Wavelet-Based Nonparametric Bayes Methods

Brani Vidakovic

ABSTRACT In this chapter, we will provide an overview of the current status of research involving Bayesian inference in wavelet nonparametric problems. 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 provide a natural terrain for both of these goals.

1 Introduction

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

In recent years statistical wavelet modeling and computational research has become a burgeoning area in both theoretical and applied statistics. Also, it is beginning to impact developments in statistical methodology and in various applied scientific fields. Wavelet ideas are developing in many statistical areas: regression; density and function estimation; factor analysis; modeling and forecasting in time series analysis; and spatial statistics; with ranges of application areas in science and engineering. The emerging interests in Bayesian statistical modeling and wavelets is generating exciting new directions for the interface of these two research areas, with significant potential for future impact on applied work.

Wavelet shrinkage is a simple and yet powerful tool in nonparametric statistical modeling. It can be described as a three step procedure. Data are transformed into a set of wavelet coefficients; a shrinkage of the coefficients is performed; and then the shrunken wavelet coefficients are transformed back in the domain of the original data.

The wavelet domain is a good modeling environment; several supporting arguments are listed below.

Discrete wavelet transformations tend to "disbalance" the data. Even though the transformations preserve the ℓ_2 norm of the data, the "energy" (an engineering term for the ℓ_2 norm) of the transformed data concentrates in only a few wavelet coefficients. That concentration narrows the class of plausible models and facilitates the thresholding. Mallat (1989) gives an interesting discussion on modeling from the signal-processing point of view. The disbalancing property also yields a variety of criteria for the best basis selection.

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 result is, in general, not true for other standard orthonormal bases which may need many "compensating" coefficients to describe discontinuity artifacts and to suppress Gibbs' effects.

Heisenberg's principle states that in modeling time-frequency phenomena one can not be precise in the time domain and in the frequency domain simultaneously. Wavelets automatically trade-off the time-frequency precision by their innate nature. The parsimony of wavelet transformations can be attributed to the ability of wavelets to handle limitations in the Heisenberg principle in a data-dependent manner.

Also, there is theoretical and empirical evidence that wavelet transformations tend to simplify the dependence structure in the original signal. It is even possible, for any given stationary dependence in the input signal, to construct a biorthogonal wavelet basis such that the corresponding wavelet coefficients become uncorrelated (a wavelet counterpart of Karhunen-Loève transformation). For a discussion and examples see Walter (1994).

These arguments identify wavelet bases as suitable tools for effective statistical modeling. More favorable arguments can be given: computational speed of the wavelet transformation, simple descriptors of self-similarity and so on.

Many nonparametric procedures are in fact infinitely parametric. An example is the orthogonal series regression or density estimator. In order to estimate such functions, shrinkage, tapering or truncation of parameter estimators from an infinite class is necessary (Chencov's orthogonal series density estimators, Stein-type estimation, and so on.). The benefits of shrinkage estimation in statistics were first explored in the mid-50's by C. Stein. In the 70's and 80's, many statisticians were interested in statistical benefits of classical and Bayesian shrinkage estimators. Since then, shrinkage estimation has been an active field in statistical research. Rapid increase of computing power and the use of Markov Chain Monte Carlo methods make shrinkage estimation a burgeoning research field in Bayesian statistics today.

2 Discrete Wavelet Transformations

Prior to describing a formal set-up for Bayesian wavelet shrinkage, we provide a brief review of discrete wavelet transformation and traditional wavelet shrinkage.

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), Walter (1994), and Ogden (1997), among others. For the sake of self-containedness of this chapter, we provide a brief overview of the discrete wavelet transformation (DWT).

2.1 Discrete Wavelet Transformations

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

Suppose that the vector $\underset{\sim}{y}$ is wavelet-transformed to a vector $\underset{\sim}{d}$. This linear and orthogonal transformation can be described by an orthogonal matrix W of dimension $n \times n$. In practice, one performs the DWT without exhibiting the matrix W explicitly but by using fast filtering algorithms. The filtering procedure is based on so-called quadrature mirror filters which are uniquely determined by the wavelet of choice. The wavelet decomposition of the vector $\underset{\sim}{y}$ can be written as a vector

$$\underset{\sim}{d} = (H^l \underset{\sim}{y}, GH^{l-1}\underset{\sim}{y}, \ldots, GH^2\underset{\sim}{y}, GH\underset{\sim}{y}, G\underset{\sim}{y}). \tag{1.1}$$

In (1.1, $\underset{\sim}{d}$ has the same length as $\underset{\sim}{y}$ and l is any fixed number between 1 and $J = \log_2 n$. The operators G and H are defined coordinate-wise via

$$(Ha)_k = \Sigma_m h_{m-2k} a_m, \text{ and } (Ga)_k = \Sigma_m g_{m-2k} a_m,$$

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

The elements of $\underset{\sim}{d}$ are called "wavelet coefficients." The sub-vectors given in (1.1) correspond to different levels in a tree-like decomposition where the wavelet coefficients are organized level-wise. For instance, the vector $G\underset{\sim}{y}$ contains $n/2$ coefficients representing the level of the finest detail. When $l = J$, the vectors $GH^{J-1} = \underset{\sim}{y} = \{d_{00}\}$ and $H^J\underset{\sim}{y} = \{c_{00}\}$ contain a single coefficient each and represent, respectively, the coarsest level of detail and the smooth part in wavelet decomposition.

In general, the level j in the wavelet decomposition of $\underset{\sim}{y}$ contains 2^j elements, and is represented as

$$GH^{J-j-1}\underset{\sim}{y} = (d_{j,0}, d_{j,1}, \ldots, d_{j,2^j-1}). \tag{1.2}$$

2.2 *Wavelet Shrinkage*

Wavelet shrinkage methods are now widely recognized as a useful tool in nonparametric function estimation and signal recovery. The simplest nonlinear wavelet shrinkage technique is thresholding. The coordinates of $\underset{\sim}{d}$ are replaced by 0 if they are smaller in absolute value than a fixed threshold λ.

