreads out the rest of the mass across larger values. Figure 1(a) depicts two such normal components. Figure 1(b) depicts the resulting mixture distribution ($p = 0.5$).

Conditional on the values of $\theta_{j,k}$ and σ^2, the empirical wavelet coefficients are then distributed as

$$\omega_{j,k}|\theta_{j,k},\sigma^2\ \sim\ \mathrm{N}(\,\theta_{j,k}\,,\,\sigma^2\,)\ \ .$$

Once data are observed, the empirical wavelet coefficients are determined, and we seek the posterior distribution on the unobserved true $\theta_{j,k}$. That is, we seek the distribution of $\theta_{j,k}|\omega_{j,k}$.

We will focus on the expected value and variance of the distribution. Closed form solutions are given in Section 4. Given the expected value, we estimate $\mathbf{f}$ by $\hat{\mathbf{f}} = \mathcal{W}^T E(\theta|\omega)$. Figures 2(a) and (b) show examples of the posterior mean and variance, as functions of the empirical wavelet coefficient, for a particular choice of the hyperparameters.

The simple form of our prior means that τ_j, c_j, and p_j have intuitive interpretations. A coefficient at level j with negligible size would have magnitude comparable to τ_j. Large coefficients at level j are c_j times larger than this. The parameter p_j for the prior on $\gamma_{j,k}$ may be thought of as the proportion of coefficients that are expected to be non-negligible at level j. Automatic methods for selecting values of these parameters based on the data and wavelet theory are discussed in Section 3.

With this basic understanding of the model and prior, we are now equipped to consider a simple example.

2.2 An Example

In this example, we compare our ABWS method to two existing methods of wavelet shrinkage. Figure 3(a) shows a 1024-length vector of nuclear magnetic resonance data. The spatially variable nature of the underlying signal and the moderate signal-to-noise ratio are characteristic of the type of data for which wavelet shrinkage has been found to be particularly appropriate. Figure 3(b) shows a reconstruction of the underlying signal, or a 'de-noising' of the data, using ABWS. In Figures 3(c) and (d) are shown the reconstructions using the VisuShrink and SureShrink methods, respectively, of Donoho and Johnstone (1994, 1995). The goal here is to remove the high-frequency background 'noise', while maintaining the integrity of the underlying signal. Note that the VisuShrink reconstruction does the most complete job of removing 'noise', but tends to attenuate the peaks. SureShrink, on the other hand, attenuates much less, but at the expense of including what appear to be more high-frequency artifacts. The ABWS method, however, strikes a happy medium between the two, as it possess nearly the same 'noise-free' visual quality of VisuShrink, but without the attenuation.

All three reconstructions used Daubechies wavelets of order 3. The VisuShrink reconstruction shrunk coefficients at resolution levels $j = 5, \ldots, 9$. The SureShrink and ABWS reconstructions shrunk coefficients at resolution levels $j = 3, \ldots, 9$, the maximum allowable, as determined by the length of the corresponding wavelet filters. As outlined in the previous subsection, the ABWS reconstruction was created using level-dependent shrinkage functions. The hyperparameters for each function were determined by the automatic method described in Section 3. To illustrate how the ABWS shrinkage function adapts to different levels, the shrinkage functions for all levels are shown in Figure 4. Each line corresponds to a different level j. In this example, as j approaches the coarser (i.e. smaller j) levels, values are shrunk less. This makes sense, since the information content is usually highest at coarser levels.

An additional feature of our method, one not possessed by the above two competitors, is the ability to quantify some degree of uncertainty in the reconstruction. Figure 5 shows the ABWS reconstruction with upper and lower uncertainty bands, obtained through usage of the posterior variance information, as detailed in Section 4. The narrowness of the bands is attributable to the high signal-to-noise ratio. In Section 5 we will see an example in which this relationship is studied in more detail.

3 Choosing the Prior

Our normal mixture prior for each $\theta_{j,k}$,

$$\theta_{j,k}|\gamma_{j,k} \sim \gamma_{j,k}\ \mathrm{N}(\,0\,,\,c_j^2\tau_j^2\,) + (1-\gamma_{j,k})\ \mathrm{N}(\,0\,,\,\tau_j^2\,)\ ,$$

depends on the constants (hyperparameters) τ_j, c_j and p_j. In order to use our prior, values for these constants must be chosen. In this section we discuss the choice of these values, and how this choice relates to the way in which the wavelet coefficients are shrunk. We give simple recommendations for choices of τ_j, c_j and p_j which we have found to work well in a variety of situations. Given our recommended choices, the normal mixture prior provides a simple, automatic approach to wavelet shrinkage. Alternatively, the simple form of the prior and the intuitive roles of τ_j, c_j and p_j make it easy to "play" with the prior in order to obtain a variety of shrinkage estimates resulting in different amounts of smoothness in the corresponding estimate of the function f.

We first discuss the general role of τ_j, c_j and p_j in determining the shrinkage of the wavelet coefficients. We then present our recommended default choices.

3.1 The Role of the Hyperparameters

The wavelet shrinkage estimation procedure will work well when the coefficient set is 'sparse': there are a few large coefficients and the rest are small. Our prior directly captures this intuition by quantifying "a few", "small", and "large". The $\mathrm{N}(0, \tau_j^2)$ component of the mixture is meant to describe a small coefficient. We choose τ_j so that if the coefficient is in the interval $(-3\tau_j, 3\tau_j)$ it is so small that for practical purposes it might as well be zero. The $\mathrm{N}(0, c_j^2\tau_j^2)$ is meant to describe "large". This normal component should be sufficiently spread out to cover the full range of plausible coefficients ($c_j \gg 1$). The parameter p_j may then be interpreted as the probability that a coefficient is non-negligible. It is the percentage of coefficients at level j that we expect to be appreciably different from zero. Small values of p_j represent the idea of sparseness.

