Statistica Sinica **9**(1999), 183-197

A BAYESIAN DECISION THEORETIC APPROACH TO THE CHOICE OF THRESHOLDING PARAMETER

Fabrizio Ruggeri and Brani Vidakovic

CNR-IAMI and Duke University

Abstract: Thresholding rules recently became of considerable interest when Donoho and Johnstone applied them in the wavelet shrinkage context. Analytically simple, such rules are very efficient in data denoising and data compression problems. In this paper we find hard thresholding decision rules that minimize Bayes risk for broad classes of underlying models. Standard Donoho-Johnstone test signals are used to evaluate performance of such rules. We show that an optimal Bayesian decision theoretic (BDT) hard thresholding rule can achieve smaller mean squared error than some standard wavelet thresholding methods, if the prior information on the noise level is precise.

Key words and phrases: Bayes rule, hard-thresholding, wavelets.

1. Introduction

Many researchers and practitioners, notably in signal and image processing, used thresholding rules in time or Fourier domains before wavelets were developed. Examples of such rules are hard and soft thresholding functions given, respectively, by $\delta^{hard}(x,\lambda) = x\ \mathbf{1}(|x| > \lambda)$ and $\delta^{soft}(x,\lambda) = (x - \text{sign}(x)\lambda)\ \mathbf{1}(|x| > \lambda)$, where λ is the threshold and $\mathbf{1}(\cdot)$ is the indicator function. In the context of nonparametric regression, Donoho and Johnstone (references in Donoho (1997)) showed that such simple rules produce regression estimates which possess asymptotic optimality properties for a variety of function spaces. An extensive simulation study of MSE properties of simple thresholding rules was performed by Bruce and Gao (1996). Hard and soft thresholding rules can be regarded as Bayesian solutions with improper and proper priors, respectively, as discussed by Fan (1997).

In this paper we find an "optimal" threshold λ by minimizing the (integrated) Bayes risk over the class of hard threshold rules δ^{hard}. Instead of finding unrestricted Bayes shrinkage rules which never threshold under the squared error loss, we find the rules that are Bayes in the class of δ^{hard}-type rules. In this respect, our approach is "restricted Bayes." When the risk-minimizing λ^* exists, the rule $\delta^{hard}(x,\lambda^*)$ will be called *Bayesian decision theoretic (BDT) rule*, and the induced shrinkage estimator, the *BDT estimator.* For unrestricted Bayesian approaches see, e.g., Vidakovic (1998), Chipman, McCulloch and Kolaczyk (1997) and Clyde, Parmigiani and Vidakovic (1998).

We are interested not only in finding the optimal rules with good denoising properties, but also in identifying the model-prior pairs which produce nontrivial risk-minimizing values for λ. Both problems are important because we want to provide both a sound theoretical background, based on decision theory, and practical guidelines on which model-based shrinkage should be performed. The discussion on model-prior pairs aims to help users choose models and priors for wavelet thresholding. Furthermore, our approach allows for the use of prior information on the unknown signal of interest and the level of noise.

Let $\underset{\sim}{y}$ be a data vector of dimension (size) N. For simplicity we choose N to be a power of 2, say 2^n, though the generalization to different sample sizes is straightforward.

Suppose that the vector $\underset{\sim}{y}$ is wavelet-transformed to a vector $\underset{\sim}{x}, \underset{\sim}{x} = W\underset{\sim}{y}$. The transformation is linear and orthogonal and can be described by an orthogonal matrix W of dimension $N \times N$. In practice, one performs the discrete wavelet transformation without finding the matrix W explicitly but by using fast filtering algorithms. These algorithms are based on so-called quadrature mirror filters that uniquely correspond to the wavelet of choice.

Basics on wavelets can be found in many texts, monographs and papers at many different levels of exposition, e.g. Daubechies (1992), Walter (1994), Hernandez and Weiss (1996).

One of the strengths of wavelet transformations in statistics is that they "unbalance" the data, meaning that most of the ℓ_2-norm of the data (which is preserved in the transformation) concentrates in only a few wavelet coefficients. Wavelets also form unconditional bases for a range of function spaces. Consequently the magnitudes of wavelet coefficients fully describe some properties of the decomposed function, such as smoothness.

Suppose the observed data $\underset{\sim}{y}$ is the sum of an unknown signal $\underset{\sim}{s}$ and random noise $\underset{\sim}{\epsilon}$, with

$$y_i = s_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{1}$$

In the wavelet domain (after applying a linear and orthogonal wavelet transformation W), expression (1) becomes $x_i = \theta_i + \eta_i,\ i = 1, \ldots, n$, where x_i, θ_i, and η_i are the ith coordinates of $W\underset{\sim}{y}, W\underset{\sim}{s}$ and $W\underset{\sim}{\epsilon}$, respectively. For notational simplicity, the double indexing typical in discrete wavelet representations is replaced by a single index.

Assuming a location model $[x_i|\theta_i] \sim f(x_i - \theta_i)$ and a prior $[\theta_i] \sim \pi(\theta_i)$ we find the Bayes risk of the decision rule δ^{hard}.

Donoho and Johnstone proposed *wavelet shrinkage,* a class of simple and efficient procedures for nonparametric estimation of functions and densities based on thresholding of wavelet coefficients. An overview is given by Donoho and Johnstone (1994, 1995).

Wavelet shrinkage can be described as a three-step procedure:

1. Data (a noisy signal, measurements, blurred image pixels, etc.) are transformed by a discrete wavelet transformation from the "time domain" to the "wavelet domain". $(\underset{\sim}{y} \rightarrow \underset{\sim}{x})$

2. The transformed data are shrunk. $(\underset{\sim}{x} \rightarrow \underset{\sim}{\hat{\theta}})$

3. The processed data are transformed back to the "time domain". $(\underset{\sim}{\hat{\theta}} \rightarrow \underset{\sim}{\hat{s}})$

The choice of a shrinkage rule in Step **2** is important. Many different thresholding methods have recently been proposed. An excellent overview of shrinkage rules based on thresholding is given by Nason (1995).

