Wavelet Bayesian Block Shrinkage via Mixtures of Normal-Inverse Gamma Priors

DANIELA DE CANDITIIS
Istituto per Applicazioni della Matematica, CNR
Naples, Italy

BRANI VIDAKOVIC
Georgia Institute of Technology
Atlanta, GA 30332-0205, USA

Abstract

In this paper we propose a block shrinkage method in the wavelet domain for estimating an unknown function in the presence of Gaussian noise. This shrinkage utilizes an empirical Bayes, block-adaptive approach that accounts for the sparseness of the representation of the unknown function. The modeling is accomplished by using a mixture of two normal-inverse gamma $(\mathcal{NIG})$ distributions as a joint prior on wavelet coefficients and noise variance in each block at a particular resolution level. This method results in explicit and fast rules. An automatic, level dependent choice for the prior hyperparameters is also suggested. Finally, the performance of the proposed method, **BBS** (Bayesian Block Shrinkage), is illustrated on the battery of standard test functions and compared to some standard wavelet-based denoising methods.

KEY WORDS: Wavelet Regression; Shrinkage; Bayesian Estimation; Normal-Inverse Gamma Priors.

Daniela De Canditiis is at Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni della Matematica, Via P. Castellino 111, I-80131 Napoly, Italy.

Brani Vidakovic is an Associate Professor at the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA.

1 Introduction

The problem of interest is to estimate the function f using the observations $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, in a standard nonparametric regression problem

$$Y_i = f(t_i) + \sigma\epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where t_i, $i = 1, \dots, n$, is a sequence of equispaced points, $\sigma > 0$ is an unknown noise level, and the errors ϵ_i are i.i.d. standard normal random variables.

Wavelet-based procedures have shown to be well suited for such settings and non-parametric estimators are obtained by applying simple shrinkage rules on the wavelet-transformed data. A rapidly growing literature on this problem is available, see Antoniadis (1997), Vidakovic (1999), and Percival and Walden (2000), among others for a general overview.

A variety of shrinkage methods based on classical and Bayesian statistical models in the wavelet domain have been proposed and studied since Donoho and his coauthors (Donoho and Johnstone, 1994; Donoho *et al.*, 1995) first introduced *VisuShrink*, *SureShrink*, and their modifications. In this broad context of function estimation, Bayesianly justified procedures have proved efficient for their capability of incorporating prior information about the unknown signal. According to the Bayesian paradigm, a location model in the wavelet domain is assumed and a prior distribution is imposed on the location and other unknown parameters of the model. Since unknown locations correspond to the signal in the time domain an estimator of the signal is obtained by inversely transforming Bayes estimators of the locations. Priors on the signal part in the wavelet domain can incorporate information about smoothness, periodicity, selfsimilarity, and some other characteristics of f. Some examples of such Bayesian procedures can be found in Chipman, Kolaczyk, and McCulloch (1997), Abramovich, Sapatinas and Silverman (1998), Clyde, Parmigiani and Vidakovic (1998), Vidakovic (1998), Vidakovic and Ruggeri (2001) and in the volume by Müller and Vidakovic (1999), among others.

Motivated by the need for spatial adaptivity, Hall, Kerkyacharian, and Picard (1998, 1999) first suggested grouping wavelet coefficients into blocks and modeling them block-wise exploiting the information that coefficients convey about the size of their *nearby* neighbors. Interesting theoretical optimality results have been obtained for such block estimators and their excellent MSE-performance has been reported. See also Cai (1999a, 1999b), Hall, Penev, Kerkyacharian and Picard (1997), Efromovich (2000), Cai and Silverman (2000). Recently, the proposed block methodology has been explored from the Bayesian point of view by

Abramovich, Besbeas and Sapatinas (2000) who develop empirical Bayes block shrinkage estimators.

We propose the Bayesian Block Shrinkage (**BBS**) method, in which non-linear block-shrinkage rules are obtained via a Bayesian modeling approach. Specifically, wavelet coefficients at each resolution level are grouped in blocks of a given size and a Bayesian model is defined on each block, by taking into account both the sparseness of wavelet representations of the noiseless signal and the magnitude of the error affecting the sample. The need to model dependence between neighboring coefficients and their spareness is accomplished by a mixture of two normal-inverse gamma (NIG) distributions with different "scales", mixed in a proportion which depends on the level in a wavelet decomposition. This is in accordance with well established Berger-Müller priors in the wavelet domain consisting of two components: point mass or almost point mass that models non-energetic coefficients, and the "spread" distribution modeling large wavelet coefficients. By considering this model in the wavelet domain we can also match the marginal (predictive) model and observed empirical distribution of wavelet coefficients. In our setting the marginal distribution of each block of wavelet coefficients turns out to be a balanced combination of two multivariate Student $\boldsymbol{t}$ distributions one of which being very concentrated around zero. Such marginal "matching" modeling is impossible if plug-in estimators of σ^2 are used; see the argument in Vidakovic and Ruggeri (2001).

The normal-inverse gamma priors have previously been used in the wavelet context because of their conjugacy with respect to normal conditional models, see Vidakovic and Müller (1995), Vannucci and Corradi (1999), and De Canditiis (2001).

The paper is organized as follows. Section 2 describes the model, prior selection and provides derivation of the shrinkage rule. Section 3 discusses a simple and automatic choice for the prior parameters, which works well for standard simulation contexts. Finally, Section 4 contains the simulational study performed on the battery of standard test functions and comparison with some standard wavelet-based methods.

