

Nonlinear wavelet shrinkage with Bayes rules and Bayes factors¹

BRANI VIDAKOVIC²
Duke University

Wavelet shrinkage, the method proposed by seminal work of Donoho and Johnstone is a disarmingly simple and efficient way of de-noising data. Shrinking wavelet coefficients was proposed from several optimality criteria. The most notable are the asymptotic minimax and cross-validation criteria. In this paper a wavelet shrinkage by imposing natural properties of Bayesian models on data is proposed. The performance of methods are tested on standard Donoho-Johnstone test functions.

Key Words and Phrases: Wavelets, Discrete Wavelet Transform, Thresholding, Bayes Model.
1991 AMS Subject Classification: 42A06, 62G07.

Contents

1	Introduction	2
2	A review of some standard thresholding policies	3
2.1	Universal threshold	3
2.2	A threshold based on Stein's unbiased estimator of risk	4
2.3	Cross-validation	5
2.4	Other methods	5
3	The model	5
3.1	Calculation of Bayes rules and examples	7
4	Thresholding via Bayesian hypotheses testing	8
5	Conclusion	11
5.1	Acknowledgment	11
6	Appendix	18
6.1	Some S functions for Nason's <code>wavethresh</code>	18
6.1.1	A code for calculating $\hat{M}(\lambda)$	18
6.1.2	BF thresholding	19

¹Part of this paper was communicated at the Joint ASA and IMS Meeting in Toronto, August 1994

²The author is an Assistant Professor at ISDS, Duke University, P.O. Box 90251, Durham, NC 27708-0251 `brani@isds.duke.edu`.
Research supported by NSF Grant DMS-9404151 at Duke University.

1 Introduction

Wavelet transforms and their modifications are becoming increasingly popular in different areas of applied and theoretical science. Data Compression, Signal Processing, Turbulence, Geophysics, Statistics, Numerical Analysis, etc., are only a few from a long list of disciplines where wavelets are used successfully.

Wavelets are the building blocks of wavelet transformations the same way that the functions e^{inx} are the building blocks of the ordinary Fourier transformation. But in contrast to sines and cosines, wavelets can be (or almost can be) supported on an arbitrarily small closed interval. Thus, wavelets are a very powerful tool in dealing with phenomena that change rapidly in time.

Statisticians are interested in wavelets as a modeling tool in the general nonlinear regression scheme. Some particular problems of interest are de-noising, density and function estimation, long range dependence, and change point detection. Though the wavelet regression is very attractive, it has limitations stemming from the intrinsic properties of wavelets:

(i) It is difficult to deal with non-uniform designs. Recently M. E. Bock and her Ph.D. student at Purdue University made progress in applying wavelets in non-uniform designs. (Bock and Lu, 1994)

(ii) No satisfactory (i.e., no fast) methods exist for dealing with data sets of size different than powers of 2. One solution is to augment the data up to the first larger degree of 2, but there is no clear optimal rule for doing this.

However, the “pluses” of wavelet methods make them one of the most interesting, applicable, and burgeoning research areas in mathematics, signal processing, and statistics today. For a nice overview of wavelet applications see Donoho (1993). Standard references on wavelets are Daubechies (1992), Meyer (1992) and Chui (1992). For an elementary introduction to wavelets see Strang (1993) and Vidakovic and Müller (1994).

Nonlinear wavelet shrinkage is the main interest of this paper. The formal statement of the problem is given next.

Let $t_i, i = 1, N (= 2^n)$ be a sequence of equally-spaced points. Let $f_i = f(t_i), i = 1, N$ be the values of the function f evaluated at t_i , and let $y_i = f(t_i) + \epsilon_i$ be the sequence of observations. It is supposed that ϵ_i are iid normal random variables with mean 0 and with unknown variance σ^2 . The problem of interest is estimating $f_i = f(t_i)$.

Donoho and Johnstone (1993) propose a class of simple and efficient procedures for estimating the unknown function f by using wavelet shrinkage. The wavelet shrinkage can be described as a procedure consisting of three main steps:

1. Raw data (a noisy signal, blurred image pixels, etc.) are transformed by a discrete wavelet transformation. Sometimes the nature of the problem calls for appropriate “pre-transformations” such as: variance stabilizing transformations, scaling transformations, and de-cumulating and de-convolving transformations.
2. The empirical wavelet coefficients are shrunk. The shrinkage may be by thresholding (small values of the coefficients are replaced by 0) or continuous.
3. The processed empirical wavelet coefficients are “returned to the time domain”.

A wavelet transformation is a linear transformation and can be defined by an orthogonal $N \times N$ matrix. However in real applications the matrices are not used. Instead, simple filters are used that make the discrete wavelet transformation extremely fast. The calculational complexity of the decomposing algorithm is $O(n)$. For instance, Daubechies’ DAUB2 wavelet filter coefficients are 0.4829629131, 0.8365163037, 0.2241438680, and -0.1294095226.

The choice of threshold, as well as the thresholding policy, is one of the most important issues in the *wavelet regression*. Donoho and Johnstone propose several thresholds (i.e., *universal, SURE*), as well as several thresholding policies. Nason (1994) adjusted the well known cross-validation method for use with wavelets. The threshold is selected by minimizing a cross-validatory estimator of ISE. Some possible generalizations of Nason’s method are described by Wang (1994b). A few other references in threshold selection and wavelet shrinkage applications are Wang (1993), Saito (1993), and Gao (1993a, 1993b).

Motivated by a body of work in nonparametric density estimation using orthogonal series and their intrinsic connections with Bayesian methods (Whittle, 1958; Watson, 1969; Brunk, 1978; Wahba, 1981; to list a few) we approached the problem of wavelet regression from a Bayesian standpoint. Bayes rules are (usually) shrinkers, and estimating wavelet coefficients in Bayesian fashion is already a wavelet shrinkage. Notice that Bayes rules

will not do thresholding, but at best a heavy shrinking of small arguments. If we are not concerned with the compactness but with other aspects of the model, such almost-threshold rules should perform well. Section 3 deals with Bayesian models that yield well behaved wavelet shrinkage rules.

The universal thresholding was heuristically explained in terms of classical testing of precise hypotheses. We shall build on that heuristic assuming a Bayesian model on the wavelet structure and test the hypothesis that wavelet coefficients of the useful signal are equal to zero. Thresholding by testing precise hypotheses is given in Section 4. Since the thresholding involves a Bayes factor we will call it Bayes Factor (BF) thresholding. It is equivalent to hard thresholding and the problem of choice of the threshold is solved through subjective elicitation of relevant parameters in the imposed model.

2 A review of some standard thresholding policies

Wavelets, well localized in time and scale, provide a useful tool in handling noisy data sets. An important feature of wavelets is that they provide unconditional bases³ for not only L^2 , but a variety of smoothness spaces such as Sobolev and Hölder spaces. As a consequence, wavelet shrinkage acts as a smoothing operator. The same can not be said about the Fourier basis. By shrinking Fourier coefficients one can get bad results in terms of mean square error. Also, some bad visual artifacts can be obtained, see Donoho (1993).

Why is thresholding good? The parsimony of wavelet transformations ensures that the signal of interest can be described by a relatively small number of coefficients. A simple Taylor series argument shows that if the mother wavelet has L vanishing moments and the unknown “signal” is in C^{L-1} , then

$$|d_{jk}| \leq \text{const} \cdot 2^{-j(L-1/2)} \int |y|^L |\psi(y)| dy.$$

For j large (fine scales) this will be negligibly small. For a nice discussion on compromise between regularity (number of vanishing moments) and the mother wavelet support see Daubechies (1992), page 244.

The coefficients corresponding to the noise are the noise themselves, because the wavelet transformation is linear and orthogonal (or almost orthogonal, depending on how the boundaries are handled).