Note that we choose different values of the parameters at different levels (as illustrated in Figure 4). For the c_j and τ_j this is natural, because whether a coefficient is small or large depends in part on the range of the corresponding wavelet, which is related to the resolution level. For the p_j, we would expect a smaller percentage of the coefficients at higher resolution levels to be large, suggesting that p_j decrease as the resolution level increases.

To understand how the choice of hyperparameters relates to the ultimate shrinkage of the coefficient we write,

$$\begin{aligned} E(\theta_{j,k}|\omega_{j,k}) &= E_{\gamma_{j,k}|\omega_{j,k}} E(\theta_{j,k}|\omega_{j,k},\gamma_{j,k}) \\ &= Pr(\gamma_{j,k}=1|\omega_{j,k})E(\theta_{j,k}|\omega_{j,k},\gamma_{j,k}=1) + Pr(\gamma_{j,k}=0|\omega_{j,k})E(\theta_{j,k}|\omega_{j,k},\gamma_{j,k}=0) \\ &= Pr(\gamma_{j,k}=1|\omega_{j,k})\frac{(c_j\tau_j)^2}{\sigma^2+(c_j\tau_j)^2}\omega_{j,k} + Pr(\gamma_{j,k}=0|\omega_{j,k})\frac{\tau_j^2}{\sigma^2+\tau_j^2}\omega_{j,k} \end{aligned}$$

The shrinkage function may be interpreted as a smooth interpolation between two lines of slope

$$\frac{\tau_j^2}{\sigma^2+\tau_j^2} \qquad \text{and} \qquad \frac{(c_j\tau_j)^2}{\sigma^2+(c_j\tau_j)^2}$$

See Figure 2(a) .

When $\omega_{j,k}$ is small this suggests that $\theta_{j,k}$ is small so that then $Pr(\gamma_{j,k}=0|\omega_{j,k})$ is large. In this case we see that the shrinkage function approximately follows a straight line with intercept zero and slope $\frac{\tau_j^2}{\sigma^2+\tau_j^2}$. When τ_j is small so is this slope. Thus, relatively small values of τ_j give us the flat portion of the shrinkage function around zero. When $\omega_{j,k}$ is large the shrinkage function approximately follows a straight line with intercept zero and slope $\frac{(c_j\tau_j)^2}{\sigma^2+(c_j\tau_j)^2}$. For large c_j this slope will be close to 1. Thus our shrinkage function is obtained as a weighted average of a linear function with a small slope and a linear function with a slope close to 1. As $\omega_{j,k}$ increases, the weight on the line with the larger slope increases to 1. The parameters c_j and τ_j determine the slopes of the two lines. Small values of p_j will increase the width of the interval about zero where the shrinkage function clings to the line with the smaller slope. Given τ_j and p_j, increasing c_j will shorten the interval in which the shrinkage function climbs from the line with the smaller slope up to the line with the larger slope (making the flat portion larger). This is because as c_j increases the two alternative components of the mixture become more sharply distinguished.

3.2 Default Choices for the Hyperparameters

In this section we describe our default choices for the hyperparameters. All of the examples we present in this paper employ the defaults. The default choices are motivated by the discussion in the previous section.

First of all, to choose τ_j we must decide what a "small" coefficient is. Our basic equation $\mathbf{f} = \mathcal{W}^T\theta$ relates the coefficients to the function $\mathbf{f}$. Let $\mathcal{W}^T_{j,k}$ be the *column* of $\mathcal{W}^T$ corresponding to $\theta_{j,k}$. The contribution of the (j,k) coefficient to $\mathbf{f}$ is then $\mathcal{W}^T_{j,k}\theta_{j,k}$. Since the average value of $\mathcal{W}^T_{j,k}$ will be zero, the contribution will be visually negligible if the maximum value of $\mathcal{W}^T_{j,k}\theta_{j,k}$ is not much different than the minimum. Let $M_j = \max \mathcal{W}^T_{j,k}$ and $m_j = \min \mathcal{W}^T_{j,k}$. Then the contribution is small if $(M_j - m_j)\theta_{j,k}$ is small. $(M_j - m_j)\theta_{j,k}$ is the maximum perturbation in $\mathbf{f}$ due the the (j,k) coefficient. Let ϵ be a perturbation in $\mathbf{f}$ which is considered to be negligible. Since our prior places $\theta_{j,k}$ in $(-3\tau_j, 3\tau_j)$ with high probability we choose τ_j as,

$$3\tau_j = \frac{\epsilon}{M_j - m_j}$$

In practice we have found that choosing ϵ to be the first percentile of the set of values $\{|y_{i+1} - y_i|\}_{i=0}^{i=n-2}$ works well.

Rather than choosing c_j directly it is easier to choose the product $c_j\tau_j$. To choose $c_j\tau_j$ we must have an idea of what the range of plausible values are for the corresponding coefficient. A simple bound from real analysis tells us that for $f \in L^2(\mathbb{R})$

$$\left|\int f(x)\psi_{j,k}(x)dx\right| \leq ||\psi_{j,k}||_{L^1}\,||f||_{L^\infty} = 2^{-j/2}\,||\psi||_{L^1}\,||f||_{L^\infty} \;\; .$$

Since $3c_j\tau_j$ represents an upper bound on what we think of as 'signal' coefficients, we would set

$$3c_j\tau_j = 2^{-j/2}\,||\psi||_{L^1}\,||f||_{L^\infty}$$

if we had such information. Instead, we settle for the estimate

$$\widehat{c_j\tau_j} = \frac{2^{-j/2}||\psi||_{L^1}\,[\max_{0\leq i\leq N-1}|y_i|]}{3} \;\; .$$

An approximation to the norm involving the underlying wavelet may be calculated numerically. Also, note that using the maximum of the absolute values tends to overestimate the sup norm of f. Despite these potential concerns, the resulting estimate has been found to work well in practice.