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

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

where $\mathbf{1}(A)$ is the indicator of A.

Graphs of hard- and soft-thresholding rules are given in Figure 1.

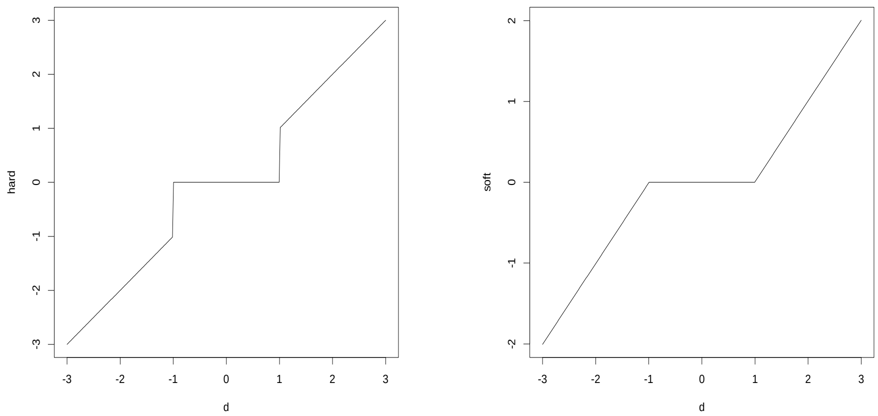

FIGURE 1. Hard and soft thresholding rules with $\lambda = 1$.

3 Bayes and Wavelets

Bayesian approaches to choosing the shrinkage method are less ad-hoc than earlier proposals, and have been shown to be effective. It is known that, in general, Bayes rules are *shrinkers* and that their shape in many cases has a desirable property for wavelet shrinkage: it can heavily shrink small arguments and only slightly shrink large arguments. If we use Bayes models for the wavelet coefficients, the resulting optimal actions can be very close to thresholding.

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

$$y_i = f_i + \epsilon_i, \quad i = 1, \dots, n, \tag{1.3}$$

or in vector notation, $\underset{\sim}{y} = \underset{\sim}{f} + \underset{\sim}{\epsilon}$. After applying the wavelet transformation W, the resulting vector $\underset{\sim}{d} = W(\underset{\sim}{y})$ is again the sum of the transformation of the signal $\underset{\sim}{\theta} = W(\underset{\sim}{f})$ and the transformation of the noise $\underset{\sim}{\eta} = W(\underset{\sim}{\epsilon})$. This outcome is a consequence of the linearity of W. If the components of the noise vector $\underset{\sim}{\epsilon}$, ϵ_i, are modeled by independent normals with mean 0 and variance σ^2, then the orthogonality of W implies that the components of $\underset{\sim}{\eta}$ have the same distribution. The case of correlated normal noise can be addressed as well, see Johnstone and Silverman (1996).

Thus, instead of estimating the function directly, we estimate its wavelet transformation. Equivalently, we estimate the means θ_i in the model $[d_i] \sim \mathcal{N}(\theta_i, \sigma^2)$. The goal of Bayesian wavelet shrinkage is to exhibit such models so that resulting optimal actions, resulting from the Bayesian inference, mimic thresholding rules.

To illustrate Bayesian shrinkage in the wavelet domain, we start with an example from Vidakovic (1998).

Example 3.1 Let

$$[d|\theta, \sigma^2] \sim \mathcal{N}(\theta, \sigma^2), \quad \sigma^2 \text{ unknown} \tag{1.4}$$

be a model for a wavelet coefficient. Because of practical (computational) reasons and noninformative properties,[1] the prior distribution on σ^2 is chosen to be exponential,

$$[\sigma^2] \sim \mathcal{E}(\mu) \qquad (f(\sigma^2|\mu) = \mu e^{-\mu\sigma^2}). \tag{1.5}$$

The marginal model (marginal likelihood) is double exponential,

$$[d|\theta] \sim \mathcal{DE}(\theta, \frac{1}{\sqrt{2\mu}}), \tag{1.6}$$

with $f(d|\theta) = \frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}$. This follows from the fact that the double exponential distribution is a scale mixture of normals.

[1] Exponential distribution minimizes the Fisher information in the class of all distributions supported on $[0, \infty)$ with a fixed first moment.

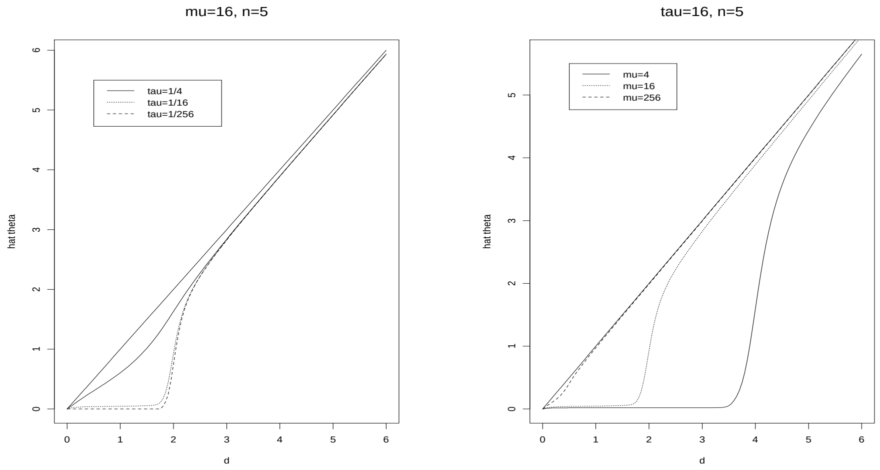

FIGURE 2. Bayes rules for selected values of τ and μ. The rules are plotted only for d positive.

The prior on θ is t with location 0, scale τ, and n degrees of freedom,

$$[\theta] \sim t_n(0, \tau). \tag{1.7}$$

Several graphs of Bayes rules, with respect to this model, are given in Figure 2. The rules are odd functions, and only positive values are plotted. The hyperparameters μ, τ, and n can be specified by empirical Bayes arguments.

