

# Bayesian Modeling in the Wavelet Domain

Fabrizio Ruggeri and Brani Vidakovic

## ABSTRACT

Wavelets are the building blocks of wavelet transforms the same way that the functions  $e^{inx}$  are the building blocks of the ordinary Fourier transform. But in contrast to sines and cosines, wavelets can be (or almost can be) supported on an arbitrarily small closed interval. This feature makes wavelets a very powerful tool in dealing with phenomena that change rapidly in time. In many statistical applications, there is a need for procedures to (i) adapt to data and (ii) use prior information. The interface of wavelets and the Bayesian paradigm provides a natural terrain for both of these goals. In this chapter, the authors provide an overview of the current status of research involving Bayesian inference in wavelet nonparametric problems. Two applications, one in functional data analysis (FDA) and the second in geoscience are discussed in more detail.

## 1 Introduction

Wavelet-based tools are now indispensable in many areas of modern statistics, for example in regression, density and function estimation, factor analysis, modeling and forecasting of time series, functional data analysis, data mining and classification, with ranges of application areas in science and engineering. Wavelets owe their initial popularity in statistics to shrinkage, a simple and yet powerful procedure in nonparametric statistical modeling. It can be described by the following three steps: (i) observations are transformed into a set of wavelet coefficients; (ii) a shrinkage of the coefficients is performed; and (iii) the processed wavelet coefficients are back transformed to the domain of the original data.

Wavelet domains form desirable modeling environments; several supporting arguments are listed below.

Discrete wavelet transforms tend to “disbalance” the data. Even though the orthogonal transforms preserve the  $\ell_2$  norm of the data (the square root of sum of squares of observations, or the “energy” in engineering terms), most of the  $\ell_2$  norm in the transformed data is concentrated in only a few wavelet coefficients. This concentration narrows the class of plausible models and facilitates the thresholding. The disbalancing property also yields a variety of criteria for the selection of best basis.

Wavelets, as building blocks in modeling, are localized well in both time and scale (frequency). Signals with rapid local changes (signals with discontinuities, cusps, sharp spikes, etc.) can be well represented with only a few wavelet coefficients. This parsimony does not, in general, hold for other standard orthonormal bases which may require many “compensating” coefficients to describe discontinuity artifacts or local bursts.

Heisenberg’s principle states that time-frequency models not be precise in the time and frequency domains simultaneously. Wavelets adaptively distribute the time-frequency precision by their innate nature. The economy of wavelet transforms can be attributed to their ability to confront the limitations of Heisenberg’s principle in a data-dependent manner.

An important feature of wavelet transforms is their whitening property. There is ample theoretical and empirical evidence that wavelet transforms simplify the dependence structure in the original signal. For example, it is possible, for any given stationary dependence in the input signal, to construct a biorthogonal wavelet basis such that the corresponding in the transform are uncorrelated (a wavelet counterpart of Karhunen-Loève transform). For a discussion and examples see Walter and Shen (2001).

We conclude this incomplete list of wavelet transform features by pointing out their sensitivity to self-similar data. The scaling laws are distinctive features of self-similar data. Such laws are clearly visible in the wavelet domain in the so called wavelet spectra, wavelet counterparts of the Fourier spectra.

More arguments can be provided: computational speed of the wavelet transform, easy incorporation of prior information about some features of the signal (smoothness, distribution of energy across scales), etc.

Basics on wavelets can be found in many texts, monographs, and papers at many different levels of exposition. The interested reader should consult monographs by Daubechies (1992), Ogden (1997), and Vidakovic (1999), and Walter and Shen (2001), among others. An introductory article is Vidakovic and Müller (1999).

With self-containedness of this chapter in mind, we provide a brief overview of the discrete wavelet transforms (DWT).

### *1.1 Discrete Wavelet Transforms and Wavelet Shrinkage*

Let  $\mathbf{y}$  be a data-vector of dimension (size)  $n$ , where  $n$  is a power of 2, say  $2^J$ . We assume that measurements  $\mathbf{y}$  belong to an interval and consider periodized wavelet bases. Generalizations to different sample sizes and general wavelet and wavelet-like transforms are straightforward.

Suppose that the vector  $\mathbf{y}$  is wavelet-transformed to a vector  $\mathbf{d}$ . This linear and orthogonal transform can be fully described by an  $n \times n$  orthogonal matrix  $\mathbf{W}$ . In practice, one performs the DWT without exhibiting the matrix  $\mathbf{W}$  explicitly, but by using fast filtering algorithms. The filtering procedures are based on so-called quadrature mirror filters which are

uniquely determined by the wavelet of choice and fast Mallat's algorithm (Mallat, 1989). The wavelet decomposition of the vector  $\mathbf{y}$  can be written as

$$\mathbf{d} = (H^\ell \mathbf{y}, GH^{\ell-1} \mathbf{y}, \dots, GH^2 \mathbf{y}, GH \mathbf{y}, G \mathbf{y}). \quad (1.1)$$

Note that in (1.1),  $\mathbf{d}$  has the same length as  $\mathbf{y}$  and  $\ell$  is any fixed number between 1 and  $J = \log_2 n$ . The operators  $G$  and  $H$  are defined coordinate-wise via

$$(Ha)_k = \sum_{m \in \mathbf{Z}} h_{m-2k} a_m, \text{ and } (Ga)_k = \sum_{m \in \mathbf{Z}} g_{m-2k} a_m, \quad k \in \mathbf{Z}$$

where  $g$  and  $h$  are high- and low-pass filters corresponding to the wavelet of choice. Components of  $g$  and  $h$  are connected via the *quadrature mirror* relationship  $g_n = (-1)^n h_{1-n}$ . For all commonly used wavelet bases, the taps of filters  $g$  and  $h$  are readily available in the literature or in standard wavelet software packages.

The elements of  $\mathbf{d}$  are called “wavelet coefficients.” The sub-vectors described in (1.1) correspond to detail levels in a levelwise organized decomposition. For instance, the vector  $G \mathbf{y}$  contains  $n/2$  coefficients representing the level of the finest detail. When  $\ell = J$ , the vectors  $GH^{J-1} \mathbf{y} = \{d_{00}\}$  and  $H^J \mathbf{y} = \{c_{00}\}$  contain a single coefficient each and represent the coarsest possible level of detail and the smooth part in wavelet decomposition, respectively.

In general,  $j$ th detail level in the wavelet decomposition of  $\mathbf{y}$  contains  $2^j$  elements, and is given as

$$GH^{J-j-1} \mathbf{y} = (d_{j,0}, d_{j,1}, \dots, d_{j,2^j-1}). \quad (1.2)$$

Wavelet shrinkage methodology consists of shrinking wavelet coefficients. The simplest wavelet shrinkage technique is thresholding. The components of  $\mathbf{d}$  are replaced by 0 if their absolute value is smaller than a fixed threshold  $\lambda$ .

The two most common thresholding policies are **hard** and **soft** thresholding with corresponding rules given by:

$$\begin{aligned} \theta^h(d, \lambda) &= d \mathbf{1}(|d| > \lambda), \\ \theta^s(d, \lambda) &= (d - \text{sign}(d)\lambda) \mathbf{1}(|d| > \lambda), \end{aligned}$$

where  $\mathbf{1}(A)$  is the indicator of relation  $A$ , i.e.,  $\mathbf{1}(A) = 1$  if  $A$  is true and  $\mathbf{1}(A) = 0$  if  $A$  is false.

In the next section we describe how the Bayes rules, resulting from the models on wavelet coefficients, can act as shrinkage/thresholding rules.

## 2 Bayes and Wavelets

Bayesian paradigm has become very popular in wavelet data processing since Bayes rules are shrinkers. This is true in general, although examples

of Bayes rules that expand can be constructed, see Vidakovic and Ruggeri (1999). The Bayes rules can be constructed to mimic the thresholding rules: to slightly shrink the large coefficients and heavily shrink the small coefficients. Bayes shrinkage rules result from realistic statistical models on wavelet coefficients and such models allow for incorporation of prior information about the *true* signal. Furthermore, most Practicable Bayes rules should be easily computed by simulation or expressed in a closed form. Reviews on early Bayesian approaches can be found in Abramovich, Bailey, and Sapatinas (2000) and Vidakovic (1998b, 1999). An edited volume on Bayesian modeling in the wavelet domain was edited by Muller and Vidakovic and appeared more than 5 years ago (Müller and Vidakovic, 1999c).

One of the tasks in which the wavelets are typically applied is recovery of an unknown signal  $\mathbf{f}$  affected by noise  $\boldsymbol{\epsilon}$ . Wavelet transforms  $\mathbf{W}$  are applied to noisy measurements  $y_i = f_i + \epsilon_i$ ,  $i = 1, \dots, n$ , or, in vector notation,  $\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}$ . The linearity of  $\mathbf{W}$  implies that the transformed vector  $\mathbf{d} = \mathbf{W}(\mathbf{y})$  is the sum of the transformed signal  $\boldsymbol{\theta} = \mathbf{W}(\mathbf{f})$  and the transformed noise  $\boldsymbol{\eta} = \mathbf{W}(\boldsymbol{\epsilon})$ . Furthermore, the orthogonality of  $\mathbf{W}$  and normality of the noise vector  $\boldsymbol{\epsilon}$  implies the noise vector  $\boldsymbol{\eta}$  is also normal.