The value p_j is the probability, at resolution level j, that a given wavelet coefficient $\theta_{j,k}$ will contain 'signal'. Donoho *et al.* (1995) have suggested the 'universal threshold value"

$$t_n = \sqrt{2\log(n)}\,\sigma$$

as a probabilistic upper bound on the size of the 'noise' over all n empirical coefficients. Interpreting this value as a cut-point which separates 'signal' from 'noise', we define

$$\hat{p}_j = \frac{\#\{\omega_{j,k} : |\omega_{j,k}| > \sqrt{2\log(2^j)}\,\hat{v}_j^{Noise}\}}{2^j} \quad ,$$

where

$$\hat{v}_j^{Noise} = \sqrt{\hat{\sigma}^2 + \hat{\tau}_j^2} \quad .$$

In other words, $\hat{p}_j$ is just the proportion of coefficients declared to be 'signal' using the '$\sqrt{2\log(n)}$-rule'. Note that we could similarly use any other rule, such as the minimax thresholds of Donoho and Johnstone (1994).

Throughout our discussion we have assumed that σ^2 is known. In practice we estimate σ and then plug in the estimate to our procedure, acting as if the estimate were the true value. Since good estimates of σ are often available from the high level coefficients we feel that the error resulting from this substitution will be small generally. Specifically, most coefficients at the highest resolution level $j = J-1$ will be 'noise' i.e.,

$$\omega_{j,k} \sim \text{Normal}(0, \sigma^2 + \tau_{J-1}^2) \quad .$$

Hence we may estimate $v_{J-1}^{Noise} \equiv \sqrt{\sigma^2 + \tau_{J-1}^2}$ by

$$\hat{v}_{J-1}^{Noise} = \frac{\text{Median}_{0 \le k \le 2^{J-1}-1} |\omega_{J-1,k}|}{0.6745} \quad ,$$

and then σ as

$$\hat{\sigma} = \sqrt{(v_{J-1}^{Noise})^2 - (\hat{\tau}_{J-1})^2} \quad .$$

In other words, our estimate of σ is similar to that of Donoho *et al.* (1995), with the obvious adjustments made for the present context.

4 Means and Variances of Coefficients and Function Values

In this section we present the details for the computation of posterior means and variances. With the assumption that σ is known, simple closed form formulas are available for $E(\theta_{j,k}|\omega_{j,k})$ and $Var(\theta_{j,k}|\omega_{j,k})$. Since the $\theta_{j,k}$ are independent in the posterior, we thus have the posterior covariance matrix of the vector of the $\theta_{j,k}$ which we denote by D. The posterior mean of $\mathbf{f}$ is then obtained from the linear wavelet transformation:$E(\mathbf{f}) = \mathcal{W}^T E(\theta|\omega)$. The covariance matrix of $\mathbf{f}$ is $\mathcal{W}^T D\mathcal{W}$. We discuss an efficient computational strategy for obtaining this matrix.

Although we concentrate on the first two moments in our inference we actually can obtain the posterior distribution as a whole. To simplify the presentation of the formulas we will write drop the j and k subscripts (so, for example, θ will refer to a single wavelet coefficient). We obtain the posterior for θ by first calculating the marginal posterior of γ and then that of $\theta|\gamma$. The marginal posterior for γ may be shown to be

$$P(\gamma = 1|\omega) = \frac{O}{O+1} \ ,$$

where

$$O = \frac{p\,\pi(\omega|\gamma = 1)}{(1-p)\pi(\omega|\gamma = 0)} \ ,$$

and $\pi(\omega|\gamma = 1) \sim N(0, \sigma^2 + c^2\tau^2)$ and $\pi(\omega|\gamma = 0) \sim N(0, \sigma^2 + \tau^2)$.

The posterior distribution of θ can then be expressed as

$$F(\theta|\omega) = F(\theta|\omega, \gamma = 1) \cdot \frac{O}{1+O} + F(\theta|\omega, \gamma = 0) \cdot \frac{1}{1+O} \ ,$$

where $F(\theta|\omega, \gamma = 1)$ is the distribution function for

$$\theta|\omega, \gamma = 1 \ \sim \ \text{Normal}\left(\frac{(c\tau)^2}{\sigma^2 + (c\tau)^2}\omega \, , \, \frac{\sigma^2 (c\tau)^2}{\sigma^2 + (c\tau)^2} \right)$$

and $F(\theta|\omega, \gamma = 0)$ is the distribution function for

$$\theta|\omega, \gamma = 0 \sim \text{Normal}\left(\frac{\tau^2}{\sigma^2 + \tau^2}\omega \, , \, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} \right) \ .$$

The posterior mean of θ is then,

$$E[\theta|\omega] = \left[\frac{(c\tau)^2}{\sigma^2 + (c\tau)^2} \cdot \frac{O}{O+1} + \frac{\tau^2}{\sigma^2 + \tau^2} \cdot \frac{1}{O+1} \right] \cdot \omega \ .$$

The result is a multiplication of ω by a shrinkage factor

$$s = \frac{(c\tau)^2}{\sigma^2 + (c\tau)^2} \cdot \frac{O}{O+1} + \frac{\tau^2}{\sigma^2 + \tau^2} \cdot \frac{1}{O+1} \ ,$$

where $|s| \leq 1$. Since s itself is a function of ω, estimation of θ via the posterior mean $E[\theta|\omega]$ is equivalent to the usage of a nonlinear shrinkage function.