It is possible to get an analytic expression for the Bayes rule under the model described, when the prior on θ is general but symmetric, i.e. $\pi(\theta) = \pi(-\theta)$. The Bayes rule with respect to the squared error loss is:

$$\delta(d) = d - \frac{\Pi_1'(c) - \Pi_2'(c)}{\Pi_1(c) + \Pi_2(c)}, \tag{1.8}$$

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

3.1 Regression Problems

The somewhat simplistic model in Example 3.1 can be replaced with more powerful and coherent Bayesian models that exploit a hierarchical structure.

In the context of wavelet regression, we will discuss two 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 σ is known.

Chipman, Kolaczyk, and McCulloch (1997) start with the model

$$[d|\theta, \sigma^2] \sim \mathcal{N}(\theta, \sigma^2). \tag{1.9}$$

The prior on θ is defined as a mixture of two normals (Figure 3, Left)

$$[\theta|\gamma_j] \sim \gamma_j \mathcal{N}(0, (c_j\tau_j)^2) + (1-\gamma_j)\mathcal{N}(0, \tau_j^2), \tag{1.10}$$

where

$$[\gamma_j] \sim \mathit{Bernoulli}(p_j). \tag{1.11}$$

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

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

$$\delta(d) = [P(\gamma_j = 1|d)\frac{(c_j\tau_j)^2}{\sigma^2 + (c_j\tau_j)^2} + P(\gamma_j = 0|d)\frac{\tau_j^2}{\sigma^2 + \tau_j^2}]\ d, \tag{1.12}$$

where

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

and

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

The shrinkage rule (1.12, Figure 3, Right) can be viewed as smooth interpolation between two lines through the origin 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 give very sophisticated empirical Bayes argument for tuning the hyperparameters level-wise. They provide a full posterior analysis on the coefficients and function values. The simulations performed demonstrate that the ABWS method is superior to the VisuShrink and SureShrink over the standard Donoho-Johnstone test functions in terms of MSE.

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 $\underset{\sim}{\theta}$ 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,

$$[\theta|\gamma_j, \sigma^2] \sim \mathcal{N}(0, (1-\gamma_j) + \gamma_j c_j \sigma^2). \tag{1.13}$$

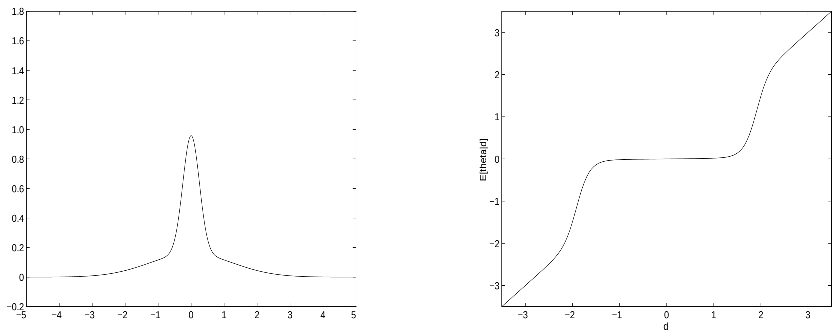

FIGURE 3. Left: Prior on θ as a mixture of two normal distributions with different variances; Right: Shrinkage rule (1.12) in Chipman, Kolaczyk, and McCulloch (1997).

The γ_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 θ belongs. The set of all possible vectors $\underset{\sim}{\gamma}$'s will be referred to as the subset space. The prior distribution for σ^2 is inverse χ^2, i.e.,

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

where λ and ν are fixed hyperparameters and the γ_j's are independently distributed as *Bernoulli* (p_j) random variables.

The posterior mean of $\underset{\sim}{\theta}|\underset{\sim}{\gamma}$ is

$$E(\underset{\sim}{\theta}|\underset{\sim}{d}, \underset{\sim}{\gamma}) = \Gamma(I_n + C^{-1})^{-1}\underset{\sim}{d}, \tag{1.14}$$

where Γ and C are diagonal matrices with γ_{jk} and c_{jk}, respectively, on the diagonal and 0 elsewhere. For a particular subset determined by the ones in $\underset{\sim}{\gamma}$, (1.14) 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 $\underset{\sim}{\theta}$:

$$E(\underset{\sim}{\theta}|\underset{\sim}{d}) = \Sigma_{\underset{\sim}{\gamma}}\pi(\underset{\sim}{\gamma}|\underset{\sim}{d})\Gamma\left(I_n + C^{-1}\right)^{-1}\underset{\sim}{d}, \tag{1.15}$$

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

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

Calculating the posterior probabilities of $\underset{\sim}{\gamma}$ and the mixture estimates for the posterior mean of $\underset{\sim}{\theta}$ above involve sums over all 2^n values of $\underset{\sim}{\gamma}$.

The calculational complexity of the mixing is prohibitive even for problems of moderate size, and either approximations or stochastic methods for selecting subsets $\underset{\sim}{\gamma}$ possessing high posterior probability must be used.

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

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

$$\begin{aligned}
\pi(\underset{\sim}{\gamma}|\underset{\sim}{d}) \approx \tilde{\pi}(\underset{\sim}{\gamma}|y) &= \prod_{j,k} \rho_{jk}^{\gamma_{jk}}(1-\rho_{jk})^{1-\gamma_{jk}} \qquad (1.16)\\
\rho_{jk}(\underset{\sim}{d},\sigma) &= \frac{a_{jk}(\underset{\sim}{d},\sigma)}{1+a_{jk}(\underset{\sim}{d},\sigma)},
\end{aligned}$$

where

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

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 θ_{jk} is approximately

$$\rho_{jk}(1+c_{jk}^{-1})^{-1}d_{jk}. \qquad (1.17)$$

Equation (1.17) 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 4, Left. Authors report good MSE performance of approximation rules.