Bayesian methods are applied in the wavelet domain, i.e. after the data have been transformed. The wavelet coefficients can be modeled in totality, as a single vector, or one by one, due to decorrelating property of wavelet transforms. Several authors consider modeling blocks of wavelet coefficients, as an intermediate solution (e.g., Abramovich, Besbeas and Sapatinas, 2002; De Canditiis and Vidakovic, 2004).

We concentrate on typical wavelet coefficient and model:  $d = \theta + \epsilon$ . Bayesian methods are applied to estimate the location parameter  $\theta$ , which will be, in the sequel, retained as the shrunk wavelet coefficient and back transformed to the data domain. Various Bayesian models have been proposed. Some models have been driven by empirical justifications, others by pure mathematical considerations; some models lead to simple, close-form rules the other require extensive Markov Chain Monte Carlo simulations to produce the estimate.

### 2.1 *An Illustrative Example*

As an illustration of the Bayesian approach we present BAMS (Bayesian Adaptive Multiresolution Shrinkage). The method, due to Vidakovic and Ruggeri (2001), is motivated by empirical considerations on the coefficients and leads to easily implementable Bayes estimates, available in closed form.

The BAMS originates from the observation that a realistic Bayes model should produce prior predictive distributions of the observations which “agree” with the observations. Other authors were previously interested in the empirical distribution of the wavelet coefficients; see, for example, Leporini and Pesquet (1998, 1999), Mallat (1989), Ruggeri (1999), Simon-

celli (1999), and Vidakovic (1998b). Quoting Vidakovic and Ruggeri (2001), their common argument can be summarized by the following statement:

For most of the signals and images encountered in practice, the empirical distribution of a typical detail wavelet coefficient is notably centered about zero and peaked at it.

In the spirit of this statement, Mallat (1989) suggested an empirically justified model for typical wavelet coefficients, the exponential power distribution

$$f(d) = C \cdot e^{-(|d|/\alpha)^\beta}, \quad \alpha, \beta > 0,$$

with  $C = \frac{\beta}{2\alpha\Gamma(1/\beta)}$ .

Following the Bayesian paradigm, prior distributions should be elicited on the parameters of the model  $d|\theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2)$  and Bayesian estimators (namely, posterior means under squared loss functions) computed. In BAMS, prior on  $\sigma^2$  is sought so that the marginal likelihood of the wavelet coefficient is a double exponential distribution  $\mathcal{DE}$ . The double exponential distribution can be obtained as scale mixture of normals with exponential as the mixing distribution. The choice of the exponential prior is additionally justified by its *maxent* property, i.e, the exponential distribution is the entropy maximizer in the class of all distributions supported on  $(0, \infty)$  with a fixed first moment.

Thus, BAMS uses the exponential prior  $\sigma^2 \sim \mathcal{E}(\mu)$ ,  $\mu > 0$ , that leads to the marginal likelihood

$$d|\theta \sim \mathcal{DE} \left( \theta, \frac{1}{\sqrt{2\mu}} \right), \quad \text{with density } f(d|\theta) = \frac{1}{2} \sqrt{2\mu} e^{-\sqrt{2\mu}|d-\theta|}.$$

The next step is elicitation of a prior on  $\theta$ . Vidakovic (1988b) considered the previous model and proposed  $t$  distribution as a prior on  $\theta$ . He found the Bayes rules with respect to the squared error loss exhibit desirable shape. In personal communication with the second author, Jim Berger and Peter Müller suggested in 1993 the use of  $\epsilon$ -contamination priors in the wavelet context pointing out that such priors would lead to rules which are smooth approximations to a thresholding.

The choice

$$\pi(\theta) = \epsilon\delta(0) + (1 - \epsilon)\xi(\theta) \tag{1.3}$$

is now standard, and reflects prior belief that some locations (corresponding to the signal or function to be estimated) are 0 and that there is non-zero spread component  $\xi$  describing “large” locations. In addition to this prior sparsity of the signal part, this prior leads to desirable shapes of their resulting Bayes rules.

In BAMS, the spread part  $\xi$  is chosen as

$$\theta \sim \mathcal{DE}(0, \tau),$$

for which the predictive distribution for  $d$  is

$$m_\xi(d) = \frac{\tau e^{-|d|/\tau} - \frac{1}{\sqrt{2\mu}} e^{-\sqrt{2\mu}|d|}}{2\tau^2 - 1/\mu}.$$

The Bayes rule with respect to the prior  $\xi$ , under the squared error loss, is

$$\delta_\xi(d) = \frac{\tau(\tau^2 - 1/(2\mu))de^{-|d|/\tau} + \tau^2(e^{-|d|\sqrt{2\mu}} - e^{-|d|/\tau})/\mu}{(\tau^2 - 1/(2\mu))(\tau e^{-|d|/\tau} - (1/\sqrt{2\mu})e^{-|d|\sqrt{2\mu}})}. \quad (1.4)$$

When

$$\pi(\theta) = \epsilon\delta_0 + (1 - \epsilon)\mathcal{DE}(0, \tau), \quad (1.5)$$

we use the marginal and rule under  $\xi$  to express the marginal and rule under the prior  $\pi$ . The marginal under  $\pi$  is

$$m_\pi(d) = \epsilon\mathcal{DE}\left(0, \frac{1}{\sqrt{2\mu}}\right) + (1 - \epsilon)m_\xi(d)$$

while the Bayes rule is

$$\delta_\pi(d) = \frac{(1 - \epsilon) m_\xi(d) \delta_\xi(d)}{(1 - \epsilon) m_\xi(d) + \epsilon \mathcal{DE}\left(0, \frac{1}{\sqrt{2\mu}}\right)}. \quad (1.6)$$

The rule (1.6) is the BAMS rule. As shown in Figure 1, the BAMS rule falls between comparable hard and soft thresholding rules.

Tuning of hyperparameters is an important implementational issue and it is thoroughly discussed in Vidakovic and Ruggeri (2001), who suggest an automatic selection based on the nature of the data.

## 2.2 Regression Problems

In the context of wavelet regression, we will discuss two early approaches in more detail. The first one is Adaptive Bayesian Wavelet Shrinkage (ABWS) proposed by Chipman, Kolaczyk, and McCulloch (1997). Their approach is based on the stochastic search variable selection (SSVS) model proposed by George and McCulloch (1997), with the assumption that  $\sigma$  is known.

The likelihood in ABWS is

$$[d|\theta] \sim \mathcal{N}(\theta, \sigma^2).$$

The prior on  $\theta$  is defined as a mixture of two normals (Figure 2(a))

$$[\theta|\gamma_j] \sim \gamma_j\mathcal{N}(0, (c_j\tau_j)^2) + (1 - \gamma_j)\mathcal{N}(0, \tau_j^2),$$

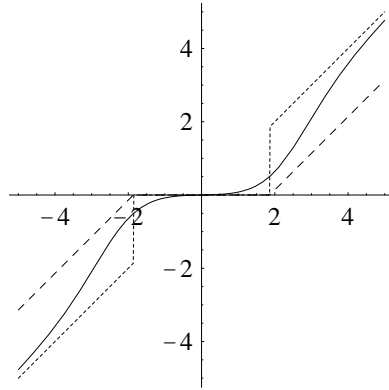


FIGURE 1. BAMS rule (1.6) and comparable hard and soft thresholding rules.

where

$$[\gamma_j] \sim \text{Ber}(p_j).$$

Because the hyperparameters  $p_j$ ,  $c_j$ , and  $\tau_j$  depend on the level  $j$  to which the corresponding  $\theta$  (or  $d$ ) belongs, and can be level-wise different, the method is date-adaptive.

The Bayes rule under squared error loss for  $\theta$  (from the level  $j$ ) has an explicit form,

$$\delta(d) = \left[ P(\gamma_j = 1|d) \frac{(c_j \tau_j)^2}{\sigma^2 + (c_j \tau_j)^2} + P(\gamma_j = 0|d) \frac{\tau_j^2}{\sigma^2 + \tau_j^2} \right] d, \quad (1.7)$$

where

$$P(\gamma_j = 1|d) = \frac{p_j \pi(d|\gamma_j = 1)}{(1 - p_j) \pi(d|\gamma_j = 0)}$$

and

$$\pi(d|\gamma_j = 1) \sim \mathcal{N}(0, \sigma^2 + (c_j \tau_j)^2) \quad \text{and} \quad \pi(d|\gamma_j = 0) \sim \mathcal{N}(0, \sigma^2 + \tau_j^2).$$

The shrinkage rule (1.7, Figure 2 (b)) can be interpreted as a smooth interpolation between two linear shrinkages with slopes  $\tau_j^2/(\sigma^2 + \tau_j^2)$  and  $(c_j \tau_j)^2/(\sigma^2 + (c_j \tau_j)^2)$ . The authors utilize empirical Bayes argument for tuning the hyperparameters level-wise. We note that most popular way to specify hyperparameters in Bayesian shrinkage is via empirical Bayes, see for example Abramovich, Besbeas, and Sapatinas (2002), Angelini and Sapatinas (2004), Clyde and George (1999, 2000), and Huang and Cressie (1999). De Canditiis and Vidakovic (2004) extend the ABWS method to multivariate case and unknown  $\sigma^2$  using a mixture of normal-inverse Gamma priors.