The process of thresholding wavelet coefficients can be divided into two steps. The first step is the policy choice, i.e., the choice of the threshold function T . Two standard choices are: **hard** and **soft** thresholding with corresponding transformations given by:

$$T^{hard}(d, \lambda) = d \mathbf{1}(|d| > \lambda), \tag{1}$$

$$T^{soft}(d, \lambda) = (d - \text{sgn}(d)\lambda) \mathbf{1}(|d| > \lambda). \tag{2}$$

The “hyperbola” function:

$$T^{hyper}(d, \lambda) = \text{sgn}(d) \sqrt{d^2 - \lambda^2} \mathbf{1}(|d| > \lambda), \tag{3}$$

is a compromise between hard and soft thresholding functions, (Vidakovic, 1994). The function T^{hyper} is an “almost” hard thresholder with the continuity property.

Another class of useful functions are shrinkage (tapering) functions. A function S from that class exhibits the following properties:

$$S(d) \sim 0, d \text{ small}; \quad S(d) \sim d, d \text{ large}.$$

Bayes rules, obtained in Subsection 3.1, are S -type functions.

The second step is the choice of a threshold. In the following subsections we briefly discuss some of the standard methods of selecting a threshold.

2.1 Universal threshold

Donoho and Johnstone (1993) propose a threshold λ based on the following result.

³Informally, a family $\{\psi_i\}$ is an unconditional basis for a space S if one can decide if the element $f = \sum_i a_i \psi_i$ belongs to S by looking only at $|a_i|_s$.

Result: Let z_i be iid standard normal random variables. Define

$$A_n = \{\max_{i=1,n} |z_i| \leq \sqrt{2 \log n}\}.$$

Then

$$\pi_n = P(A_n) \rightarrow 0, n \rightarrow \infty.$$

In addition, if

$$B_n(t) = \{\max_{i=1,n} |z_i| > t + \sqrt{2 \log n}\}.$$

then $P(B_n(t)) < e^{-\frac{t^2}{2}}$. That motivates the following threshold:

$$\lambda^U = \sqrt{2 \log n} \hat{\sigma}, \quad (4)$$

which Donoho and Johnstone call *universal*. This threshold is one of the first proposed and provides an easy, fast, and automatic thresholding. The rationale is to remove all wavelet coefficients that are smaller than the expected maximum of an assumed iid normal noise sequence of given size. There are several possibilities for the estimator $\hat{\sigma}$. Almost all methods involve the wavelet coefficients of the finest scale. The signal-to-noise ratio is smallest at high resolutions in a wavelet decomposition for almost all reasonably behaved signals.

Some standard estimators are:

$$(i) \quad \hat{\sigma} = \frac{1}{N/2 - 1} \sum_{i=1}^{N/2} (d_{n-1,i} - \bar{d})^2,$$

or a more robust

$$(ii) \quad \hat{\sigma} = 1/0.6745 \text{ MAD}(\{d_{n-1,i}, i = 1, N/2\}),$$

where $n - 1$ is the highest level.

In some problems, especially with large data sets and when the σ is over-estimated, the universal thresholding gives underfitted models. For contaminated Doppler data set the universal threshold is found to be $\lambda^U = 1.17$ (Theoretically, it should be 0.94.).

2.2 A threshold based on Stein's unbiased estimator of risk

Donoho and Johnstone (1994) developed a technique of selecting a threshold by minimizing Stein's unbiased estimator of risk.

Result: Let $x_i \stackrel{iid}{\sim} N(\mu_i, 1), i = 1, k$. Let $\hat{\mu}$ be an estimator of $\underline{\mu} = (\mu_1, \dots, \mu_k)$. If the function $\mathbf{g} = \{g_i\}_{i=1}^k$ in representation $\hat{\mu}(\underline{x}) = \underline{x} + \mathbf{g}(\underline{x})$ is weakly differentiable, then

$$E^\mu \|\hat{\mu} - \underline{\mu}\|^2 = k + E^\mu \|\mathbf{g}(\underline{x})\|^2 + 2 \nabla \mathbf{g}(\underline{x}), \quad (5)$$

where $\nabla \mathbf{g} = \frac{\partial}{\partial x_i} g_i$. It is interesting that estimator $\hat{\mu}$ can be nearly arbitrary; for instance, biased and non-linear.

The application of (5) to $T^{soft}(\underline{x}, \lambda)$ gives:

$$\text{SURE}(\underline{x}, \lambda) = k - 2 \sum_{i=1}^k \mathbf{1}(|x_i| \leq \lambda) + \sum_{i=1}^k (|x_i| \wedge \lambda)^2. \quad (6)$$

The SURE is an unbiased estimator of risk, i.e.,

$$E \|T^{soft}(\underline{x}, \lambda) - \underline{\mu}\|^2 = E \text{SURE}(\underline{x}, \lambda).$$

The LLN argument motivates the following threshold selection:

$$\lambda^{sure} = \arg \min_{0 \leq \lambda \leq \lambda^U} \text{SURE}(\underline{x}, \lambda). \quad (7)$$

It is possible to derive a SURE-type threshold for T^{hard} and T^{hyper} but the simplicity of the representation (6) is lost.

2.3 Cross-validation

Nason (1994) proposed a very interesting cross-validatory threshold selection procedure. From the original noisy data set y_i , $i = 1, N (= 2^n)$, two subsequences are formed:

$$\bar{y}_i^{ODD} = \frac{y_{2i-1} + y_{2i+1}}{2}, i = 1, N/2; y_{N+1} = y_{N-1}, \quad (8)$$

and

$$\bar{y}_i^{EVEN} = \frac{y_{2i} + y_{2i+2}}{2}, i = 1, N/2; y_{N+2} = y_N. \quad (9)$$

The cross-validatory threshold λ^C is a minimizer of

$$\hat{M}(\lambda) = \sum_{j,k} (T^{soft}(d_{jk}^{EVEN}; \lambda) - d_{jk}^{ODD})^2 + \sum_{j,k} (T^{soft}(d_{jk}^{ODD}; \lambda) - d_{jk}^{EVEN})^2, \quad (10)$$

multiplied by the correction factor $(1 - \frac{\log 2}{\log N})^{-\frac{1}{2}}$, where d_{jk}^{ODD} and d_{jk}^{EVEN} are discrete wavelet transformations of the sequences \bar{y}^{ODD} and \bar{y}^{EVEN} .

Nason (1994) showed that almost always one can find a unique minimizer of $\hat{M}(\lambda)$ and compared the performance of the cross-validatory threshold to the Donoho-Johnstone universal and SURE methods. An SPLUS function for calculating Nason's crossvalidatory threshold is given in Appendix. Our experience with crossvalidatory thresholds is that they are very close to the optimal (minimizing L^2 distance between the true signal and the estimate) threshold. However, they tend to overfit the model.

The left panel on figure (2) is a plot of $\hat{M}(\lambda)$ for *contaminated Doppler data* [1024 Doppler points + $N(0, 0.25^2)$ noise]. The minimizing λ is 0.453. Since we know the original signal, the performance of Nason's procedure could be evaluated.

The right panel in Figure (2) shows the *Doppler* signal recovered by the cross-validation threshold: $\lambda^C = 1.054 * 0.453$. Small spikes along the signal show the general phenomenon of overfitting for the cross-validatory threshold selection.

2.4 Other methods

At the expense of a slight increase of computational complexity (up to $O(n \log n)$), Donoho and Johnstone (1993) propose the *SUREShrink* method. The idea is to shrink wavelet coefficients level-wise. The SURE is used only if the level has a significant signal present. Otherwise universal thresholding is used. The proposed method has excellent smoothness adaptation properties.

Wang (1994b) generalizes Nason's crossvalidation technique by removing more than half of the data each time. The motivation is to robustify the threshold selection procedure against the effect of a correlated noise (with a long range dependence). Saito (1994) incorporates the hard thresholding into a minimum description length procedure.