Clyde and George (1998) propose a model in which the distributions for the error ϵ and θ are scale mixtures of normals, thus justifying Mallat's paradigm (Mallat, 1989). They use an empirical Bayes approach to estimate the prior hyperparamters, 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.

3.2 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*

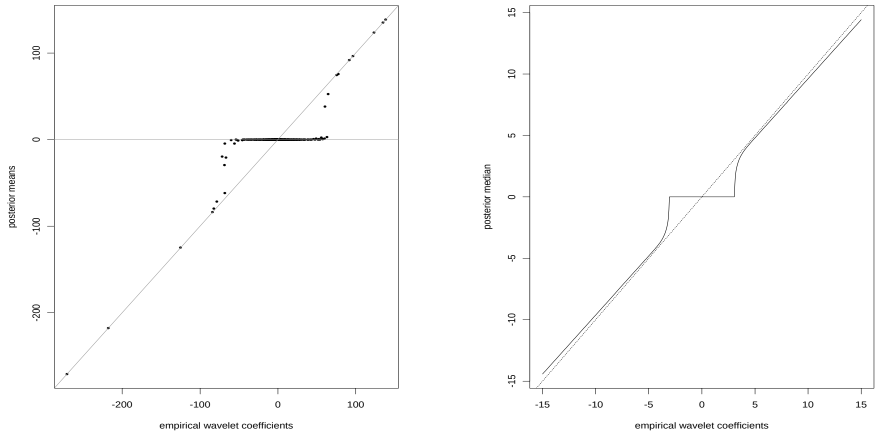

FIGURE 4. Left: Shrinkage rule from Clyde, Parmigiani, and Vidakovic (1998) based on independence approximation (1.17); Right: Posterior median thresholding rule (1.23) from Abramovich, Sapatinas, and Silverman (1998).

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 (1994).

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, θ is estimated by d. Let

$$[\theta] \sim \pi(\theta) = \pi_0 \delta_0 + \pi_1 \xi(\theta), \tag{1.18}$$

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

The resulting Bayesian procedure is:

$$\hat{\theta} = d \; \mathbf{1}(P(H_0|d) < \frac{1}{2}), \tag{1.19}$$

where

$$P(H_0|d) = (1 + \frac{\pi_1}{\pi_0} \frac{1}{B})^{-1}, \tag{1.20}$$

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.

For instance, let, as in Example 3.1,

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

and

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

then d will be thresholded if

$$\frac{\pi_0 e^{-c|d|}}{\pi_0 e^{-c|d|} + \pi_1 (\Pi_1(c) + \Pi_2(c))} \geq \frac{1}{2}, \tag{1.21}$$

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

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

$$[\theta] \sim \pi_j \mathcal{N}(0, \tau_j^2) + (1 - \pi_j)\delta(0) \tag{1.22}$$

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

$$Med(\theta|d) = \mathrm{sign}(d) \max(0, \zeta). \tag{1.23}$$

Here

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

The index j, as before, points to the level containing θ (or d) facilitating adaptivity. The plot of the thresholding function (1.23) is given in Figure 4, Right.

Abramovich, Sapatinas, and Silverman (1998) assume

$$\tau_j^2 = C_1 2^{-\alpha j} \quad \text{and} \quad \pi_j = \min(1, C_2 2^{-\beta j}), \tag{1.24}$$

where C_1, C_2, α, and β are non-negative hyperparameters. The hyperparameters α and β are determined from the assumption that the function to be estimated belongs to a particular Besov space, while C_1 and C_2 are determined in an empirical Bayes fashion.

The authors compare their BayesThresh rule (1.23) with several existing methods (Cross-Validation, False Discovery Rate, VisuShrink and GlobalSure) and report very good MSE performance.

3.3 The Density Estimation Problem

Delyon and Juditsky (1993), Donoho *et al.* (1996), Hall and Patil (1995), Vannucci (1996), and Walter (1994), 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

$$\hat{f}(x) \quad = \quad \Sigma_k c_{j_0 k}\phi_{j_0 k}(x) + \Sigma_{j_0 \le j \le j_1}\Sigma_k d_{jk}\psi_{jk}(x), \tag{1.25}$$

where the coefficients c_{jk} and d_{jk} constituting the vector $\underset{\sim}{d}$ 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

$$c_{jk} = \frac{1}{n}\Sigma_{i=1}^n \phi_{jk}(X_i), \quad \text{and} \quad d_{jk} = \frac{1}{n}\Sigma_{i=1}^n \psi_{jk}(X_i). \tag{1.26}$$

The local nature of wavelet functions makes the wavelet estimator superior to projection estimators that use classical orthonormal bases (Fourier, Hermite, etc.).

Brunk (1978) first proposed the shrinkage of coefficients in the projection density estimate based on linear Bayesian estimation. When an $\mathbb{L}_2$ density f has the Fourier expansion $f(x) = \Sigma_n a_n \xi(x)$ in some orthonormal basis $\xi_n(x), n \in \mathbb{N}$, Brunk (see also Wahba, 1981) assumed normal $\mathcal{N}(0, b_n)$ and independent priors on the coefficients a_n. The suggested choice for b_n was C/m^n, where C is a constant and $m > 1$.

Brunk's linear Bayes shrinkage paradigm, in the wavelet setup, was explored by Vidakovic and Müller (1995). They assumed

$$[\underset{\sim}{d}|\underset{\sim}{\theta}, \sigma^2] \sim \mathcal{MVN}(\underset{\sim}{\theta}, \sigma^2 I), \tag{1.27}$$

where the components of $\underset{\sim}{d}$ are empirical wavelet coefficients from (1.26).

The rationale behind the model in (1.27) is that both c_{jk} and d_{jk} are averages of n iid random variables, see (1.26). Given $\underset{\sim}{\theta}$, the components of $\underset{\sim}{d}$ are independent. The model in (1.27) was completed by specifying the prior distribution on the location $\underset{\sim}{\theta}$ and scale σ.