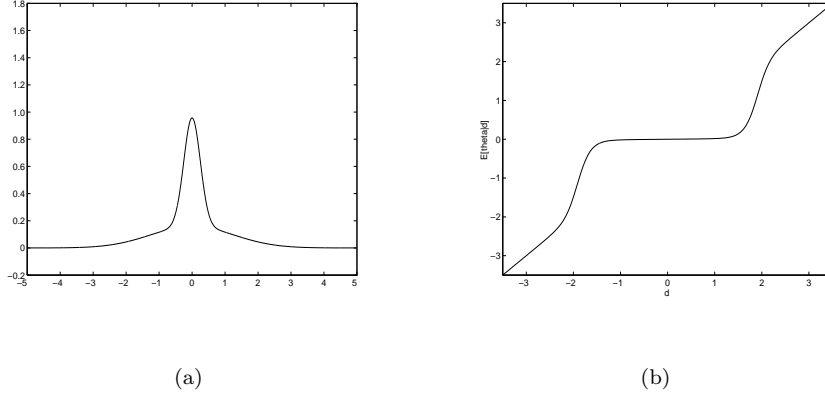


FIGURE 2. (a) Prior on  $\theta$  as a mixture of two normal distributions with different variances; (b) Bayes rule (1.7) in Chipman, Kolaczyk, and McCulloch (1997).

The approach used by Clyde, Parmigiani, and Vidakovic (1998) is based on a limiting form of the conjugate SSVS prior in George and McCulloch (1997). A similar model was used before in Müller and Vidakovic (1994) but in the context of density estimation.

Clyde, DeSimone, and Parmigiani (1996) consider a prior for  $\theta$  which is a mixture of a point mass at 0 if the variable is excluded from the wavelet regression and a normal distribution if it is included,

$$[\theta|\gamma_j, \sigma^2] \sim \mathcal{N}(0, (1 - \gamma_j) + \gamma_j c_j \sigma^2).$$

The  $\gamma_j$  are indicator variables that specify which basis element or column of  $W$  should be selected. As before, the subscript  $j$  points to the level to which  $\theta$  belongs. The set of all possible vectors  $\gamma$ 's will be referred to as the subset space. The prior distribution for  $\sigma^2$  is inverse  $\chi^2$ , i.e.,

$$[\lambda\nu/\sigma^2] \sim \chi_\nu^2,$$

where  $\lambda$  and  $\nu$  are fixed hyperparameters and the  $\gamma_j$ 's are independently distributed as Bernoulli  $Ber(p_j)$  random variables.

The posterior mean of  $\theta|\gamma$  is

$$E(\theta|\mathbf{d}, \gamma) = \Gamma(I_n + C^{-1})^{-1} \mathbf{d}, \quad (1.8)$$

where  $\Gamma$  and  $C$  are diagonal matrices with  $\gamma_{jk}$  and  $c_{jk}$ , respectively, on the diagonal and 0 elsewhere. For a particular subset determined by the ones in  $\gamma$ , (1.8) corresponds to thresholding with linear shrinkage.

The posterior mean is obtained by averaging over all models. Model averaging leads to a multiple shrinkage estimator of  $\theta$ :

$$E(\theta|\mathbf{d}) = \Sigma_\gamma \pi(\gamma|\mathbf{d}) \Gamma(I_n + C^{-1})^{-1} \mathbf{d},$$

where  $\pi(\gamma|\mathbf{d})$  is the posterior probability of a particular subset  $\gamma$ .

An additional nonlinear shrinkage of the coefficients to 0 results from the uncertainty in which subsets should be selected.

Calculating the posterior probabilities of  $\gamma$  and the mixture estimates for the posterior mean of  $\boldsymbol{\theta}$  above involve sums over all  $2^n$  values of  $\gamma$ . The calculational complexity of the mixing is prohibitive even for problems of moderate size, and either approximations or stochastic methods for selecting subsets  $\gamma$  possessing high posterior probability must be used.

In the orthogonal case, Clyde, DeSimone, and Parmigiani (1996) obtain an approximation to the posterior probability of  $\gamma$  which is adapted to the wavelet setting in Clyde, Parmigiani, and Vidakovic (1998). The approximation can be achieved by either conditioning on  $\sigma$  (plug-in approach) or by assuming independence of the elements in  $\gamma$ .

The approximate model probabilities, for the conditional case, are functions of the data through the regression sum of squares and are given by:

$$\begin{aligned}\pi(\gamma|\mathbf{d}) &\approx \prod_{j,k} \rho_{jk}^{\gamma_{jk}} (1 - \rho_{jk})^{1-\gamma_{jk}} \\ \rho_{jk}(\mathbf{d}, \sigma) &= \frac{a_{jk}(\mathbf{d}, \sigma)}{1 + a_{jk}(\mathbf{d}, \sigma)},\end{aligned}$$

where

$$\begin{aligned}a_{jk}(\mathbf{d}, \sigma) &= \frac{p_{jk}}{1 - p_{jk}} (1 + c_{jk})^{-1/2} \cdot \exp \left\{ \frac{1}{2} \frac{S_{jk}^2}{\sigma^2} \right\} \\ S_{jk}^2 &= d_{jk}^2 / (1 + c_{jk}^{-1}).\end{aligned}$$

The  $p_{jk}$  can be used to obtain a direct approximation to the multiple shrinkage Bayes rule. The independence assumption leads to more involved formulas. Thus, the posterior mean for  $\theta_{jk}$  is approximately

$$\rho_{jk}(1 + c_{jk}^{-1})^{-1} d_{jk}. \quad (1.9)$$

Equation (1.9) can be viewed as a level dependent wavelet shrinkage rule, generating a variety of nonlinear rules. Depending on the choice of prior hyperparameters, shrinkage may be monotonic, if there are no level dependent hyperparameters, or non-monotonic; see Figure 3 (a).

Clyde and George (1999, 2000) propose a model in which the distributions for the error  $\epsilon$  and  $\theta$  are scale mixtures of normals, thus justifying Mallat's paradigm (Mallat, 1989). They use an empirical Bayes approach to estimate the prior hyperparameters, and provide analytic expressions for the shrinkage estimator based on Bayesian model averaging. They report an excellent denoising performance of their shrinkage method for a range of noise distributions.

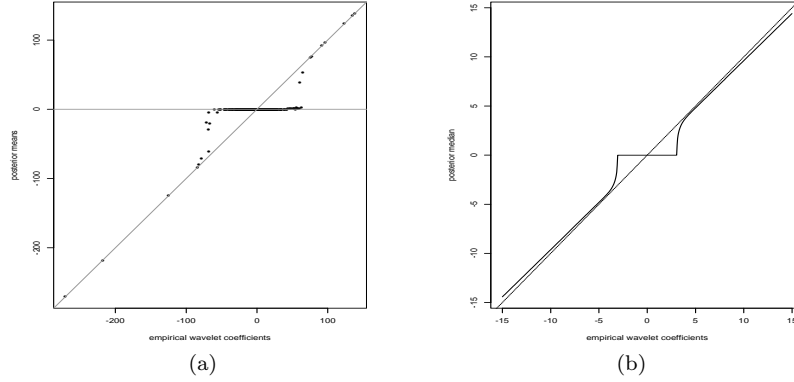


FIGURE 3. (a) Shrinkage rule from Clyde, Parmigiani, and Vidakovic (1998) based on independence approximation (1.9); (b) Posterior median thresholding rule (1.11) from Abramovich, Sapatinas, and Silverman (1998).

### 2.3 Bayesian Thresholding Rules

Bayes rules under the squared error loss and regular models are never thresholding rules. We discuss two possible approaches for obtaining *bona fide* thresholding rules in a Bayesian manner. The first one is via hypothesis testing, while the second one uses weighted absolute error loss.

Donoho and Johnstone (1994, 1995) gave a heuristic for the selection of the universal threshold via rejection regions of suitable hypotheses tests. Testing a precise hypothesis in Bayesian fashion requires a prior which has a point mass component. A method based on Bayes factors is discussed first. For details, see Vidakovic (1998a).

Let

$$[d|\theta] \sim f(d|\theta).$$

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

$$[\theta] \sim \pi(\theta) = \pi_0 \delta_0 + \pi_1 \xi(\theta), \quad (1.10)$$

where  $\pi_0 + \pi_1 = 1$ ,  $\delta_0$  is a point mass at 0, and  $\xi(\theta)$  is a prior that describes distribution of  $\theta$  when  $H_0$  is false.