Here we narrow our attention to hard thresholding rules δ^{hard}. Note that δ^{hard} is fully specified by the threshold λ.

We compare our method with the nonadaptive shrinkage methods of Donoho and Johnstone in (1994), (1995) (hard thresholding with the universal threshold, UNIV) and that of Bruce and Gao (1996) (hard thresholding with an optimal minimax threshold, OPT). We also provide a comparison with the adaptive Bayesian wavelet shrinkage (ABWS) method of Chipman, McCulloch and Kolaczyk (1997).

The paper is organized as follows. In Section 2, we give the decision theoretic background necessary for optimality considerations, and identify models that have nontrivial minimizers of the Bayes risk for a general class of loss functions. In Section 3, we discuss the selection of models and priors and elaborate in detail on the normal-double exponential setup. Section 4 contains some simulation results where we use the standard Donoho-Johnstone test functions for mean squared error comparisons of our method with the aforementioned methods.

2. Results

Given the real space $\mathcal{R}$ and a subset Θ, let X be a random variable on a dominated statistical space denoted by $(\mathcal{R}, \mathcal{B}_X, \{P_\theta, \theta \in (\Theta, \mathcal{B}_\Theta)\})$, with density $f(x|\theta)$. Let Π denote a probability measure on the parameter space $(\Theta, \mathcal{B}_\Theta)$, with density $\pi(\theta)$.

Let $\delta(x)$, $\mathcal{A}$ and $L(\theta, a)$ denote, respectively, a decision rule, the action space and a loss function, with $x \in X$, $\theta \in \Theta$ and $a \in \mathcal{A}$. By $R(\theta,\delta) = E^{X|\theta}(L(\theta, \delta(X)))$ we denote the frequentist risk, while the Bayes risk is given by $r(\pi,\delta) = \int R(\theta,\delta)\pi(\theta)d\theta = E^\pi(R(\theta,\delta))$. Here we consider the class $\mathcal{D}$ of decision rules, $\mathcal{D} = \{\delta_\lambda(x) = x \cdot \mathbf{1}(|x| > \lambda), \lambda \in [0,\infty)\}$, which corresponds, in the wavelet world, to the class of hard thresholding rules.

2.1. General results

Suppose that $f(x|\theta) = f(x-\theta)$ for any $(x, \theta) \in \mathcal{B}_X \otimes \mathcal{B}_\Theta$, and that both f and π are symmetric functions. Consider a symmetric loss function $L(\theta, a) = L(\theta - a)$ and the decision rule $\delta_\lambda(x)$ from the class $\mathcal{D}$. It follows that $L(\theta, \delta_\lambda) = L(x-\theta)$ (for $|x| > \lambda$) or $L(\theta)$ (for $|x| \leq \lambda$). It can be readily shown that the risk function and the Bayes risk, with respect to the prior π, of the decision rule δ_λ are given, respectively, by $R(\theta, \delta_\lambda) = C + \int_{-\lambda}^{\lambda}[L(\theta) - L(\theta - x)] f(x-\theta) dx$ and $r(\pi, \delta_\lambda) = C + \int_{\mathcal{R}}[\int_{\theta-\lambda}^{\theta+\lambda}[L(\theta) - L(t)] f(t) dt] \pi(\theta) d\theta$, Here $C = \int_{\mathcal{R}} L(t) f(t) dt$ is the risk function (and the Bayes risk) when considering the decision rule $\delta(x) = x$ for any real x. We are looking for λ^* which improves upon the Bayes risk, $r(\pi, \delta_0) = C$.

Searching for the optimal λ, we first consider the derivative of $r(\pi, \delta_\lambda)$ to find λ^*, i.e. the rule δ_{λ^*} which minimizes the Bayes risk. Suppose the usual conditions for the differentiability under the integral sign are fulfilled to get following lemma.

Lemma 2.1.

$$\frac{\partial r(\pi, \delta_\lambda)}{\partial \lambda} = r'(\lambda) = 2\int_{\mathcal{R}}[L(\theta) - L(\theta-\lambda)] f(\theta-\lambda)\pi(\theta) d\theta.$$

Proof. By differentiating $r(\pi, \delta_\lambda)$ with respect to λ we get

$$r'(\lambda) = \int_{\mathcal{R}}[L(\theta) - L(\theta+\lambda)] f(\theta+\lambda)\pi(\theta) d\theta + \int_{\mathcal{R}}[L(\theta) - L(\theta-\lambda)] f(\theta-\lambda)\pi(\theta) d\theta.$$

The result then follows because the symmetry of f, π and L implies that $\int_{\mathcal{R}} L(\theta) f(\theta+\lambda)\pi(\theta) d\theta = \int_{\mathcal{R}} L(\theta) f(\theta-\lambda)\pi(\theta) d\theta$ and $\int_{\mathcal{R}} L(\theta+\lambda) f(\theta+\lambda)\pi(\theta) d\theta = \int_{\mathcal{R}} L(\theta-\lambda) f(\theta-\lambda)\pi(\theta) d\theta$.

It should be noted that $r'(0) = 0$, while $\lim_{\lambda\to\infty} r'(\lambda) = 0$ under very mild conditions.

A useful tool in finding models and priors which result in nontrivial λ^* (i.e., away from 0 and ∞) is given by the following

Remark 2.1. If λ^* maximizes Bayes risk when f is the model for X, and π is the prior on θ, then λ^* minimizes Bayes risk when π is the model and f is the prior. Indeed, by the transformation $\theta - \lambda = -t$ and the symmetry of f and π, it follows that $\int_{\mathcal{R}}[L(\theta) - L(\theta-\lambda)]\pi(\theta-\lambda) f(\theta) d\theta = -\int_{\mathcal{R}}[L(\theta) - L(\theta-\lambda)] f(\theta-\lambda)\pi(\theta) d\theta$.

