Chapter 14
Bayesian Wavelet Shrinkage Strategies: A Review

Norbert Reményi and Brani Vidakovic

Abstract In this chapter the authors overview recent developments and current status of use of Bayesian paradigm in wavelet shrinkage. The paradigmatic problem where wavelet shrinkage is employed is that of nonparametric regression where data are modeled as observations from an unknown signal contaminated with a Gaussian noise. Bayes rules as general shrinkers provide a formal mechanism to implement shrinkage in the wavelet domain that is model based and adaptive. New developments including dependence models, complex wavelets and MCMC strategies are described. Applications include inductance plethysmography data and curve classification procedure applied in botany. The chapter features an extensive set of references consisting of almost 100 entries.

14.1 Introduction

Wavelet-based tools became standard methodology 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, and 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. Wavelet shrinkage is a

N. Reményi (✉)
School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, NW Atlanta, GA 30332-0205, USA
e-mail: nremenyi@gatech.edu

B. Vidakovic
School of Biomedical Engineering, Georgia Institute of Technology, 765 Ferst Drive, NW Atlanta, GA 30332-0205, USA
e-mail: brani@gatech.edu

X. Shen and A.I. Zayed (eds.), *Multiscale Signal Analysis and Modeling*,
DOI 10.1007/978-1-4614-4145-8_14,

three-step procedure: (1) data are transformed into a set of wavelet coefficients; (2) a shrinkage of the coefficients is performed; and (3) the processed wavelet coefficients are transformed back to the domain of the original data.

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

Discrete wavelet transformations tend to "disbalance" the data. Even though the orthogonal transforms preserve the ℓ_2 norm of the data (the square root of sum of squares of observations, or the "energy" as engineers like to say), most of the ℓ_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 modeling building blocks, are well localized in both time and scale (frequency). Signals with rapid local changes (signals with discontinuities, cusps, sharp spikes, etc.) can be 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 cannot be arbitrarily precise in the time and frequency domains simultaneously, rather this precision is bounded from the below by a universal constant. 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 data. 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 [91].

We conclude this incomplete list of features of wavelet transforms 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 given: computational speed of the wavelet transformation, easy incorporation of prior information about some features of the signal (smoothness, distribution of energy across scales), etc.

Prior to describing a formal setup for Bayesian wavelet shrinkage, we provide a brief review of discrete wavelet transforms 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 [33, 68, 87, 91], among others. An introductory article is [88].

14.1.1 Discrete Wavelet Transformations and Wavelet Shrinkage

Let $\boldsymbol{y}$ be a data vector of dimension (size) n. For the simplicity we choose n to be a power of 2, say 2^J. We assume that measurements $\boldsymbol{y}$ 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 $\boldsymbol{y}$ is wavelet transformed to a vector $\boldsymbol{d}$. This linear and orthogonal transform can be fully described by an $n \times n$ orthogonal matrix $\boldsymbol{W}$. The use of the matrix $\boldsymbol{W}$ is possible when n is not large (of order of a few thousand, at most), but for large n, fast filtering algorithms are employed. The filtering procedures are based on so-called quadrature mirror filters which are uniquely determined by the choice of wavelet and fast Mallat's algorithm [63]. The wavelet decomposition of the vector $\boldsymbol{y}$ can be written as

$$\boldsymbol{d} = (H^{\ell}\boldsymbol{y}, GH^{\ell-1}\boldsymbol{y}, \ldots, GH^2\boldsymbol{y}, GH\boldsymbol{y}, G\boldsymbol{y}). \tag{14.1}$$

Note that in (14.1), $\boldsymbol{d}$ has the same length as $\boldsymbol{y}$ and ℓ is any fixed number between 1 and $J = \log_2 n$. The operators G and H acting on data sequences are defined coordinate-wise via

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

where g and h are high- and low-pass wavelet filters. 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 software packages.

The elements of $\boldsymbol{d}$ are called "wavelet coefficients." The subvectors described in (14.1) correspond to detail levels. For instance, the vector $G\boldsymbol{y}$ contains $n/2 = 2^{J-1}$ coefficients representing the level of the finest detail. When $\ell = J$, the vectors $GH^{J-1}\boldsymbol{y} = \{d_{00}\}$ and $H^J\boldsymbol{y} = \{c_{00}\}$ contain a single coefficient each and represent the coarsest possible level of detail and the smooth part in wavelet decomposition, respectively.

In general, jth detail level in the wavelet decomposition (14.1) contains 2^j elements, and can be written as

$$GH^{J-j-1}\boldsymbol{y} = (d_{j,0}, d_{j,1}, \ldots, d_{j,2^j-1}). \tag{14.2}$$

Wavelet shrinkage methodology consists of shrinking the magnitudes of wavelet coefficients. The simplest wavelet shrinkage technique is thresholding. The components of $\boldsymbol{d}$ are replaced by 0 if their absolute value does not exceed a fixed threshold λ.

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

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

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 coefficient, can act as shrinkage/thresholding rules.

14.2 Wavelets and Bayes

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 found, see [89]. The Bayes rules can be constructed to mimic the thresholding rules: to slightly shrink the large coefficients and heavily shrink the small coefficients. In addition, Bayes rules result from realistic statistical models on wavelet coefficients and such models allow for incorporation of prior information about the *true* signal. Furthermore, most Bayes rules can be easily either computed by simulation or expressed in a closed form. Reviews of early Bayesian approaches can be found in [3, 78, 86, 87]. An edited volume on Bayesian modeling in the wavelet domain appeared 12 years ago [65].

A paradigmatic task in which the wavelets are typically applied is recovery of an unknown signal $\boldsymbol{f}$ observed with noise $\boldsymbol{e}$. In statistical terms this would be a task of nonparametric regression. Wavelet transformations $\boldsymbol{W}$ are applied to noisy measurements $y_i = f_i + e_i,\ i = 1,\ldots,n$, or, in vector notation, $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{e}$. The linearity of $\boldsymbol{W}$ implies that the transformed vector $\boldsymbol{d} = \boldsymbol{W}(\boldsymbol{y})$ is the sum of the transformed signal $\boldsymbol{\theta} = \boldsymbol{W}(\boldsymbol{f})$ and the transformed noise $\boldsymbol{\varepsilon} = \boldsymbol{W}(\boldsymbol{e})$. Furthermore, the orthogonality of $\boldsymbol{W}$ and Gaussianity of $\boldsymbol{e}$ implies Gaussianity of $\boldsymbol{\varepsilon}$ as well.

Bayesian methods are applied in the wavelet domain, that is, 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. Block-modeling approaches are also possible.

When the model is on individual wavelet (detail) coefficients $d_i \sim N(\theta_i, \sigma^2),\ i = 1,\ldots,n$, the interest relies in the estimation of the θ_i. Usually we concentrate on typical wavelet coefficient and model: $d = \theta + \varepsilon$. Bayesian methods are applied to estimate the location parameter θ, which will be, in the sequel, argument in the inverse wavelet transform. A prior on θ, and possibly on other parameters of the distribution of ε, is elicited, and the corresponding Bayes estimators are back-transformed. Various choices of Bayesian models have been motivated by different,

often contrasting, interests. Some models were driven by empirical justifications, others by pure mathematical considerations; some models lead to simple closed-form rules, the other require extensive Markov Chain Monte Carlo (MCMC) simulations to produce the estimate. Bayes rules with respect to absolute or 0-1 loss functions are capable of producing bona fide thresholding rules.

14.2.1 An Illustrative Example

As an illustration of the Bayesian approach we present BAMS (Bayesian adaptive multiresolution shrinkage). The method, due to [90], 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, [57, 58, 63, 77, 81, 86]. 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 accordance with the spirit of this statement, [63] suggested to fit empirical distributions of wavelet coefficients by the exponential power model

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

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

Following the Bayesian paradigm, prior distributions should be elicited on the parameters of the model $d|\theta, \sigma^2 \sim N(\theta, \sigma^2)$ and Bayesian estimators (namely, posterior means under squared loss) computed. In BAMS, priors on θ and σ^2 are set such that the marginal (prior predictive) distribution of the wavelet coefficients is a double exponential distribution DE, that is, an exponential power one with $\beta = 1$. The double exponential distribution can be obtained by marginalizing the normal likelihood by adopting exponential prior on its variance σ^2. The choice of an exponential prior can be justified by its *maxent* property, that is, exponential distribution is the entropy maximizer in the class of all distributions supported on $(0, \infty)$ with a fixed first moment, and in that sense is noninformative.

Thus, BAMS uses the exponential prior $\sigma^2 \sim E(\mu), \quad \mu > 0$, which leads to the marginal likelihood

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

Vidakovic [86] considered the previous marginal likelihood but with a t distribution as the prior on θ. The Bayes rules with respect to the squared error loss under general but symmetric priors $\pi(\theta)$ can be expressed using the Laplace transforms of $\pi(\theta)$.

In personal communication with the second author, Jim Berger and Peter Müller suggested in 1993 the use of ε-contamination priors in the wavelet context pointing out that such priors would lead to rules which are smooth approximations to a thresholding.

The choice

$$\pi(\theta) = \varepsilon\delta(0) + (1-\varepsilon)\xi(\theta) \tag{14.3}$$

also reflects prior belief that some locations (corresponding to the signal or function to be estimated) are 0 and that there is a nonzero spread component ξ describing "large" locations. In addition to this prior sparsity of the signal part, this prior leads to desirable shapes of the resulting Bayes rules. Note that here $0 \le \varepsilon \le 1$ denotes the mixing weight, not the random error component, and will be used throughout this chapter in contamination priors.