Let

$$[\underset{\sim}{\theta}, \sigma^2] \sim \mathcal{NIG}(\alpha, \delta, 0, \Sigma), \tag{1.28}$$

where $\mathcal{NIG}(\alpha, \delta, m, \Sigma)$ is the normal-inverse-gamma distribution with density function $f(\underset{\sim}{\theta}, \sigma^2|\alpha, \delta, m, \Sigma) = C \cdot (\sigma^2)^{\frac{\delta+p+2}{2}} \exp[-\{(\underset{\sim}{\theta} - \underset{\sim}{m})'\Sigma^{-1}(\underset{\sim}{\theta} - \underset{\sim}{m}) + \alpha\}/(2\sigma^2)]$.

The posterior is straightforward because of the conjugate structure. The choice of a prior covariance matrix Σ corresponds to the choice of variances b_n in the traditional (Fourier) problems discussed in Brunk (1978) and

Wahba (1981). The authors consider a more general case in which the wavelet coefficients are not independent. More precisely, the coefficients in different levels are assumed independent, but the coefficients in the same level are assumed correlated. If the coefficients between levels j_0 and j_1 are considered in the estimator (1.25), covariance structure can be described by the following covariance matrix:

$$\Sigma = \lambda_{j_0,\phi}\Sigma_{j_0,\phi} \oplus \bigoplus_{j=j_0}^{j_1} \lambda_{j,\psi}\Sigma_{j,\psi}, \tag{1.29}$$

where $\oplus$ is direct sum operation. Σ in (1.29), is a block diagonal matrix in which the block sub-matrices $\Sigma_{j_0,\phi}$ and $\Sigma_{j,\psi}, j = j_0, \ldots, j_1$ describe correlations within their corresponding levels.

The posterior for $[\underset{\sim}{\theta}, \sigma^2]$ is again the normal-inverse-gamma distribution

$$[\underset{\sim}{\theta}, \sigma^2 | \underset{\sim}{d}] \sim \mathcal{NIG}(\alpha^\star, \delta^\star, m^\star, \Sigma^\star),$$

with $\Sigma^\star = (I+\Sigma^{-1})^{-1}$, $\hat{\underset{\sim}{\theta}} = \Sigma^\star \underset{\sim}{d}\ \ (= E(\underset{\sim}{\theta} \mid \underset{\sim}{d}))$, $\alpha^\star = \alpha + ||\underset{\sim}{d}|| + (\theta^\star)'(\Sigma^\star)^{-1}\theta^\star$, and $\delta^\star = \delta + n$.

The Bayes estimator of $\underset{\sim}{\theta}$ is $\hat{\underset{\sim}{\theta}} = \Sigma^\star \underset{\sim}{d}$. Naturally, we estimate the density in (1.25) by replacing c_{jk} and d_{jk} from $\underset{\sim}{d}$ by their affine Bayes estimators.

Example 3.2 In practical implementations of the method, the matrix Σ was a block diagonal matrix consisting of Laurent submatrices $\Sigma_{j_0,\phi}$, $\Sigma_{j_0,\psi}$, $\ldots$, $\Sigma_{j_1,\psi}$ of appropriate dimensions with entries $\sigma_{i,j} = \rho^{|i-j|},\ |\rho| < 1$.

Figure 5 shows shrinkage of empirical wavelet coefficients $\underset{\sim}{d}$ for the *galaxy velocities* data set (Roeder, 1990). There were 176 empirical coefficients between levels $j_0 = 0$ and $j_1 = 6$. The following values for λ's and ρ were adopted: $\lambda_{0,\phi} = 1000, \lambda_{0,\psi} = 1000, \lambda_{1,\psi} = 100, \lambda_{2,\psi} = 10, \lambda_{3,\psi} = 1, \lambda_{4,\psi} = 0.1, \lambda_{5,\psi} = 0.01, \lambda_{6,\psi} = 0.001$, and $\rho = 0.8$.

Building on (1.27)-(1.29), Huerta (1997) proposed the following model

$$\begin{aligned} \left[\underset{\sim}{d}|\underset{\sim}{\theta}, \sigma^2\right] &\sim \mathcal{MVN}(\underset{\sim}{\theta}, \sigma^2 I), \quad \left[\sigma^2\right] \sim \mathcal{IG}(\alpha_1, \delta_1), \\ \left[\underset{\sim}{\theta}|\tau^2\right] &\sim \mathcal{MVN}(\underset{\sim}{0}, \tau^2\Sigma), \quad \left[\tau^2\right] \sim \mathcal{IG}(\alpha_2, \delta_2). \end{aligned}$$

The closed form for $E(\underset{\sim}{\theta}|\underset{\sim}{d})$ in the above model is not possible, however the full conditionals can be specified. The shrinkage is performed by Gibbs sampling. The prior covariance matrix Σ was defined as in (1.29). Obtained simulation results are promising.

Reducing the number of hyperparameters in (1.27)-(1.29) and imposing a prior structure on the remaining hyperparameters in a hierarchical fashion was suggested by Vannucci and Corradi (1996). They applied the shrinkage on a regression problem and called their method BayesShrink. Assuming that $\underset{\sim}{\theta}$ corresponds to an autoregressive process in the time domain, the