Example 2.1. We will say that the random variable X has a double-exponential $\mathcal{DE}(\theta, \beta)$ distribution if the density of X has the form: $f(x) = \beta \exp\{-\beta|x - \theta|\}/2,\ \beta > 0$.

For $X \sim \mathcal{DE}(\theta, 1)$ and $\theta \sim \mathcal{N}(0, 1)$, it can be shown (Section 3.1) that $\lambda^* = .94055$ maximizes the Bayes risk under a squared loss function, while it minimizes the Bayes risk when $X \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{DE}(0, 1)$.

2.2. Case $f = \pi$

The next theorem gives important practical advice on which models and priors are not to be chosen, if interested in nontrivial λ^*.

Theorem 2.1. *Given $f(x|\theta) = f(x-\theta)$, let $f(t) = \pi(t)$ for all $t \in \mathcal{R}$. It follows that the Bayes risk is constant, i.e. $r(\pi, \delta_\lambda) = C$, where $C = \int_{\mathcal{R}} L(t)f(t)dt$.*

Proof. Consider $r(\pi, \delta_\lambda) - C = \int_{\mathcal{R}} L(\theta)[\int_{\theta-\lambda}^{\theta+\lambda} f(t)dt]f(\theta)d\theta - \int_{\mathcal{R}}[\int_{\theta-\lambda}^{\theta+\lambda} L(t)f(t)dt]$ $f(\theta)d\theta$. The two terms in the right-hand side of the equation coincide; this follows by applying Fubini's Theorem to the second term.

Under the above conditions, any number in $[0, \infty]$ can be chosen as λ^*. Therefore, the choice of the same model and prior is to be avoided in the search of a threshold λ^*, as shown by the following example involving the normal model and prior.

Example 2.2. (Normal) Assume that $X \sim \mathcal{N}(\theta, \sigma^2)$, $\theta \sim \mathcal{N}(0, \sigma^2)$, with σ^2 known. It can be shown that $r(\pi, \delta_\lambda) = \sigma^2$ for $L(t) = t^2$. Similarly, $r(\pi, \delta_\lambda)$ equals $\sqrt{2/\pi}\sigma$ for $L(t) = |t|$ and $2\Phi(-\mu)$ for $L(t) = \mathbf{1}(|t| > \mu)$.

2.3. Case $f \neq \pi$

Consider the case when the model $f(x-\theta)$ and the prior $\pi(\theta)$ have the same functional form, apart from a scale parameter. We will see that λ^* cannot be a positive, finite number when the ratio of model and prior is decreasing.

Many standard distributions satisfy the conditions of the next Theorem, whose proof is in the Appendix.

Theorem 2.2. *Consider a density $f(x|\theta) = f(x-\theta)$, the prior $\pi(\theta)$, and a symmetric loss function L such that $L(t)$ is nondecreasing for nonnegative t. If $f(t)/\pi(t)$ is decreasing, then $\lambda^* = 0$; if it is increasing, then $\lambda^* = \infty$.*

If the noise $\underset{\sim}{\epsilon}$ in $\underset{\sim}{y} = \underset{\sim}{s} + \underset{\sim}{\epsilon}$ is i.i.d. normal, then orthogonality and linearity of the wavelet transformation ensure that the noise $\underset{\sim}{\eta}$ in $\underset{\sim}{x} = \underset{\sim}{\theta} + \underset{\sim}{\eta}$ is i.i.d. normal as well. When σ^2 in $x_i \sim N(\theta_i, \sigma^2)$ is known, one may proceed by specifying a prior on θ_i. However, if σ^2 is not known, it should be integrated out by eliciting an appropriate prior $\pi(\sigma^2)$, before proceeding with the marginal likelihood $f(x_i|\theta_i) = \int \phi_\sigma(x_i - \theta_i)\pi(\sigma^2)d\sigma^2$. The function $\phi_\sigma(x_i - \theta_i)$ is the density function of the normal $N(\theta_i, \sigma^2)$ law and $\pi(\sigma^2)$ is a prior on σ^2.

For instance, if $\pi(\sigma^2)$ is an exponential distribution, the resulting marginal likelihood $f(x_i - \theta_i)$ is a double exponential; whereas if $\pi(\sigma^2)$ is an inverse gamma, then $f(x_i - \theta_i)$ is distributed as t.

We consider only three types of models on x_i: the normal, double exponential, and t. Since the three distributions are symmetric and exhibit different tail behavior, we use them as candidates for the prior on θ_i as well.

For any possible combination of the above distributions, we check whether the ratio of model and prior is strictly monotone, looking for conditions under which λ^* equals either 0 or ∞. In absence of monotonicity, the proposed decision theoretic approach to thresholding could be applicable, since it could lead to $0 < \lambda^* < \infty$. Notice that the number of pairs of interest can be reduced because of Remark 2.1 and Theorem 2.2. For example, if the pair $(f(x-\theta), \pi(\theta))$ gives a trivial λ^*, then the pair $(\pi(x-\theta), f(\theta))$ also gives a trivial λ^*.

The next example shows that, once the model and the prior have the same functional form, $\lambda^* = 0$ or ∞, depending upon whether the variance of the model is smaller or larger than the variance of the prior. Similar result holds for double exponential and t distributions.

Example 2.3. (Normal - Normal) Let $f(x-\theta) \propto \exp\{-(x-\theta)^2/(2\sigma_f^2)\}$ and $\pi(\theta) \propto \exp\{-\theta^2/(2\sigma_\pi^2)\}$. For any $0 < x_1 < x_2$, it follows that $\Delta(x_1, x_2) \propto \exp\{-x_1^2/(2\sigma_f^2) - x_2^2/(2\sigma_\pi^2)\} - \exp\{-x_2^2/(2\sigma_f^2) - x_1^2/(2\sigma_\pi^2)\}$, which is positive if and only if $\sigma_f^2 < \sigma_\pi^2$. Therefore, applying Theorems 2.1 and 2.2, it follows that λ^* equals 0 for $\sigma_f^2 < \sigma_\pi^2$, ∞ for $\sigma_f^2 > \sigma_\pi^2$ or it can take any value for $\sigma_f^2 = \sigma_\pi^2$.