In BAMS, the spread part ξ is chosen as $\theta \sim DE(0,\tau)$. The Bayes rule under the squared error loss is

$$\delta_\pi(d) = \frac{(1-\varepsilon)\, m_\xi(d)\, \delta_\xi(d)}{(1-\varepsilon)\, m_\xi(d) + \varepsilon\, DE\left(0, \frac{1}{\sqrt{2\mu}}\right)}, \tag{14.4}$$

where

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

and

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

are the prior predictive distribution and the Bayes rule for the spread part of the prior, ξ. Rule (14.4) is the BAMS rule, which falls between comparable hard and soft thresholding rules.

Bayes rules under the squared error loss and regular models are never thresholding rules. To extend this motivating example, we consider the posterior median as an estimator for θ. It is well known that under the absolute error loss $L(\theta,d) = |\theta - d|$, the posterior risk is minimized by the posterior median. The posterior median was first considered by Abramovich et al. [7] in the context of wavelet shrinkage. It could be a thresholding rule, which is preferable to smooth shrinkage rules in many applications, like model selection, data compression, dimension reduction, and related statistical tasks in which it is desirable to replace by zero a majority of the processed coefficients.

For the model above the posterior distribution is $\pi^\star(\theta|d) = f(d|\theta)\pi(\theta)/m_\pi(d)$, where

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

In order to find the median of the posterior distribution, the solution of the following equation, with respect to u, is needed:

$$\int_{-\infty}^{u} \pi^\star(\theta|d)\mathrm{d}\theta = \frac{1}{2}. \tag{14.5}$$

It is easy to show with simple calculus that if $d \geq 0$,

$$\max \int_{-\infty}^{0^-} \pi^\star(\theta|d)\mathrm{d}\theta = \frac{1}{2}, \tag{14.6}$$

and in case $d < 0$,

$$\min \int_{-\infty}^{0} \pi^\star(\theta|d)\mathrm{d}\theta = \frac{1}{2}. \tag{14.7}$$

Because $\pi^\star(\theta|d)$ is a probability density, the integral in (14.5) is non-decreasing in u. Therefore, by using results (14.6) and (14.7), the posterior median is always greater than equal to zero, when $d \geq 0$, and less than equal to zero, when $d < 0$.

To find the posterior median, first consider the case $d \geq 0$. We know that the solution u satisfies $u \geq 0$. The equation in (14.5) becomes

$$\frac{\varepsilon \frac{\sqrt{2\mu}}{2} \mathrm{e}^{-\sqrt{2\mu}d} + (1-\varepsilon)\frac{\sqrt{2\mu}}{4\tau}\mathrm{e}^{-\sqrt{2\mu}d}\left\{\frac{1}{\sqrt{2\mu}+1/\tau} + \frac{1}{\sqrt{2\mu}-1/\tau}\left[\mathrm{e}^{(\sqrt{2\mu}-1/\tau)u} - 1\right]\right\}}{m_\pi(d)} = \frac{1}{2}.$$

Next, assume $d < 0$. Then the solution satisfies $u \leq 0$ and (14.5) becomes:

$$\frac{(1-\varepsilon)\frac{\sqrt{2\mu}}{4\tau}\left\{\frac{1}{\sqrt{2\mu}+1/\tau}\mathrm{e}^{d/\tau} + \frac{1}{\sqrt{2\mu}-1/\tau}\mathrm{e}^{d/\tau} - \frac{1}{\sqrt{2\mu}-1/\tau}\mathrm{e}^{-(\sqrt{2\mu}-1/\tau)u}\right\}}{m_\pi(x)} = \frac{1}{2}.$$

From the above, the algorithm for finding the posterior median $\delta_M(d)$ is:

For $d > 0$,

$$\text{if } \frac{\varepsilon \frac{\sqrt{2\mu}}{2} e^{-\sqrt{2\mu}d} + (1-\varepsilon)\frac{\sqrt{2\mu}}{4\tau} e^{-\sqrt{2\mu}d} \frac{1}{\sqrt{2\mu}+1/\tau}}{m_\pi(d)} > \frac{1}{2}, \quad \delta_M(d) = 0$$

$$\text{else } \delta_M(d) = \frac{1}{\sqrt{2\mu}-1/\tau} \log\left\{ \left[\frac{m_\pi(d)/2 - \varepsilon \frac{\sqrt{2\mu}}{2} e^{-\sqrt{2\mu}d}}{(1-\varepsilon)\frac{\sqrt{2\mu}}{4\tau} e^{-\sqrt{2\mu}d}} + \frac{2/\tau}{2\mu - 1/\tau^2} \right] (\sqrt{2\mu} - 1/\tau) \right\}.$$

For $d < 0$,

$$\text{if } \frac{(1-\varepsilon)\frac{\sqrt{2\mu}}{4\tau}\left[\frac{1}{\sqrt{2\mu}+1/\tau} e^{d/\tau} + \frac{1}{\sqrt{2\mu}-1/\tau} e^{d/\tau} - \frac{1}{\sqrt{2\mu}-1/\tau} e^{(\sqrt{2\mu}-1/\tau)d}\right]}{m_\pi(d)} < \frac{1}{2},$$

$$\delta_M(d) = 0$$

$$\text{else } \delta_M(d) = -\frac{1}{\sqrt{2\mu}-1/\tau} \log\left\{ -\left[\frac{\frac{m_\pi(d)/2}{(1-\varepsilon)\frac{\sqrt{2\mu}}{4\tau}} - \frac{1}{\sqrt{2\mu}+1/\tau} e^{d/\tau}}{\frac{1}{\sqrt{2\mu}-1/\tau} e^{\sqrt{2\mu}d}} - e^{-(\sqrt{2\mu}-1/\tau)d} \right] \right\}.$$

For $d = 0$,

$$\delta_M(d) = 0. \tag{14.8}$$

The rule $\delta_M(d)$ based on algorithm (14.8) is the BAMS-MED rule. As evident from Fig. 14.1, the BAMS-MED rule is a thresholding rule.

14.3 Bayesian Wavelet Regression

14.3.1 Term-by-Term Shrinkage

As we indicated in the introduction, the most popular application of wavelets is the nonparametric regression problem

$$y_i = f(x_i) + e_i, \quad i = 1, \ldots, n.$$

The usual assumptions are that x_i, $i = 1, \ldots, n$ are equispaced (e.g., time points), and the random errors e_i are i.i.d. normal, with zero mean and variance σ^2. The interest

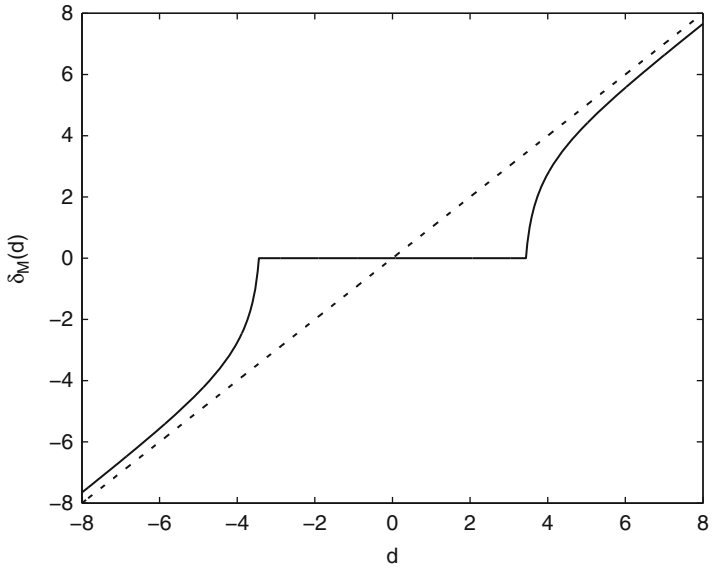

Fig. 14.1 BAMS-MED rule (14.8) for $\varepsilon = 0.9$, $\mu = 1$, and $\tau = 2$

is to estimate the function f using the observations y. After applying a linear and orthogonal wavelet transform, the problem becomes

$$d_{jk} = \theta_{jk} + \varepsilon_{jk},$$

where d_{jk}, θ_{jk}, and ε_{jk} are the wavelet coefficients (at resolution j and position k) corresponding to y, f, and e, respectively.

Due to the whitening property of wavelet transforms [39], many existing methods assume independence of the wavelet coefficients and model the wavelet coefficients one by one using notation for a generic wavelet coefficient, $d = \theta + \varepsilon$. Shrinkage is performed term by term, which is sometimes referred to as diagonal shrinkage.

An early example of the diagonal Bayesian approach to wavelet regression is the adaptive Bayesian wavelet shrinkage (ABWS) proposed by Chipman et al. [27]. Their approach is based on the stochastic search variable selection (SSVS) proposed by George and McCulloch [41], with the assumption that σ is known.

Chipman et al. [27] start with the model

$$d|\theta, \sigma^2 \sim N(\theta, \sigma^2).$$

The prior on θ is defined as a mixture of two normals

$$\theta|\gamma_j \sim \gamma_j N(0, (c_j \tau_j)^2) + (1 - \gamma_j) N(0, \tau_j^2),$$

where

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

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,

$$\delta(d) = \left[P(\gamma_j = 1|d) \frac{(c_j\tau_j)^2}{\sigma^2 + (c_j\tau_j)^2} + P(\gamma_j = 0|d) \frac{\tau_j^2}{\sigma^2 + \tau_j^2} \right] d, \quad (14.9)$$

where

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

and

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

For other early examples of the Bayesian approach to wavelet regression see papers, for example, by Abramovich et al. [7, 28, 31, 85].

A more recent paper by Johnstone and Silverman [51] presents a class of empirical Bayes methods for wavelet shrinkage. The hyperparameters of the model are estimated by marginal maximum likelihood; therefore, the threshold is estimated from the data. The authors consider different level-dependent priors, all of which are a mixture of point mass at zero and a heavy-tailed density. One of the choices for the heavy-tailed density is the double exponential (Laplace) prior, for which we present the posterior mean to exemplify their methodology.