3 The model

Suppose, as in Introduction, that y_i are 'noisy' measurements, i.e., the sums of an unknown signal f_i and noise ϵ_i :

$$y_i = f_i + \epsilon_i, \quad i = 1, N (= 2^n). \quad (11)$$

After applying the wavelet transformation W , the resulting sequence $d_i = W(f_i)$ is again the sum of the transformation of the signal $\theta_i = W(f_i)$ and the transformation of the noise $\eta_i = W(\epsilon)$. This is a consequence of the linearity of W . If the noise sequence ϵ_i is modeled by independent normals with mean 0 and variance σ^2 , then the orthogonality of W implies that the sequence η_i has the same distributional properties as the sequence ϵ_i .

Thus, instead of estimating the signal directly, we estimate its wavelet transformation. Equivalently, we estimate the means θ_i in $d_i \sim N(\theta_i, \sigma^2)$.

Our goal is to exhibit a Bayesian parametric model such that the resulting optimal actions in the proposed inference problem mimic desirable thresholding rules.

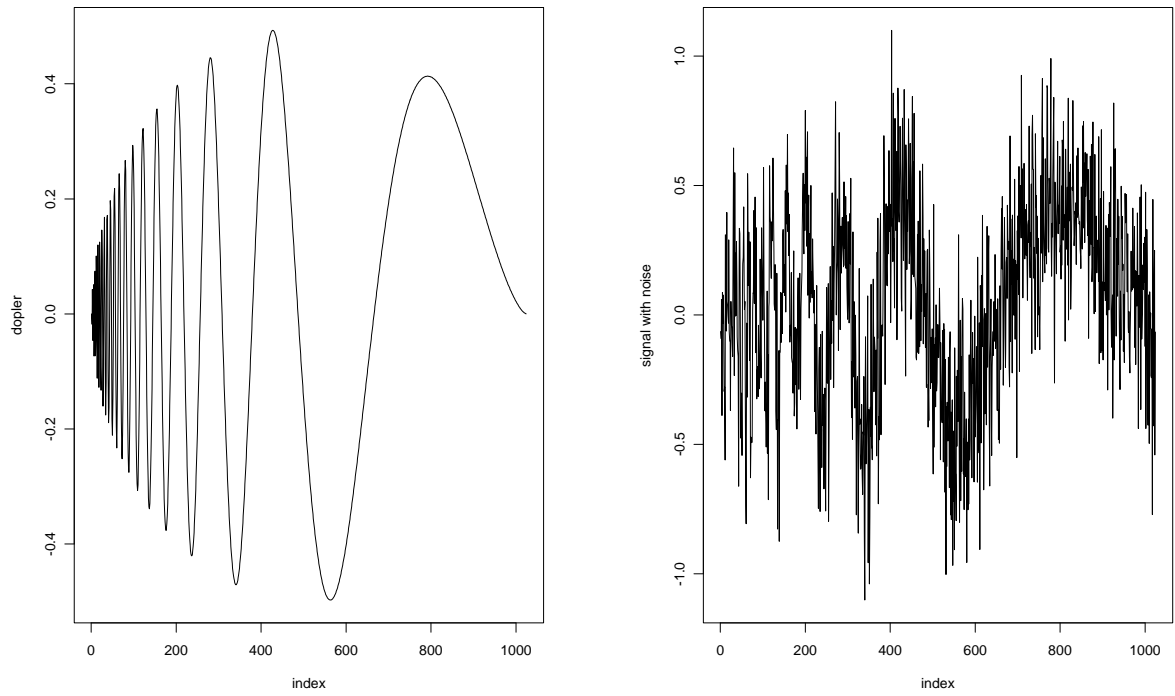


Figure 1: Doppler signal: original and with $\mathcal{N}(0, 0.25^2)$ noise

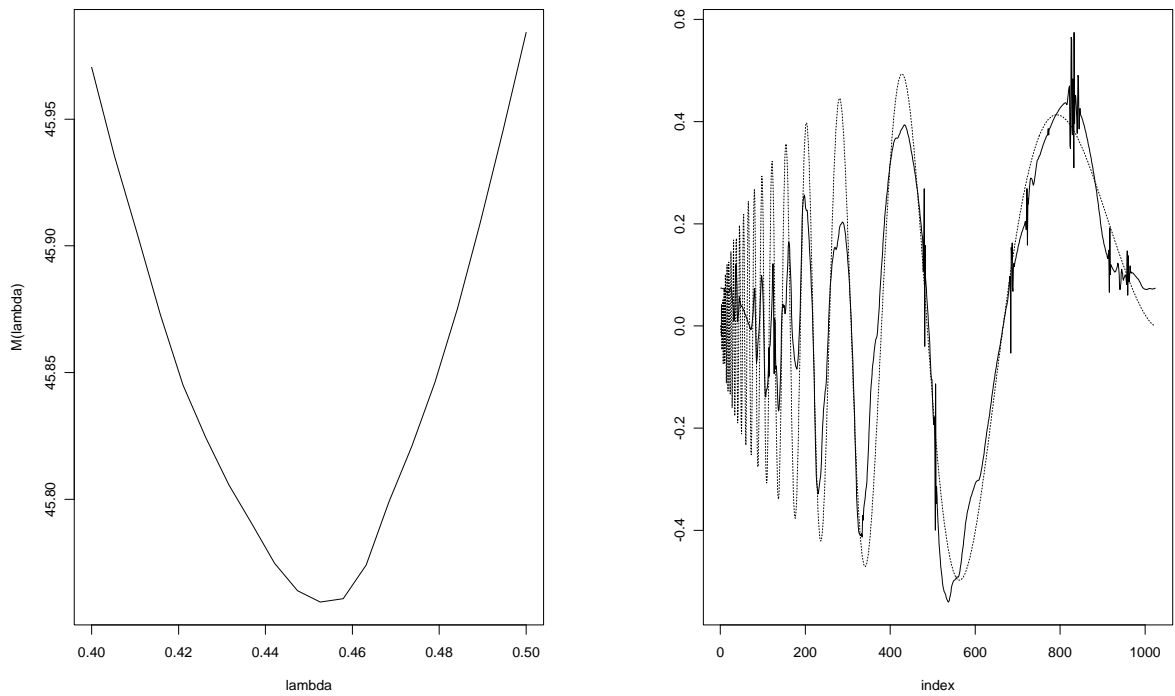


Figure 2: Function $\hat{M}(\lambda)$ and Cross-validation denoising of *Doppler*

Let

$$d|\theta, \sigma^2 \sim N(\theta, \sigma^2), \quad \sigma^2 \text{ unknown.} \quad (12)$$

The first problem is the choice of a prior on σ^2 . Because of practical (computational) reasons and informational properties the prior distribution on σ^2 is chosen to be exponential. As it is shown by Zellner (1994), the exponential distribution is the entropy maximizer among all distributions supported on $(0, \infty)$ with a fixed first moment. Thus,

$$\sigma^2 \sim \mathcal{E}(\mu). \quad (f(\sigma^2|\mu) = \mu e^{-\mu\sigma^2}). \quad (13)$$

There are several other standard ways of integrating out σ^2 ; for an account on that issue see Berger (1985) and Robert (1994).

The marginal model (marginal likelihood) is double exponential,

$$d|\theta \sim \mathcal{DE}(\theta, \frac{1}{\sqrt{2\mu}}), \quad (14)$$

with $f(d|\theta) = \frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}$. Equation (14) follows from the fact that the double exponential distribution is a scale mixture of normals (Special case of West, 1987, page 111).

For the prior on θ there are many different distributions. The choice of normal distribution is not recommended for robustness reasons (see Berger, 1985). In addition, the resulting rule will not have a desirable shape. Berger (personal communication) suggested the use of flat-tailed priors, such as the t family, and anticipated that the shape of Bayes rule will be similar to thresholding functions.