The choice of different models and priors could give non trivial λ^*, as in the following case (the pairs t/double exponential and t/normal can be treated similarly).

Example 2.4. (Normal - Double Exponential) Let $f(x-\theta) \propto \exp\{-(x-\theta)^2/(2\sigma_f^2)\}$ and $\pi(\theta) \propto \exp\{-\beta|\theta|\}$. For any $0 < x_1 < x_2$, it follows that $\Delta(x_1, x_2) \propto \exp\{-x_1^2/(2\sigma_f^2) - \beta x_2\} - \exp\{-x_2^2/(2\sigma_f^2) - \beta x_1\}$, which is positive if and only if $x_1 + x_2 > 2\sigma^2\beta$. Therefore, the ratio $f(t)/\pi(t)$ is not strictly monotone and it is possible that $0 < \lambda^* < \infty$, as shown later in Section 3.1.

It is worth mentioning that the conditions in Theorem 2.2 do not depend on the choice of the loss function $L(t)$, provided that it is nondecreasing in $|t|$ and symmetric.

We might consider other priors, for example uniform distributions and two-point masses, both symmetric around 0. Our numerical experience leads us to conjecture that λ^* equals either 0 or ∞. Improper priors, even though strongly questioned by many Bayesians, are often considered in literature. A typical choice of an improper prior about the location parameter θ, defined all over $\mathcal{R}$, is given by $\pi(\theta) = 1$ (or any other constant). It can be shown that such a prior leads to $\lambda^* = \infty$.

In Subsection 3.1 we thoroughly discuss a situation (Normal model, Double exponential prior) where $0 < \lambda^* < \infty$, which is suitable for wavelet thresholding. Based on that particular model we perform the simulations discussed in Section 4.

3. Selection of Models and Priors

In this section we provide practical guidance on how to choose models and priors.

As discussed in Section 2, we assume a normal model with a location parameter θ and the nuisance parameter σ^2. We will use our prior knowledge to specify a distribution on σ^2, necessary for obtaining the marginal likelihood on θ.

If σ^2 is known or estimable, then the prior is a point mass and the marginal likelihood becomes normal. An example with the normal likelihood will be discussed in detail in Subsection 3.1.

When information on σ^2 is vague, we consider two possible cases:

(i) If we want to be noninformative about σ^2, it is reasonable to specify an exponential prior. This is justified since an exponential distribution minimizes the Fisher information in the class of all distributions with a fixed first moment supported on $[0, \infty)$. The corresponding marginal likelihood is a double exponential, being an exponential scale mixture of normals.

(ii) When we have more information about σ^2, then it is reasonable and mathematically convenient to specify an inverse gamma prior. The inverse gamma priors can model a variety of prior beliefs on σ^2, such as information about the most probable values, some moments, etc. The corresponding marginal likelihood is t.

The choice of the prior on θ is another key issue. It is natural to assume that the prior is symmetric and unimodal about zero since the detail coefficients should not contain any systematic aberration. Also, our analysis suggests that the functional forms of the prior and the model should be different. We consider only three priors: normal, double exponential and t. Apart from their mathematical convenience, these three priors describe a wide variety of tail behaviors.

For example, if the marginal likelihood is a double exponential, we suggest using either normal or t priors depending on the prior information available about the signal itself. Smooth signals tend to give fine-scale wavelet coefficients that can be modeled by light-tailed distributions, like the normal one. On the other hand, if it is believed that the signal has discontinuities, some of the fine-scale wavelet coefficients will be large. In such cases, modeling by heavy tailed distributions, such as the t, is suggested.

Finally, one has to specify the scale parameter of the prior distribution. Scale parameters that contribute to the small variability of θ reflect our prior willingness to substantially shrink the corresponding wavelet coefficient.

3.1. Normal model and double exponential prior

In this subsection we give in detail one illustrative example. The optimal threshold is derived as a function of the hyperparameter of the prior, and comparisons are made with the universal threshold.

Consider the normal model $X \sim \mathcal{N}(\theta, \sigma^2)$, in which σ^2 is assumed known (or estimable), and the double exponential prior $\theta \sim \mathcal{DE}(0, \beta)$. Let the loss be the squared error, $L(t) = t^2$. Under such assumptions, it follows that

$$\begin{aligned}
r^{'}(\lambda)/\lambda \propto & \int_{\mathcal{R}} (2\theta - \lambda) \exp\{-(\theta - \lambda)^2/(2\sigma^2)\} \exp\{-\beta|\theta|\} d\theta \\
\propto & \int_{\mathcal{R}} (2\theta\sigma + \lambda) \exp\{-\theta^2/2 - \beta|\sigma\theta + \lambda|\} d\theta \\
\propto & \exp\{\beta\lambda + (\beta\sigma)^2/2\} \int_{-\infty}^{-\lambda/\sigma} (2\theta\sigma + \lambda) \exp\{-(\theta - \beta\sigma)^2\} d\theta \\
& + \exp\{-(\beta\lambda + (\beta\sigma)^2/2\} \int_{-\lambda/\sigma}^{\infty} (2\theta\sigma + \lambda) \exp\{-(\theta + \beta\sigma)^2\} d\theta \\
\propto & \exp\{\beta\lambda + (\beta\sigma)^2/2\}\{-2\sigma\phi(-\lambda/\sigma - \beta\sigma) + (2\beta\sigma^2 + \lambda)\Phi(-\lambda/\sigma - \beta\sigma)\} \\
& + \exp\{-\beta\lambda + (\beta\sigma)^2/2\}\{2\sigma\phi(-\lambda/\sigma + \beta\sigma) \\
& + (-2\beta\sigma^2 + \lambda)[1 - \Phi(-\lambda/\sigma + \beta\sigma)]\}.
\end{aligned}$$

Thus, λ^* can be found by solving $r^{'}(\lambda) = 0$ and checking that it is a minimum. Figure 1 (left) depicts $r'(\lambda)$ for $\sigma = \beta = 1$. The optimal λ^* is 0.94055. For this specific case, the Bayesian solution is the soft thresholding rule with $\lambda = 1$ (see e.g. Fan (1997)).