At level j of the wavelet decomposition, define the sequence $z_k = d_{jk}/\sigma_j$, where σ_j is the standard deviation of the noise at level j, which is estimated from the data. Therefore, $z_k = \mu_k + \varepsilon_k$, where the ε_k are i.i.d. $N(0,1)$ random variables. The authors model parameters μ_k with independent mixture prior distributions

$$\pi(\mu) = (1-w)\delta_0(\mu) + w\gamma(\mu),$$

where $\delta_0(\mu)$ denotes a point mass at zero. Using the double exponential distribution $\gamma_a(\mu) = \frac{1}{2}\exp\{-a|\mu|\}$, with scale parameter $a > 0$, the marginal distribution of z becomes

$$m(z) = (1-w)\varphi(z) + wg(z),$$

where φ denotes the standard normal density and

$$g(z) = \tfrac{1}{2} a \exp\left\{\tfrac{1}{2}a^2\right\}\left[\mathrm{e}^{-az}\Phi(z-a) + \mathrm{e}^{az}\tilde{\Phi}(z+a)\right].$$

In the above equation Φ denotes the cumulative distribution of the standard normal and $\tilde{\Phi} = 1 - \Phi$. The posterior distribution of μ becomes

$$\pi^{\star}(\mu|z) = (1 - w_{\text{post}})\delta_0(\mu) + w_{\text{post}} f_1(\mu|z),$$

where the posterior probability w_{post} is

$$w_{\text{post}}(z) = w g(z) \big/ \left[w g(z) + (1-w)\varphi(z)\right]$$

and

$$f_1(\mu|z) = \begin{cases} \mathrm{e}^{az}\varphi(\mu - z - a) \big/ \left[\mathrm{e}^{-az}\Phi(z-a) + \mathrm{e}^{az}\tilde{\Phi}(z+a)\right], & \mu \le 0 \\ \mathrm{e}^{-az}\varphi(\mu - z + a) \big/ \left[\mathrm{e}^{-az}\Phi(z-a) + \mathrm{e}^{az}\tilde{\Phi}(z+a)\right], & \mu > 0, \end{cases}$$

which is a weighted sum of truncated normal distributions. Detailed derivations of $g(z)$ and $f_1(\mu|z)$ are provided by Pericchi and Smith [72]. It can be shown that the posterior mean is

$$\mathbb{E}(\mu|z) = w_{\text{post}}(z)\left[z - \frac{a\left[\mathrm{e}^{-az}\Phi(z-a) - \mathrm{e}^{az}\tilde{\Phi}(z+a)\right]}{\mathrm{e}^{-az}\Phi(z-a) + \mathrm{e}^{az}\tilde{\Phi}(z+a)}\right]. \qquad (14.10)$$

A schematic picture of the posterior mean (14.10) is presented in Fig. 14.2 for $w = 0.1$ and $a = 0.5$. It exhibits a desirable shrinkage pattern slightly shrinking large and heavily shrinking small coefficients in magnitude.

The mixing weight w and scale parameter a are estimated by marginal maximum likelihood for each dyadic level j. The authors also provide the posterior median for the above model and closed-form equations for the posterior mean and median in case $\gamma(\mu)$ is a quasi-Cauchy distribution. For more details and related theoretical results the reader is referred to [51], and for more examples using the method, see [52].

Several more recent papers have considered term-by-term Bayesian wavelet shrinkage. Angelini and Sapatinas [10] consider an empirical Bayes approach to wavelet regression by eliciting the ε-contamination class of prior distributions and using type II maximum likelihood approach to prior selection. Angelini and Vidakovic [11] show that Γ-minimax shrinkage rules are Bayes with respect to a least favorable contamination prior with a uniform spread distribution. Their method allows for incorporation of information about the energy in the signal of interest. Cutillo et al. [32] consider thresholding rules induced by a variation of the Bayesian

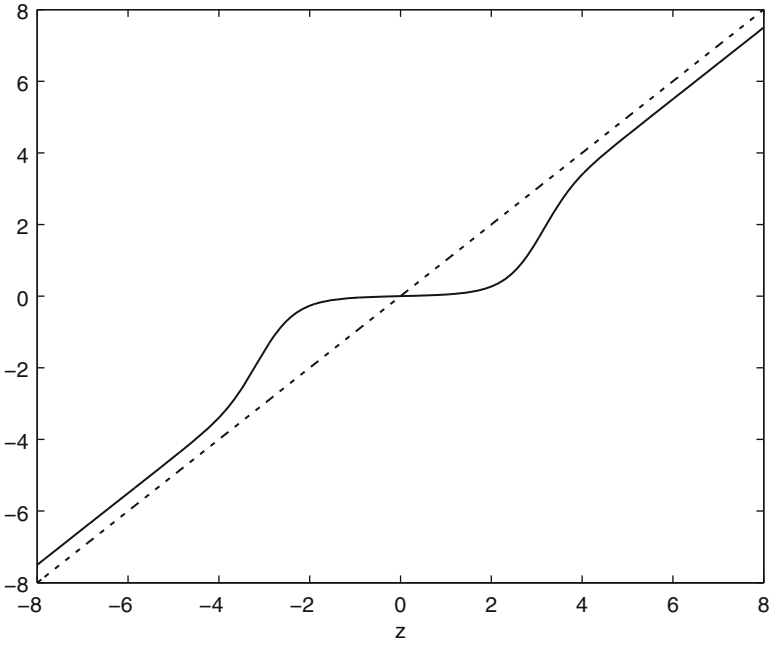

Fig. 14.2 Posterior mean rule (14.10) for $w = 0.1$ and $a = 0.5$

MAP principle in a properly set Bayesian model. The rule proposed is called larger posterior mode (LPM) because it always picks the mode of the posterior larger in absolute value. Ter Braak (2006) extends the normal Bayesian linear model by specifying a flat prior on the δth power of the variance components of the regression coefficients. In the orthonormal case, easy-to-compute analytic expressions are derived, and the procedure is applied in a simulation study of wavelet denoising.

14.3.2 Bayesian Block Shrinkage

Methods considered above are called diagonal, since the wavelet coefficients are assumed independent. In reality the wavelet coefficients are dependent, but this dependence is weak and decreases with increasing the separation distance between them and the number of vanishing moments of the decomposing wavelet. Many authors argued that shrinkage performance can be improved by considering the neighborhoods of wavelet coefficients (blocks, parent-child relations, cones of influence, etc.) and report improvements over the diagonal methods. Examples include classical block thresholding methods by Hall et al. [19–21, 44–46] where wavelet coefficients are thresholded based on block sums of squares.

Abramovich et al. [4] considered an empirical Bayes approach to incorporating information on neighboring wavelet coefficients into function estimation. The authors group wavelet coefficients d_{jk} into m_j nonoverlapping blocks b_{jK} $(K = 1,\ldots,m_j)$ of length l_j at each resolution level j. The block of observed wavelet coefficients will be denoted as $\hat{b}_{jK}$. They consider the following prior model for blocks b_{jK}:

$$b_{jK}|\gamma_{jK} \sim N(0, \gamma_{jK} V_j),$$
$$\gamma_{jK} \sim \text{Ber}(\pi_j).$$

Independence of blocks across different resolution levels is assumed. This prior model allows for a covariance structure between neighboring coefficients in the same block, supporting the fact that wavelet coefficients are more likely to contain signal if this is true for their neighbors as well. The covariance matrix V_j is specified at each level j by two hyperparameters τ_j and ρ_j, where the correlation between the coefficients, ρ_j, decreases as the distance between the coefficients increases. Combining the prior model with the likelihood $\hat{b}_{jK} \sim N(b_{jK}, \sigma^2 I)$ leads to the posterior mean of b_{jK} as

$$\mathbb{E}(b_{jK}|\hat{b}_{jK}) = \frac{1}{1+O_{jK}} A_j \hat{b}_{jK}, \tag{14.11}$$

where

$$O_{jK} = \frac{1-\pi_j}{\pi_j} \left(\frac{\det(V_j)}{\sigma^{2l_j} \det(A_j)} \right)^{1/2} \exp\left\{ -\frac{\hat{b}'_{jK} A_j \hat{b}_{jK}}{2\sigma^2} \right\},$$
$$A_j = (\sigma^2 V_j^{-1} + I)^{-1}.$$

Rule (14.11) is a nonlinear block shrinkage rule, by which the observed wavelet coefficients in block jK are shrunk by the same factor determined by all the coefficients within the block. The authors also provide details for the posterior median and the Bayes factor procedure, which are individual and block thresholding rules, respectively.

Hyperparameters π_j, τ_j, and ρ_j are estimated by marginal maximum likelihood method for each level j, and hyperparameter σ is estimated by the standard median absolute deviation suggested by Donoho and Johnstone [35]. After plugging in the estimate $\hat{\sigma}$ and some reparametrization, the negative log-likelihood function $-l_j(\pi_j, \tau_j, \rho_j, \hat{\sigma})$ was minimized by the Nelder–Mead simplex search method.

The authors present detailed simulation study of the method and an application to inductance plethysmography data. For details the reader is referred to [4].