2 The Model

We consider the following regression model:

$$\boldsymbol{Y} = \boldsymbol{f} + \sigma\boldsymbol{\epsilon}\,, \tag{2}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is a vector of equispaced observations, $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))'$ is the vector of values of the regression function at the design points t_i and $\boldsymbol{\epsilon}$ is a vector of i.i.d. normal random errors. In the sequel, bold variables stand for vectors. Without loss of generality, we assume that the sample size is $n = 2^N$. The standard probability model for the data is

$$[\boldsymbol{Y} \,|\boldsymbol{f}, \sigma^2] \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 I_n),$$

where $\sim$ means 'distributed as'.

Let $\mathbf{W}$ represent the orthogonal matrix of size $n \times n$ performing a discrete wavelet transformation down to the level J. If $c_{j,k}$ and $d_{j,k}$ are scaling and wavelet coefficients respectively at level j and shift k then

$$\mathbf{W} \begin{pmatrix} c_{N,0} \\ \vdots \\ c_{N,2^N-1} \end{pmatrix} = (c_{J,0}, \ldots, c_{J,2^J-1}, \quad (3)$$

$$d_{J,0}, \ldots, d_{J,2^J-1}, \ldots, d_{N-1,0}, \ldots, d_{N-1,2^{N-1}-1})' = \boldsymbol{d},$$

where $(c_{N,0} \ldots c_{N,2^N-1})' = \mathbf{Y}$. Here, $J \geq 0$ is fixed and its selection depends on the application. Since $\mathbf{W}$ is an orthogonal matrix, transforming the regression model (2) into the time-frequency (wavelet) domain by applying $\mathbf{W}$ does not change the structure of the model.

Regression problem (2) becomes

$$\boldsymbol{d} = \boldsymbol{\theta} + \sigma\boldsymbol{\epsilon}'$$

in the wavelet domain. The noise $\boldsymbol{\epsilon}'$ is a vector of i.i.d. standard normal variables, $\boldsymbol{\theta} = \mathbf{W}\boldsymbol{f}$ is the location vector, and σ^2 is the variance of the noise. Thus, distribution of $\boldsymbol{d}$, given $\boldsymbol{\theta}$ and σ^2 remains normal,

$$[\boldsymbol{d} \,|\boldsymbol{\theta}, \sigma^2] \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 I_n), \quad (4)$$

and estimating unknown $\boldsymbol{f}$ amounts to estimating $\boldsymbol{\theta}$ in (4) and transforming the estimator by $\mathbf{W}^{-1}$. We first discuss completion of model (4) by normal-inverse gamma prior on $(\boldsymbol{\theta}, \sigma^2)$.

2.1 Normal-Inverse Gamma Prior

From a Bayesian viewpoint, the model (4) is completed with a prior distribution for parameters $(\boldsymbol{\theta}, \sigma^2)$. When the joint prior is given by

$$[\boldsymbol{\theta}, \sigma^2] \sim \mathcal{NIG}(\alpha, \delta, \boldsymbol{m}, \Sigma), \text{ with } \alpha, \delta \in \mathbb{R}, \ \boldsymbol{m} \in \mathbb{R}^n, \ \Sigma \in \mathbb{R}^n \times \mathbb{R}^n, \quad (5)$$

where $\mathcal{NIG}(\alpha, \delta, \boldsymbol{m}, \Sigma)$ is the normal-inverse gamma distribution defined by

$$[\boldsymbol{\theta} \,|\sigma^2] \sim \mathcal{N}(\boldsymbol{m}, \sigma^2\Sigma) \text{ and } [\sigma^2] \sim IG(\alpha, \delta), \tag{6}$$

then, by the conjugate structure, the posterior is still normal-inverse gamma

$$p(\boldsymbol{\theta}, \sigma^2) = \mathcal{NIG}(\alpha, \delta, \boldsymbol{m}, \Sigma)p(\boldsymbol{d} \,|\boldsymbol{\theta}, \phi)/p(\boldsymbol{y}) = \mathcal{NIG}(\alpha^\star, \delta^\star, \boldsymbol{m}^\star, \Sigma^\star) \tag{7}$$

with

$$\begin{array}{rcl} \Sigma^{\star -1} & = & (\Sigma^{-1} + I) \\ \boldsymbol{m}^\star & = & \Sigma^\star(\Sigma^{-1}\boldsymbol{m} + \boldsymbol{d}) \\ \alpha^\star & = & \boldsymbol{d}'\boldsymbol{d} + \boldsymbol{m}'\Sigma^{-1}\boldsymbol{m} + \alpha - \boldsymbol{m}^{\star T}\Sigma^{\star -1}\boldsymbol{m}^\star \\ \delta^\star & = & \delta + n. \end{array} \tag{8}$$

Note that the estimator of $\boldsymbol{\theta}$ (posterior location $\mathbf{m}^\star$) in (8) is an affine transformation of $\boldsymbol{d}$. Vidakovic and Müller (1995) and Vannucci and Corradi (1999) utilize this estimator in wavelet shrinkage with $\boldsymbol{m}$ =0 and conveniently defined Σ.

The marginal (predictive) distribution of the data is

$$p(\boldsymbol{d}) = \frac{|\Sigma^\star|^{1/2}\alpha^{\delta/2}\Gamma(\delta^\star/2)}{|\Sigma|^{1/2}(2\pi)^{n/2}\Gamma(\delta/2)}(\alpha^\star)^{-\delta^\star/2}. \tag{9}$$

Note that $\boldsymbol{d}$ appears in (9) through $\alpha^\star$. See O'Hagan (1994) for a detailed discussion.

Although the $\mathcal{NIG}$ prior is suitable for modeling correlation structure among wavelet coefficients, a shortcoming of resulting rules is that they shrink "linearly", which is believed to be suboptimal in the wavelet context. Also, unlike the empirical distributions of wavelet coefficients, the marginal distribution (9) is not peaked at 0.