Thus, we assume

$$\theta \sim t_n(0, \tau). \quad (15)$$

In general, Bayes rules are “shrinkers,” and their shape in many cases has a desirable property for wavelet shrinkage: it shrinks small arguments heavily and large arguments only slightly. Some examples of Bayes rules, with respect to this model, are given in Figure 3. The hyperparameters μ and τ should be elicited as follows:

- μ is our subjective estimator of the precision $\frac{1}{\sigma^2}$. ($E\sigma^2 = 1/\mu$)
- τ regulates the way we shrink. Small values of τ make Bayes rules to be almost zero for small values of its argument.

3.1 Calculation of Bayes rules and examples

It is possible to get an analytic expression for the Bayes rule under the model described in this section, up to a Laplace transformation.

Theorem 3.1 *Let $f(d|\theta) \sim \mathcal{DE}(\theta, \frac{1}{\sqrt{2\mu}})$, and let $\pi(\theta)$ be a prior on θ satisfying the symmetry condition:*

$$\pi(\theta) = \pi(-\theta), \quad \theta \in R. \quad (16)$$

Then the Bayes rule with respect to the squared error loss is:

$$\delta(d) = d - \frac{\Pi_1'(c) - \Pi_2'(c)}{\Pi_1(c) + \Pi_2(c)}, \quad (17)$$

where Π_1 and Π_2 are the Laplace transforms of functions $\pi(\theta + d)$ and $\pi(\theta - d)$, $\theta \in (0, \infty)$, and $c = \sqrt{2\mu}$.

Proof: Bayes rule with respect to the squared error loss is

$$\delta(d) = \frac{\int_R \theta f(d|\theta)\pi(\theta)d\theta}{\int_R f(d|\theta)\pi(\theta)d\theta}. \quad (18)$$

The denominator is:

$$\begin{aligned}
\int_{\mathcal{R}} f(d|\theta)\pi(\theta)d\theta &= \int_{-\infty}^{\infty} e^{-c|\theta-d|}\pi(\theta)d\theta \\
&= \int_{-\infty}^0 e^{cu}\pi(u+d)du + \int_0^{\infty} e^{-cu}\pi(u+d)du \\
&= \int_0^{\infty} e^{-cu}[\pi(u+d) + \pi(u-d)]du \\
&= \Pi_1(c) + \Pi_2(c).
\end{aligned}$$

The numerator is:

$$\begin{aligned}
\int_{\mathcal{R}} \theta f(d|\theta)\pi(\theta)d\theta &= \int_0^{\infty} (d-t)e^{-ct}\pi(t-d)dt + \int_0^{\infty} (d+t)e^{-ct}\pi(t+d)dt \\
&= d(\Pi_1(c) + \Pi_2(c)) - (\Pi_1'(c) + \Pi_2'(c)) \quad \square
\end{aligned}$$

Functions Π_1 and Π_2 are finite for only some special densities from the t family. It is a problem of separate interest to find good analytic approximations or fast numerical procedures for calculating $\Pi_i^{(j)}$, $i, j + 1 = 1, 2$.

To make our rules effective, we applied the MC method on

$$\Pi(c) = \int_0^{\infty} e^{-ct}g(t)dt. \tag{19}$$

An approximation is

$$\Pi(c) \approx \frac{1}{cM} \sum_{i=1}^M g(T_i/c), \tag{20}$$

where T_i are iid random variables with the exponential $\mathcal{E}(1)$ distribution and M is large.

In applications that follow, $M = 10,000$ was used. For $N = 1024$, MN exponential were generated and the whole thresholding procedure took only a few seconds on DEC 5000/25.

On figures 4 and 5 the thresholding rule and its application on *Doppler* signal are shown.

4 Thresholding via Bayesian hypotheses testing

Bayes rules (17) in an estimation contest are never thresholding rules. One way to obtain bona fide thresholding rules in a Bayesian framework is by testing. Donoho and Johnstone gave a heuristic for the selection of the universal threshold via rejection regions of suitable hypothesis tests. Testing a precise hypothesis in Bayesian fashion requires a prior that has a point mass component. Otherwise, the testing is impossible since any continuous prior density will give the prior (and hence the posterior) probability of 0 to a precise hypothesis. For discussion on testing precise hypotheses in Bayesian fashion see Berger (1985) and Berger and Delampady (1987).

Let

$$d|\theta \sim f(d|\theta),$$

After observing d , we test the hypothesis $H_0 : \theta = 0$, versus $H_1 : \theta \neq 0$. If the hypothesis H_0 is rejected, θ is estimated by d . Let

$$\theta \sim \pi(\theta) = \pi_0\delta_0 + \pi_1\xi(\theta), \tag{21}$$

where $\pi_0 + \pi_1 = 1$, δ_0 is a point mass at 0, and $\xi(\theta)$ is a prior that describes spread of θ when H_0 is false.

More formally, the procedure is:

$$\hat{\theta} = d \mathbf{1}(P(H_0|d) < \frac{1}{2}), \tag{22}$$

where

$$P(H_0|d) = (1 + \frac{\pi_1}{\pi_0} \frac{1}{B})^{-1}, \tag{23}$$

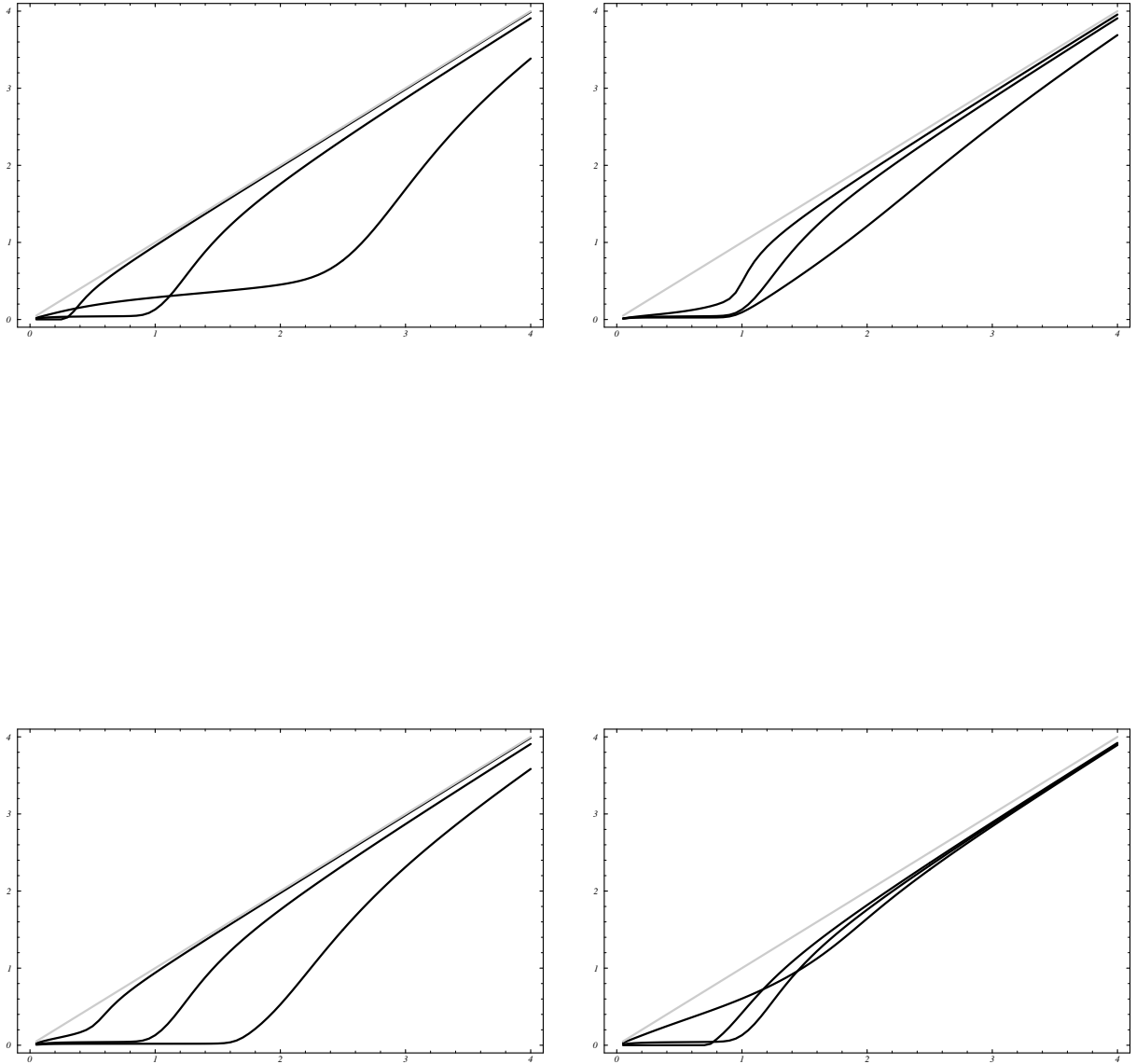


Figure 3: Left top: Bayes rule for cases: (i) $\mu = 4, \tau = 1/4$, (ii) $\mu = 16, \tau = 1/16$, and (iii) $\mu = 128, \tau = 1/128$. Right top: Bayes rule for cases: $\mu = 16, \tau = 1/16$, and (i) $df=2$, (ii) $df=5$, and (iii) $df=15$. Left bottom: Bayes rule for cases: (i) $\mu = 4, \tau = 1/16$, (ii) $\mu = 16, \tau = 1/16$, and (iii) $\mu = 128, \tau = 1/16, df=5$. Right bottom: Bayes rule for cases: (i) $\mu = 16, \tau = 1/4$, (ii) $\mu = 16, \tau = 1/16$, and (iii) $\mu = 16, \tau = 1/128, df=5$.

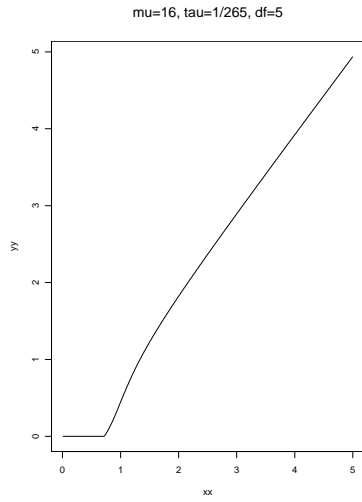


Figure 4: Bayes rule for $\mu = 16$, $\tau = 1/256$, and $df = 5$.

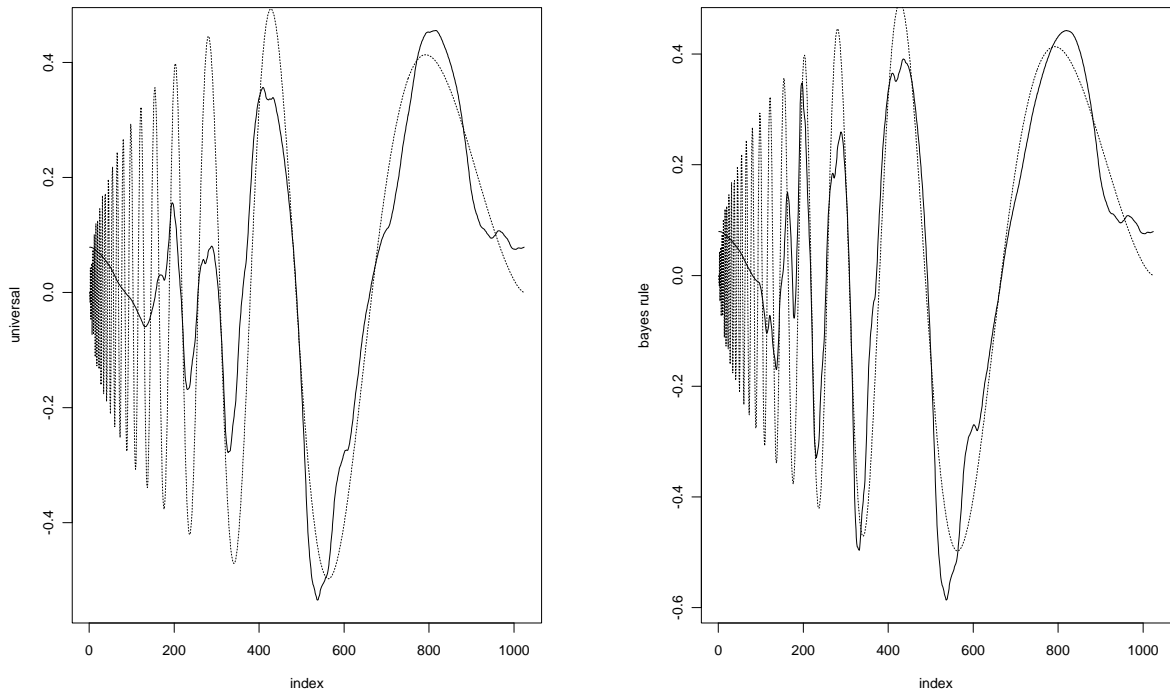


Figure 5: Doppler signal comparison of Universal and Bayes methods

is the posterior probability of H_0 hypothesis, and

$$B = \frac{f(d|0)}{\int_{\theta \neq 0} f(d|\theta)\xi(\theta)d\theta} \quad (24)$$

is the Bayes factor in favor of H_0 . That motivates the name: Bayes factor (BF) thresholding.

Theorem 4.1 *If*

$$d|\theta \sim \mathcal{DE}(\theta, \frac{1}{\sqrt{2\mu}})$$

and

$$\pi(\theta) = \pi_0\delta_0 + \pi_1\xi(\theta),$$

then d will be “thresholded” if

$$\frac{\pi_0 e^{-c|d|}}{\pi_0 e^{-c|d|} + \pi_1(\Pi_1(c) + \Pi_2(c))} \geq \frac{1}{2}, \quad (25)$$

where Π_1 and Π_2 are Laplace transformations of $\xi(\theta - d)$ and $\xi(\theta + d)$.

Proof: Straightforward.

We have chosen ξ to be from the t family. By construction, the above thresholding procedure is a hard thresholding with the choice of the threshold λ incorporated into the imposed model, or more precisely, in the parameters of the model.

As one may expect, the threshold is sensitive with respect to the choice of μ . Since μ is our subjective elicitation of precision ($1/\sigma^2$), the threshold decreases with an increase in μ . The procedure is surprisingly robust with respect to the choice of the scale parameter τ and the number of degrees of freedom of the t prior. The procedure can be made automatic by an empirical Bayes argument: Take $\mu = \frac{1}{\hat{\sigma}^2}$, for some estimate of variance, $\hat{\sigma}^2$.

The numbers π_0 and π_1 in (25) are the prior probabilities of hypotheses H_0 and H_1 . Since the parsimony of the wavelet model is expected it is reasonable to choose $\pi_0 = 1 - \pi_1$ close to 1.

Figure 6 gives results of BF thresholding on an artificial data set. The sequence `seq(-3,3, length=512)` was imputed as the 9th level in the DAUB4-wavelet decomposition of 1024 zeroes (The wavelet coefficients were also zeroes, of course.). For $\mu = 1/3, \tau = 1/16$ and $\pi_0 = 0.5, 0.7$, and 0.9 , the plots of thresholded “wavelet coefficients” are given.

It is possible to modify the above described “kill-or-keep” procedure to a “kill-or-shrink” policy. In the process of BF thresholding the procedure constructs a 0-1 vector of a size equal of the size of data, in which 0 corresponds to “kill” and 1 corresponds to “keep.” By multiplying the vector with a constant between 0 and 1, one will replace “keep” with “shrink.”