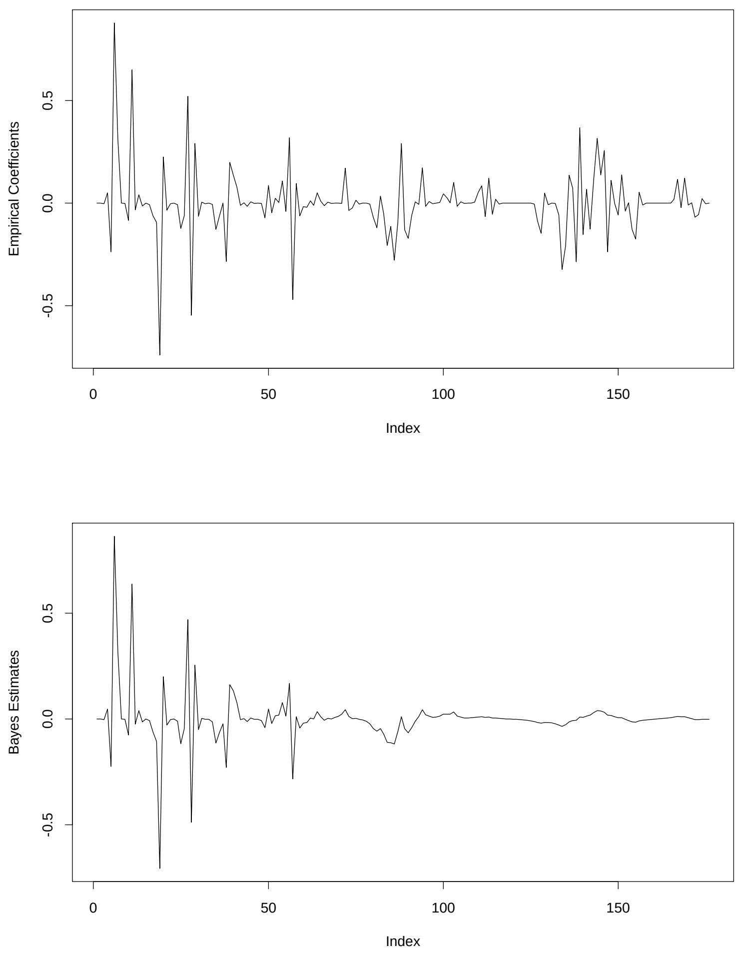

FIGURE 5. Empirical coefficients (1.26) for the `galaxy` data and their affine Bayes shrinkage. The coefficients are arranged as a concatenation of levels in the decomposition, starting with scaling coefficients and ending with coefficients of fine detail. Notice that the Bayes estimate preserves coarse structure and heavily shrinks detail coefficients that cause overfitting.

authors demonstrate that the matrix Σ depends on only two hyperparameters, λ and ρ. The parameter ρ is the "autocovariance index," and λ is the precision parameter. The covariance matrix $\Sigma(\lambda, \rho) = \lambda\Sigma(\rho)$ has an interesting "finger-like" structure.

Vannucci and Corradi (1996) suggest

$$[\lambda] \sim \mathcal{IG}(p/2, q/2), \text{ and}$$

$$[\rho] \propto (C-\rho)^{r_1-1}(C+\rho)^{r_2-1}, \ |\rho| < C,$$

and simulate from the posterior distribution of $\underset{\sim}{\theta}$ using MCMC method.

3.4 Full Bayesian Model

Müller and Vidakovic (1994) 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:

$$\log f(x) = \Sigma_{k\in Z}\xi_{j_0 k}\phi_{j_0,k}(x) + \Sigma_{j\geq j_0, k\in Z} s_{jk}\theta_{jk}\psi_{jk}(x) - \log K, \quad (1.30)$$

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 $\theta = (\xi_{j_0,k}, s_{jk}, \theta_{jk},\ j = j_0, \ldots, j_1,\ k \in Z)$ of wavelet coefficients and indicators is expressed by $f(x) = p(x|\theta)$. The sample $X = \{X_1, \ldots, X_n\}$ 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 θ. Without loss of generality, $j_0 = 0$ can be assumed. Also, any particular application will determine a maximum level of detail j_1.

$$\begin{aligned}
[\xi_{0k}] &\sim \mathcal{N}(0, \tau r_0), \\
[\theta_{jk}|s_{jk}=1] &\sim \mathcal{N}(0, \tau r_j), \quad r_j = 2^{-j}, \\
[s_{jk}] &\sim \textit{Bernoulli } (\alpha^j), \\
[\alpha] &\sim \textit{Beta } (a, b), \\
[1/\tau] &\sim \textit{Gamma } (a_\tau, b_\tau).
\end{aligned} \quad (1.31)$$

The wavelet coefficients θ_{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 θ 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, α^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 θ_{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 θ_{jk} which are consistent with the data.

The particular MCMC simulation scheme used to estimate the model (1.30, 1.31) is described. Starting with some initial values for θ_{jk}, $j = 0, \ldots, j_1$, ξ_{00}, α, and τ, the following Markov chain was implemented.

1. For each $j = 0, \ldots, j_1$ and $k = 1, \ldots, 2^j - 1$ go over the steps **2** and **3.**

2. Update s_{jk}. Let θ_0 and θ_1 indicate the current parameter vector θ 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 θ_{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 σ_{jk} is some rough estimate of the posterior standard deviation of θ_{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 θ with θ_{jk} replaced by $\tilde{\theta}_{jk}$, and $p(\theta_{jk})$ is the p.d.f. of the normal prior distribution given in (1.31).
With probability $a(\theta_{jk}, \tilde{\theta}_{jk})$ replace θ_{jk} by $\tilde{\theta}_{jk}$; else keep θ_{jk} unchanged.

3b. If $s_{jk} = 0$, generate θ_{jk} from the full conditional posterior $p(\theta_{jk}|\ldots, X) = p(\theta_{jk}|s_{jk} = 0) = h(\theta_{jk})$.

4. Update ξ_{00}. Generate $\tilde{\xi_{00}} \sim g(\tilde{\xi_{00}}|\xi_{00})$. Use, for example, $g(\tilde{\xi_{00}}|\xi_{00}) = \mathcal{N}(\xi_{00}, 0.25\rho_{00})$, where ρ_{00} is some rough estimate of the posterior standard deviation of ξ_{00}. Analogously to step **3a,** compute an acceptance probability a and replace ξ_{00} with probability a.