Instead of a single $\mathcal{NIG}$ prior for locations and variances of all detail coefficients, we consider a mixture of two $\mathcal{NIG}$ priors on small, non-overlapping blocks of coefficients. Motivation and description for this selection is provided next.

2.2 Prior Selection

We restrict our modeling only to coefficients corresponding to wavelets (detail coefficients in $\boldsymbol{d}$). The scaling coefficients (2^J of them, corresponding to scaling functions) carry the information mostly about the underlining signal and in the process of wavelet shrinkage these coefficients are usually left unchanged.

Bayesian shrinkage is applied only for the wavelet coefficients, $d_{j,k}, j = J, \ldots, N-1; \;\; k = 0, \ldots, 2^j - 1$, in the decomposition $\boldsymbol{d}$. At each resolution level j in $\boldsymbol{d}$ ($j = J, \ldots, N-1$) the wavelet coefficients are grouped into $l_j = 2^j/l$ non-overlapping blocks, $\boldsymbol{d}_{j,H} = \boldsymbol{\theta}_{j,H} + \boldsymbol{\epsilon}', \; (H = 1, \ldots, l_j)$, where l is the size block (a power of 2). The blocks are assumed independent of each other, and for each $j = J, \ldots, N-1$ and $H = 1, \ldots, l_j$ we assume normal likelihood and prior distribution over $(\boldsymbol{\theta}_{j,H}, \sigma^2)$ designed to capture the sparseness of the signal in the wavelet domain.

Our prior choice is a level-dependent mixture of two $\mathcal{NIG}$ distributions with different variances balanced by random weight γ_j:

$$[\boldsymbol{\theta}_{j,H}, \sigma^2 | \gamma_j] \sim \gamma_j \mathcal{NIG}(\alpha, \delta, \mathbf{0}, \Sigma_j) + (1-\gamma_j)\mathcal{NIG}(\alpha, \delta, \mathbf{0}, \Delta_j), \tag{10}$$

$$[\gamma_j] \sim \textit{Bernoulli}(p_j)$$

with $\Sigma_j = c_j^2 I_l$ and $\Delta_j = \tau_j^2 I_l$. Parameters $\alpha > 0$, $\delta > 2$, $0 \leq p_j \leq 1$, $\tau_j > 0$ and $c_j > 1$ are selected depending on the problem (data); an automatic selection of parameters is possible and will be discussed later.

The first component in the mixture (11) is a spread distribution that models large wavelet coefficients ($c_j \gg 1$) while the second component describes small coefficients (τ_j is small). Parameter p_j is the proportion of large wavelet coefficients at each resolution level j, and should decrease with increase of j. The proposed model can be thought as a generalization of the model used in Chipman, Kolaczyk and McCulloch (1997).

2.3 Derivation of the Shrinkage Rule.

Utilizing equation (9), we find the marginal distribution for each data block $\boldsymbol{d}_{j,H}$

$$\begin{aligned} p(\boldsymbol{d}_{j,H}) &= p_j p(\boldsymbol{d}_{j,H}|\gamma_j = 1) + (1-p_j)p(\boldsymbol{d}_{j,H}|\gamma_j = 0) \\ &= p_j \frac{|\Sigma_j^\star|^{1/2}\alpha^{\delta/2}\Gamma(\delta_j^\star/2)}{|\Sigma_j|^{1/2}(2\pi)^{l/2}\Gamma(\delta/2)}(\alpha_{j,H}^\star)^{-\delta_j^\star/2} \\ &\quad + (1-p_j)\frac{|\Delta_j^{\star\star}|^{1/2}\alpha^{\delta/2}\Gamma(\delta_j^{\star\star}/2)}{|\Delta_j|^{1/2}(2\pi)^{l/2}\Gamma(\delta/2)}(\alpha_{j,H}^{\star\star})^{-\delta_j^{\star\star}/2}, \end{aligned} \tag{11}$$

where symbols * and ** indicate the associated posterior parameters of the components corresponding to "large" and "small" coefficients respectively, as given in (8).

By using the marginal distribution (9) and the conjugate structure of $\mathcal{NIG}$ priors discussed earlier, we obtain the posterior. For related results see Berger, 1985, page 207.

$$
\begin{aligned}
& p_j \, \frac{p(\boldsymbol{d}_{j,H}|\gamma_j = 1)}{p(\boldsymbol{d}_{j,H})} \; \frac{\mathcal{NIG}(\alpha, \delta, \mathbf{0}, \Sigma_j) p(\boldsymbol{d}_{j,H}|\boldsymbol{\theta}_{j,H}, \sigma^2)}{p(\boldsymbol{d}_{j,H}|\gamma_j = 1)} + \\
& (1 - p_j) \, \frac{p(\boldsymbol{d}_{j,H}|\gamma_j = 0)}{p(\boldsymbol{d}_{j,H})} \; \frac{\mathcal{NIG}(\alpha, \delta, \mathbf{0}, \Delta_j) p(\boldsymbol{d}_{j,H}|\boldsymbol{\theta}_{j,H} \sigma^2)}{p(\boldsymbol{d}_{j,H}|\gamma_j = 0)} \qquad (12)
\end{aligned}
$$