5 Conclusion

In this paper we have shown that wavelet shrinkage can be obtained through appropriate Bayesian model selection. The models considered are simple and illustrative and there is ample space for future research in applying more general Bayesian models, such as nonparametric models (through mixtures of Dirichlet Processes and applications of MCMC methods).

5.1 Acknowledgment

Professor Iain Johnstone kindly provided the test functions (*Doppler, HeaviSine, Bumps, and Blocks*), which have become a standard for testing performance of wavelet methods.

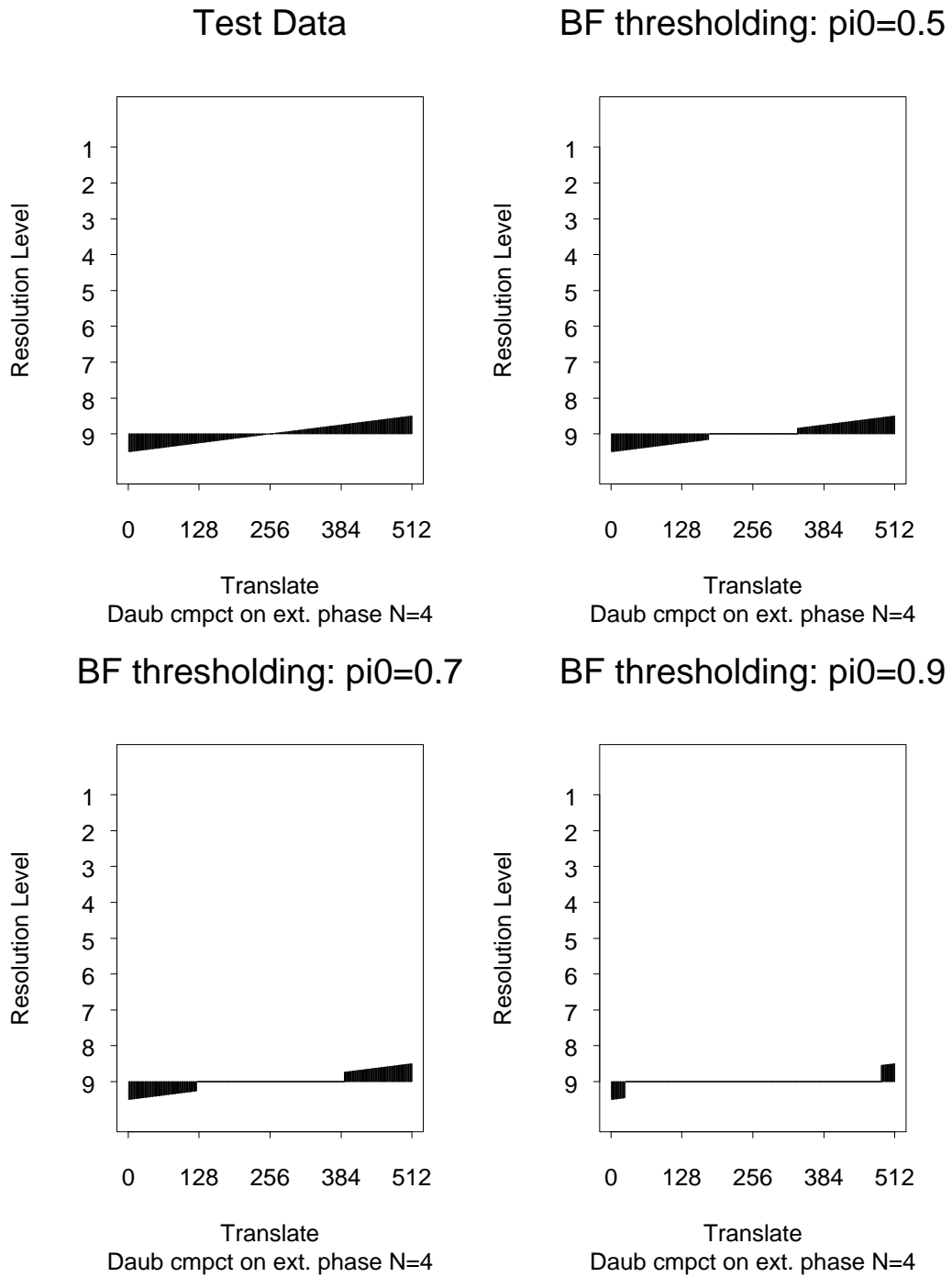


Figure 6: Bayes Factor Hard Thresholding for some values of π_0

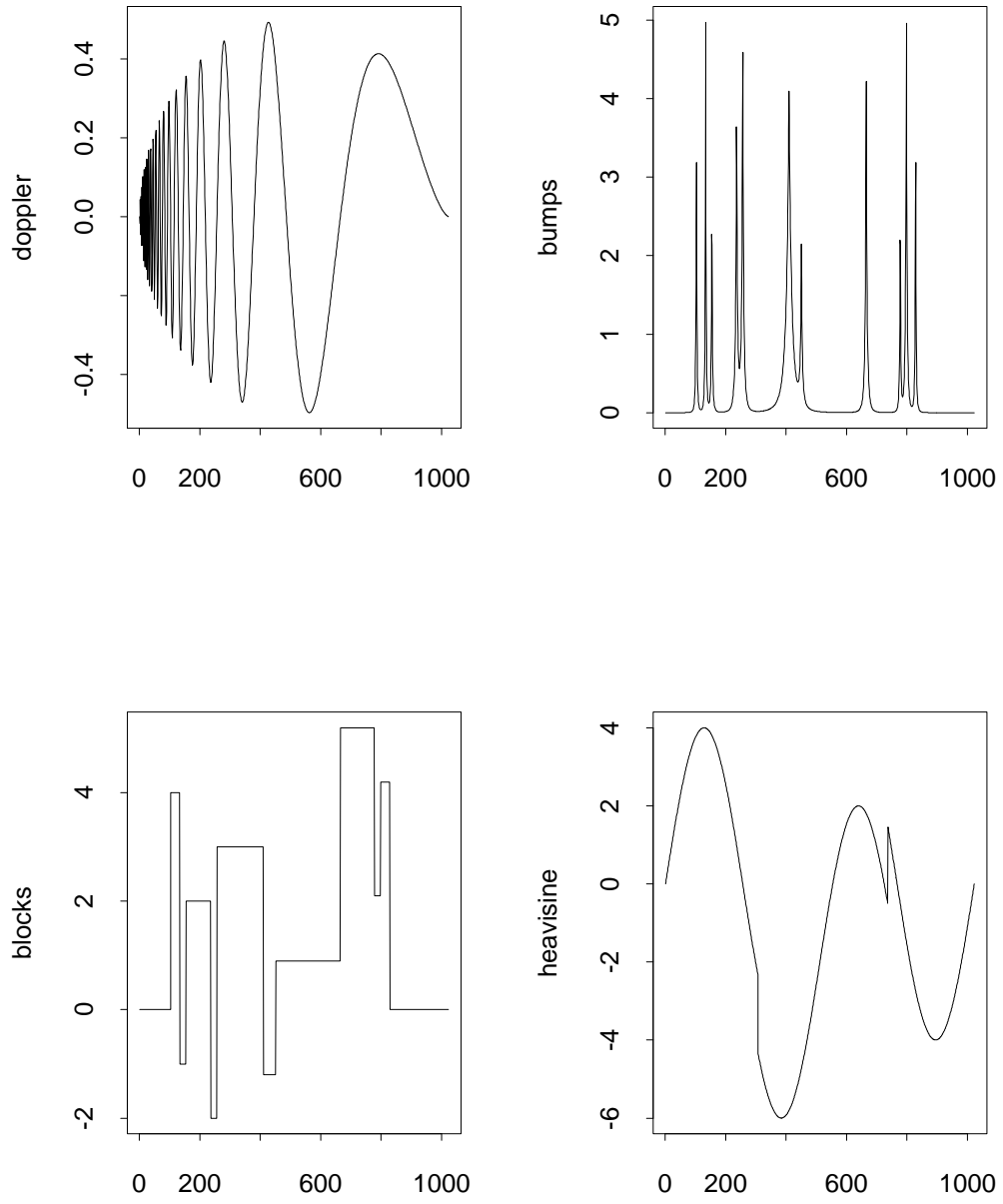


Figure 7: Four Donoho and Johnstone test Functions: *Doppler*, *Bumps*, *Blocks* and *HeaviSine*