A paper by De Canditiis and Vidakovic [34] proposed the BBS (Bayesian block shrinkage) method, which also allows for dependence between the

wavelet coefficients. The modeling is accomplished by using a mixture of two normal-inverse-gamma (NIG) distributions as a joint prior on wavelet coefficients and noise variance within each block. In this sense it is a generalization of the ABWS method by Chipman et al. [27]. The authors group the wavelet coefficients into nonoverlapping, mutually independent blocks $\boldsymbol{d}_{jH}$ of size l_j. Assuming a normal likelihood $\boldsymbol{d}_{jH} \sim N(\boldsymbol{\theta}_{jH}, \sigma^2 I)$, the prior model is specified as

$$\begin{aligned} \boldsymbol{\theta}_{jH}, \sigma^2 | \gamma_j &\sim \gamma_j NIG(\alpha, \delta, \mathbf{0}, \Sigma_j) + (1-\gamma_j) NIG(\alpha, \delta, \mathbf{0}, \Delta_j), \\ \gamma_j &\sim Ber(p_j), \end{aligned}$$

where the covariance matrices are specified as $\Sigma[s,t] = c_j^2 \rho^{|s-t|}$ and $\Delta[s,t] = \tau_j^2 \rho^{|s-t|}$, which is in the same fashion as in [4]. The first part of the above mixture prior models wavelet coefficients with large magnitude ($c_j \gg 1$) and the second part captures small coefficients (τ_j is small), similarly to the ABWS method. The posterior distribution for the model above remains a mixture of NIG distribution with mixing weights updated by the observed wavelet coefficients. The posterior and marginal distributions are derived in the paper. The posterior mean of $\boldsymbol{\theta}_{jH}$ becomes

$$\mathbb{E}(\boldsymbol{\theta}_{jH} | \boldsymbol{d}_{jH}) = A_{jH}(\boldsymbol{d}_{jH}) \boldsymbol{m}^{\star}_{jH} + (1 - A_{jH}(\boldsymbol{d}_{jH})) \boldsymbol{m}^{\star\star}_{jH}, \tag{14.12}$$

where

$$A_{jH}(\boldsymbol{d}_{jH}) = \frac{p_j \frac{|\Sigma_j^{\star}|^{1/2}}{|\Sigma_j|^{1/2}}}{p_j \frac{|\Sigma_j^{\star}|^{1/2}}{|\Sigma_j|^{1/2}} + (1-p_j) \frac{|\Delta_j^{\star\star}|^{1/2}}{|\Delta_j|^{1/2}} + \left[\frac{\alpha + \boldsymbol{d}_{jH}^T (I - \Delta_j^{\star\star}) \boldsymbol{d}_{jH}}{\alpha + \boldsymbol{d}_{jH}^T (I - \Sigma_j^{\star}) \boldsymbol{d}_{jH}} \right]^{-(\delta + l_j)/2}}$$

and

$$\begin{aligned} \Sigma_j^{\star} &= (\Sigma_j^{-1} + I)^{-1}, \\ \Delta_j^{\star\star} &= (\Delta_j^{-1} + I)^{-1}, \\ \boldsymbol{m}^{\star}_{jH} &= \Sigma_j^{\star} \boldsymbol{d}_{jH}, \\ \boldsymbol{m}^{\star\star}_{jH} &= \Delta_j^{\star\star} \boldsymbol{d}_{jH}. \end{aligned}$$

The posterior mean (14.12) is a linear combination of two affine shrinkage estimators $\boldsymbol{m}^{\star}_{jH}$ and $\boldsymbol{m}^{\star\star}_{jH}$, which preserve the smooth part and remove the noise, respectively. The weight $A_{jH}(\boldsymbol{d}_{jH})$ depends on the observed wavelet coefficients in a nonlinear fashion. For more details on hyperparameter selection, simulations, and performance the reader is referred to [34].

Huerta [47] proposed a multivariate Bayes wavelet shrinkage method which allows for correlations between wavelet coefficients corresponding to the same level of detail. The paper assumes the multivariate normal likelihood for the observed wavelet coefficients, that is,

$$d|\theta,\sigma^2 \sim N(\theta,\sigma^2 I_n).$$

Note that the wavelet coefficients are not grouped into blocks, as opposed to the methods discussed before. The prior structure is specified as

$$\begin{aligned}\theta|\tau^2 &\sim N(0,\tau^2\Sigma),\\ \sigma^2 &\sim \mathrm{IG}(\alpha_1,\delta_1),\\ \tau^2 &\sim \mathrm{IG}(\alpha_2,\delta_2),\end{aligned}$$

where Σ is an $n\times n$ matrix defining the prior correlation structure among wavelet coefficients. The matrix is specified as a block diagonal matrix, where each block defines the correlation structure for different wavelet decomposition level. The building blocks of matrix Σ are defined in the same way as in the methods discussed above.

Since there is no closed-form expression for the marginal posterior $\pi^\star(\theta|d)$, a standard Gibbs sampling procedure is adopted to obtain posterior inferences on the vector of wavelet coefficients d. For further details and applications of the method the reader is referred to [47].

Wang and Wood [93] considered a different approach for Bayesian block shrinkage, based directly on the block sum of squares. The sum of squares of the coefficients in the block forms a noncentral chi-square random variable, on which the Bayesian model is formulated. Let $\hat{c}_B$ denote the block of empirical wavelet coefficients, B representing the labels and $n(B)$ the number of labels, in general. Then the assumed likelihood function is $\hat{c}_B \sim N_{n(B)}(c_B,\sigma^2 I_{n(B)})$. Define $z=\|\hat{c}_B\|^2=\Sigma_{i\in B}\hat{c}_i^2$, the sum of squares of the coefficients in the block. It follows that $z\sim\chi^2_m(z|\rho,\sigma^2)$, that is, z has noncentral χ^2 distribution with $m=n(B)$ degrees of freedom, noncentrality parameter $\rho=\|c_B\|^2$, and scale parameter σ^2. The authors formulate the prior model on the noncentrality parameter as

$$\begin{aligned}\rho|\beta &\sim \chi^2_m(\rho|0,\beta^{-1}),\\ \beta|\sigma^2,\theta &\sim F(\beta|\sigma^2,\theta).\end{aligned}$$

In other words this specifies a central χ^2 density with m degrees of freedom and scale parameter β^{-1} as a prior for ρ and specifies a prior for β with cumulative distribution function $F(\beta|\sigma^2,\theta)$. Their article focuses on a mixture structure

$$F(\beta|\sigma^2,\theta)=pF(\beta|\sigma^2,\lambda,J=1)+(1-p)F(\beta|\sigma^2,\lambda,J=0),$$

where

$$F(\beta|\sigma^2,\lambda,J=1)=I_{\{\beta=\infty\}}(\beta).$$

Here J is a Bernoulli random variable, with $J=0$ corresponding to a distribution on the right side of the mixture, and $J=1$ referring to a point mass at infinity distribution. Using an identity satisfied by the noncentral χ^2 density the authors provide closed-form equations for the marginal distribution and the posterior mean of ρ for the model setup above. The equations are the function of $F(\beta|\sigma^2,\lambda,J=0)$, which is to be specified. The authors consider four particular cases of this prior, the point mass prior, the power prior, the exponential prior, and general discrete prior. For the power prior—on which the paper focuses on—the marginal distribution and posterior mean of ρ is derived as

$$f\left(z|\sigma^2,\theta\right)=p\chi^2_m\left(\rho|0,\sigma^2\right)+(1-p)\frac{(\lambda+1)\left(2\sigma^2\right)^{\lambda+1}}{\Gamma\left(\frac{1}{2}m\right)z^{\lambda+2}}\gamma\left(\eta,\frac{z}{2\sigma^2}\right),$$

$$\mathbb{E}\left(\rho|z,\sigma^2,\theta\right)=(1-\pi)\left\{m\sigma^2+z-\frac{m\sigma^2+2z}{z/\left(2\sigma^2\right)}\mathscr{C}_{\eta,1}\left(\frac{z}{2\sigma^2}\right)+\frac{4\sigma^4}{z}\mathscr{C}_{\eta,2}\left(\frac{z}{2\sigma^2}\right)\right\},$$

where

$$\begin{aligned}\pi&=\frac{p\chi^2_m(\rho|0,\sigma^2)}{f(z|\sigma^2,\theta)},\\ \mathscr{C}_{\eta,j}(x)&=\gamma(\eta+j,x)/\gamma(\eta,x),\\ \eta&=1+\lambda+\frac{1}{2}m,\\ \gamma(a,x)&=\int_0^x t^{a-1}\mathrm{e}^{-t}\mathrm{d}t.\end{aligned}$$

Hyperparameter σ^2 is estimated analogously to the median absolute deviation estimator suggested by Donoho and Johnstone [35], hyperparameter λ is estimated by a "quick-and-dirty" heuristics, and finally hyperparameter p is estimated by marginal maximum likelihood. Given values of hyperparameters σ^2 and $\theta=(p,\lambda)$, the authors propose to estimate wavelet coefficients c_B by the shrinkage procedure

$$c_B=\hat{c}_B\{\mathscr{B}_{\sigma^2,\theta}(z)/z\}^{\frac{1}{2}}, \tag{14.13}$$

where $\mathscr{B}_{\sigma^2,\theta}(z)$ denotes the posterior mean or posterior median of ρ. The authors report good MSE results based on simulations on well-known test functions. For more details the reader is referred to [93].

There is a wide range of other articles considering Bayesian modeling of neighboring wavelet coefficients. To name a few, [76] use a Bayesian hidden Markov tree (HMT) to model the structure of wavelet coefficients in images. Jansen and Bultheel [48] introduce a geometrical prior model for configurations of wavelet coefficients and combine this with local characterization of a classical thresholding into a Bayesian framework. Sendur and Selesnick [80] use parent–child neighboring relation and Laplacian bivariate prior to derive MAP estimators for wavelet coefficients. Pižurica et al. [73] use a Markov random field (MRF) prior model to incorporate inter- and intrascale dependencies of wavelet coefficients. Portilla et al. [74] models neighborhoods of image wavelet coefficients at adjacent positions and scales using scale mixtures of Gaussians.

A recent non-Bayesian development was proposed by Fryzlewitz [40] in a form of fast, hard-thresholding algorithm based on coupling parents and children in the wavelet coefficient tree.