The resulting Bayesian procedure is:

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

where

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

is the posterior probability of the  $H_0$  hypothesis, and  $B = \frac{f(d|0)}{\int_{\theta \neq 0} f(d|\theta)\xi(\theta)d\theta}$  is the Bayes factor in favor of  $H_0$ . The optimality of Bayes Factor shrinkage was recently explored by Pensky and Sapatinas (2004) and Abramovich, Amato, and Angelini (2004). They show that Bayes Factor shrinkage rule is optimal for wide range smoothness spaces and can outperform the posterior mean and the posterior median.

Abramovich, Sapatinas, and Silverman (1998) use weighted absolute error loss and show that for a prior on  $\theta$

$$[\theta] \sim \pi_j \mathcal{N}(0, \tau_j^2) + (1 - \pi_j)\delta(0)$$

and normal  $\mathcal{N}(\theta, \sigma^2)$  likelihood, the posterior median is

$$\text{Med}(\theta|d) = \text{sign}(d) \max(0, \zeta). \quad (1.11)$$

Here

$$\begin{aligned} \zeta &= \frac{\tau_j^2}{\sigma^2 + \tau_j^2} |d| - \frac{\tau_j \sigma}{\sqrt{\sigma^2 + \tau_j^2}} \Phi^{-1} \left( \frac{1 + \min(\omega, 1)}{2} \right), \text{ and} \\ \omega &= \frac{1 - \pi_j}{\pi_j} \frac{\sqrt{\tau_j^2 + \sigma^2}}{\sigma} \exp \left\{ -\frac{\tau_j^2 d^2}{2\sigma^2(\tau_j^2 + \sigma^2)} \right\}. \end{aligned}$$

The index  $j$ , as before, points to the level containing  $\theta$  (or  $d$ ). The plot of the thresholding function (1.11) is given in Figure 3 (b).

The authors compare the rule (1.11), they call BayesThresh, with several methods (Cross-Validation, False Discovery Rate, VisuShrink and Global-Sure) and report very good MSE performance.

#### 2.4 Bayesian Wavelet Methods in Functional Data Analysis

Recently wavelets have been used in functional data analysis as a useful tool for dimension reduction in the modeling of multiple curves. This has also led to important contributions in interdisciplinary fields, such as chemometrics, biology and nutrition.

Brown, Fearn, and Vannucci (2001) and Vannucci, Brown, and Fearn (2001) considered regression models that relate a multivariate response to functional predictors, applied wavelet transforms to the curves, and used Bayesian selection methods to identify features that best predict the responses. Their model in the data domain is

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\alpha}' + \mathbf{X} \mathbf{B} + \mathbf{E}, \quad (1.12)$$

where  $\mathbf{Y}(n \times q)$  are  $q$ -variate responses and  $\mathbf{X}(n \times p)$  the functional predictors data, each row of  $\mathbf{X}$  being a vector of observations of a curve at  $p$

equally spaced points. In the practical context considered by the authors, the responses are given by the composition (by weight) of the  $q = 4$  constituents of 40 biscuit doughs made with variations in quantities of fat, flour, sugar and water in a recipe. The functional predictors are near infrared spectral data measured at  $p = 700$  wavelengths (from  $1100nm$  to  $2498nm$  in steps of  $2nm$ ) for each dough piece. The goal is to use the spectral data to predict the composition. With  $n \ll p$  and  $p$  large, wavelet transforms are employed as an effective tool for dimension reduction that well preserves local features.

When a wavelet transform is applied to each row of  $\mathbf{X}$ , the model (1.12) becomes

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\alpha}' - \mathbf{Z} \tilde{\mathbf{B}} + \mathbf{E}$$

with  $\mathbf{Z} = \mathbf{X} \mathbf{W}'$  a matrix of wavelet coefficients and  $\tilde{\mathbf{B}} = \mathbf{W} \mathbf{B}$  the transformed regression coefficients. Shrinkage mixture priors are imposed on the regression coefficients. A latent vector with  $p$  binary entries serves to identify one of two types of regression coefficients, those close to zero and those not. The authors use results from Vannucci and Corradi (1999) to specify suitable prior covariance structures in the domain of the data that nicely transform to modified priors on the wavelet coefficients domain. Using a natural conjugate Gaussian framework, the marginal posterior distribution of the binary latent vector is derived. Fast algorithms aid its direct computation, and in high dimensions these are supplemented by a Markov Chain Monte Carlo approach to sample from the known posterior distribution. Predictions are then based on the selected coefficients. The authors investigate both model averaging strategies and single models predictions.

In Vannucci, Brown, and Fearn (2003) an alternative decision theoretic approach is investigated where variables have genuine costs and a single subset is sought. The formulation they adopt assumes a joint normal distribution of the  $q$ -variate response and the full set of  $p$  regressors. Prediction is done assuming quadratic losses with an additive cost penalty non-decreasing in the number of variables. Simulated annealing and genetic algorithms are used to maximize the expected utility.

Morris *et al.* (2003) extended wavelet regression to the nested functional framework. Their work was motivated by a case study investigating the effect of diet on *O*<sup>6</sup>-methylguanine-DNA-methyltransferase (MGMT), an important biomarker in early colon carcinogenesis. Specifically, two types of dietary fat (fish oil or corn oil) were investigated as potentially important factors that affect the initiation stage of carcinogenesis, i.e. the first few hours after the carcinogen exposure. In the experiment 30 rats were fed one of the 2 diets for 14 days, exposed to a carcinogen, then sacrificed at one of 5 times after exposure (0, 3, 6, 9, or 12 hours). Rat's colons were removed and dissected, and measurements of various biomarkers, including MGMT, were obtained. Each biomarker was measured on a set of 25 crypts in the distal and proximal regions of each rat's colon. Crypts are

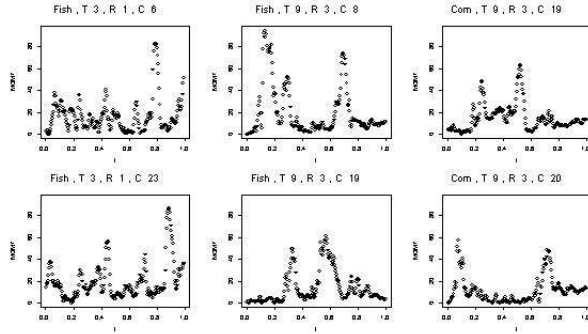


FIGURE 4. DNA repair enzyme for selected crypts.

fingerlike structures that extend into the colon wall. The procedure yielded observed curves for each crypt consisting of the biomarker quantification as a function of relative cell position within the crypt, the position being related to cell age and stage in the cell cycle. Due to the image processing used to quantify the measurements, these functions may be very irregular, with spikes presumably corresponding to regions of the crypt with high biomarker levels, see Figure 4.

The primary goal of the study was to determine whether diet has an effect on MGMT levels, and whether this effect depends on time and/or relative depth within the crypt. Another goal was to assess the relative variability between crypts and between rats. The authors model the curves in the data domain using a nonparametric hierarchical model of the type

$$\begin{aligned} Y_{abc} &= g_{abc}(t) + \epsilon_{abc}, \\ g_{abc}(t) &= g_{ab}(t) + \eta_{abc}(t), \\ g_{ab}(t) &= g_a(t) + \zeta_{ab}(t), \end{aligned}$$

with  $Y_{abc}$  the response vector for crypt  $c$ , rat  $b$  and treatment  $a$ ,  $g_a(t)$  the true crypt, rat or treatment profile,  $\epsilon_{abc}$  the measurement error and  $\eta_{abc}(t), \zeta_{ab}(t)$  the crypt/rat level error.

The wavelet-based Bayesian method suggested by the authors leads to adaptively regularized estimates and posterior credible intervals for the mean function and random effects functions, as well as the variance components of the model. The approach first applies DWT to each observed curve to obtain the corresponding wavelet coefficients. This step results

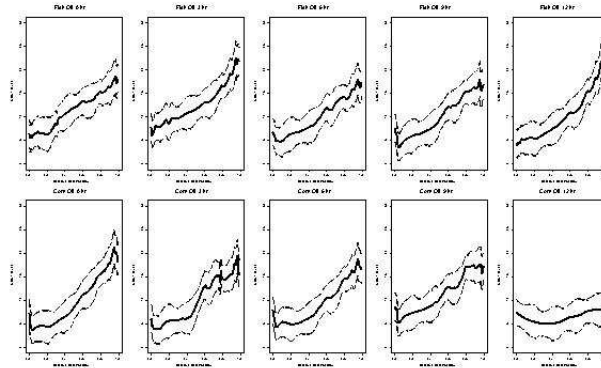


FIGURE 5. Estimated mean profiles by diet/time with 90% posterior bounds.

in the projection of the original curves into a transformed domain, where modeling can be done in a more parsimonious way. A Bayesian model is then fit to each wavelet coefficient across curves using an MCMC procedure to obtain posterior samples of the wavelet coefficients corresponding to the functions at each hierarchical level and the variance components. The inverse DWT is applied to transform the obtained estimates back to the data domain.

The Bayesian modeling adopted is such that the function estimates at all levels are adaptively regularized using a multiple shrinkage prior imposed at the top level of the hierarchy. The treatment level functions are directly regularized by this shrinkage, while the functions at the lower levels of the hierarchy are subject to some regularization induced by the higher levels, as modulated by the variance components. The authors provide guidelines for selecting these regularization parameters, together with empirical Bayes estimates, introduced during the rejoinder to the discussion.