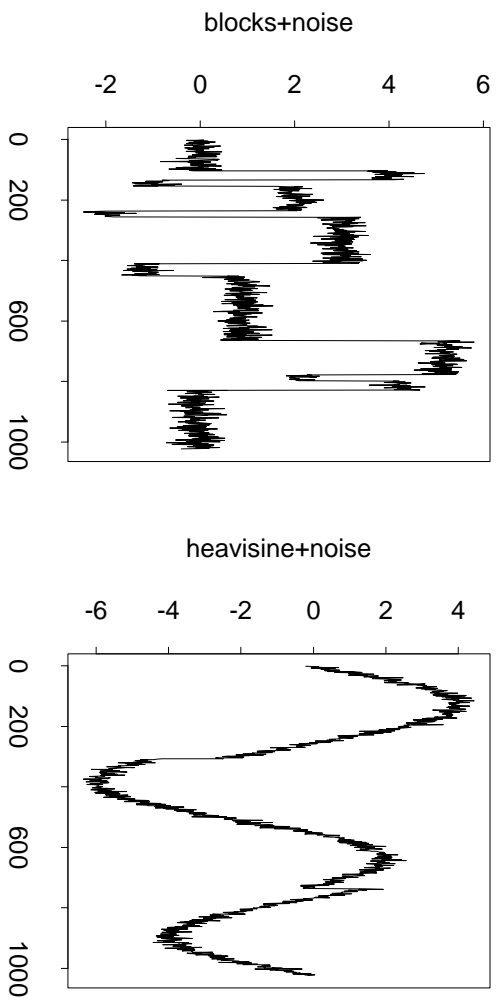
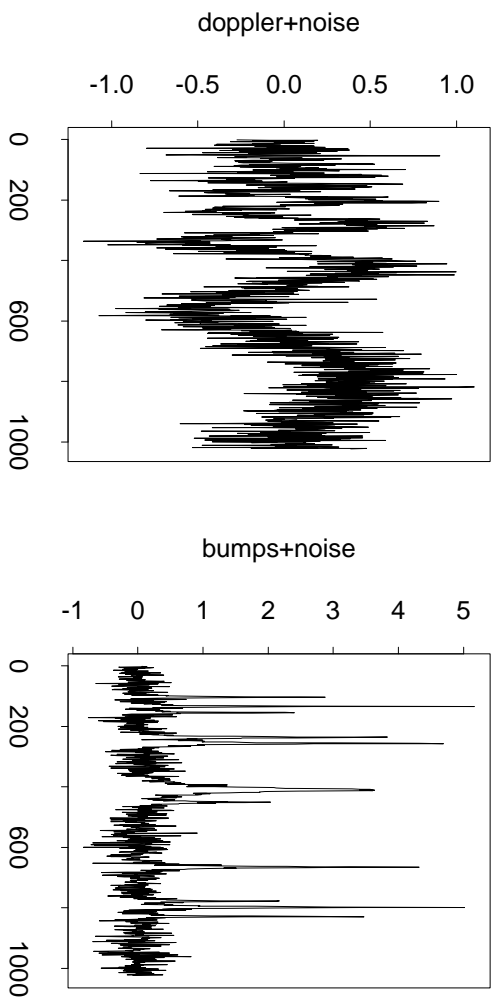