14.3.3 Complex Wavelet Shrinkage

Wavelet shrinkage methods using complex-valued wavelets provide additional insights to shrinkage process. Lina and Mayrand [61] describes the complex-valued Daubechies' wavelets in detail. Both complex- and real-valued Daubechies' wavelets are indexed by the number of vanishing moments, N. For a given N, there are 2^{N-1} solutions to the defining equations of Daubechies' wavelets, of which not all are distinct. For example, in case $N = 3$, there are four possible solutions to the defining equations, but only two are distinct. Two solutions give the real-valued extremal-phase wavelet and the other two are a complex-valued conjugate pair, giving equivalent complex-valued wavelets. This complex wavelet was also derived by Lawton [56] through "zero-flipping"; he notes that apart from the Haar wavelet, complex wavelets with an odd number of vanishing moments are the only compactly supported wavelets which are symmetric. The complex-valued wavelet transformation can also be represented by a complex-valued matrix W, which is unitary; therefore, $\bar{W}^T W = W\bar{W}^T = I$. Here $\bar{W}$ denotes the complex conjugate of W.

After taking complex wavelet transformation of a real-valued signal, our model becomes

$$d_{jk} = \theta_{jk} + \varepsilon_{jk},$$

where the observed wavelet coefficients d_{jk} are complex numbers at resolution j and location k.

Several papers considering Bayesian wavelet shrinkage with complex wavelets are available. For example, [59, 60, 62] focus on image denoising, in which the phase of the observed wavelet coefficients is preserved, but the modulus of the coefficients is shrunk by the Bayes rule.

Here we summarize the complex empirical Bayes (CEB) procedure proposed by Barber and Nason [14], which modifies both the phase and modulus of wavelet

coefficients by a bivariate shrinkage rule. The authors assume a common i.i.d. normal noise model $\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2 I_n)$; however, after taking complex wavelet transform, the real and imaginary parts of the transformed noise $\boldsymbol{\varepsilon} = W\boldsymbol{e}$ become correlated. The authors demonstrate that

$$\begin{aligned}
\mathrm{cov}\{\mathrm{Re}(\boldsymbol{\varepsilon}), \mathrm{Im}(\boldsymbol{\varepsilon})\} &= -\sigma^2 \mathrm{Im}(WW^T)/2, \\
\mathrm{cov}\{\mathrm{Re}(\boldsymbol{\varepsilon}), \mathrm{Re}(\boldsymbol{\varepsilon})\} &= \sigma^2 \{I_n + \mathrm{Re}(WW^T)\}/2, \\
\mathrm{cov}\{\mathrm{Im}(\boldsymbol{\varepsilon}), \mathrm{Im}(\boldsymbol{\varepsilon})\} &= \sigma^2 \{I_n - \mathrm{Re}(WW^T)\}/2. \qquad (14.14)
\end{aligned}$$

Representing the complex-valued wavelet coefficients as a bivariate real-valued random variables, the model for the observed wavelet coefficients becomes

$$d_{jk}|\theta_{jk} \sim N_2(\theta_{jk}, \Sigma_j),$$

where Σ_j is determined by (14.14) for each dyadic level j. Noise variance σ^2 is estimated by the usual median absolute deviation by Donoho and Johnstone [35].

The authors consider a bivariate mixture prior of the form

$$\theta_{jk} \sim p_j N_2(\mathbf{0}, V_j) + (1 - p_j)\delta_0,$$

where δ_0 is the usual point mass probability at $(0,0)^T$. This prior is the bivariate extension of the prior considered by Abramovich et al. [7]. Conjugacy of the normal distribution results in the posterior distribution

$$\theta_{jk}|d_{jk} \sim \tilde{p}_{jk} N_2(\mu_{jk}, \tilde{V}_j) + (1 - \tilde{p}_{jk})\delta_0,$$

where

$$\begin{aligned}
\tilde{p}_{jk} &= \frac{p_j f(d_{jk}|p_j = 1)}{p_j f(d_{jk}|p_j = 1) + (1 - p_j) f(d_{jk}|p_j = 0)}, \\
f(d_{jk}|p_j = 1) &= \frac{1}{2\pi\sqrt{|V_j + \Sigma_j|}} \exp\left\{-\frac{1}{2} d_{jk}^T (V_j + \Sigma_j)^{-1} d_{jk}\right\}, \\
f(d_{jk}|p_j = 0) &= \frac{1}{2\pi\sqrt{|\Sigma_j|}} \exp\left\{-\frac{1}{2} d_{jk}^T \Sigma_j^{-1} d_{jk}\right\}, \\
\tilde{V}_j &= \left(V_j^{-1} + \Sigma_j^{-1}\right)^{-1} \text{ and } \mu_{jk} = \tilde{V}_j \Sigma_j^{-1} d_{jk}.
\end{aligned}$$

The posterior mean of θ_{jk} becomes

$$\mathbb{E}(\theta_{jk}) = \tilde{p}_{jk}\mu_{jk}, \qquad (14.15)$$

which is denoted as "CEB-Posterior mean." The authors consider two additional estimation rules, the phase-preserving "CEB-Keep or kill" and the hybrid "CEB-MeanKill" procedure.

Estimation of the prior parameters p_j and V_j is employed by the data-driven empirical Bayes approach maximizing the logarithm of the marginal likelihood. However, optimizing the bivariate likelihood is more involved because we have more parameters compared to the real-valued case.

Barber and Nason [14] present an extensive simulation study of the CEB method alongside with the phase-preserving CMWS hard-thresholding method also developed in their paper. Simulations show that complex-valued denoising is very effective and dominates existing real-valued wavelet shrinkage methods.

14.3.4 Complex Wavelet Shrinkage via Gibbs Sampling

In this section, we describe a new adaptive wavelet denoising methodology using complex wavelets. The method is based on a fully Bayesian hierarchical model that uses a bivariate mixture prior. The crux of the procedure is computational in which the posterior mean is computed through MCMC simulations.

We build on the results of [14] and formulate a bivariate model in the complex wavelet domain, representing the wavelet coefficients as bivariate real-valued random variables. As standardly done in Bayesian modeling, we formulate a hierarchical model which accounts for the uncertainty of the prior parameters by adopting hyperpriors on them. Since a closed-form solution to the Bayes estimator does not exist, MCMC methodology is applied and an approximate estimator (posterior mean) from the output of simulational runs is computed. Although the simplicity of a closed-form solution is lost, the procedure is fully Bayesian, adaptive to the underlying signal and the estimation of the hyperparameters is automatic via the MCMC sampling algorithm. The estimation is governed by the data and hyperprior distributions on the parameters.

We start with the following hierarchical bivariate Bayesian model on the observed complex-valued wavelet coefficients d_{jk}:

$$\begin{aligned} d_{jk}|\theta_{jk},\sigma^2 &\sim N_2(\theta_{jk},\sigma^2\Sigma_j), \\ \theta_{jk}|\varepsilon_j,C_j &\sim (1-\varepsilon_j)\delta_0+\varepsilon_j EP_2(\mu,C_j,\beta), \end{aligned} \tag{14.16}$$

where EP_2 denotes the bivariate exponential power distribution. The multivariate exponential power distribution is an extension of the class of normal distributions in which the heaviness of tails can be controlled. Its definition and properties can be found in [42]. The prior on the location θ_{jk} is a bivariate extension of the standard mixture prior in the Bayesian wavelet shrinkage literature, consisting of a point mass at zero and a heavy-tailed distribution. As a prior, [14] considered a mixture of point

mass and bivariate normal distribution. A heavy-tailed mixture prior can probably better capture the sparsity of wavelet coefficients; however, in the bivariate case, a closed-form solution is infeasible, and we rely on MCMC simulation.

To specify the general case exponential power prior in (14.16), we use $\mu = 0$, because the wavelet coefficients are centered around zero by their definition. We also fix $\beta = 1/2$, which gives our prior the following form:

$$\pi(\theta|C) = \frac{1}{8\pi|C|^{1/2}} \exp\left\{-\frac{1}{2}\left(\theta' C^{-1}\theta\right)^{1/2}\right\}. \tag{14.17}$$

The prior in (14.17) is equivalent to the bivariate double exponential distribution. The univariate double exponential prior was extensively used in the real-valued wavelet context, hence it is natural to extend it to the bivariate case.

From model (14.16) it is apparent that the mixture prior on θ_{jk} is set level-wise, for each dyadic level j, which ensures the adaptivity of the method. Quantity $\sigma^2\Sigma_j$ represents the scaled covariance matrix of the noise for each decomposition level, and C_j represents the level-wise scale matrix in the exponential power prior. Explicit expression for the covariance (Σ_j) induced by white noise in complex wavelet shrinkage can be found in [14] and mentioned above in (14.14). We adopt the approach described in their paper to model the covariance structure of the noise.

Instead of estimating hyperparameters σ^2, ε_j, and C_j, we specify hyperprior distributions on them in a fully Bayesian manner. We specify a conjugate inverse gamma prior on the noise variance σ^2 and an inverse-Wishart prior on the matrix C_j describing the covariance structure of the spread prior of θ_{jk}. Mixing weight ε_j regulates the strength of shrinkage of a wavelet coefficient to zero. We specify a "noninformative" uniform prior on this parameter, allowing the estimation to be fully governed by the data.