5. Update α. 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 α by $\tilde{\alpha}$, else keep α unchanged.

6. Update τ. Resample τ from the complete inverse Gamma conditional posterior.

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

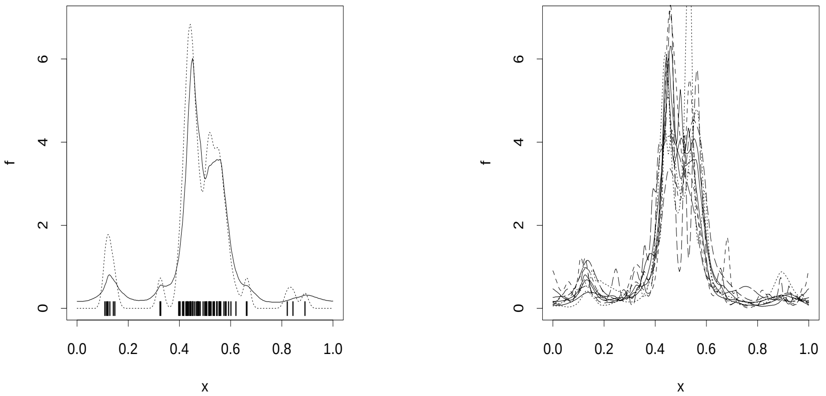

FIGURE 6. Left: The estimated p.d.f. $\hat{f}(x) = \int p(x|\theta)dp(\theta|X)$. The dotted line plots a conventional kernel density estimate for the same data. Right: The posterior distribution of the unknown density $f(x) = p(x|\theta)$ induced by the posterior distribution $p(\theta|X)$. The lines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

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, $g(\tilde{\theta}_{jk}|\theta_{jk}) = \mathcal{N}(\theta_{jk}, 0.25\sigma_{jk})$ with $\sigma_{jk} = 2^{-j}$, was used.

Example 3.3 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 α is reasonably noninformative compared to the likelihood based on $n = 82$ observations.

Initially, all s_{jk} are set to one, and α 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.

4 Other Problems

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

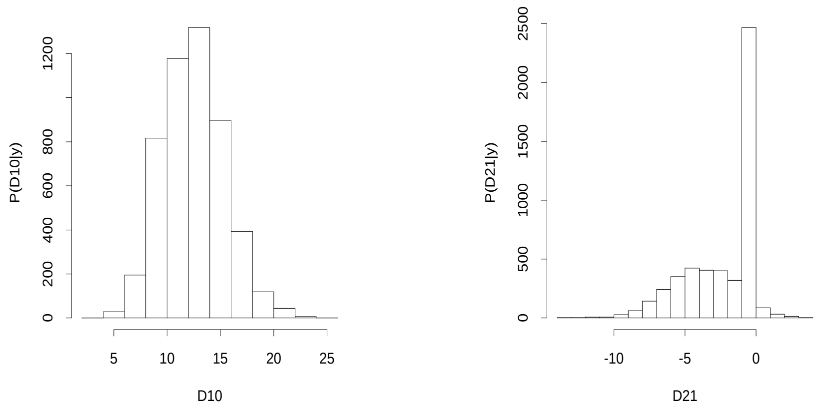

FIGURE 7. Posterior distributions $p((s_{10}\theta_{10})|X)$ and $p((s_{21}\theta_{21})|X)$. While $s_{10}\theta_{10}$ is non-zero with posterior probability close to one, the posterior distribution $p((s_{21}\theta_{21})|X)$ is a mixture of a point mass at zero and a continuous part.

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.

Malfait and Roose (1995) and Malfait, Jansen, and Roose (1996) employed the theory of Markov random fields in wavelet denoising of images. The idea is that the thresholding of a coefficient in a 2-D array depends not only on its own magnitude but also on the magnitudes of its neighbors. This dependence is modeled by using masks consisting of 0 and 1 and exhibiting properties of a Markov field. Crouse, Nowak, and Baraniuk (1997) also consider hidden Markov fields in a similar setup. They develop the Efficient Expectation Maximization (EEM) algorithm to fit their model.

Green, Sapatinas, and Vidakovic (1998) consider a fully Bayesian adaptive wavelet shrinkage. Their model is in fact hierarchical in two respects: the wavelet coefficients θ_{jk} are modeled conditionally on hyperparameters η, that are in turn drawn from hyperpriors $p(\eta)$, and, further, $p(\underset{\sim}{\theta} \mid \eta)$ is specified hierarchically in j. Shrinkage induced by Bayesian models in which the hyperparameters of the prior are made time dependent in an empirical Bayes fashion is considered in Bielza and Vidakovic (1996).

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. Gendron (1997) builds on the Bayesian hypothesis testing idea in the wavelet domain. He uses the Bayesian paradigm to construct a cost function which points to the best wavelet packet. Using the best basis he develops a filtering procedure for seismic signals by applying posterior expectations when the prior is a mixture of a uniform distribution and a point mass at 0.

Ogden (1996) develops 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 τ is given in the wavelet domain, and it is updated by observed empirical wavelet coefficients.

Ruggeri and Vidakovic (1997) 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. They identify model-prior pairs that work well (in sense of exhibiting a hard thresholding rule that minimizes Bayes risk) and show that in the presence of prior information on noise and signal their procedure outperforms SureShrink and VisuShrink methods. Rios and Vidakovic (1997) propose algorithms for construction of wavelet-based random densities and discuss fast Bayesian updating of random priors.

Lu, Huang, and Tung (1997) performed 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 ($y = f(x_i) + \sigma\epsilon_i,\ i = 1, \ldots, n;\ 0 \leq x \leq 1;\ \sigma > 0;\ Cov(\epsilon_1, \ldots, \epsilon_n) = R$) is given a prior of the form $f(x) = \Sigma_k \alpha_{Jk}\phi_{Jk}(x) + \delta Z(x);\quad Z(x) \sim \Sigma_{j \geq J}\Sigma_k \theta_{jk}\psi_{jk}(x)$, where θ_{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.

I would like to thank the editors, especially Peter Müller, for the invitation to compile a chapter on Bayesian wavelet nonparametrics. Peter's contributions to this chapter have been numerous.