A useful approximation to λ^* for large β is given by $\hat{\lambda} = 2\beta\sigma^2$. This follows by observing that $r^{'}(\lambda)/\lambda$ is a continuous function and

$$0 < r^{'}(\hat{\lambda}) < 4\beta^2\sigma^2 \exp\{-2\beta^2\sigma^2\}/3\pi. \tag{2}$$

The inequality (2) follows from computing $r^{'}(\hat{\lambda}) = \hat{\lambda}\beta^2\sigma/\pi \exp\{5(\beta\sigma)^2/2\} \int_{-\infty}^{-3\beta\sigma} \exp\{-t^2/2\} dt$, observing that $r^{'}(\hat{\lambda})$ is positive, and from the fact that the integral can be bounded from above by $\int_{-\infty}^{-3\beta\sigma} \exp\{3\beta\sigma t/2\} dt$. The approximation result relies on the continuity of $r^{'}(\lambda)$, and the fact that the right-hand side in (2) is approximately equal to zero for large values of $\beta\sigma$.

Numerical computations of λ^* for fixed $\sigma = 1$ are summarized in Figure 1 (right). Non-monotonicity of $f(t)/\pi(t)$ in Theorem 2.2 does not ensure existence of non-trivial λ^*; note that $\lambda^* = 0$ for $\beta < 0.9$ (approx.). Eliciting extreme (small or large) values of β may cause unsatisfactory MSE performance of BDT rules. This is evident from the dependence of λ^* from β, as illustrated in Figure 1 (right). No thresholding is performed for small β and severe oversmoothing occurs for large β.

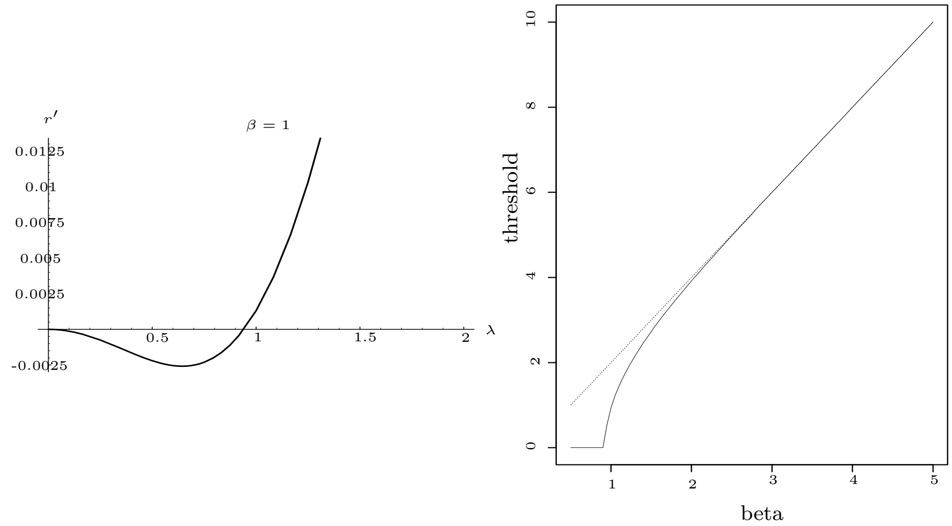

Figure 1. $\mathcal{N}(\theta, 1)$ model and $\mathcal{DE}(0, \beta)$ prior. (left) The derivative of Bayes risk r_λ for $\beta = 1$; (right) Threshold λ^* (continuous line) and approximation $\hat{\lambda}$ (dotted line) as functions of β.

Figure 1 (left) gives the plot of r' for $\beta = 1$ and $\sigma = 1$. Similar results can be obtained by keeping β fixed and allowing σ to vary. We noticed that λ^* is sensitive to the choice of the prior and that the user must be careful in choosing the parameter values. Sensitivity to a prior is a well studied problem in Bayesian analysis. Discussion on this issue exceeds the scope of this paper. See Berger, Betrò, Moreno, Pericchi, Ruggeri, Salinetti, and Wasserman (1996) for a thorough discussion.

Remark 3.1. In the normal-double exponential case we can give a single heuristic rule for eliciting the parameter β based on inspecting the minimum of the AMSE graphs (as in Figure 2). For a variety of signals and images we have found that, in the case of unknown σ^2, the choice of $\beta = \sqrt{0.4 \log N}/\hat{\sigma}$, where N is a sample size and $\hat{\sigma}$ is an estimator of the standard deviation of the noise, gives a good thresholding rule whenever the signal-to-noise ratio (SNR), defined as the ratio of standard deviations of the signal and the noise, is moderate. Notice that the corresponding threshold $\lambda \approx \sqrt{1.6 \log N}\ \hat{\sigma}$ is smaller than the standard Donoho and Johnstone threshold λ^U. This is in agreement with findings of Bruce and Gao (1996) in the context of optimal minimax thresholding.