Figure 8: Signals contaminated by random $N(0, 0.25^2)$ noise

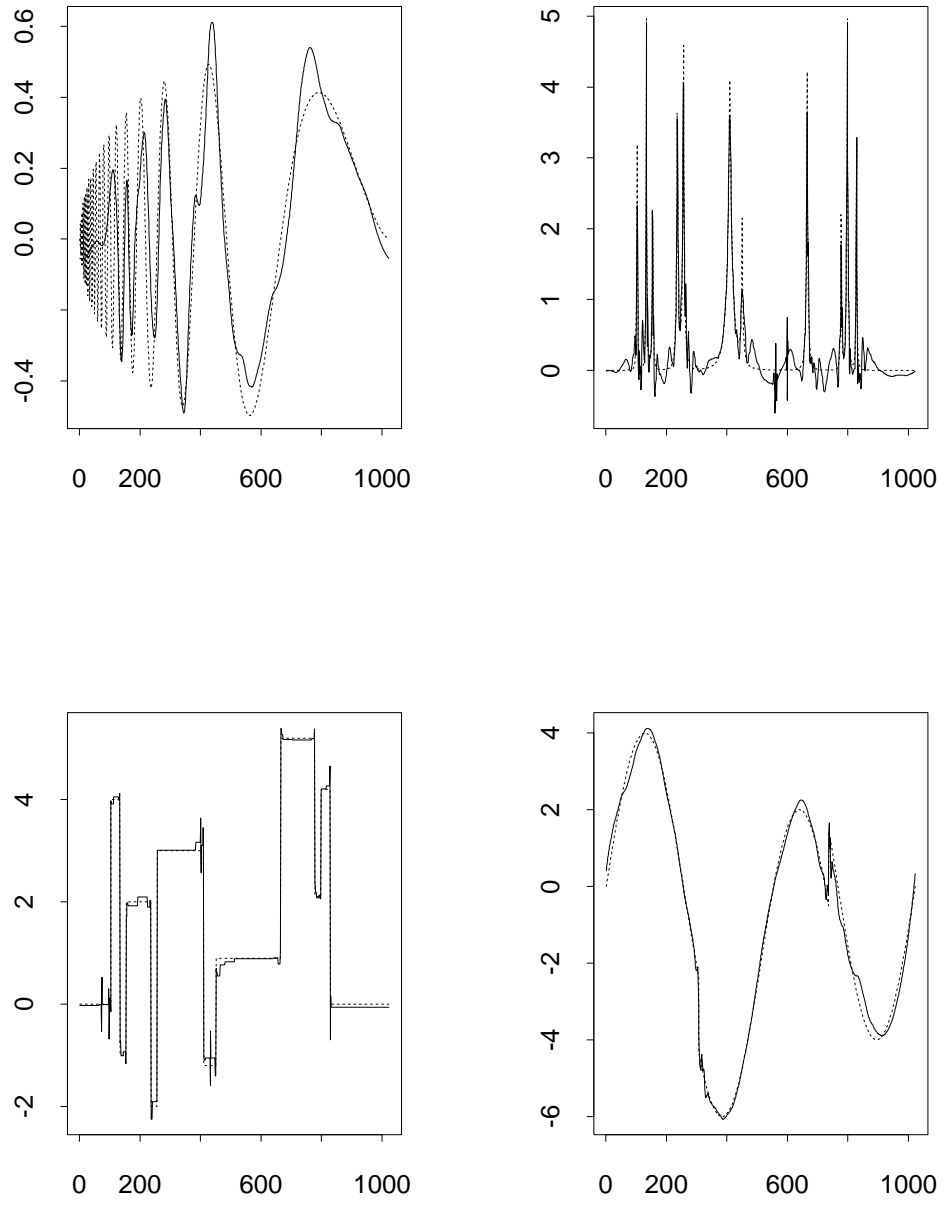


Figure 9: Bayes Factor recovery

References

- [1] BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag.
- [2] BERGER, J. and DELAMPADY, M. (1987). Testing precise hypothesis. *Statistical Science, vol 2, Num. 3, 317-352*.
- [3] BOCK, M.E. (1994) An introduction to the Daubechies D4 wavelet with statistical applications, *IMS Special Invited Paper, 230th Meeting of IMS, Cleveland, April 1994*.
- [4] BOCK, M.E. and LU, M. (1994) Wavelet transformations for non-uniform designs. *Interface '94- 26th Symposium on the Interface Computing Science and Statistics, June 15-18, 1994 RTP, NC*.
- [5] BRUNK, H. (1978). Univariate density estimation by orthogonal series. *Biometrika*, **65**, 3, 521-528.
- [6] CHUI, C. (1992). *Wavelets*. Academic Press.
- [7] DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. S.I.A.M., Philadelphia.
- [8] DONOHO, D. (1993). Nonlinear Wavelet Methods for Recovery of Signals, Densities, and Spectra from Indirect and Noisy Data. *Proceedings of Symposia in Applied Mathematics*, American Mathematical Society.
- [9] DONOHO, D. and JOHNSTONE, I. (1992a) Minimax Estimation via wavelet shrinkage. *Technical Report No 402*, Department of Statistics, Stanford University.
- [10] DONOHO, D. and JOHNSTONE, I. (1992b) Minimax Risk over l_p -Balls for l_q -error. *Technical Report*. Department of Statistics, Stanford University.
- [11] DONOHO, D. and JOHNSTONE, I. (1992c) Minimax estimation via wavelet shrinkage. *Technical Report*. Department of Statistics, Stanford University.
- [12] DONOHO, D. and JOHNSTONE, I. (1992d) New minimax theorems, thresholding and adaptation. *Technical Report*. Department of Statistics, Stanford University.
- [13] DONOHO, D. and JOHNSTONE, I. (1993) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*. to appear.
- [14] DONOHO, D. and JOHNSTONE, I. (1994). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Ass.*, to appear.
- [15] DONOHO, D., JOHNSTONE, I., KERKYACHARIAN, G., and PICARD, D. (1994). Wavelet shrinkage: Asymptopia? *J. R. Statis. Soc.* to appear.
- [16] FELLNER, W. (1974). Heuristic estimation of probability densities. *Biometrika*, **61**, 3, 485-492.
- [17] GAO, H-Y. (1993). Choice of thresholds for wavelet estimation of the log-spectrum. *Tech. Report*, Statistics, Stanford University.
- [18] GAO, H-Y. (1993). Spectral density estimation via wavelet shrinkage. *Tech. Report*, Statistics, Stanford University.
- [19] MEYER, Y. (1992) *Wavelets and Operators*. Cambridge Studies in Advanced Mathematics **37**. Cambridge University Press.
- [20] NASON, G. (1994). Wavelet regression by cross-validation. *Technical Report 447*. Department of Statistics, Stanford University.
- [21] NASON, G. and SILVERMAN, B. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*. to appear.
- [22] ROBERT, C. (1994). *The Bayesian Choice*. Springer-Verlag.
The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*. to appear.

- [23] SAITO N. (1994) Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. In: *Wavelets in Geophysics*, Foufoula-Georgiou and Kumar (eds.), Academic Press.
- [24] VIDAKOVIC, B. (1993). Random densities via wavelets. *Discussion Paper 94-06*. ISDS, Duke University. Submitted.
- [25] VIDAKOVIC, B. (1994). Wavelets and Bayesian statistics. Invited talk at: *Interface '94- 26th Symposium on the Interface Computing Science and Statistics, June 15-18, 1994 RTP, NC*.
- [26] VIDAKOVIC, B. and MÜLLER, P. (1994). Wavelets for kids, Tutorial Introduction. *Discussion Paper 94-13*. ISDS, Duke University.
- [27] WANG, Y. (1994). Jump and sharp cusp detection by wavelets - One dimensional case. *Tech. Report*. Department of Statistics, University of Missouri-Columbia.
- [28] WANG, Y. (1994). Function estimation via wavelets for data with long-range dependence. *Tech. Report*. Department of Statistics, University of Missouri-Columbia.
- [29] WANG, Z. (1993). Estimating a Hölder continuous function from a noisy sample via shrinkage and truncation of wavelet coefficients. *Tech. Report. 93-9* Department of Statistics, Purdue University.
- [30] WAHBA, G. (1981). Data based optimal smoothing or orthogonal series density estimates. *Ann. Statist.* **9**, 1, 146-156.
- [31] WATSON, G. (1969). Density estimation by orthogonal series. *Ann. Math. Statist.* **40**. 4, 1496-1498.
- [32] WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 3, 646-648.
- [33] WHITTLE, P. (1958). On the smoothing of probability density function. *J. R. Statist. Soc.* **B 20**, 334-343.
- [34] ZELLNER, A. (1994). Bayesian Method of Moments/Instrumental Variable (BMOM/IV) Analysis of Mean and Regression Models, Presented at ISBA 1994, Toronto.

6 Appendix

6.1 Some S functions for Nason's wavethresh

6.1.1 A code for calculating $\hat{M}(\lambda)$.

```
>Cross <- function(x, filter = 4, lambda = 1)
{
  N <- length(x)
  y <- c(x, x[N], x[N - 1])
  yodd <- NULL
  yeven <- NULL
  xodd <- NULL
  xeven <- NULL
  for(i in 1:(N/2)) {
    yodd <- c(yodd, (y[2 * i - 1] + y[2 * i + 1])/2)
    yeven <- c(yeven, (y[2 * i] + y[2 * i + 2])/2)
    xodd <- c(xodd, x[2 * i - 1])
    xeven <- c(xeven, x[2 * i])
  }
  dodd <- wd(xodd, filter.number = filter)
  deven <- wd(xeven, filter.number = filter)
  doddt <- threshold(dodd, policy = "manual", type = "soft", value =
    lambda)
  devent <- threshold(deven, policy = "manual", type = "soft", value =
    lambda)
  fhodd <- wr(doddt)
  fheven <- wr(devent)
  m <- sum((yodd - fheven)^2) + sum((yeven - fhodd)^2)
  return(1/sqrt(1-log(2)/log(N)) * m)
}
```

6.1.2 BF thresholding

```
> BFThr <- function(wd.structure, levels = 3:(wd.structure$nlevels - 1),
  verbose = F, pi0 = 0.5, mu = 1, tau = 1/16, df = 5, N = 10000)
{
  n <- 2^wd.structure$nlevels
  if(verbose == T)
    cat("Total n=", n, "\n")
  nthresh <- length(levels)
  if(verbose == T)
    cat("nthresh=", nthresh, "\n")
  d <- NULL
  for(i in 1:nthresh)
    d <- c(d, accessD(wd.structure, level = levels[i]))
  for(i in 1:nthresh) {
    djk <- accessD(wd.structure, level = levels[i])
    if(verbose == T)
      cat("For level i=", levels[i], "djk=(", djk, ")\n")
    djk <- Test(djk, pi0, mu, tau, df, N)
    if(verbose == T)
      cat("For level i=", levels[i], "new djk=(", djk, ")\n")
    wd.structure <- putD(wd.structure, level = levels[i], djk)
  }
  wd.structure
}

> Test <- function(x, pi0, mu, tau, df, N)
{
  c <- 1/sqrt(2 * mu)
  logic <- (pi0 * exp(- c * abs(x)))/(pi0 * exp(- c * abs(x)) + (1 -
    pi0) * (Simul(N, c, x, tau, df) + Simul(N, c, - x, tau, df))) > 0.5
  lon <- 1 - as.numeric(logic)
  x <- x * lon
  return(invisible(x))
}

> Simul <- function(N, c, d, tau, df)
{
  return(1/c * mean(hh(rexp(N, 1)/c, d, tau, df)))
}

> hh <- function(x, loc, sca, df) #prior under H_1
{
  1/sca * dt((x - loc)/sca, df)
}
```