For computational purposes, we represent our exponential power prior as a scale mixtures of multivariate normal distributions, which is an essential step for efficient Monte Carlo simulation. From [43], the bivariate exponential power distribution with $\mu = 0$ and $\beta = 1/2$ can be represented as

$$EP_2(\mu = 0, C_j, \beta = 1/2) = \int_0^\infty N_2(0, vC_j) \frac{1}{\Gamma(3/2)8^{3/2}} v^{1/2} \mathrm{e}^{-v/8} \mathrm{d}v,$$

which is a scale mixtures of bivariate normal distributions with mixing distribution gamma. Using the specified hyperpriors and the mixture representation, the model in (14.16) extends to

$$\begin{aligned} d_{jk}|\theta_{jk}, \sigma^2 &\sim N_2(\theta_{jk}, \sigma^2\Sigma_j), \\ \sigma^2 &\sim \mathrm{IG}(a, b), \\ \theta_{jk}|z_{jk}, v_{jk}, C_j &\sim (1 - z_{jk})\delta_0 + z_{jk} N_2(0, v_{jk}C_j), \end{aligned}$$

$$\begin{aligned} z_{jk}|\varepsilon_j &\sim \text{Ber}(\varepsilon_j), \\ \varepsilon_j &\sim U(0,1), \\ v_{jk} &\sim \text{Ga}(3/2,8), \\ C_j &\sim \text{IW}(A_j,w). \end{aligned} \tag{14.18}$$

Note that, for computational purposes, we also introduced a latent variable z_{jk} in the above model. Variable z_{jk} is a Bernoulli variable indicating whether our parameter θ_{jk} comes from a point mass at zero ($z_{jk} = 0$) or from a bivariate normal distribution ($z_{jk} = 1$). By representing the exponential power prior as a scale mixtures of normals, the hierarchical model in (14.18) becomes tractable, because the full conditional distributions of all the parameters become explicit. Therefore, we can develop a Gibbs sampling algorithm to update all the necessary parameters. We used the sample average $\hat{\theta}_{jk} = \Sigma_i \theta_{jk}^{(i)}/N$ of the simulational runs, as the standard estimator for the posterior mean. To apply the Gibbs sampling algorithm we only need to specify hyperparameters a, b, A_j, and w, which influence lower level of the hierarchical model. The rest of the parameters are updated via the Gibbs sampling procedure. The method is called complex Gibbs sampling wavelet smoother (CGSWS). For more details about the implementation, contact the authors.

Application to Inductance Plethysmography Data

For illustration we apply the described CGSWS method to a real-world data set from anesthesiology collected by inductance plethysmography. The recordings were made by the Department of Anaesthesia at the Bristol Royal Infirmary and represent measure of flow of air during breathing. The data set was analyzed by several authors, for example, [4, 7, 66]. For more information about the data, refer to these papers.

The top part of Fig. 14.3 shows a section of plethysmograph recording lasting approximately 80 s ($n = 4{,}096$ observations), while the bottom part shows the reconstruction of the signal with the CGSWS method. In the reconstruction process we applied $N = 5{,}000$ iterations of the Gibbs sampler of which the first 2,000 was burn-in. The aim of smoothing is to preserve features such as peak heights while eliminating spurious rapid variation. The result provided by the proposed method satisfies these requirements providing a very smooth result. Abramovich et al. [4] report the heights of the first peak while analyzing this data set. In our case the height is 0.8389, which is quite close to the result 0.8433, obtained by Abramovich et al. [4], and better compared to the results obtained by other established methods analyzed in their paper.

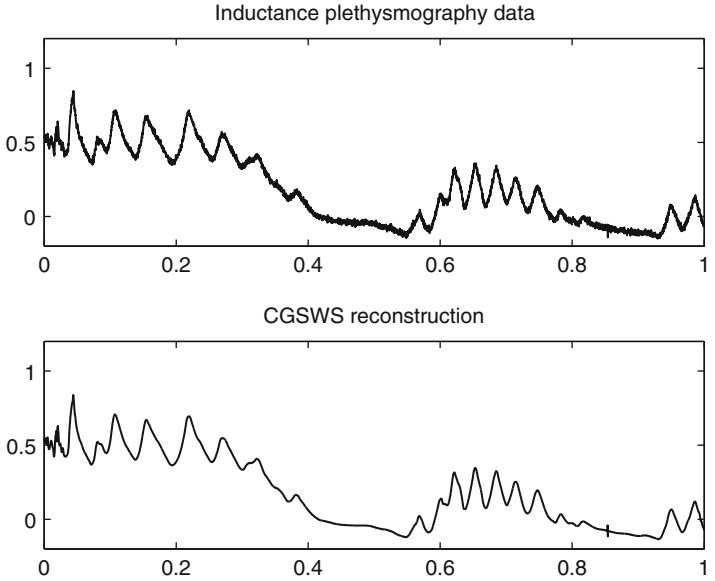

Fig. 14.3 Reconstruction of the (*IPD*) inductance plethysmography data by CGSWS

14.3.5 Bayesian Wavelet Shrinkage in Curve Classification

We consider the paper by Wang et al. [92] to give an application of Bayesian wavelet shrinkage in curve classification. The authors consider Bayesian wavelet-based classification models for binary and multicategory data where the predictor is a random function.

Functional data analysis deals with the analysis of data sets where the units are curves that are ordered measurements on a regular grid. Functional data is frequently encountered in scientific research. Classification of functional data is a relatively new problem, and there are several approaches, from using simple summary quantiles to nonparametric methods using splines. Wang et al. [92] propose a Bayesian wavelet-based classification method, because wavelets are known to have nice properties for representing a wide range of functional spaces including functions with sharp-localized changes. The proposed method unifies wavelet-based regression with logistic classification models, representing functional data using wavelet basis functions and using the wavelet coefficients for classification within a logistic model.

Consider data set $\{\boldsymbol{Y_i}, z_i\},\ i = 1, \ldots, n$, where $\boldsymbol{Y_i}$ is a vector of m measurements and z_i is a binary classification variable. We represent the vector of measurements as $\boldsymbol{Y_i} = \boldsymbol{f_i} + \boldsymbol{\varepsilon_i}$, where $\boldsymbol{f_i}$ is an underlying nonparametric function and $\boldsymbol{\varepsilon_i} \sim N(0, \sigma^2 I)$. Representing functions $\boldsymbol{f_i}$ in wavelet basis we get $\boldsymbol{Y_i} = \boldsymbol{X}\boldsymbol{\beta_i} + \boldsymbol{\varepsilon_i}$, where $\boldsymbol{X}$ is the discrete wavelet transformation matrix and $\boldsymbol{\beta_i}$ is the vector of wavelet coefficients.

The authors consider the following unified hierarchical Bayesian model for wavelet regression and classification:

$$\begin{aligned}
\text{Random function } \boldsymbol{Y_i} &\sim N(\boldsymbol{X\beta_i}, \sigma^2 I),\\
\boldsymbol{\beta_i}, \sigma^2 | \boldsymbol{\eta_i}, \boldsymbol{g} &\sim \text{NIG}(0, \text{diag}(\boldsymbol{\eta_i})\text{diag}(\boldsymbol{g}), a_\sigma, b_\sigma),\\
g_j &\sim \text{IG}(u_j, v_j),\\
\eta_{ijk} &\sim \text{Ber}(\rho_j).\\
\text{Binary outcome } z_i &\sim \text{Ber}(p_i),\\
T_i &\sim N(\boldsymbol{\beta_i^t\theta}, \tau^2), \quad \text{where } T_i = \text{logit}(p_i),\\
\boldsymbol{\theta}, \tau^2 | \boldsymbol{\gamma}, \boldsymbol{h} &\sim \text{NIG}(0, \text{diag}(\boldsymbol{\gamma})\text{diag}(\boldsymbol{h}), a_\tau, b_\tau),\\
h_j &\sim \text{IG}(c_j, d_j),\\
\gamma_{jk} &\sim \text{Ber}(\pi_j),
\end{aligned} \tag{14.19}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, \log_2 m$, and $k = 0, \ldots, 2^j - 1$.

The first part in (14.19) is a model for the observed random functions $\boldsymbol{Y_i}$, where variable selection priors for the wavelet coefficients are adopted from the Bayesian wavelet modeling literature similar to [34]. Parameter g_j is a scaling parameter, and parameter η_{ijk} is the usual latent indicator variable to model the sparsity of the wavelet representation. The second part in (14.19) is a classification model for variable $z_i \in \{0, 1\}$ taking unit value with unknown probability p_i. The logistic classification model relates the wavelet coefficients $\boldsymbol{\beta_i}$ to the latent variable $T_i = \text{logit}(p_i)$ through a linear model $T_i = \boldsymbol{\beta_i^t\theta} + \delta_i$, where $\delta_i \sim N(0, \tau^2)$ and where $\boldsymbol{\theta}$ is a vector of regression coefficients. Similar variable selection prior for $\boldsymbol{\theta}$ is assumed as for $\boldsymbol{\beta_i}$ to reduce the dimensionality of the problem.

For functional data with binary outcomes the model in (14.19) is an extension of a standard classification model with an additional layer of functional regression model. Because the posterior distribution of the parameters is not available in a standard form, posterior inference has to rely on MCMC methods. Wang et al. [92] derive the full conditional distributions for the parameters, which allow for implementation of a Gibbs sampling algorithm. The model in (14.19) is also extended to multicategory classification by the authors.

14.3.5.1 Application to Leaf Data

Wang et al. [92] analyzed a data set from [53] that contains leaf images of six different species. The data was converted into a pseudo-time series by measuring local angle and trace of the leaf images. For a purpose of binary classification analysis one maple (*Circinatum*) and one oak (*Garryana*) species were selected with 150 instances. Example curves adopted from [92] can be seen in Fig. 14.4.