After straightforward, but tedious algebra we obtain that the posterior is a mixture of $\mathcal{NIG}$ distributions, as the prior was, but with data-dependent mixing weights,

$$
\begin{aligned}
[\boldsymbol{\theta}_{j,H}, \sigma^2 | \boldsymbol{d}_{j,H}] \sim \quad & A_{j,H} \mathcal{NIG}(\alpha^\star_{j,H}, \delta^\star, \boldsymbol{m}^\star_{j,H}, \Sigma^\star_j) + \\
& (1 - A_{j,H}) \mathcal{NIG}(\alpha^{\star\star}_{j,H}, \delta^\star, \boldsymbol{m}^{\star\star}_{j,H}, \Delta^{\star\star}_j), \qquad (13)
\end{aligned}
$$

where

$$
\begin{aligned}
\Sigma^\star_j &= \tfrac{c_j^2}{1+c_j^2} I_l & \Delta^{\star\star}_j &= \tfrac{\tau_j^2}{1+\tau_j^2} I_l \\
\boldsymbol{m}^\star_{j,H} &= \Sigma^\star_j \boldsymbol{d}_{j,H} & \boldsymbol{m}^{\star\star}_{j,H} &= \Delta^{\star\star}_j \boldsymbol{d}_{j,H} \\
\alpha^\star_{j,H} &= \alpha + \boldsymbol{d}'_{j,H} \boldsymbol{d}_{j,H} & \alpha^{\star\star}_{j,H} &= \alpha + \boldsymbol{d}'_{j,H} \boldsymbol{d}_{j,H} \\
& \quad - \boldsymbol{m}^{\star T}_{j,H} \Sigma^{\star -1}_j \boldsymbol{m}^\star_{j,H} & & \quad - \boldsymbol{m}^{\star\star T}_{j,H} \Delta^{\star\star -1}_{j,H} \boldsymbol{m}^{\star\star}_{j,H} \\
\delta^\star &= \delta + l, & \delta^{\star\star} &= \delta + l \;\; (= \delta^\star),
\end{aligned}
\qquad (14)
$$

and

$$
A_{j,H} = \frac{p_j / \sqrt{1 + c_j^2}}{\frac{p_j}{\sqrt{1+c_j^2}} + \frac{1-p_j}{\sqrt{1+\tau_j^2}} \left[(\alpha + \frac{\boldsymbol{d}'_{j,H} \boldsymbol{d}_{j,H}}{1+\tau_j^2}) / (\alpha + \frac{\boldsymbol{d}'_{j,H} \boldsymbol{d}_{j,H})}{1+c_j^2}) \right]^{-(\delta+l)/2}}
$$

In inference about $\boldsymbol{\theta}_{j,H}$, σ^2 is a nuisance parameter and after integrating it out from (13) we obtain

$$
\boldsymbol{\theta}_{j,H} | \boldsymbol{d}_{j,H} \sim A_{j,H} t_{\delta^\star}(\boldsymbol{m}^\star_{j,H}, \alpha^\star_{j,H} \Sigma^\star_j) + (1 - A_{j,H}) t_{\delta^\star}(\boldsymbol{m}^{\star\star}_{j,H}, \alpha^{\star\star}_{j,H} \Delta^{\star\star}_j), \quad (15)
$$

According to Bayesian paradigm, the posterior distribution carries all the information about the modeled parameters. The Bayes rule (when the underlying loss is the squared error) is the mean of the posterior

$$\mathbb{E}(\boldsymbol{\theta}_{j,H}|\boldsymbol{d}_{j,H}) = A_{j,H}(\boldsymbol{d}_{j,H})\boldsymbol{m}^{\star}_{j,H} + (1 - A_{j,H}(\boldsymbol{d}_{j,H}))\boldsymbol{m}^{\star\star}_{j,H}. \quad (16)$$

The posterior variance, important in assessing the variability of the estimator is also explicit

$$\begin{aligned} var(\boldsymbol{\theta}_{j,H}|\boldsymbol{d}_{j,H}) &= A_{j,H}\frac{\alpha^{\star}_{j,H}}{\delta^{\star}-2}\Sigma^{\star}_{j} + (1-A_{j,H})\frac{\alpha^{\star\star}_{j,H}}{\delta^{\star\star}-2}\Sigma^{\star\star}_{j} + \\ &\quad A_{j,H}(1-A_{j,H})(\boldsymbol{m}^{\star}_{j,H} - \boldsymbol{m}^{\star\star}_{j,H})(\boldsymbol{m}^{\star}_{j,H} - \boldsymbol{m}^{\star\star}_{j,H})'. \end{aligned}$$

3 Eliciting the Hyper-parameters

Given the potential vast variability of functions f we propose an automatic procedure for determining the hyperparameters and thus, the shrinkage rule. It is commonly believed that the selection of hyperparameters must depend on data, and a formal way of incorporating such an information is the empirical Bayes paradigm.

In order to give a default selection of hyperparameters procedure for the estimates in (16), here we propose a data dependent specifications of the prior parameters, α, δ, p_j, τ_j and c_j ($j = J, \ldots, N-1$). Purely subjective elicitations of prior could be done only when the extensive knowledge about the underlining signal is available but, even in this case, "well scrambled" wavelet domain makes the elicitation difficult.

Next, we propose an empirical parameter specification that works well for standard test cases and discuss the sensitivity with respect to this specification.

Parameters α and δ describe our prior knowledge about the variance of the noise, σ^2. Independently of the level j, σ^2 is modeled a priori via $IG(\alpha,\delta)$ for both components in the mixture. To assess the hyperparameters α and δ, we propose to estimate σ^2 using the coefficients from the finest level of detail, $\hat{\sigma} = \text{median}_{0\le k\le 2^{N-1}}|d_{N-1,k}|\ /0.6745$, and then link α and δ such that the mode $(\alpha/(\delta+2))$ of the distribution $IG(\alpha,\delta)$ coincides with the estimate $\hat{\sigma^2}$ obtained from the sample. This condition does not specify α exactly. We fix $\alpha = 8$ since our simulations indicate that the shrinkage estimator remains robust when α is between 3 and 12.