We now derive the posterior variance. Begin with the relation

$$\begin{aligned} \text{Var}(\theta|\omega) &= E_\gamma[\text{Var}(\theta|\omega, \gamma)] + \text{Var}_\gamma(E[\theta|\omega, \gamma]) \\ &= E_\gamma[\text{Var}(\theta|\omega, \gamma)] + E_\gamma[(E[\theta|\omega, \gamma])^2] - (E_\gamma[E[\theta|\omega, \gamma]])^2 \ . \end{aligned}$$

The above results yield the expressions

$$E_\gamma[\text{Var}(\theta|\omega, \gamma)] = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} \cdot \frac{1}{1+O} + \frac{\sigma^2(c\tau)^2}{\sigma^2 + (c\tau)^2} \cdot \frac{O}{1+O}$$

and

$$E_\gamma[(E[\theta|\omega, \gamma])^2] = \left(\frac{\tau^2}{\sigma^2 + \tau^2}\omega \right)^2 \cdot \frac{1}{1+O} + \left(\frac{(c\tau)^2}{\sigma^2 + (c\tau)^2}\omega \right)^2 \cdot \frac{O}{1+O}$$

and, of course,

$$(E_\gamma[E[\theta|\omega, \gamma]])^2 = (E[\theta|\omega])^2 \ .$$

For small values of ω,

$$\text{Var}(\theta|\omega) \approx \tau^2 \ ,$$

and for large values of ω,

$$\text{Var}(\theta|\omega) \approx \sigma^2 \ .$$

In between these two extremes is where the variance is greatest, which makes intuitive sense: This is precisely the range in which we are *least* sure whether an underlying coefficient contains 'signal' or not. See Figure 2(b).

Given the first two moments of θ (now let θ and ω denote the vectors of all wavelet coefficients) it is straightforward to compute the first two moments of $\mathbf{f}$. Our reconstruction, $\hat{\mathbf{f}}$, is the result of inverting the posterior mean vector $E[\theta|\omega]$ i.e.,

$$\hat{\mathbf{f}} = \mathcal{W}^T \left(E[\theta|\omega] \right) \ .$$

The covariance matrix for this reconstruction vector is just

$$\Delta \equiv \mathcal{W}^T D \mathcal{W} \ ,$$

where $D \equiv \text{diag}\{\text{Var}(\theta|\omega)\}$. Calculation of Δ may be accomplished in $O(n^2)$ operations, instead of the expected $O(n^3)$ operations, by noting that

$$\Delta \equiv \mathcal{W}^T D \mathcal{W} = \mathcal{W}^T (\mathcal{W}^T D)^T$$

suggests that we compute the $O(n)$ inverse wavelet transform of each of the n columns of D, transpose the result, and repeat.

The matrix Δ may be viewed as a gray-scale image, and perhaps serve to give a visual representation of the pattern of associations within $\mathbf{f}$. Also, since typically Δ will be *roughly* a diagonal matrix, we can augment our plot of $\mathbf{f}$ by plotting upper and lower bands

$$\mathbf{f} \pm 3\sqrt{\text{diag}(\Delta)} \ .$$

Of course these intervals are not frequentist confidence intervals but Bayesian posterior intervals based on the assumed form of prior. They should still serve to give some indication of variability. Note that for simplicity we use the bands based on the posterior variances even though the posteriors of the individual wavelet coefficients are bimodal.

5 Uncertainty Bands: An example.

In this section we consider a brief example to illustrate the quality of the uncertainty bands discussed in Section 4. In Section 2 we saw that the high signal-to-noise ratio of the nuclear magnetic resonance data lead to extremely tight uncertainty bands. In general, we would expect that higher signal-to-noise ratios lead to more accurate reconstructions and tighter uncertainty bands.

Figure 6(a) shows data and the ABWS reconstruction for the 'Blocks' signal of Donoho *et al.* (1995). The signal-to-noise ratio, $SD(f)/\sigma$, is 7. In Figure 6(b) is shown the upper and lower uncertainty bands for this reconstruction, as well as the true underlying signal. As expected, the reconstruction is quite accurate, displaying a minimum amount of noise and Gibbs phenomena. Accordingly, the uncertainty bands are quite tight.

Similarly, Figure 7(a) shows data and the ABWS reconstruction for the 'Blocks' signal with a signal-to-noise ratio of 1, while Figure 7(b) shows the uncertainty bands and the 'Blocks' signal itself. Considering that the underlying signal is difficult to distinguish with the human eye, the ABWS reconstruction still does rather well. Even more encouraging is the fact that the true signal is contained almost entirely within the uncertainty bands.

6 Simulations.

The standard VisuShrink (Donoho and Johnstone, 1994) approach to de-noising with wavelets uses the soft-threshold function and the universal threshold. This approach is *not* intended to minimize mean squared error, but rather to achieve a type of 'near-minimax' optimality. The result is an estimator which achieves a low variance, at the expense of bias. In the SureShrink approach (Donoho and Johnstone, 1995), the soft-threshold function is retained, but the universal threshold is used only at the highest resolution levels where the coefficients contain primarily noise. At lower resolution levels the thresholds are obtained by selecting a value which minimizes an estimate of expected mean squared error (in the space of wavelet coefficients). Hence the bias of the resulting estimator should be less than that of VisuShrink, at the expense of an increase in variance.

Similarly, our Bayesian approach (ABWS), by definition, seeks to minimize mean squared error, using shrinkage functions which typically appear to be smooth interpolations of threshold functions. Hence we might expect properties similar to those of SureShrink. A simulation study was conducted to compare the performance of these three estimators.

Figure 8 shows four standard test functions, created by Donoho and Johnstone. Each function was sampled at $N = 1024$ points. Noise distributed as $N(0,1)$ was added, and reconstructions were created using VisuShrink, SureShrink, and ABWS. A total of 1000 trials were conducted for each of the four signals. In the VisuShrink reconstructions, de-noising was done only down through resolution level $j = 5$. For the other two methods, de-noising was done through the lowest possible resolution level, as dictated by the lengths of the corresponding wavelet filters. For the function 'Bumps', Daubechies wavelets of order 3 were used; for 'Blocks', Haar wavelets; and for 'Doppler' and 'HeaviSine', most nearly symmetric Daubechies wavelets of order 8.