Results from the analysis of the case study reveal that there is more MGMT expressed at the luminal surface of the crypt, and suggest a diet difference in the MGMT expression at this location 12 hours after exposure to the carcinogen, see Figure 5. Also, the multiresolution wavelet analysis highlights features present in the crypt-level profiles that may correspond to individual cells, suggesting the hypothesis that MGMT operates on a largely cell-by-cell basis.

## 2.5 The Density Estimation Problem

Donoho *et al.* (1996), Hall and Patil (1995), and Walter and Shen (2001), among others, applied wavelets in density estimation from a classical and data analytic perspective.

Chencov (1962) proposed projection type density estimators in terms of an arbitrary orthogonal basis. In the case of a wavelet basis, Chencov's estimator has the form

$$\hat{f}(x) = \sum_k c_{j_0 k} \phi_{j_0 k}(x) + \sum_{j_0 \leq j \leq j_1} \sum_k d_{jk} \psi_{jk}(x), \quad (1.13)$$

where the coefficients  $c_{jk}$  and  $d_{jk}$  constituting the vector  $d$  are defined via the standard empirical counterparts of  $\langle f, \phi_{jk} \rangle$  and  $\langle f, \psi_{jk} \rangle$ . Let  $X_1, \dots, X_n$  be a random sample from  $f$ . Then

$$c_{jk} = \frac{1}{n} \sum_{i=1}^n \phi_{jk}(X_i), \quad \text{and} \quad d_{jk} = \frac{1}{n} \sum_{i=1}^n \psi_{jk}(X_i). \quad (1.14)$$

Müller and Vidakovic (1999a, b) parameterize an unknown density  $f(x)$  by a wavelet series on its logarithm, and propose a prior model which explicitly defines geometrically decreasing prior probabilities for non-zero wavelet coefficients at higher levels of detail.

The unknown probability density function  $f(\cdot)$  is modeled by:

$$\log f(x) = \sum_{k \in Z} \xi_{j_0 k} \phi_{j_0, k}(x) + \sum_{j \geq j_0, k \in Z} s_{jk} \theta_{jk} \psi_{jk}(x) - \log K, \quad (1.15)$$

where  $K = \int f(x) dx$  is the normalization constant and  $s_{jk} \in \{0, 1\}$  is an indicator variable that performs model induced thresholding.

The dependence of  $f(x)$  on the vector  $\theta = (\xi_{j_0, k}, s_{jk}, \theta_{jk}, j = j_0, \dots, j_1, k \in Z)$  of wavelet coefficients and indicators is expressed by  $f(x) = p(x|\theta)$ . The sample  $X = \{X_1, \dots, X_n\}$  defines a likelihood function  $p(X|\theta) = \prod_{i=1}^n p(X_i|\theta)$ .

The model is completed by a prior probability distribution for  $\theta$ . Without loss of generality,  $j_0 = 0$  can be assumed. Also, any particular application will determine the finest level of detail  $j_1$ .

$$\begin{aligned} [\xi_{0k}] &\sim \mathcal{N}(0, \tau r_0), \\ [\theta_{jk} | s_{jk} = 1] &\sim \mathcal{N}(0, \tau r_j), \quad r_j = 2^{-j}, \\ [s_{jk}] &\sim \text{Bernoulli}(\alpha^j), \\ [\alpha] &\sim \text{Beta}(a, b), \\ [1/\tau] &\sim \text{Gamma}(a_\tau, b_\tau). \end{aligned} \quad (1.16)$$

The wavelet coefficients  $\theta_{jk}$  are non-zero with geometrically decreasing probabilities. Given that a coefficient is non-zero, it is generated from a normal distribution. The parameter vector  $\theta$  is augmented in order to include all model parameters, i.e.  $\theta = (\theta_{jk}, \xi_{jk}, s_{jk}, \alpha, \tau)$ .

The scale factor  $r_j$  contributes to the adaptivity of the method. Wavelet shrinkage is controlled by both: the factor  $r_j$  and geometrically decreasing prior probabilities for non-zero coefficient,  $\alpha^j$ .

The conditional prior  $p(\theta_{jk}|s_{jk} = 0) = h(\theta_{jk})$  is a pseudo-prior as discussed in Carlin and Chib (1995). The choice of  $h(\cdot)$  has no bearing on the inference about  $f(x)$ . In fact, the model could be alternatively formulated by dropping  $\theta_{jk}$  under  $s_{jk} = 0$ . However, this model would lead to a parameter space of varying dimension. Carlin and Chib (1995) argue that the pseudo-prior  $h(\theta_{jk})$  should be chosen to produce values for  $\theta_{jk}$  which are consistent with the data.

The particular MCMC simulation scheme used to estimate the model (1.15, 1.16) is described. Starting with some initial values for  $\theta_{jk}$ ,  $j = 0, \dots, j_1$ ,  $\xi_{00}$ ,  $\alpha$ , and  $\tau$ , the following Markov chain was implemented.

1. For each  $j = 0, \dots, j_1$  and  $k = 1, \dots, 2^j - 1$  go over the steps **2** and **3**.
2. Update  $s_{jk}$ . Let  $\theta_0$  and  $\theta_1$  indicate the current parameter vector  $\theta$  with  $s_{jk}$  replaced by 0 and 1, respectively. Compute  $p_0 = p(y|\theta_0) \cdot (1 - \alpha^j)h(\theta_{jk})$  and  $p_1 = p(y|\theta_0) \cdot \alpha^j p(\theta_{jk}|s_{jk} = 1)$ . With probability  $p_1/(p_0 + p_1)$  set  $s_{jk} = 1$ , else  $s_{jk} = 0$ .
- 3a. Update  $\theta_{jk}$ . If  $s_{jk} = 1$ , generate  $\tilde{\theta}_{jk} \sim g(\tilde{\theta}_{jk}|\theta_{jk})$ . Use, for example,  $g(\tilde{\theta}_{jk}|\theta_{jk}) = \mathcal{N}(\theta_{jk}, 0.25\sigma_{jk})$ , where  $\sigma_{jk}$  is some rough estimate of the posterior standard deviation of  $\theta_{jk}$ . We will discuss alternative choices for the probing distribution  $g(\cdot)$  below.  
Compute

$$a(\theta_{jk}, \tilde{\theta}_{jk}) = \min \left[ 1, \frac{p(y|\tilde{\theta})p(\tilde{\theta}_{jk})}{p(y|\theta)p(\theta_{jk})} \right],$$

where  $\tilde{\theta}$  is the parameter vector  $\theta$  with  $\theta_{jk}$  replaced by  $\tilde{\theta}_{jk}$ , and  $p(\theta_{jk})$  is the p.d.f. of the normal prior distribution given in (1.16).

With probability  $a(\theta_{jk}, \tilde{\theta}_{jk})$  replace  $\theta_{jk}$  by  $\tilde{\theta}_{jk}$ ; else keep  $\theta_{jk}$  unchanged.

- 3b. If  $s_{jk} = 0$ , generate  $\theta_{jk}$  from the full conditional posterior  $p(\theta_{jk}|\dots, X) = p(\theta_{jk}|s_{jk} = 0) = h(\theta_{jk})$ .
4. Update  $\xi_{00}$ . Generate  $\xi_{00} \sim g(\xi_{00}|\xi_{00})$ . Use, for example,  $g(\xi_{00}|\xi_{00}) = \mathcal{N}(\xi_{00}, 0.25\rho_{00})$ , where  $\rho_{00}$  is some rough estimate of the posterior standard deviation of  $\xi_{00}$ . Analogously to step **3a**, compute an acceptance probability  $a$  and replace  $\xi_{00}$  with probability  $a$ .
5. Update  $\alpha$ . Generate  $\tilde{\alpha} \sim g_\alpha(\tilde{\alpha}|\alpha)$  and compute

$$a(\alpha, \tilde{\alpha}) = \min \left[ 1, \frac{\prod_{jk} \tilde{\alpha}^{j s_{jk}} (1 - \tilde{\alpha}^j)^{s_{jk}}}{\prod_{jk} \alpha^{j s_{jk}} (1 - \alpha^j)^{s_{jk}}} \right].$$

With probability  $a(\alpha, \tilde{\alpha})$  replace  $\alpha$  by  $\tilde{\alpha}$ , else keep  $\alpha$  unchanged.