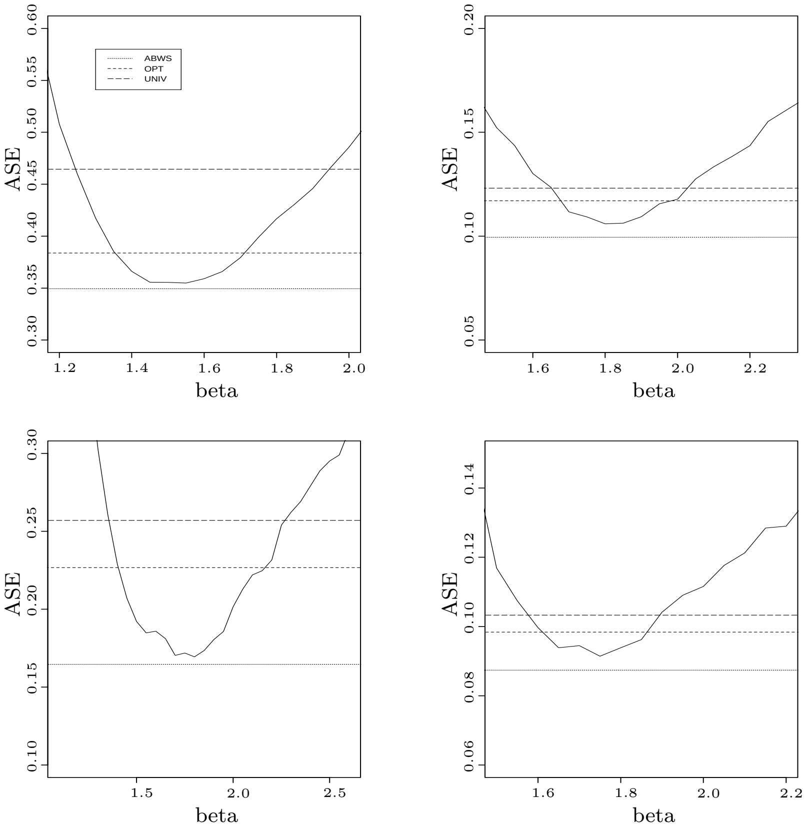

Figure 2. AMSE as a function of β compared with ABWS, optimal minimax, and universal AMSE errors. Panel (upper left) is for `bumps`, (upper right) is for `blocks`, (lower left) is for `doppler`, and (lower right) is for `heavisine` test functions. The legend is provided in Panel (upper left).

4. Simulations

In this section we consider BDT thresholding estimators for the standard test signals (`bumps, blocks, doppler` and `heavisine`). The scaled signals are affected by i.i.d. normal noise with unit variance. The length of signals is $n = 1024$ and decomposing wavelets are Haar for `blocks`, and least asymmetric Daubechies' of order 8 for `bumps, doppler`, and `heavisine.` The signal-to-noise ratio SNR is 7 and the variance of noise is chosen to match the conditions used by Bruce and Gao (1996) and Chipman, McCulloch and Kolaczyk (1997). The number of levels in the wavelet decompositions is the maximum (yielding a single coarse detail coefficient and a single "smooth" coefficient). Shrinkage has been applied only to the detail coefficients and the "smooth" coefficient has been left intact.

We evaluate the performance of BDT estimators by comparing the averaged mean squared errors, AMSE $= \sum_{i=1}^{n} \sum_{j=1}^{N} (\theta_i - \hat{\theta}_{i,j})^2/(Nn)$, where N is the number of simulated runs, θ_i is the true signal value, $\hat{\theta}_i$ is the estimate of the signal value from a simulation, and $\hat{\theta}_{i,j}$ is $\hat{\theta}_i$ in the jth simulation run.

The simulations are illustrated in Figure 2 where, for a range of β's, we depict the AMSE obtained from $N = 10$ simulation runs. The AMSE's from ABWS, OPT, and UNIV methods found in Bruce and Gao (1996) and Chipman, McCulloch and Kolaczyk (1997) are provided for comparison purposes. The panel (upper left) depicts the AMSE for the `bumps` signal, (upper right) for the `blocks`, (lower left) for the `doppler`, and (lower right) panel for the `heavisine` test functions.

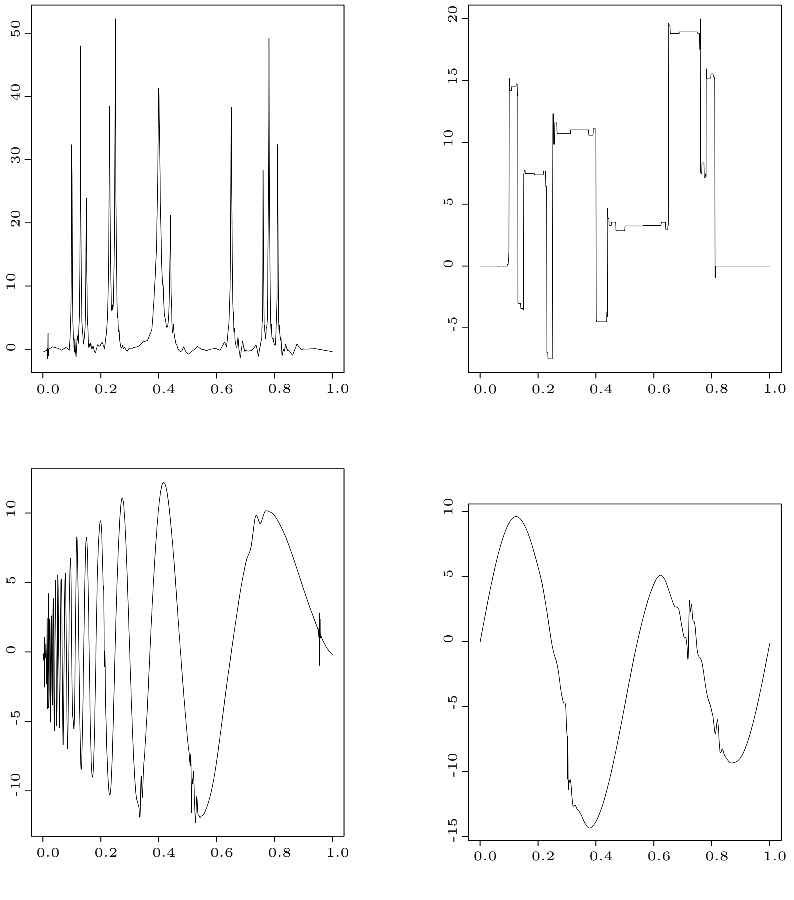

Figure 3. BDT estimators of the standard Donoho and Johnstone test signals.