In order to evaluate performance, we used estimates of the integrated mean squared error $\int E[(\hat{f}(x) - f(x))^2]\,dx$, and its decomposition into bias and variance components i.e.,

$$\begin{aligned} \int E\left[\left(\hat{f}(x) - f(x)\right)^2\right] dx &= \int E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] dx + \int \left(E[\hat{f}(x)] - f(x)\right)^2 dx \\ &= \int Var(\hat{f}(x))\,dx + \int \mathrm{Bias}^2(\hat{f}(x))\,dx \ . \end{aligned}$$

In Table 1 we give estimates of these three values (approximations of the integrals by sums over the 1000 trials) for the four signals, and the three methods of reconstruction.

The simulations show that ABWS has uniformly smaller mean squared error over the four test signals, with SureShrink second and VisuShrink third in the rankings. As expected, both

ABWS and SureShrink show marked improvements over VisuShrink in bias, but at the expense of an increase in variance. Note that the proportional change in both bias and in variance varies considerably across signals. Except for the case of the 'HeaviSine' signal, SureShrink and ABWS tend to roughly double the variance to achieve reductions in bias ranging from about 50% to less than 10% of the bias of VisuShrink. The least striking improvements over VisuShrink occur in the 'HeaviSine' signal, the least spatially variable in some sense.

Compared to its more appropriate competitor, SureShrink, the ABWS method still exhibits a smaller bias across all four test functions, and a similar variance. The reconstruction of the 'Blocks' signal is especially striking, however, where ABWS has approximately 12% the bias of SureShrink *and* only about 67% the variance. In fact, its variance is quite close to that of VisuShrink, approximately 20% larger, while its bias is only 2% the size of that of VisuShrink !

7 Conclusion

In this paper, we have developed what amounts to a class of shrinkage functions for wavelet shrinkage by approaching the standard context from a Bayesian point of view. An automatic method has been proposed whereby a set of level-dependent shrinkage functions may be chosen adaptively for a given dataset. Using this method, our experience has been that as the resolution level decreases, the chosen shrinkage functions approach, and sometimes become, the identity function i.e., no shrinkage is done. This result indicates that not only are the shrinkage functions chosen adaptively, but where we stop shrinking is also chosen adaptively. In principle, our method even might choose to not shrink at a given resolution level, and yet shrink slightly at a lower level if deemed necessary.

Additionally, we have offered a method by which the uncertainty in a reconstruction may be quantified and displayed. The method is simple, both conceptually and computationally. The posterior variance functions upon which this method is based are also interesting in and of themselves, as they indicate where in the process of shrinkage one is more or less sure of a particular value.

In general, we believe the Bayesian approach to wavelet shrinkage offers a conceptually simple way to obtain intuitively appealing shrinkage functions. Our goal in this paper is to choose a prior in a way that is simple and yet reflects the structure of the wavelet problem. Other prior choices may also lead to reasonable results. For example, Clyde, Parmigiani, and Vidakovic (1996) specify a mixture prior similar to ours, but the small component (our $N(0, \tau^2)$) is simply a point mass at zero. Although we could effectively achieve this point mass by taking the limit as $\tau \to 0$, this would fix the slope of the shrinkage function at zero for small coefficient values. With regards to the larger $N(0, c\tau)$ component, Clyde *et al.* base their choice of $c\tau$ (which they call c) on general model selection results. We feel our automatic choice is more closely adapted to the wavelet problem. Another basic difference from that of Clyde *et al.* is that we assume a reasonably good estimate of σ is available. This gives the simple closed form of our posterior mean and variance for the wavelet coefficients. Clyde *et al.* put a prior on σ and infer its value directly as part of the Bayesian calculation. In order to compute the posterior with σ unknown, they are forced to use Monte Carlo or approximation methods. In situations where a reasonably good estimate of σ is not available, their method may be preferable. We also note that the resulting shrinkage functions do not have the smooth, monotonic character displayed in Figure 2(a).

An interesting extension to this work would be to incorporate the fact that wavelets at different levels, *but at the same relative position*, are similarly large or small, depending on the underlying function f. One should be able to incorporate such information into the prior distribution. Rather than the current independent priors on γ's and θ's, correlations could express beliefs about such relations. This would borrow strength from different levels of coefficients, and might prove useful in, for example, edge detection in image processing applications. We note, however, that priors that are not independent will likely increase the level of computation required.

References.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1995), "Curve Fitting with Wavelets using Model Mixing," Technical Report, ISDS, Duke University.

Daubechies, I. (1992), *Ten Lectures on Wavelets,* Philadelphia, Pennsylvania: SIAM.

DeVore, R.A. and Lucier, B.J. (1992), "Fast wavelet techniques for near-optimal image processing," pgs. 48.3.1-48.3.7., IEEE-ICASSP-93. IEEE Military Communications Conference, New York, NY, 1992.