6. Update  $\tau$ . Resample  $\tau$  from the complete inverse Gamma conditional posterior.

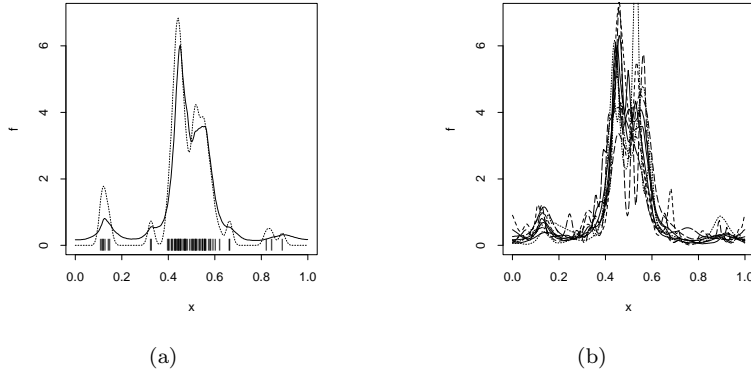


FIGURE 6. (a) The estimated p.d.f.  $\hat{f}(x) = \int p(x|\theta)dp(\theta|X)$ . The dotted line plots a conventional kernel density estimate for the same data. (b) The posterior distribution of the unknown density  $f(x) = p(x|\theta)$  induced by the posterior distribution  $p(\theta|X)$ . The lines plot  $p(x|\theta_i)$  for ten simulated draws posterior  $\theta_i \sim p(\theta|X), i = 1, \dots, 10$ .

**7.** Iterate over steps **1** through **6** until the chain is judged to have practically converged.

The algorithm implements a Metropolis chain changing one parameter at a time in the parameter vector. See, for example, Tierney (1994) for a description and discussion of Metropolis chains for posterior exploration. For a practical implementation,  $g$  should be chosen such that the acceptance probabilities  $a$  are neither close to zero, nor close to one. In the implementations,  $g(\hat{\theta}_{jk}|\theta_{jk}) = \mathcal{N}(\theta_{jk}, 0.25\sigma_{jk})$  with  $\sigma_{jk} = 2^{-j}$ , was used.

Described wavelet based density estimation model is illustrated on the galaxy data set (Roeder, 1990). The data is rescaled to the interval  $[0, 1]$ . The hyperparameters were fixed as  $a = 10$ ,  $b = 10$ , and  $a_\tau = b_\tau = 1$ . The  $Beta(10, 10)$  prior distribution on  $\alpha$  is reasonably non-informative compared to the likelihood based on  $n = 82$  observations.

Initially, all  $s_{jk}$  are set to one, and  $\alpha$  to its prior mean  $\alpha = 0.5$ . The first 10 iterations as burn-in period were discarded, then 1000 iterations of steps 1 through 6 were simulated. For each  $j, k$ , Step 3 was repeated three times. The maximum level of detail selected was  $j_1 = 5$ .

Figures 6 and 7 describe some aspects of the analysis.

## 2.6 An Application in Geoscience

Wikle *et al.* (2001) considered a problem common in the atmospheric and ocean sciences, in which there are two measurement systems for a given process, each of which is imperfect. Satellite-derived estimates of near-surface

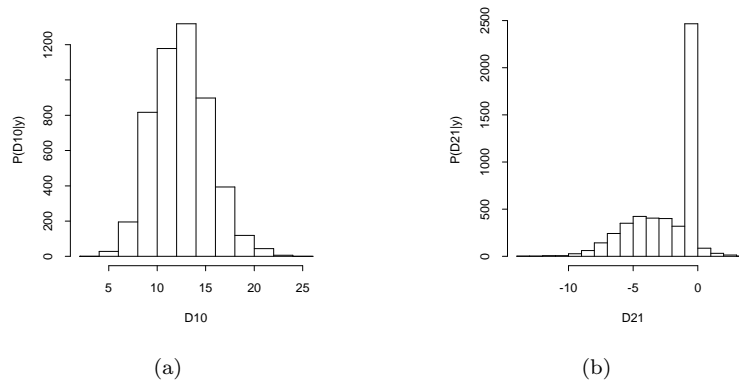


FIGURE 7. Posterior distributions  $p((s_{10}\theta_{10})|X)$  and  $p((s_{21}\theta_{21})|X)$ . While  $s_{10}\theta_{10}$  is non-zero with posterior probability close to one, the posterior distribution  $p((s_{21}\theta_{21})|X)$  is a mixture of a point mass at zero and a continuous part.

winds over the ocean are available, and although they have very high spatial resolution, their coverage is incomplete (e.g., top panel of Figure 8; the direction of the wind is represented by the direction of the arrow and the wind speed is proportional to the length of the arrow). In addition, wind fields from the major weather centers are produced through combinations of observations and deterministic weather models (so-called “analysis” winds). These winds provide complete coverage, but are have relatively low spatial resolution (e.g., bottom panel of Figure 8 shows such winds from the National Centers for Environmental Prediction (NCEP)). Wikle *et al.* (2001) were interested in predicting spatially distributed wind fields at intermediate spatial resolutions and regular time intervals (every six hours) over the tropical oceans. Thus, they sought to combine these data sets (over multiple time periods) in such a way as to incorporate the space-time dynamics inherent in the surface wind field. They utilized the fact that to a first approximation, tropical winds can be described by so-called “linear shallow-water” dynamics. In addition, previous studies (e.g., Wikle, Milliff, and Large, 1999) showed that tropical near surface winds exhibit turbulent scaling behavior in space. That is, the system can be modeled as a fractal self-similar process that is nonstationary in space, but when considered through certain spatially-invariant filters, appears stationary (i.e., a  $1/f$  process). In the case of tropical near-surface winds, the energy spectrum is proportional to the inverse of the spatial frequency taken to the  $5/3$  power (Wikle, Milliff, and Large, 1999). Thus, the model for the underlying wind process must consider the shallow-water dynamics and the spectral scaling relationship.

Wikle *et al.* (2001) considered a Bayesian hierarchical approach that considered models for the data conditioned on the desired wind spatio-temporal

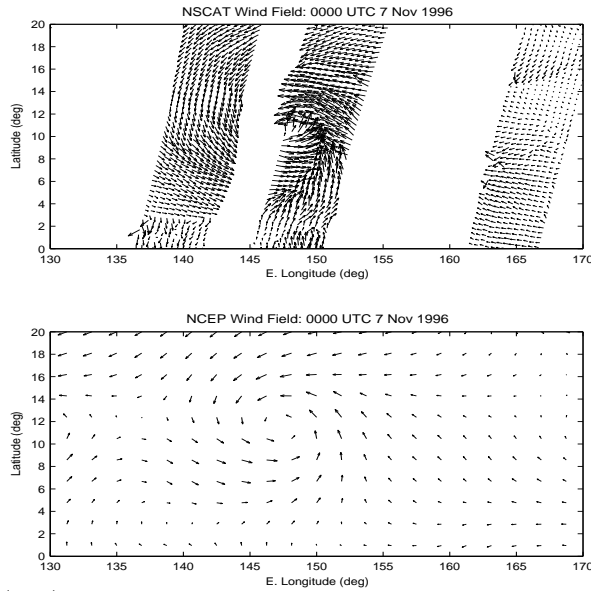


FIGURE 8. (Top) Incomplete satellite-derived estimates of near-surface winds over the ocean; (Bottom) The “analysis” winds from the National Centers for Environmental Prediction (NCEP).

wind process and parameters, models for the process given parameters and finally models for the parameters. The data models considered the change of support issues associated with the two disparate data sets. More critically, the spatio-temporal wind process was decomposed into three components, a spatial mean component representative of climatological winds over the area, a large-scale dynamical component representative of linear shallow-water dynamics, and a multiresolution (wavelet) component representative of medium to fine-scale processes in the atmosphere. The dynamical evolution models for these latter two components then utilized quite informative priors which made use of the aforementioned prior theoretical and empirical knowledge about the shallow-water and multiresolution processes. In this way, information from the two data sources, regardless of level of resolution, could impact future prediction times through the evolution equations. An example of the output from this model is shown in Figure 9.

The top panel shows the wind field from the NCEP weather center analysis winds (arrows) along with the implied convergence/divergence field (shaded area). More intense convergence (darker shading) often implies strong upward vertical motion and thus is suggestive of higher clouds and stronger precipitation development. This location and time corresponds to be a period when tropical cyclone “Dale” (centered near 146 E, 12 N) was the dominant weather feature. Note that the this panel does not show much definition to the storm, relative to panel 3 which shows a satellite

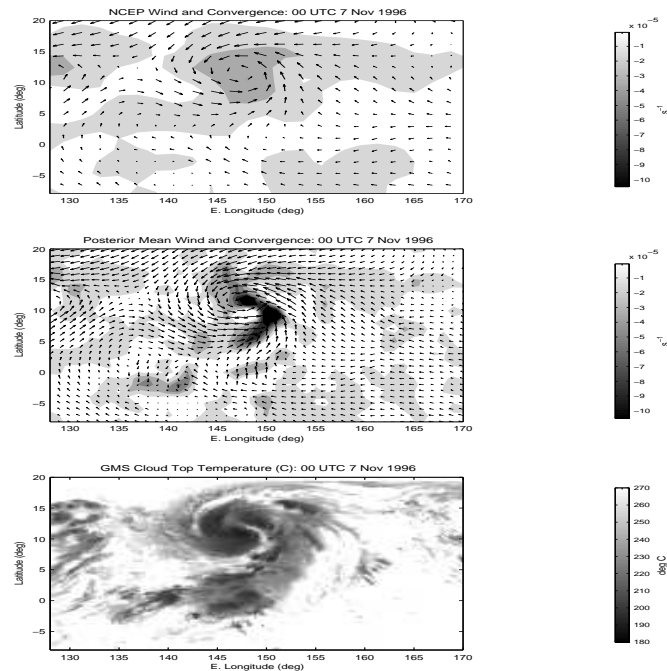


FIGURE 9. (Top) The wind field from the NCEP weather center analysis winds (arrows) along with the implied convergence/divergence field (shaded area). More intense convergence (darker shading) often implies strong upward vertical motion and thus is suggestive of higher clouds and stronger precipitation development; (Middle) The posterior mean wind and convergence fields from the model of Wikle *et al.* (2001); (Bottom) Satellite cloud image (cloud top temperature) for the same period (darker colors imply higher clouds and thus more intense precipitation).

cloud image (cloud top temperature) for the same period (darker colors imply higher clouds and thus more intense precipitation). The posterior mean wind and convergence fields from the model of Wikle *et al.* (2001) are shown in the middle panel of Figure 9. Note in this case that the convergence fields match up more closely with features in the satellite cloud image. In addition, the multiresolution (turbulent) nature of the cloud image is better represented in the posterior mean fields as well. Although not shown here, another advantage of this approach is that one can quantify the uncertainty in the predictions of the winds and derivative fields (see Wikle *et al.* 2001).