We have found that the hard thresholding rule using the $\lambda^{\star}$ threshold (BDT rule) can achieve AMSE comparable to that of the adaptive method, ABWS.

The AMSE performance clearly depends on the choice of the parameter β. For instance in the case of the `blocks` signal, better AMSE than for the UNIV method can be expected for β in the approximate range of [1.4, 2.2].

Figure 3 gives the graphs of the processed signals obtained for the risk-minimizing β. We also noticed that repeated simulations gave stable and comparable reconstructions.

We also provide summaries of AMSE of BDT estimators for different sample sizes and signal-to-noise ratios. For sample sizes $n = 256, n = 512$, and $n = 1024$, SNR$=$ 3 and 7, and the four Donoho and Johnstone test functions, we found means and standard deviations of MSE based on 100 simulation runs, see Table 1.

Table 1. Descriptive statistics of MSE for BDT thresholding estimators based on $N = 100$ runs.

	size	SNR	Ave. MSE BDT (Ave. MSE Univ)	SD MSE BDT (SD MSE Univ.)	Ave. λ^* (λ^*/λ_{univ})	SD λ^*
bumps (symm 8)	256	3	0.6716 (1.0284)	0.1027 (0.1119)	2.0921 (0.628)	0.3107
		7	0.6614 (0.8553)	0.0678 (0.0943)	2.3093 (0.693)	0.3019
	512	3	0.4673 (0.6060)	0.0505 (0.0640)	2.6395 (0.747)	0.2779
		7	0.4814 (0.6449)	0.0521 (0.0658)	2.5350 (0.718)	0.2789
	1024	3	0.3222 (0.0395)	0.0318 (0.0395)	2.8411 (0.763)	0.2456
		7	0.3784 (0.4812)	0.0320 (0.0417)	2.7464 (0.737)	0.2746
blocks (haar)	256	3	0.3827 (0.5127)	0.0780 (0.0976)	2.6326 (0.791)	0.3234
		7	0.2646 (0.3100)	0.0536 (0.0673)	3.0113 (0.904)	0.4046
	512	3	0.2282 (0.3215)	0.0349 (0.0493)	2.8226 (0.799)	0.2707
		7	0.1543 (0.1808)	0.0323 (0.0369)	3.1573 (0.894)	0.3348
	1024	3	0.1397 (0.1728)	0.0192 (0.0263)	3.1518 (0.846)	0.2886
		7	0.1046 (0.1228)	0.0187 (0.0248)	3.3565 (0.901)	0.3669
doppler (symm 8)	256	3	0.3917 (0.5074)	0.0672 (0.0731)	2.6737 (0.803)	0.3417
		7	0.4205 (0.5075)	0.0652 (0.0828)	2.7109 (0.814)	0.3455
	512	3	0.2303 (0.2574)	0.0335 (0.0354)	3.1984 (0.906)	0.3315
		7	0.3186 (0.4094)	0.0403 (0.0624)	2.7806 (0.787)	0.2713
	1024	3	0.1554 (0.1695)	0.0163 (0.0177)	3.3977 (0.912)	0.4355
		7	0.2108 (0.2527)	0.0258 (0.0292)	3.1265 (0.840)	0.2715
heavisine (symm 8)	256	3	0.1580 (0.1786)	0.0316 (0.0420)	3.0248 (0.908)	0.5424
		7	0.2352 (0.2914)	0.0573 (0.0598)	2.9018 (0.871)	0.4058
	512	3	0.1199 (0.1394)	0.0256 (0.0292)	3.3658 (0.953)	0.5290
		7	0.1527 (0.1877)	0.0278 (0.0378)	3.1214 (0.884)	0.3728
	1024	3	0.0518 (0.0561)	0.0079 (0.0105)	3.6740 (0.986)	0.5578
		7	0.0886 (0.0998)	0.0153 (0.0178)	3.5218 (0.946)	0.3959

We compare AMSE of BDT estimators with those obtained by universal thresholding. Standard deviations of MSE's are compared as well. The numbers in brackets correspond to UNIV estimators. In the last two columns we provide average BDT thresholds, their ratios with the universal threshold (numbers in brackets) and the sample standard deviations of BDT thresholds. It is evident from Table 1 that AMSE for BDT estimators are in all cases smaller than those for the universal threshold. The corresponding standard deviations of MSE are smaller as well. We found that the size of the BDT threshold is, usually, 60-90% of the size of a corresponding universal threshold. The size and performance of a BDT threshold is comparable to an optimal minimax threshold of Bruce and Gao (1996), Table 3, page 739.

5. Discussion and Conclusions

In this paper, we have outlined an approach for Bayesian wavelet shrinkage based on the decision theoretic paradigm. The shrinkage is simple: a unique hard thresholding rule is applied to all the detail wavelet coefficients. When there is precise information about the parameters of the model, decision theoretic shrinkage can lead to mean squared error performance exceeding that of some of the popular shrinkage techniques.

We conclude by discoursing on some potential developments of our approach. Similar work can be done when the distribution depends on a scale parameter. Such models are interesting when the noise in the wavelet domain is multiplicative. Theoretical results about scale-parameter models, similar to those presented in Section 3 of this paper, can be found in Ruggeri and Vidakovic (1996).

The BDT approach may be further generalized by specifying the models in the wavelet domain which are level dependent. For instance, if the prior model is normal, the variance may be a decreasing function of the level, as suggested by Wahba (1981), in the context of density estimation by classical orthogonal series.