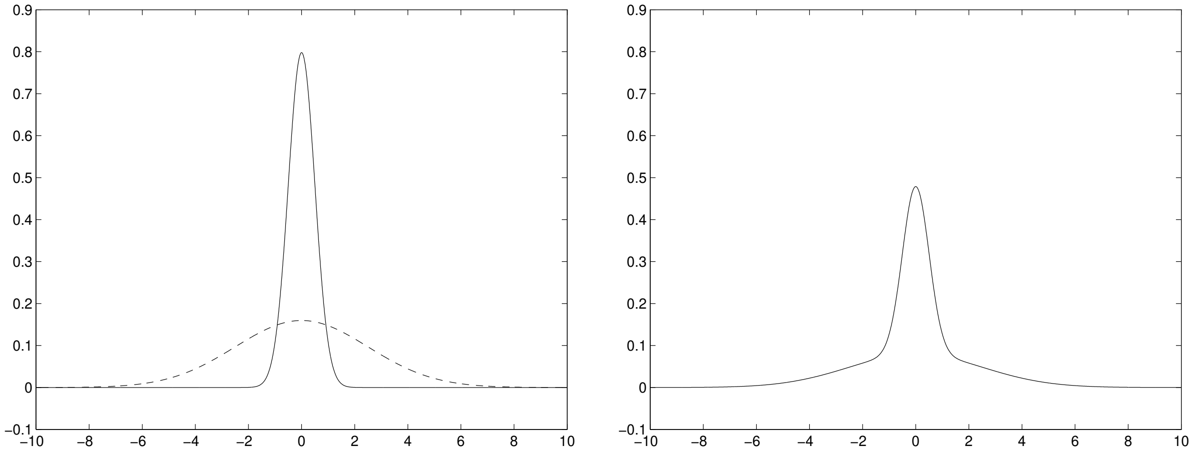
Donoho, D.L., and Johnstone, I.M. (1994), "Ideal spatial adaptation via wavelet shrinkage," *Biometrika,* 81, 425-455.

—— (1995), "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association,* 90, 1200-1224.

Donoho, D.L, Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995), "Wavelet Shrinkage: Asymptopia ?" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 57, 301-370.

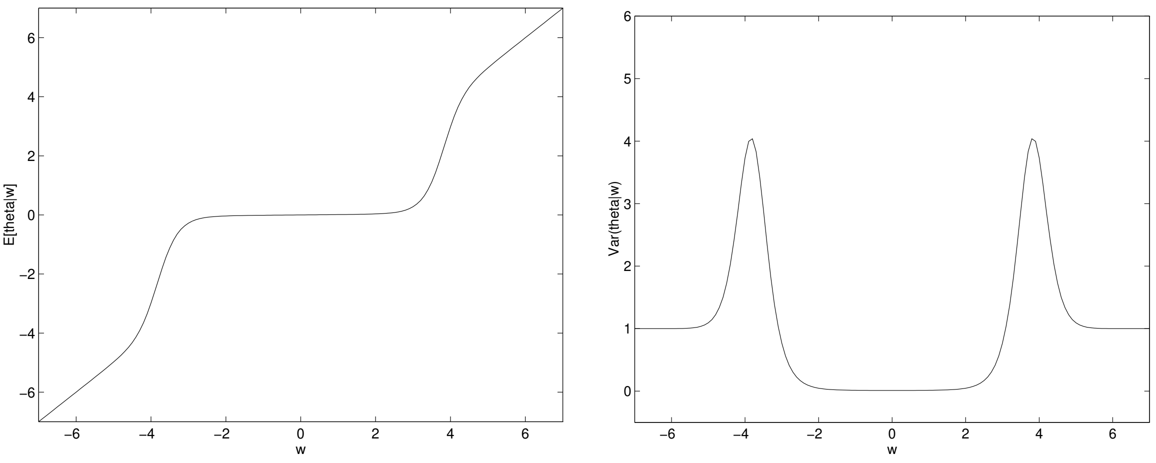
Mallat, S. (1989), "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11, 674-693.

Vidakovic, B. (1994), "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," Discussion Paper 94-24, ISDS, Duke University. (Under revision for the *Journal of the American Statistical Association.*)



(a) (b)

Figure 1: (a): Two zero-mean normal density functions, one concentrated (—) and the other diffuse (−−). (b): Mixture of the normal distributions in (a), with $p = 0.5$.



(a) (b)

Figure 2: (a) and (b): Posterior mean, $E[\theta|\omega]$, and variance, $Var(\theta|\omega)$, as functions of the empirical wavelet coefficient ω. Hyperparameters were chosen as $\tau = 0.1$, $c = 500$, and $p = 0.05$, while σ was fixed at 1.