### 3 Other Problems

The field of Bayesian wavelet modeling deserves an extensive monograph and this chapter is only highlighting the field. In this section we add some more references on various Bayesian approaches in wavelet data processing.

Lina and MacGibbon (1997) apply a Bayesian approach to wavelet regression with complex valued Daubechies wavelets. To some extent, they exploit redundancy in the representation of real signals by the complex wavelet coefficients. Their shrinkage technique is based on the observation that the modulus and the phase of wavelet coefficients encompass very different information about the signal. A Bayesian shrinkage model is constructed for the modulus, taking into account the corresponding phase.

Simoncelli and Adelson (1996) discuss Bayes “coring” procedure in the context of image processing. The prior on the signal is Mallat’s model, see Mallat (1989), while the noise is assumed normal. They implement their noise reduction scheme on an oriented multiresolution representation - known as the *steerable pyramid*. They report that Bayesian coring outperforms classical Wiener filtering. See also Simoncelli (1999) and Portilla *et al.* (2003) for related research. A comprehensive comparison of Bayesian and non-Bayesian wavelet models applied to neuroimaging can be found in Fadili and Bullmore (2004).

Crouse, Nowak, and Baraniuk (1998) consider hidden Markov fields in a problems of image denoising. They develop the Efficient Expectation Maximization (EEM) algorithm to fit their model. See also Figueiredo and Nowak (2001). Shrinkage induced by Bayesian models in which the hyperparameters of the prior are made time dependent in an empirical Bayes fashion is considered in Vidakovic and Bielza Lozoya (1998). Kolaczyk (1999) and Novak and Kolaczyk (2000) apply Bayesian modeling in the 2-D wavelet domains where Poisson counts are of interest.

Leporini and Pesquet (1998) explore cases for which the prior is an exponential power distribution  $[\mathcal{EPD}(\alpha, \beta)]$ . If the noise also has an  $\mathcal{EPD}(a, b)$  distribution with  $0 < \beta < b \leq 1$ , the maximum a posteriori (MAP) solu-

tion is a hard-thresholding rule. If  $0 < \beta \leq 1 < b$  then the resulting MAP rule is

$$\delta(d) = d - \left( \frac{\beta a^b}{b \alpha^\beta} \right)^{1/(b-1)} |d|^{(\beta-1)/(b-1)} + o(|d|^{(\beta-1)/(b-1)}).$$

The same authors consider the Cauchy noise as well and explore properties of the resulting MAP rules. When the priors are hierarchical (mixtures) Leporini, Pesquet, and Krim (1999) demonstrated that the MAP solution can be degenerated and suggested Maximum Generalized Marginal Likelihood method. Some related derivations can be found in Chambolle *et al.* (1998) and Leporini and Pesquet (1999).

Pesquet *et al.* (1996) develop a Bayesian-based approach to the best basis problem, while preserving the classical tree search efficiency in wavelet packets and local trigonometric bases. Kohn and Marron (1997) use a model similar to one in Chipman, Kolaczyk, and McCulloch (1997) but in the context of the best basis selection.

Ambler and Silverman (2004a, b) allow for the possibility that the wavelet coefficients are locally correlated in both location (time) and scale (frequency). This leads to an analytically intractable prior structure. However, they show that it is possible to draw independent samples from a close approximation to the posterior distribution by an approach based on *Coupling From The Past*, making it possible to take a simulation-based approach to wavelet shrinkage.

Angelini and Vidakovic (2004) show that  $\Gamma$ -minimax shrinkage rules are Bayes with respect to a least favorable contamination prior with a uniform spread distribution  $\xi$ . Their method allows for incorporation of information about the energy in the signal of interest.

Ogden and Lynch (1999) describe a Bayesian wavelet method for estimating the location of a change-point. Initial treatment is for the standard change-point model (that is, constant mean before the change and constant mean after it) but extends to the case of detecting discontinuity points in an otherwise smooth curve. The conjugate prior distribution on the change point  $\tau$  is given in the wavelet domain, and it is updated by observed empirical wavelet coefficients.

Ruggeri and Vidakovic (1998) discuss Bayesian decision theoretic thresholding. In the set of all hard thresholding rules, they find restricted Bayes rules under a variety of models, priors, and loss functions. When the data are multivariate, Chang and Vidakovic (2002) propose a wavelet-based shrinkage estimation of a single data component of interest using information from the rest of multivariate components. This incorporation of information is done via Stein-type shrinkage rule resulting from an empirical Bayes standpoint. The proposed shrinkage estimators maximize the predictive density under appropriate model assumptions on the wavelet coefficients.

Lu, Huang, and Tung (1997) suggested linear Bayesian wavelet shrinkage in a nonparametric mixed-effect model. Their formulation is conceptually inspired by the duality between reproducing kernel Hilbert spaces and random processes as well as by connections between smoothing splines and Bayesian regressions. The unknown function  $f$  in the standard nonparametric regression formulation ( $y = f(x_i) + \sigma\epsilon_i$ ,  $i = 1, \dots, n$ ;  $0 \leq x \leq 1$ ;  $\sigma > 0$ ;  $Cov(\epsilon_1, \dots, \epsilon_n) = R$ ) is given a prior of the form  $f(x) = \sum_k \alpha_{Jk} \phi_{Jk}(x) + \delta Z(x)$ ;  $Z(x) \sim \sum_{j \geq J} \sum_k \theta_{jk} \psi_{jk}(x)$ , where  $\theta_{jk}$  are uncorrelated random variables such that  $E\theta_{jk} = 0$  and  $E\theta_{jk}^2 = \lambda_j$ . The authors propose a linear, empirical Bayes estimator  $\hat{f}$  of  $f$  that enjoys Gauss-Markov type of optimality. Several non-linear versions of the estimator are proposed, as well. Independently, and by using different techniques, Huang and Cressie (1997) consider the same problem and derive a Bayesian estimate.

### *Acknowledgment.*

We would like to thank Dipak Dey for the kind invitation to compile this chapter. This work was supported in part by NSA Grant E-24-60R at the Georgia Institute of Technology.

## REFERENCES

- Abramovich F., Amato U., and Angelini C. (2004). On optimality of Bayesian wavelet estimators. *Scandinavian Journal of Statistics*, 31, 217–234.
- Abramovich, F., Bailey, T. C. and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *The Statistician*, 49, 1–29.
- Abramovich, F., Besbeas, P., and Sapatinas, T. (2002). Empirical Bayes approach to block wavelet function estimation. *Comput. Statist. Data Anal.*, 39, 435–451.
- Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via Bayesian approach. *Journal of the Royal Statistical Society, Ser. B*, 60, 725–749.
- Ambler, G. K. and Silverman, B. W. (2004a). Perfect simulation of spatial point processes using dominated coupling from the past with application to a multiscale area-interaction point process. Manuscript, Department of Mathematics, University of Bristol.
- Ambler, G. K. and Silverman, B. W. (2004b). Perfect simulation for wavelet thresholding with correlated coefficients. Technical Report 04:01, Department of Mathematics, University of Bristol.
- Angelini, C. and Sapatinas, T. (2004). Empirical Bayes approach to wavelet regression using  $\epsilon$ -contaminated priors. *Journal of Statistical Computation and Simulation*, 74, 741–764.
- Angelini, C. and Vidakovic, B. (2004).  $\Gamma$ -Minimax Wavelet Shrinkage: A Robust Incorporation of Information about Energy of a Signal in Denoising