Three additional hyperparameters have to be specified for each resolution level $j = J, \ldots, N-1$. Hyperparameters τ_j^2 and c_j^2 are variances of the "concentration" and "spread" parts respectively in the prior mixture distribution, while p_j is the level-dependent probability that a given block is of high energy . Our recommendations are the following:

$$c_j^2 = 3 \max |\boldsymbol{d}_{j,\cdot}| \quad \text{and} \quad \tau_j^2 = \max \left\{ 10^{-6} max|d_{j,\cdot}|, min|d_{j,\cdot}| \right\}, \tag{17}$$

where $d_{j,\cdot}$ is an arbitrary wavelet coefficient at level j. Finally, for p_j we follow recommendation of Chipman, Kolaczyk, and McCulloch (1997); p_j is defined as the proportion of wavelet coefficients with locations corresponding to the "spread" part,

$$p_j = \frac{\#\{d_{j,k} : |d_{j,k}| > \sqrt{2\log(2^j)}\hat{\sigma}^2\}}{2^j}.$$

The "universal threshold value", $\sqrt{2\log(2^j)}\hat{\sigma}^2$, is the probabilistic upper-bound for the size of normal noise at resolution level j, as given in Donoho, Johnstone, Kerkyacharian and Picard (1995). The main advantages of these selections, fully defining the **BBS** method, are their intuitive appeal and computational simplicity.

Described prior selection works well when the signal f is normalized such that the added noise with variance $\sigma^2 = 1$ makes a prescribed signal-to-noise ratio (SNR). In real-life examples, when the level of noise is not known, some information on SNR will be useful for rescaling the problem so that the noise level is close to 1.

Alternative proposals can be given for a general case. An example is the choice of c_j and τ_j as in Chipman, Kolaczyk and McCulloch (1997). We tested the **BBS** method using their hyperparameter proposals and with slightly increased computational time obtained comparable results.

4 Simulations

In this section we show simulated numerical results obtained by applying the estimator given in (16) on a standard battery of test functions `blocks`, `bumps`, `doppler`, and `heavisine`. We denote $\hat{\boldsymbol{f}} = (\hat{f}(t_1), \ldots, \hat{f}(t_n))'$ the estimator evaluated on the grid design and measure its performance by an average mean square error AMSE, calculated as

$$\frac{1}{Mn}\sum_{k=1}^{M}\sum_{i=1}^{n}\left(\hat{f}_k(t_i) - f(t_i)\right)^2,$$

where M is the number of simulational runs, $f(t_i)$ are exact values to be estimated and $\hat{f}_k(t_i)$ are corresponding estimates in the k-th simulational run, $k = 1, \ldots, M$. To assess the performance of the procedure we compared the AMSE, variance and squared bias of the proposed estimator to those tabulated in Vidakovic and Ruggeri (2001) (Tab.1 VisuShrink, SureShrink, ABWS, BAMS). The standard test functions are re-scaled so that an added standard normal noise produces a signal-to-noise ratio (SNR) of 7; the sample size was n=1024, and the wavelets used were: Symmlet 8 for `heavisine` and `doppler`, Haar for `blocks` and Daubechies with 3 vanishing moments for `bumps`. Results are presented in Tab.1, where M=1000 simulation runs are summarized. Improvements in AMSE are notable for functions possessing significant spatial variability such as `bumps` and `doppler`; however, to achieve better results for `blocks` and `heavisine` a reduction of block lengths is necessary. All simulations summarized in Tab. 1 are performed using block size $l = 4$ and parameter $\alpha = 8$. The size of blocks plays an important rule in estimation and related theoretical results have been addressed in Cai (1999b). By empirically tuning the block-size, we found that $l = 4$ is an overall best choice. Better values for AMSE can be achieved for `blocks` and `heavisine` test functions by reducing l to 2. Block-sizes exceeding 4 are not recommended.

The **BBS** method is robust with respect to the choice of parameter α. In Fig.1, for the standard test functions and for $M = 1000$ simulation runs, the AMSE's are plotted for each choice of α in the interval (1,25). For $\alpha \in (3, 12)$ the value of AMSE is not changing much. Finally, in Tab.2 we summarize an extensive simulational study of the estimator. For $M = 1000$ runs, we present AMSE for four different SNR's (3,5,7,10) and five sample sizes n (256, 512, 1024, 2048, 4096), using the choice $\alpha = 8$ and block size $l = 4$. Typical reconstructions are plotted for each test function in Figs.2 and 3.

5 Conclusions

In this paper we proposed and explored wavelet shrinkage method capable of modeling dependence of neighboring wavelet coefficients and addressing local changes in the data. The proposed non-linear Bayesian Block Shrinkage (**BBS**)is adaptive in scale and time since the prior distribution on the signal part is (i) level-dependent and (ii) defined in coefficients belonging to short blocks. An additional time-adaptivity can be incorporated by specifying different priors block-wise via the parameters $c^2_{j,H}$ and $\tau^2_{j,H}$ influencing the component means and their weights

	block	bumps
VisuShrink	0.6840 (0.6122+0.0719)	1.5707 (1.4543+0.1165)
SureShrink	0.2225 (0.0856+0.1369)	0.6827 (0.4167+0.1165)
ABWS	0.0995 (0.0121+0.0874)	0.3495 (0.1267+0.2228)
BAMS	0.1107 (0.0142+0.0965)	0.3404 (0.1428+0.1976)
BBS	0.2034 (0.0061+0.1973)	0.2961 (0.0373+0.2588)
	doppler	heavisine
VisuShrink	0.4850 (0.4327+0.0523)	0.1204 (0.0864+0.0339)
SureShrink	0.2285 (0.1340+0.0946)	0.0949 (0.0534+0.0416)
ABWS	0.1646 (0.0640+0.1006)	0.0874 (0.0433+0.0442)
BAMS	0.1482 (0.0584+0.0899)	0.0815 (0.0304+0.0511)
BBS	0.1185 (0.0288+0.0897)	0.0860 (0.0394+0.0466)

Table 1: AMSE (Bias^2 + Variance) for VisuShrink, SureShrink, ABWS, BAMS and BBS on standard test functions. Test signals are rescaled so that the noise variance $\sigma^2 = 1$ gives SNR=7.

in the posterior $\mathcal{NIG}$ mixture.