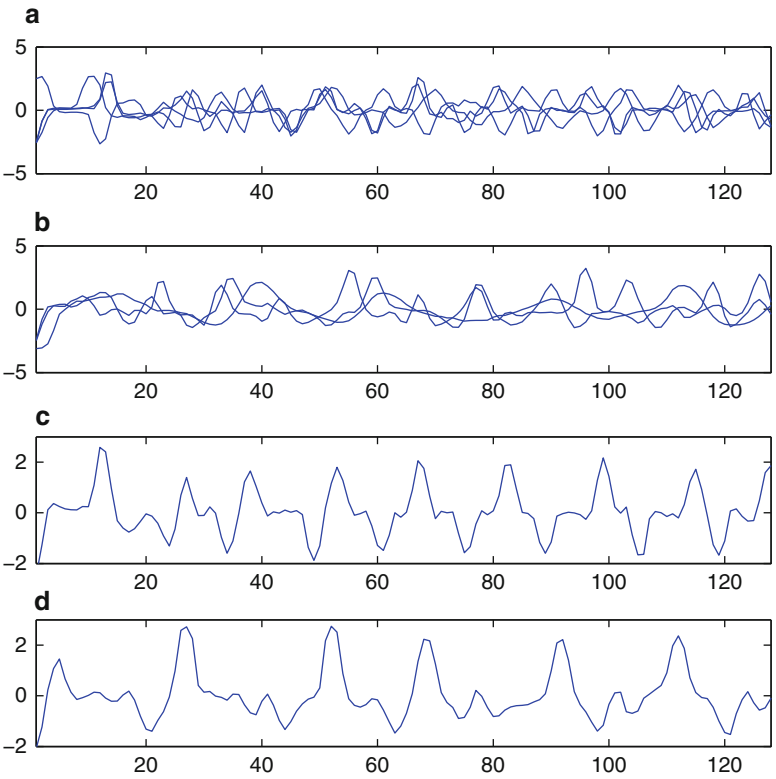

Fig. 14.4 Adopted from [92]: "Pseudo-time series curves from leaf images. (**a**) and (**b**) Every other curve in two species in the data set, 33 of *Circinatum* and 42 of *Garryana*. (**c**) and (**d**) Example of single curve from two species, *Circinatum* and *Garryana*"

The classification was carried out by randomly selecting 140 curves from the training and ten curves from the testing set. This was repeated 20 times, and the correct classification rate (CCR) was reported. The proposed wavelet-based classification method had CCR=94% and outperformed all other methods considered, including empirical Bayes thresholding plugged into a support vector machine (SVM) classifier. The authors carried out analysis for other existing and simulated data sets, including nonequispaced and multicategory data, and reported good performance. For more details the reader is referred to [92].

14.3.6 Related Work

There are numerous papers related to wavelet shrinkage and wavelet regression. Here we list some additional references related to the topics discussed in this chapter, as a repository for researchers interested in the area.

For related overview summaries about wavelet methods see [3, 12, 67], for example. An excellent critical overview and simulation study comparing different wavelet shrinkage methods can be found in [13]. Articles focusing only on Bayesian wavelet-based modeling include [65, 78, 86].

Some recent results about theoretical properties and optimality of Bayesian wavelet estimators can be found in [1, 2, 16, 17, 51, 70, 71].

There are several papers on Bayesian wavelet estimation in the signal and image processing community. These papers usually specify a single, nonmixture prior on the wavelet coefficients and compute a Bayes estimator. Posterior mode is a popular choice, which is used, for example, by Figueiredo and Nowak [38, 64], who use generalized Gaussian and complexity priors to model wavelet coefficients. Other articles in this group include [18] using approximate α-stable prior, [23] using generalized Gaussian distribution (GCD) as a prior, [37] using Bessel K forms (BKF) densities, and [58] using Besov norm priors for modeling wavelet coefficients. Achim and Kuruoğlu [8] develop a bivariate maximum a posteriori estimator using a bivariate α-stable distribution to model wavelet coefficients in the complex wavelet domain.

Some non-Bayesian improvements related to block thresholding include [20, 22, 24–26, 36], to name a few. More general theoretical results about block empirical Bayes estimation appear in [95].

All Bayesian estimators depend on hyperparameters that have to be specified. Purely subjective elicitation is only possible when considerable knowledge about the underlying signal is available. The empirical Bayes method is an efficient, completely data-driven procedure to estimate the hyperparameters based on marginal maximum likelihood method. Several papers in the literature used this method to estimate hyperparameters of the model. For more information about the method see, for example, papers by Clyde and George [29, 30, 50, 51].

The usual assumptions for wavelet regression are equispaced sampling points with a sample size being a power of 2, i.i.d. normal random errors with zero mean and constant variance. Extension of these assumptions has been considered in several articles. To name a few non-Bayesian procedures, [49] consider wavelet thresholding with stationary correlated noise, and [55] extend wavelet thresholding to irregularly spaced data, to equally spaced data sets of arbitrary size, to heteroscedastic and correlated data, and to data which contains outliers. An early example of a Bayesian wavelet shrinkage method incorporating theoretical results on the covariance structure of wavelet coefficients is by Vannucci and Corradi [84]. Ambler and Silverman [9] 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. Wang and Wood [94] consider a Bayesian wavelet shrinkage method which includes both time and wavelet domain methods to estimate the correlation structure of the noise and a Bayesian block shrinkage procedure based on [93]. Ray and Mallick [75] develop a Bayesian

wavelet shrinkage method to accommodate broad class of noise models for image processing applications. The method is based on the Box-Cox family of power transformations.

Kohn et al. [54] develop a wavelet shrinkage method which incorporates a Bayesian approach for automatically choosing among wavelet bases and averaging of the regression function estimates over different bases.

Barber et al. [15, 79] derive Bayesian credible intervals for Bayesian wavelet regression estimates based on cumulants and saddlepoint approximation, respectively.

Olhede and Walden [69] discuss an "analytic" wavelet thresholding which incorporates information from the discrete Hilbert transform of the signal, creating a complex-valued "analytic" vector. A recent paper describing a data-adaptive thresholding by controlling the false discovery rate (FDR) is by Abramovich et al. [5]. A Bayesian interpretation of the FDR procedure and application to wavelet thresholding can be found in [82].

Application of the Bayesian maximum a posteriori multiple testing (testimation) procedure to wavelet thresholding can be found in [6].

References

1. Abramovich F, Amato U, Angelini C (2004) On optimality of Bayesian wavelet estimators. Scand J Stat 31:217–234
2. Abramovich F, Angelini C, De Canditiis D (2007) Pointwise optimality of Bayesian wavelet estimators. Ann Inst Stat Math 59:425–434
3. Abramovich F, Bailey TC, Sapatinas T (2000) Wavelet analysis and its statistical applications. Statist 49:1–29
4. Abramovich F, Besbeas P, Sapatinas T (2002) Empirical Bayes approach to block wavelet function estimation. Comput Stat Data Anal 39:435–451
5. Abramovich F, Benjamini Y, Donoho DL, Johnstone IM (2006) Adapting to unknown sparsity by controlling the false discovery rate. Ann Stat 34:584–653
6. Abramovich F, Grinshtein V, Petsa A, Sapatinas T (2010) On Bayesian "testimation" and its application to wavelet thresholding. Biometrika 97:181–198
7. Abramovich F, Sapatinas T, Silverman BW (1998) Wavelet thresholding via Bayesian approach. J Roy Stat Soc Ser B 60:725–749
8. Achim A, Kuruoğlu EE (2005) Image denoising using bivariate α-stable distributions in the complex wavelet domain. IEEE Signal Process Lett 12:17–20
9. Ambler GK, Silverman BW (2004) Perfect simulation for wavelet thresholding with correlated coefficients. In: Technical Report 04:01. Department of Mathematics, University of Bristol
10. Angelini C, Sapatinas T (2004) Empirical Bayes approach to wavelet regression using ε-contaminated priors. J Stat Comput Simul 74:741–764
11. Angelini C, Vidakovic B (2004) Γ-minimax wavelet shrinkage: a robust incorporation of information about energy of a signal in denoising applications. Stat Sinica 14:103–125
12. Antoniadis A (2007) Wavelet methods in statistics: some recent developments and their applications. Stat Surv 1:16–55
13. Antoniadis A, Bigot J, Sapatinas T (2001) Wavelet estimators in nonparametric regression: a comparative simulation study. J Stat Softw 6:1–83
14. Barber S, Nason GP (2004) Real nonparametric regression using complex wavelets. J Roy Stat Soc Ser B 66:927–939

15. Barber S, Nason GP, Silverman BW (2002) Posterior probability intervals for wavelet thresholding. J Roy Stat Soc Ser B 64:189–205
16. Bochkina N, Sapatinas T (2006) On pointwise optimality of Bayes factor wavelet regression estimators. Sankhyā 68:513–541
17. Bochkina N, Sapatinas T (2009) Minimax rates of convergence and optimality of Bayes factor wavelet regression estimators under pointwise risks. Stat Sinica 19:1389–1406
18. Boubchir L, Fadili JM (2006) A closed-form nonparametric Bayesian estimator in the wavelet domain of images using an approximate α-stable prior. Pattern Recogn Lett 27:1370–1382
19. Cai T (1999) Adaptive wavelet estimation: a block thresholding and oracle inequality approach. Ann Stat 27:898–924
20. Cai T (2002) On block thresholding in wavelet regression: adaptivity, block size, and threshold level. Stat Sinica 12:1241–1273
21. Cai T, Silverman BW (2001) Incorporating information on neighboring coefficients into wavelet estimation. Sankhyā, Ser B 63:127–148
22. Cai T, Zhou H (2009) A data-driven block thresholding approach to wavelet estimation. Ann Stat 37:569–595
23. Chang SG, Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression. IEEE Trans Image Process 9:1532–1546
24. Chicken E (2003) Block thresholding and wavelet estimation for nonequispaced samples. J Stat Plan Inf 116:113–129
25. Chicken E (2005) Block-dependent thresholding in wavelet regression. J Nonparametr Stat 17:467–491
26. Chicken E (2007) Nonparametric regression with sample design following a random process. Commun Stat Theor Meth 36:1915–1934
27. Chipman H, McCulloch R, Kolaczyk E (1997) Adaptive Bayesian wavelet shrinkage. J Am Stat Assoc 92:1413–1421
28. Clyde M, George E (1998) Robust empirical Bayes estimation in wavelets. ISDS discussion paper, Duke University, Institute of Statistics and Decision Sciences
29. Clyde M, George E (1999) Empirical Bayes estimation in wavelet nonparametric regression. In: Müller P, Vidakovic B (eds.) Bayesian inference in wavelet based models. Lecture notes in statistics, vol 141. Springer-Verlag, New York, pp 309–322
30. Clyde M, George E (2000) Flexible empirical Bayes estimation for wavelets. J Roy Stat Soc Ser B 62:681–698
31. Clyde M, Parmigiani G, Vidakovic B (1998) Multiple shrinkage and subset selection in wavelets. Biometrika 85:391–402
32. Cutillo L, Jung YY, Ruggeri F, Vidakovic B (2008) Larger posterior mode wavelet thresholding and applications. J Stat Plan Inference 138:3758–3773
33. Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia
34. De Canditiis D, Vidakovic B (2004) Wavelet Bayesian block shrinkage via mixtures of normal-inverse gamma priors. J Comput Graph Stat 13:383–398
35. Donoho D, Johnstone I (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81:425–455
36. Efromovich S (2004) Analysis of blockwise shrinkage wavelet estimates via lower bounds for no-signal setting. Ann Inst Stat Math 56:205–223
37. Fadili J, Boubchir L (2005) Analytical form for a Bayesian wavelet estimator of images using the Bessel K form densities. IEEE Trans Image Process 14:231–240
38. Figueiredo M, Nowak R (2001) Wavelet-based image estimation: an empirical bayes approach using Jeffreys' noninformative prior. IEEE Trans Image Process 10:1322–1331
39. Flandrin P (1992) Wavelet analysis and synthesis of fractional Brownian motion. IEEE Trans Inf Theor 38:910–917
40. Fryzlewicz P (2007) Bivariate hard thresholding in wavelet function estimation. Stat Sinica 17:1457–1481