- Applications. *Statistica Sinica*, 14, 103–125.
- Antoniadis, A., Bigot, J., and Sapatinas, T. (2001). Wavelet estimators in non-parametric regression: A comparative simulation study. *Journal of Statistical Software*, 6, 1–83.
- Berliner, L. M., Wikle, C. K., and Milliff, R. F. (1999). Multiresolution wavelet analyses in hierarchical Bayesian turbulence models. Bayesian Inference in Wavelet Based Models, In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 341–359. Springer-Verlag, New York.
- Brown, P. J., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96, 398–408.
- Carlin, B. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo, *Journal of the Royal Statistical Society, Ser. B*, 57, 473–484.
- Chambolle, A., DeVore, R. A., Lee, N-Y, and Lucier, B. J. (1998). Nonlinear wavelet image processing: variational problems, compression and noise removal through wavelet shrinkage, *IEEE Trans. Image Processing*, 7, 319–335.
- Chang, W. and Vidakovic, B. (2002). Wavelet estimation of a baseline signal from repeated noisy measurements by vertical block shrinkage. *Computational Statistics and Data Analysis*, 40, 317–328.
- Chencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Doklady*, 3, 1559–1562.
- Chipman, H., McCulloch, R., and Kolaczyk, E. (1997). Adaptive Bayesian Wavelet Shrinkage. *Journal of the American Statistical Association*, 92, 1413–1421.
- Clyde, M., DeSimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91, 1197–1208.
- Clyde, M. and George, E. (1999). Empirical Bayes Estimation in Wavelet Non-parametric Regression. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 309–322. Springer-Verlag, New York.
- Clyde, M. and George, E. (2000). Flexible Empirical Bayes Estimation for Wavelets, *Journal of the Royal Statistical Society, Ser. B*, 62, 681–698.
- Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets, *Biometrika*, 85, 391–402.
- Crouse, M., Nowak, R., and Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Processing*, 46, 886–902.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. S.I.A.M., Philadelphia.
- De Canditiis, D. and Vidakovic, B. (2004). Wavelet Bayesian Block Shrinkage via Mixtures of Normal-Inverse Gamma Priors. *Journal of Computational and Graphical Statistics*, 13, 383–398.
- 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. *Journal of the American Statistical Association*, 90, 1200–1224.

- Donoho, D., Johnstone, I., Kerkycharian, G., and Pickard, D. (1996). Density Estimation by Wavelet Thresholding. *The Annals of Statistics*, 24, 508–539.
- Fadili, M. J. and E.T. Bullmore, E. T. (2004). A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps. *NeuroImage*, 23, 1112–1128.
- Figueiredo, M. and Robert Nowak, R. (2001). Wavelet-Based Image Estimation: An empirical Bayes approach using Jeffreys’ noninformative prior. *IEEE Transactions on Image Processing*, 10, 1322–1331.
- George, E.I., and McCulloch, R. (1997). Approaches to Bayesian variable selection. *Statistica Sinica*, 7, 339–373.
- Hall, P. and Patil, P. (1995). Formulae for the mean integrated square error of non-linear wavelet based density estimators. *The Annals of Statistics*, 23, 905–928.
- Huang, H.-C. and Cressie, N. (1999). Empirical Bayesian spatial prediction using wavelets. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 203–222. Springer-Verlag, New York.
- Huang, H.-C. and Cressie, N. (2000). Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, 42, 262–276. (Matlab code is available from the Wavelet Denoising software written by Antoniadis, Bigot, and Sapatinas, 2001)
- Huang, S. Y. (2002). On a Bayesian aspect for soft wavelet shrinkage estimation under an asymmetric linex loss. *Statistics and Probability Letters*, 56, 171–175.
- Huang, S. Y. and Lu, H. S. (2000). Bayesian wavelet shrinkage for nonparametric mixed-effects models. *Statistica Sinica*, 10, 1021–1040.
- Johnstone, I. and Silverman B. W. (1996). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Ser. B*, 59, 319–351.
- Kohn, R., Marron, J.S., and Yau, P. (2000). Wavelet estimation using Bayesian basis selection and basis averaging. *Statistica Sinica*, 10, 109 – 128.
- Kolaczyk, E. D. (1999). Bayesian multi-scale models for Poisson processes. *Journal of the American Statistical Association*, 94, 920–933.
- Leporini, D., and Pesquet, J.-C. (1998). Wavelet thresholding for a wide class of noise distributions, EUSIPCO’98, Rhodes, Greece, 993–996.
- Leporini, D., and Pesquet, J.-C. (1999). Bayesian wavelet denoising: Besov priors and non-gaussian noises, *Signal Processing*, 81, 55–67.
- Leporini, D., Pesquet, J.-C., Krim, H. (1999). Best basis representations with prior statistical models. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 109–113. Springer-Verlag, New York.
- Lina, J.-M., and MacGibbon, B. (1997). Non-Linear shrinkage estimation with complex Daubechies wavelets. Proceedings of SPIE, Wavelet Applications in Signal and Image Processing V, vol. 3169, 67–79.
- Lu, H. S., Huang, S. Y. and Lin, F. J. (2003). Generalized cross-validation for wavelet shrinkage in nonparametric mixed-effects models. *J. Computational and Graphical Statistics*, 12, 714–730.
- Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 11, 674–693.

- Morris, J.S., Vannucci, M., Brown, P.J. and Carroll, R.J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association*, 98, 573–597.
- Müller, P. and Vidakovic, B. (1999a). Bayesian inference with wavelets: Density estimation. *Journal of Computational and Graphical Statistics*, 7, 456–468.
- Müller, P. and Vidakovic, B. (1999b). MCMC methods in wavelet shrinkage: Non-equally spaced regression, density and spectral density estimation. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 187–202. Springer-Verlag, New York.
- Müller, P. and Vidakovic, B. (Editors) (1999c). *Bayesian Inference in Wavelet Based Models*. Lecture Notes in Statistics 141, Springer-Verlag, New York.
- Nowak, R. and Kolaczyk, E. (2000). A Bayesian multiscale framework for Poisson inverse problems. *IEEE Transactions on Information Theory, Special Issue on Information-Theoretic Imaging*, 46, 1811–1825.
- Ogden, T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston.
- Ogden, R. T. and Lynch, J. D. (1999). Bayesian analysis of change-point models. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 67–82. Springer-Verlag, New York.
- Pensky, M. and Sapatinas, T. (2004). Frequentist optimality of Bayes factor thresholding estimators in wavelet regression models. Technical Report at University of Cyprus.
- Pesquet, J., Krim, H., Leporini, D., and Hamman, E. (1996). Bayesian approach to best basis selection. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5, 2634–2637.
- Portilla, J., Strela, V., Wainwright, M., and Simoncelli, E. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12, 1338–1351.
- Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85, 617–624.
- Ruggeri, F. (1999). Robust Bayesian and Bayesian decision theoretic wavelet shrinkage. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 139–154. Springer-Verlag, New York.
- Ruggeri, F. and Vidakovic, B. (1999). A Bayesian decision theoretic approach to the choice of thresholding parameter. *Statistica Sinica*, 9, 183–197.
- Simoncelli, E. (1999). Bayesian denoising of visual images in the wavelet domain. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 291–308. Springer-Verlag, New York.
- Simoncelli, E. and Adelson, E. (1996). Noise removal via Bayesian wavelet coring. Presented at: 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22, 1701–1728.

- Vannucci, M., Brown, P.J., and Fearn, T. (2001). Predictor selection for model averaging. Bayesian methods with applications to science, policy and official statistics. (Eds E. I. George and P. Nanopoulos), Eurostat: Luxemburg, 553–562.
- Vannucci, M., Brown, P.J., and Fearn, T. (2003). A decision theoretical approach to wavelet regression on curves with a high number of regressors. *Journal of Statistical Planning and Inference*, 112, 195–212.
- Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective. *Journal of the Royal Statistical Society, Ser. B*, 61, 971–986.
- Vidakovic, B. (1998a). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, 93, 173–179.
- Vidakovic, B. (1998b). Wavelet-based nonparametric Bayes methods. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller and D. Sinha, Lecture Notes in Statistics, vol. 133, 133–155. Springer-Verlag, New York.
- Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley & Sons, Inc., New York, 384 pp.
- Vidakovic, B. and Bielza Lozoya, C. (1998). Time-adaptive wavelet denoising. *IEEE Transactions on Signal Processing*, 46, 2549–2554.
- Vidakovic, B. and Müller, P. (1999). An introduction to wavelets. In *Bayesian Inference in Wavelet Based Models*, eds. P. Müller and B. Vidakovic, Lecture Notes in Statistics, vol. 141, 1–18. Springer-Verlag, New York.
- Vidakovic, B. and Ruggeri, F. (1999). Expansion estimation by Bayes rules. *Journal of Statistical Planning and Inference*, 79, 223–235.
- Vidakovic, B. and Ruggeri, F. (2001). BAMS method: Theory and simulations. *Sankhyā, Series B*, 63, 234–249. (Matlab code is available from the Wavelet Denoising software written by Antoniadis, Bigot, and Sapatinas, 2001)
- Walter, G.G. and Shen X. (2001). *Wavelets and Others Orthogonal Systems*, Second Edition. Chapman & Hall/CRC, Boca Raton.
- Wikle, C. K., Milliff, R. F., and Large, W. G. (1999). Observed wavenumber spectra from 1 to 10,000 km of near-surface winds during the TOGA COARE IOP. *Journal of Atmospheric Science*, 56, 2222–2231.
- Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, 96, 382–397.