This flexibility can be utilized when appropriate prior knowledge on the underlying function f is available; in the case when such prior information is absent, the **BBS** procedure, with an automatic selection of hyperparameters, performs well. Implementation of the method and simulations are done in MATLAB. In the spirit of Donoho's initiative for reproducible research, all m-files used in producing the tables and figures are available on request from the authors.

Acknowledgements: This work was done while the first author was visiting the Georgia Institute of Technology with support from Agenzia Spaziale Italiana. The work of the second author is supported in part by the NSF grant DMS 0004131 at GaTech.

FUNCTION	n	SNR=3	SNR=5	SNR=7	SNR=10
BLOCKS	256	0.4309	0.4211	0.4142	0.4100
	512	0.2944	0.2977	0.2825	0.2755
	1024	0.2324	0.2147	0.2034	0.1968
	2048	0.1398	0.1369	0.1299	0.1206
	4096	0.0812	0.0769	0.0699	0.0657
BUMPS	256	0.5141	0.5597	0.5747	0.5824
	512	0.4267	0.4308	0.4331	0.4354
	1024	0.2706	0.2889	0.2961	0.3066
	2048	0.1673	0.1883	0.1917	0.1924
	4096	0.0952	0.1030	0.1097	0.1162
DOPPLER	256	0.2885	0.2925	0.3208	0.3332
	512	0.1726	0.1874	0.1840	0.2020
	1024	0.0969	0.1098	0.1185	0.1322
	2048	0.0566	0.0659	0.0694	0.0702
	4096	0.0322	0.0337	0.0345	0.0371
HEAVISINE	256	0.1231	0.1777	0.2241	0.2675
	512	0.0778	0.1139	0.1425	0.1745
	1024	0.0480	0.0664	0.0860	0.1063
	2048	0.0301	0.0423	0.0519	0.0626
	4096	0.0190	0.0262	0.0311	0.0363

Table 2: AMSE obtained with 1000 simulation runs for a variety of sample sizes and SNR's.

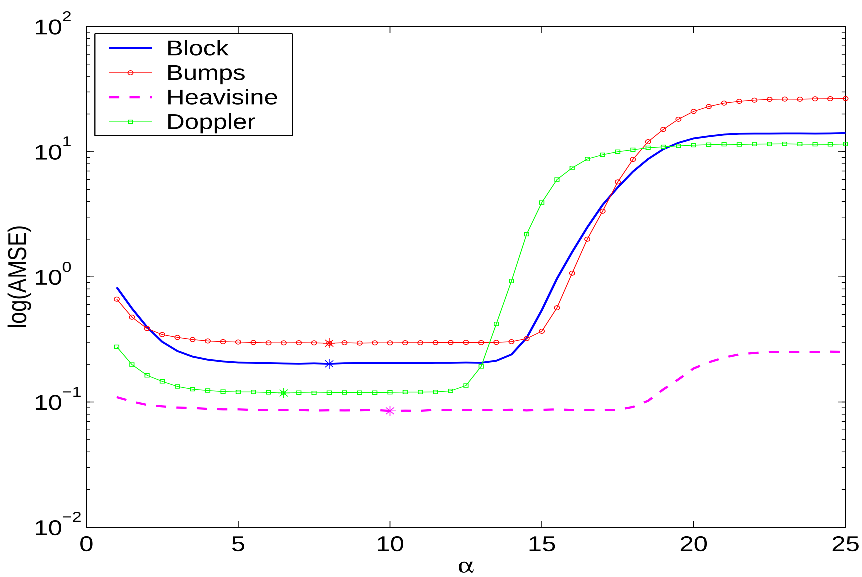

Figure 1: Plot of $\log_{10}(AMSE)$, obtained from $M = 1000$ simulational runs, with parameter α in the interval $(1, 25)$ with step 0.5.