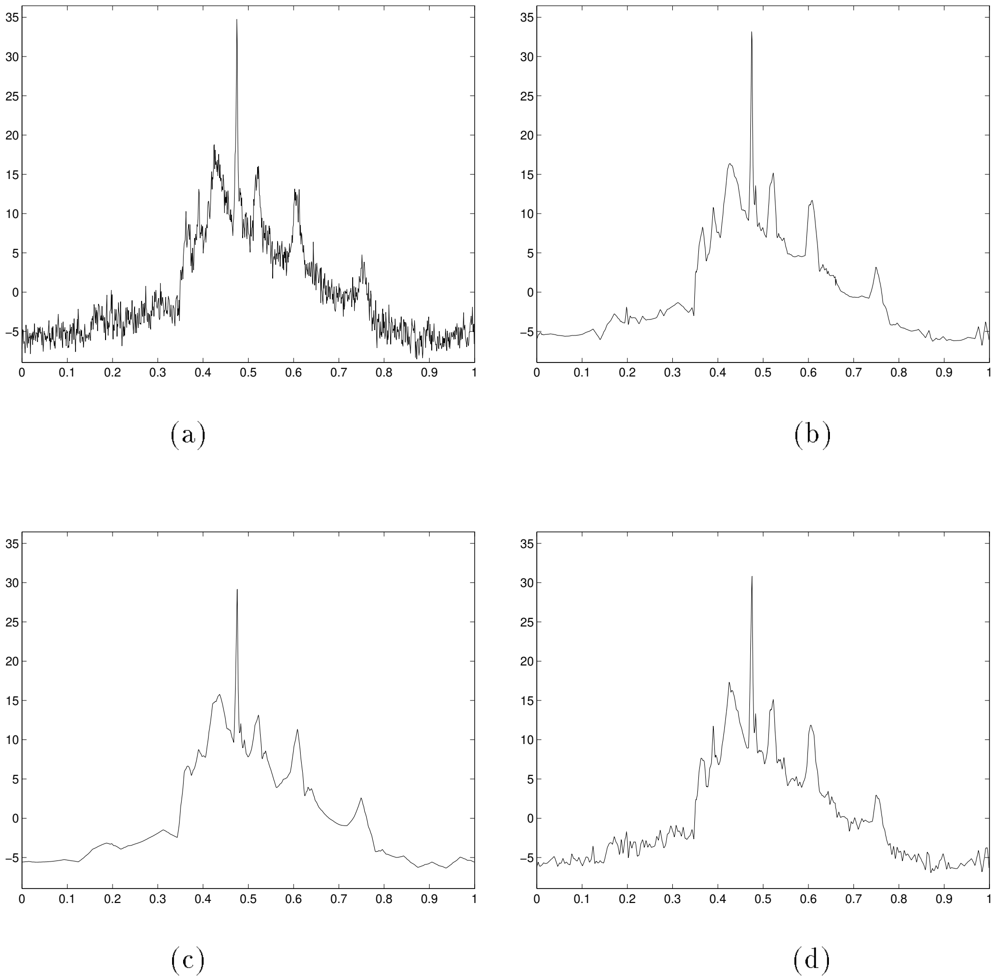


Figure 3: (a): Nuclear magnetic resonance data. (Provided in the WaveLab software package, courtesy of Chris Raphael.) (b-d): Reconstructions using ABWS, VisuShrink, and SureShrink, respectively.

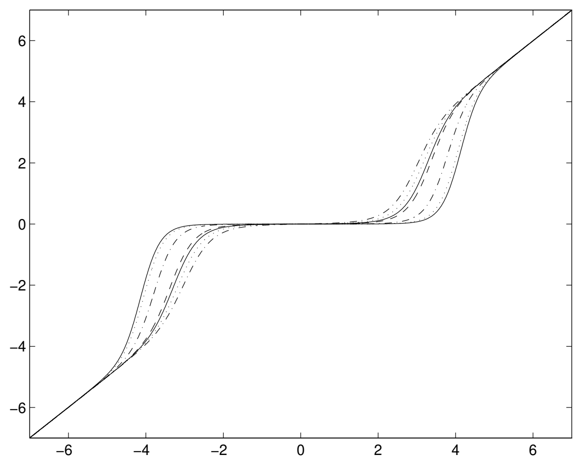


Figure 4: Shrinkage functions for nuclear magnetic resonance data, for resolution levels $j = 3, \ldots, 9$.

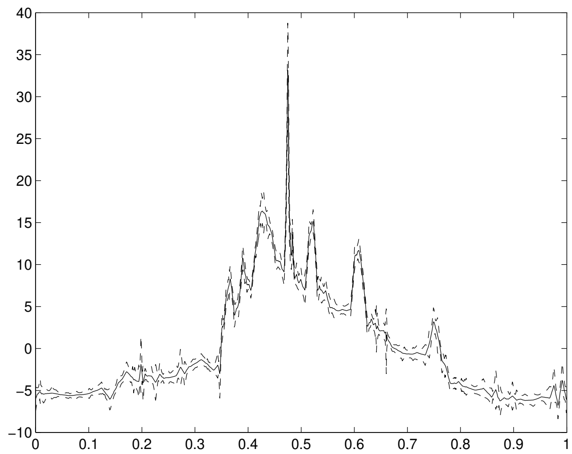
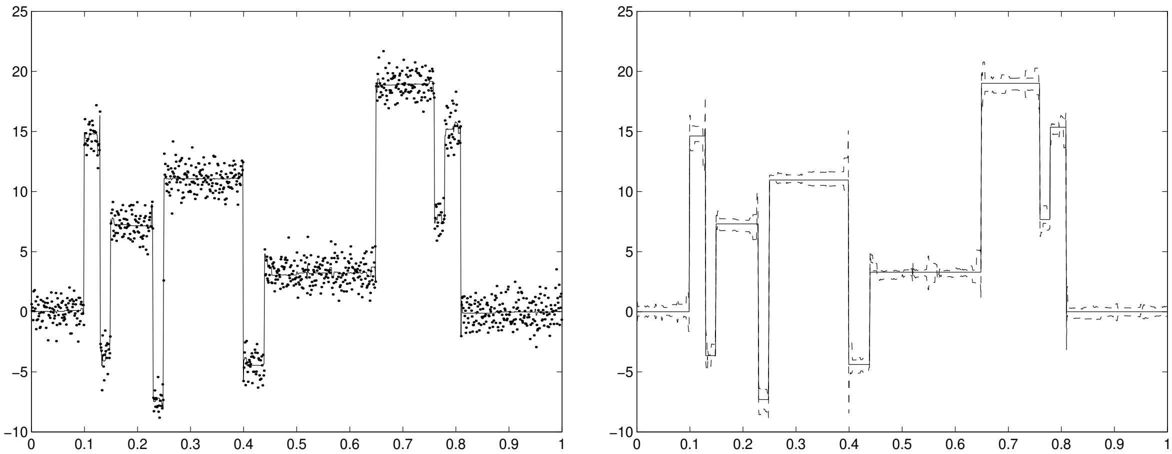
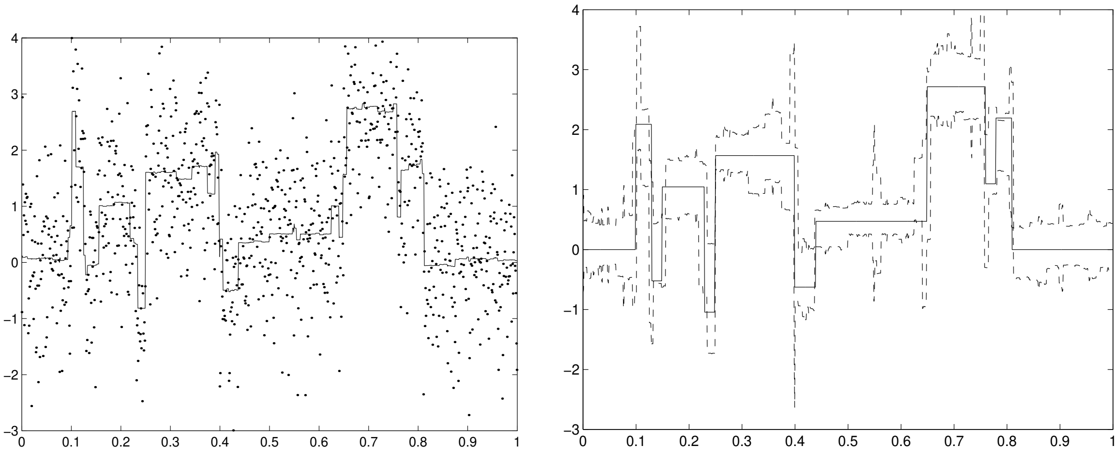