Acknowledgements

The research of the second author was supported by NSF Grant DMS-9626159 at Duke University and ONR contract N00014-93-C-0106 at StatSci. The research was developed during a visit of the first author at Duke University. The authors wish to thank David Rios Insua and Bruno Sansò for their comments. Comments of the editor Ching-Shui Cheng, the associate editor and two anonymous referees further improved clarity and focus of the paper. S+Wavelets software was used in the simulation part. The S-Plus codes used in simulations and figures are available at request from the authors.

Appendix

Proof of Theorem 2.2. By applying the symmetry properties of f, π and L and the transformation $\theta - \lambda = -t$, it follows that

$$\begin{aligned}
&\int_{\mathcal{R}} [L(\theta) - L(\theta - \lambda)] f(\theta - \lambda)\pi(\theta) d\theta \\
&= -\int_{\lambda}^{\infty} [L(\theta) - L(\theta - \lambda)] f(\theta)\pi(\theta - \lambda) d\theta + \int_0^{\lambda} [L(\theta) - L(\theta - \lambda)] f(\lambda - \theta)\pi(\theta) d\theta \\
&\quad + \int_{\lambda}^{\infty} [L(\theta) - L(\theta - \lambda)] f(\theta - \lambda)\pi(\theta) d\theta \\
&= \int_0^{\lambda} [L(\theta) - L(\theta - \lambda)] f(\lambda - \theta)\pi(\theta) d\theta \\
&\quad + \int_{\lambda}^{\infty} [L(\theta) - L(\theta - \lambda)][f(\theta - \lambda)\pi(\theta) - f(\theta)\pi(\theta - \lambda)] d\theta \\
&= I_1(\lambda) + I_2(\lambda).
\end{aligned}$$

Let $\tau_\lambda(\theta) = f(\theta - \lambda)\pi(\theta) - f(\theta)\pi(\theta - \lambda)$. Suppose that $f(t)/\pi(t)$ is decreasing. It then holds that $\Delta(x_1, x_2) = f(x_1)\pi(x_2) - f(x_2)\pi(x_1) > 0$ for all $0 < x_1 < x_2$.

Consider $I_2(\lambda) = \int_{\lambda}^{\infty} [L(\theta) - L(\theta - \lambda)]\tau_\lambda(\theta) d\theta$. Since $\theta > \theta - \lambda \geq 0$ for all $\theta \geq \lambda$ and since $L(t)$ is nondecreasing for nonnegative t, observe that $[L(\theta) - L(\theta - \lambda)] \geq 0$.

Similarly, $\Delta(\theta - \lambda, \theta) = \tau_\lambda(\theta) > 0$, so that $I_2(\lambda) > 0$.

Consider now $I_1(\lambda)$. It follows that

$$\begin{aligned}
I_1(\lambda) &= \int_0^{\lambda} L(\theta) f(\lambda - \theta)\pi(\theta) d\theta + \int_0^{\lambda} L(\theta) f(\theta)\pi(\theta - \lambda) d\theta \\
&= \int_0^{\lambda} L(\theta)\tau_\lambda(\theta) d\theta \\
&= \int_0^{\lambda/2} [L(\theta) - L(\theta - \lambda)]\tau_\lambda(\theta) d\theta.
\end{aligned}$$

By an argument similar to the previous one, it follows that $I_1(\lambda) > 0$, since both $[L(\theta) - L(\theta - \lambda)]$ and $\tau_\lambda(\theta)$ are negative. Also $\lambda^* = 0$, since $r'(\lambda) > 0$ for all $\lambda > 0$.

When $f(t)/\pi(t)$ is increasing, it can be similarly proved that $\lambda^* = \infty$ since $r'(\lambda) < 0$ for all $\lambda > 0$.

References

Berger, J., Betrò, B., Moreno, E., Pericchi, L. R., Ruggeri, F., Salinetti, G. and Wasserman, L. (1996). *Bayesian Robustness.* Lecture Notes **29**, IMS. Hayward, California.

Bruce, A. and Gao, H-Y. (1996). Understanding WaveShrink: Variance and bias estimation. *Biometrika* **83**, 727-745.

Chipman, H., McCulloch, R. and Kolaczyk, E. (1997). Adaptive Bayesian shrinkage. *J. Amer. Statist. Assoc.* **92**, 1413-1421.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391-402.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia, Pennsylvania.

Donoho, D. (1997). CART and best-ortho-basis: a connection. *Ann. Statist.* **25**, 1870-1911.

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

Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

Fan, J. (1997). Comments on *Wavelets in Statistics: A Review* by A. Antoniadis, to appear in *Journal of Italian Statistical Society.*

Hernandez, E. and Weiss, G. (1996). *A First Course on Wavelets.* CRC Press, Boca Raton, Florida.

Nason, G. (1995). Choice of the threshold parameter in wavelet function estimation. In *Wavelets in Statistics* (Edited by A. Antoniadis and G. Oppenheim), 261-280. Lecture Notes in Statistics, Springer-Verlag, New York,

Ruggeri, F. and Vidakovic, B. (1996). A Bayesian decision theoretic approach to wavelet thresholding: scale parameter models. *Proceedings Joint Statistical Meetings*, Chicago, Illinois.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* **93**, 173-179.

Wahba, G. (1981). Data based optimal smoothing of orthogonal series density estimates. *Ann. Statist.* **9**, 146-156.

Walter, G. (1994). *Wavelets and Others Orthogonal Systems with Applications.* CRC Press, Boca Raton, Florida.

Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni, della Matematica e dell'Informatica, Via A. M. Ampere, 56, I-120131 Milano, Italy.

E-mail: fabrizio@iami.mi.cnr.it

Institute of Statistics and Decision Sciences, Old Chem Building 223 B, Duke University, Durham, NC 27708-0251, U.S.A.

E-mail: brani@isds.duke.edu

(Received January 1997; accepted September 1998)