Several colleagues and two anonymous referees also commented on an early version of this chapter. I am grateful to Fabrizio Ruggeri, Gabriel Huerta, Fanis Sapatinas, Marina Vannucci, and Todd Ogden for their suggestions.

This work was supported in part by NSF Grant DMS-9626159 at Duke University.

REFERENCES

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998), "Wavelet thresholding via Bayesian approach," *Journal of the Royal Statistical Society,* 60, No.3.

Bielza, C., and Vidakovic, B. (1996), "Time-dependent wavelet shrinkage," Discussion Paper, 96-24, ISDS, Duke University.

Brunk, H. (1978), "Univariate density estimation by orthogonal series," *Biometrika,* 65, 3, 521-528.

Carlin, B. and Chib, S. (1995), "Bayesian model choice via Markov chain Monte Carlo," *Journal of the Royal Statistical Society B*, 57, 473-484.

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. To appear *Journal of the American Statistical Association.*

Clyde, M., and George, E. (1998), "Empirical wavelet estimation in wavelets," Discussion Paper, ISDS, Duke University

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998), "Multiple shrinkage and subset selection in wavelets," Discussion Paper, 95-37, ISDS, Duke University. To appear in *Biometrika.*

Crouse, M., Nowak, R., and Baraniuk, R. (1997), "Statistical signal processing using wavelet-domain hidden Markov models," Proceedings of SPIE, Wavelet Applications in Signal and Image Processing V, vol. 3169, 248-259.

Daubechies, I. (1992), *Ten Lectures on Wavelets.* S.I.A.M., Philadelphia.

Delyon, B., and Juditsky, A. (1993), "Wavelet estimators, global error measures: Revisited," Publication interne, 782, IRISA-INRIA, France.

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., Kerkyacharian, G., and Pickard, D. (1996), "Density Estimation by Wavelet Thresholding," *The Annals of Statistics*, 24, 508–539.

"Wavelet shrinkage: Asymptopia? (With discussion)" *Journal of the Royal Statistical Society B,* 57, 301-369.

George, E.I., and McCulloch, R. (1997), "Approaches to Bayesian variable selection," *Statistica Sinica,* 7, 339-373.

Green, P.J., Sapatinas, T., and Vidakovic, B. (1998), "Fully Bayesian adaptive wavelet shrinkage," Manuscript in Preparation.

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.

Hernández, E., and Weiss, G. (1996), *A First Course on Wavelets*, Boca Raton: CRC Press Inc.

Huerta, G. (1997), "Bayes wavelet shrinkage and applications to data denoising," In electronic proceedings of International Workshop on Wavelets in Statistics, Duke University, 12-13 October 1997.

Huang, H-C., and Cressie, N. (1997), "Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data," Technical Report, 97-23, Department of Statistics, Iowa State University.

Johnstone, I., and Silverman B. W. (1996), "Wavelet threshold estimators for data with correlated noise," *Journal of the Royal Statistical Society B,* 59, 319-351.

Kohn R, and Marron J. S. (1997), "Bayesian wavelet shrinkage," International Workshop on Wavelets in Statistics, Duke University, 12-13 October 1997.

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. H-S., Huang, S-Y., and Tung Y-C. (1997), "Wavelet shrinkage for nonparametric mixed-effects models," Technical Report, Institute of Statistics, National Chiao Tung University. Also in: Electronic proceedings of International Workshop on Wavelets in Statistics, Duke University, 12-13 October 1997.

Malfait, M., and Roose, D. (1995), "Wavelets and Markov random fields in a Bayesian framework," In *Wavelets and Statistics*, Eds: A. Antoniadis and G. Oppenheim. Lecture Notes in Statistics, Springer-Verlag 103 225-238.

Malfait, M., Jansen, M., and Roose, D. (1996), "Bayesian approach to wavelet-based image processing," In Proceedings of The Joint Statistical Meetings, Chicago, August 1996.

Mallat, S. (1989), "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, 11, 674-693.

Meyer, Y. (1992), *Wavelets and Operators,* Cambridge.

Müller, P., and Vidakovic, B. (1995), "Bayesian inference with wavelets:

Density estimation," Discussion Paper, 95-33, ISDS, Duke University.

Ogden, T. (1996), "Wavelets in Bayesian change-point analysis," Technical Report, Department of Statistics, University of South Carolina.

Ogden, T. (1997), *Essential Wavelets for Statistical Applications and Data Analysis,* Birkhäuser, Boston.

Pesquet, J., Krim, H., Leporini, D., and Hamman, E. (1996), "Bayesian approach to best basis selection," In: *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 5, 2634-2637.

Rios, D., and Vidakovic, B. (1997), "Wavelet-based random densities," Discussion Paper, 97-05, ISDS, Duke University. To appear in: *Computational Statistics.*

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., and Vidakovic, B. (1998), "A Bayesian decision theoretic approach to wavelet thresholding," Discussion Paper, 95-35, ISDS, Duke University. To appear in *Statistica Sinica.*

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. (1996), *Sull'Applicazione delle "Wavelets" in Statitstica,* Tesi di dottorato, Dipartimento di Statistica G. Parenti, Università degli Studi di Firenze.

Vannucci, M., and Corradi, F. (1997), "Some findings on the covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective," Report UKC/IMS, 97-05, Institute of Maths and Stats, University of Kent at Canterbury.

Vannucci, M., and Vidakovic, B. (1995), "Preventing the Dirac disaster: Wavelet based density estimation," Discussion Paper, 95-27, ISDS, Duke University.

Vidakovic, B. (1998), "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *Journal of the American Statistical Association*, 93 173-179.

Vidakovic, B., and Müller, P. (1995), "Wavelet shrinkage with affine Bayes rules with applications," Discussion Paper, 95-34, ISDS, Duke University.

Wahba, G. (1981), "Data-based optimal smoothing of orthogonal series density estimates," *The Annals of Statistics,* 9, 146-156.

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