Figure 5: ABWS reconstruction of nuclear magnetic resonance data, with upper and lower uncertainty bands.



(a) (b)

Figure 6: (a): Data and ABWS reconstruction (using Haar wavelets) for the signal 'Blocks' (SNR=7). (b): True function 'Blocks' (−), with upper and lower uncertainty bands (−−).



(a) (b)

Figure 7: (a): Data and ABWS reconstruction for the signal 'Blocks' (SNR=1). (b): True function 'Blocks' (−), with upper and lower uncertainty bands (−−).

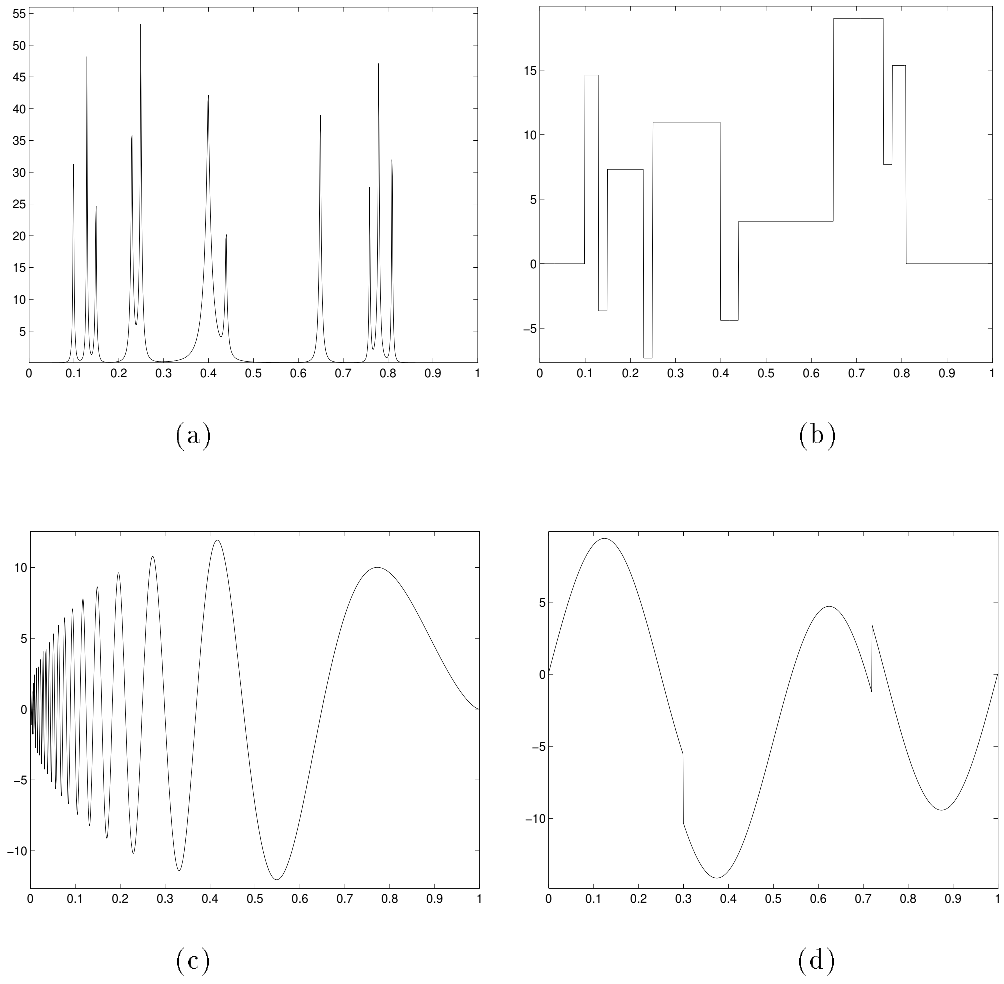


Figure 8: Test signals from Donoho *et al.* (1995). (a) Bumps, (b) Blocks, (c) Doppler, and (d) HeaviSine.

	BUMPS			BLOCKS		
Method	Variance	Bias^2	MSE	Variance	Bias^2	MSE
Visu	0.1165	1.4543	1.5707	0.0719	0.6122	0.6840
Sure	0.2660	0.4167	0.6827	0.1369	0.0856	0.2225
ABWS	0.2228	0.1267	0.3495	0.0874	0.0121	0.0995

	DOPPLER			HEAVISINE		
Method	Variance	Bias^2	MSE	Variance	Bias^2	MSE
Visu	0.0523	0.4327	0.4850	0.0339	0.0864	0.1204
Sure	0.0946	0.1340	0.2285	0.0416	0.0534	0.0949
ABWS	0.1006	0.0640	0.1646	0.0442	0.0433	0.0874

Table 1: Simulation results comparing ABWS, VisuShrink, and SureShrink with respect to mean squared error performance and its decomposition into bias and variance components. Numerical values are based on 1000 trials.