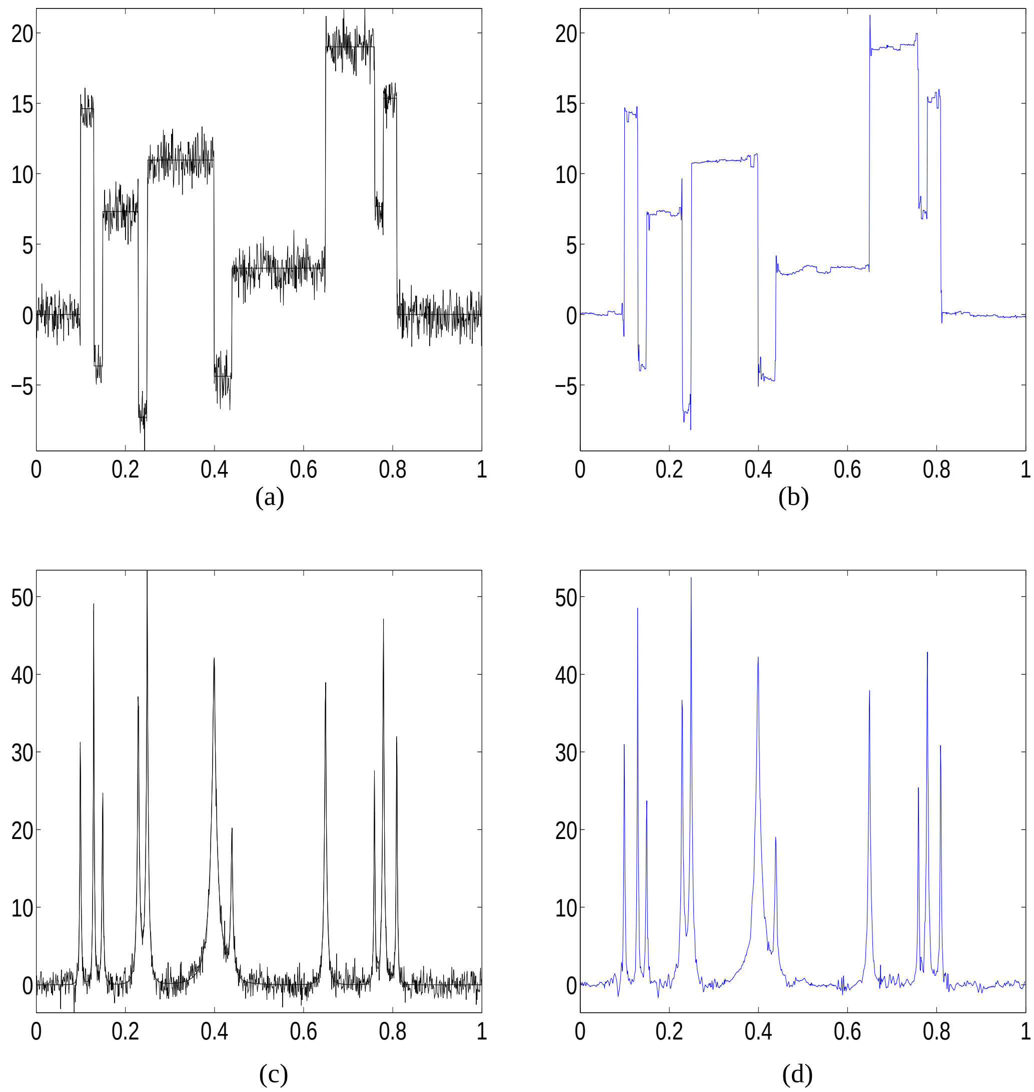

Figure 2: (a) A noisy block signal [SNR=7, n=1024, noise variance=1]; (b) Signal reconstructed using block size $l = 2$; (c) A noisy bumps signal [SNR=7, n=1024, noise variance=1]; (d) Signal reconstructed.