41. George EI, McCulloch R (1997) Approaches to Bayesian variable selection. Stat Sinica 7:339–373
42. Gomez E, Gomez-Villegas MA, Marin JM (1998) A multivariate generalization of the power exponential family of distributions. Commun Stat Theor Meth 27:589–600
43. Gomez E, Gomez-Villegasa MA, Marnb JM (2008) Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. Commun Stat Theor Meth 37:972–985
44. Hall P, Kerkyacharian G, Picard D (1998) Block threshold rules for curve estimation using kernel and wavelet methods. Ann Stat 26:922–942
45. Hall P, Kerkyacharian G, Picard D (1999) On the minimax optimality of block thresholded wavelet estimators. Stat Sinica 9:33–50
46. Hall P, Penev S, Kerkyacharian G, Picard D (1997) Numerical performance of block thresholded wavelet estimators. Stat Comput 7:115–124
47. Huerta G (2005) Multivariate Bayes wavelet shrinkage and applications. J Appl Stat 32:529–542
48. Jansen M, Bultheel A (2001) Empirical Bayes approach to improve wavelet thresholding for image noise reduction. J Am Stat Assoc 96:629–639
49. Johnstone I, Silverman BW (1997) Wavelet threshold estimators for data with correlated noise. J Roy Stat Soc Ser B 59:319–351
50. Johnstone I, Silverman BW (1998) Empirical Bayes approaches to mixture problems and wavelet regression. In: Technical report. Department of Statistics, Stanford University
51. Johnstone IM, Silverman BW (2005a) Empirical Bayes selection of wavelet thresholds. Ann Stat 33:1700–1752
52. Johnstone IM, Silverman BW (2005b) Ebayesthresh: R programs for empirical Bayes thresholding. J Stat Softw 12:1–38 (With accompanying software and manual.)
53. Keogh E, Folias T (2002) UCR time series classification/clustering page. Computer Science and Engineering Department, University of California Riverside Available at http://www.cs.ucr.edu/eamonn/time_series_data
54. Kohn R, Marron JS, Yau P (2000) Wavelet estimation using Bayesian basis selection and basis averaging. Stat Sinica 10:109–128
55. Kovac A, Silverman BW (2000) Extending the scope of wavelet regression methods by coefficient-dependent thresholding. J Am Stat Assoc 95:172–183
56. Lawton W (1993) Applications of complex valued wavelet transforms to subband decomposition. IEEE Trans Signal Process 41:3566–3569
57. Leporini D, Pesquet J-C (1998) Wavelet thresholding for a wide class of noise distributions. EUSIPCO'98, Rhodes, Greece, pp 993–996
58. Leporini D, Pesquet J-C (2001) Bayesian wavelet denoising: Besov priors and non-gaussian noises. Signal Process 81:55–67
59. Lina J-M (1997) Image processing with complex Daubechies wavelets. J Math Imag Vis 7:211–223
60. Lina J-M, MacGibbon B (1997) Non-linear shrinkage estimation with complex Daubechies wavelets. Proc SPIE, Wavelet Appl Signal Image Process V 3169:67–79
61. Lina J-M, Mayrand M (1995) Complex Daubechies wavelets. Appl Comput Harmon Anal 2:219–229
62. Lina J-M, Turcotte P, Goulard B (1999) Complex dyadic multiresolution analysis. In: Advances in imaging and electron physics, vol 109. Academic Press
63. Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11:674–693
64. Moulin P, Liu J (1999) Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors. IEEE Trans Inf Theor 45:909–919
65. Müller P, Vidakovic B (eds.) (1999c) Bayesian inference in wavelet based models. Lecture notes in statistics, vol 141. Springer-Verlag, New York
66. Nason GP (1996) Wavelet shrinkage using cross-validation. J Roy Stat Soc Ser B 58:463–479

67. Nason GP (2008) Wavelet methods in statistis with R. Springer-Verlag, New York
68. Ogden T (1997) Essential wavelets for statistical applications and data analysis. Birkhäuser, Boston
69. Olhede S, Walden A (2004) 'Analytic' wavelet thresholding. Biometrika 91:955–973
70. Pensky M (2006) Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise. Ann Stat 34:769–807
71. Pensky M, Sapatinas T (2007) Frequentist optimality of Bayes factor thresholding estimators in wavelet regression models. Stat Sinica 17:599–633
72. Pericchi LR, Smith AFM (1992) Exact and approximate posterior moments for a normal location parameter. J Roy Stat Soc Ser B 54:793–804
73. Pižurica A, Philips W, Lemahieu I, Acheroy M (2002) A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising. IEEE Trans Image Process 11:545–557
74. Portilla J, Strela V, Wainwright M, Simoncelli E (2003) Image denoising using scale mixtures of Gaussians in the wavelet domain. IEEE Trans Image Process 12:1338–1351
75. Ray S, Mallick BK (2003) A Bayesian transformation model for wavelet shrinkage. IEEE Trans Image Process 12:1512–1521
76. Romberg JK, Choi H, Baraniuk RG (2001) Bayesian tree structured image modeling using wavelet-domain hidden Markov models. IEEE Trans Image Process 10:1056–1068
77. Ruggeri F (1999) Robust Bayesian and Bayesian decision theoretic wavelet shrinkage. In: Müller P, Vidakovic B (eds.) Bayesian inference in wavelet based models. Lecture Notes in Statistics, vol 141. Springer-Verlag, New York, pp 139–154
78. Ruggeri F, Vidakovic B (2005) Bayesian modeling in the wavelet domain. In: Dey DK, Rao CR (eds.) Bayesian thinking: modeling and computation. Handbook of Statistics, vol 25. North-Holland, Amsterdam, pp 315–338
79. Semadeni C, Davison AC, Hinkley DV (2004) Posterior probability intervals in Bayesian wavelet estimation. Biometrika 91:497–505
80. Sendur L, Selesnick IW (2002) Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. IEEE Trans Signal Process 50:2744–2756
81. Simoncelli E (1999) Bayesian denoising of visual images in the wavelet domain. In: Müller P, Vidakovic B (eds.) Bayesian inference in wavelet based models. Lecture Notes in Statistics, vol 141. Springer-Verlag, New York, pp 291–308
82. Tadesse MG, Ibrahim JG, Vannucci M, Gentleman R (2005) Wavelet thresholding with Bayesian false discovery rate control. Biometrics 61:25–35
83. Ter Braak CJF (2010) Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising. Comput Stat Data Anal 51: 1232–1242
84. Vannucci M, Corradi F (1999) Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. J Roy Stat Soc Ser B 61: 971–986
85. Vidakovic B (1998a) Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. J Am Stat Assoc 93:173–179
86. Vidakovic B (1998b) Wavelet-based nonparametric Bayes methods. In: Dey D, Müller P, Sinha D (eds.) Practical nonparametric and semiparametric bayesian statistics. Lecture Notes in Statistics, vol 133. Springer-Verlag, New York, pp 133–155
87. Vidakovic B (1999) Statistical modeling by wavelets. John Wiley & Sons, Inc., New York
88. Vidakovic B, Müller P (1999) An introduction to wavelets. In: Müller P, Vidakovic B (eds.) Bayesian inference in wavelet based models. Lecture Notes in Statistics, vol 141. Springer-Verlag, New York, pp 1–18
89. Vidakovic B, Ruggeri F (1999) Expansion estimation by Bayes rules. J Stat Plan Inf 79:223–235
90. Vidakovic B, Ruggeri F (2001) BAMS method: theory and simulations. Sankhyā, Ser B 63:234–249
91. Walter GG, Shen X (2001) Wavelets and other orthogonal systems, 2ns edn. Chapman & Hall/CRC, Boca Raton

92. Wang X, Ray S, Mallick BK (2007) Bayesian curve classification using wavelets. J Am Stat Assoc 102:962–973
93. Wang X, Wood ATA (2006) Empirical Bayes block shrinkage of wavelet coefficients via the non-central χ^2 distribution. Biometrika 93:705–722
94. Wang X, Wood ATA (2010) Wavelet estimation of an unknown function observed with correlated noise. Commun Stat Simul Comput 39:287–304
95. Zhang C-H (2005) General empirical Bayes wavelet methods and exactly adaptive minimax estimation. Ann Stat 33:54–100