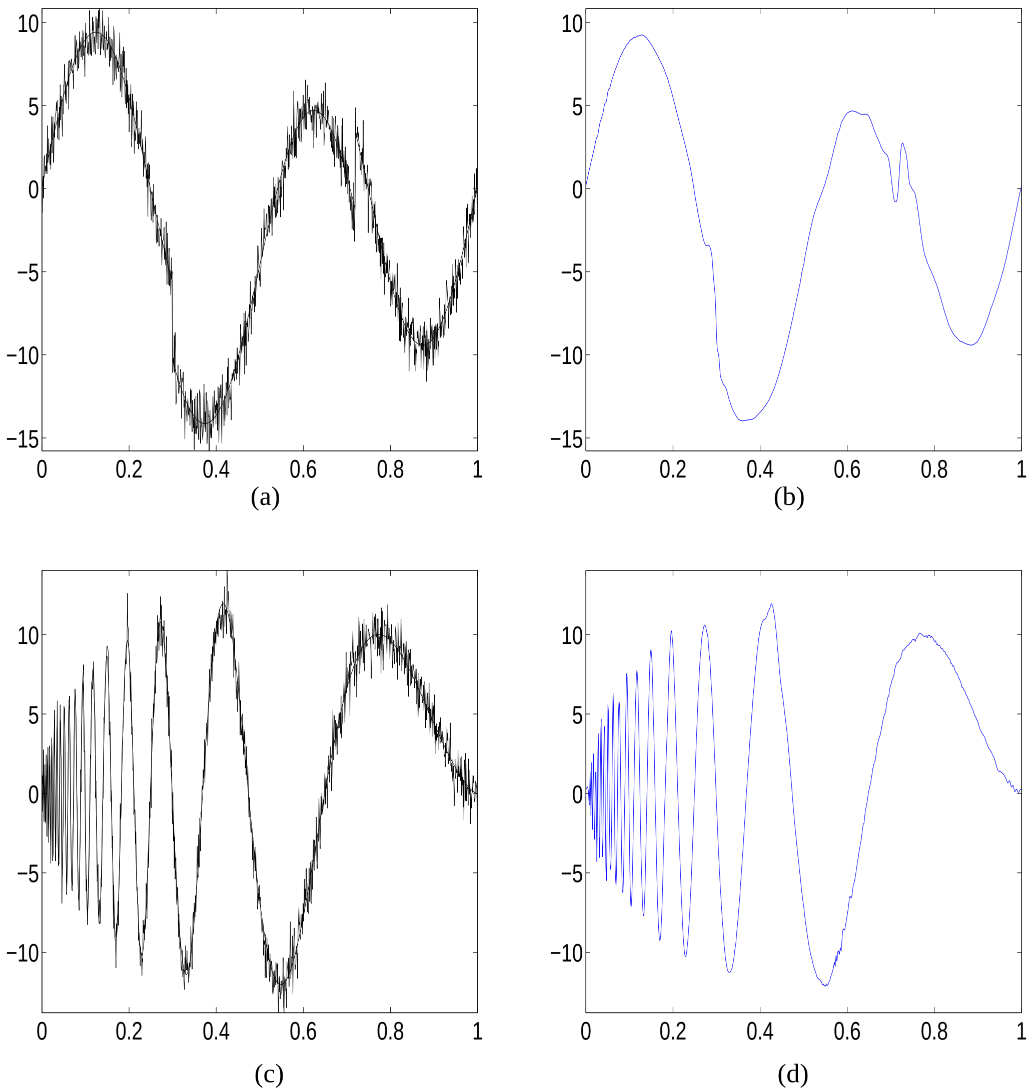

Figure 3: (a) A noisy heavisine signal [SNR=7, n=1024, noise variance=1]; (b) Signal reconstructed; (c) A noisy doppler signal [SNR=7, n=1024, noise variance=1]; (d) Signal reconstructed.

References

ABRAMOVICH, F., BESBEAS, P. and SAPATINAS, T (2000). Empirical Bayes Approach to Block Wavelet Function Estimation, *Technical report*, Department of Mathematics and Statistics, University of Cyprus, Cyprus.

ABRAMOVICH, F., SAPATINAS, T and SILVERMAN, B.W. (1998). Wavelet Thresholding via a Bayesian Approach, *J. Roy. Statist. Soc.*, B,**60**, 725–749.

ANTONIADIS, A. (1997). Wavelet in Statistics: a review (with discussion), *J. Ital. Statist. Soc.*, **6**, 97-144.

ANTONIADIS, A., BIGOT, J. and SAPATINAS, T. (2001). Wavelet Estimators in Nonparametric Regression: Description and Simulative Comparison. Technical report IMAG, France. `http://www-lmc.imag.fr/SMS/software/wavden/wavden.pdf.zip.`

BERGER, J.O. (1985). Statistical decision Theory and Bayesian Analysis, *Springer Series in Statistics*, Springer-Verlag New York Inc.

CAI, T.T., (1999a). Adaptive Wavelet Estimator: a block thresholding and oracle inequality approach, *Annals of Statistics*, **27**, n.3, 898–924.

CAI, T.T., (1999b). On Block Thresholding in Wavelet Regression: adaptivity, block size and threshold level, *Thec. Repor*, available http://www-stat.wharton.upenn.edu/ tcai/

CAI, T.T. and SILVERMAN, B.W., (2000). Incorporating information on neighboring coefficients into wavelet estimation, *Technical Report*, Department of Statistics, Purdue University,USA.

CHIPMAN, H. A., KOLACZYK, E. D. and MCCULLOCH, R. E. (1997). Adaptive Byesian Wavelet Shrinkage, *Journal of American Statistical Association*, **92**, n.440 1413–1421.

CLYDE, M., PARMIGIANI, G and VIDAKOVIC, B. (1998). Multiple Shrinkage and Subset Selection in Wavelets, *Biometrica*, **85**, 391–401.

DE CANDITIIS, D. (2001). Posteriori Confidence Intervals for the “Approximated” Wavelet Estimator. Technical Report 220/2001, IAM.

DONOHO, D. and JOHNSTONE, I. (1994). Ideal spatial Adaptation by Wavelet Shrinkage. *Biometrica*, **81**, 425–455.

DONOHO, D., JOHNSTONE, I., KERKYACHARIAN, G and PICARD, D (1995). Wavelet Shrinkage: Asymptotia?. *Jour.Roy. Stat. Soc.*, B, **57**, 301–369.

EFROMOVICH, S.(2000). Sharp linear and block shrinkage wavelet estimation, *Statist. probab. Lett.*, **49**, 323–329.

MÜLLER, P. and VIDAKOVIC, B. (Eds.) (1999). Bayesian Inference in Wavelet Based Models. *Lec. Notes Statistics*, **141**, New York: Springer-Verlag.

O’HAGAN, A (1994). Bayesian Inference, *Kendall’s Advanced Theory of Statistics*, **2B**.

HALL, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using Kernel and Wavelet methods, *Annals of Statistics*, **26**, 922–942.

HALL, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimator, *Statist. Sinica*, **9**, 33–50.

HALL, P., PENEV, S., Kerkyacharian, G. and Picard, D. (1997). Numerical performance of block thresholded wavelet estimators, *Statist. Comput.*, **7**, 115–124.

PERCIVAL, D.B. and WALDEN, A.T. (2000). Wavelet Methods for Time Series Analysis, *Cambridge:* Cambridge University Press.

VANNUCCI, M. and CORRADI, F. (1999). Modeling Dependence in the Wavelet Domain. In *Bayesian Inference in Wavelet Based Models.* Lecture Notes in Statistics **141**, Mller, P. and Vidakovic, B. (Eds). New York: Springer-Verlag.

VIDAKOVIC, B (1998). Nonlinear Wavelet Shrinkage with Bayes rules and Bayes factors, *J. Am. Statist. Ass.*, **93**, 173–179.

VIDAKOVIC, B (1999). Statistical Modeling by Wavelets, *Wiley Series in probability and Statistics* , John Wiley & Sons, INC.

VIDAKOVIC, B and MÜLLER, P. (19995). Wavelet Shrinkage with Affine Bayes Rules with Applications, Discussion Papers, 34-95, Institute of Statistics and Decision Sciences, Duke University, `http://www.isds.duke.edu/papers/working-papers-95.html`

VIDAKOVIC, B. AND RUGGERI, F. (2001). BAMS Methods: Theory and Simulations, In Print: *Sankhya: The Indian Journal of Statistics. Special Iusse on Wavelet Methods.*