CONTRIBUTIONS TO BAYESIAN WAVELET SHRINKAGE

A Thesis
Presented to
The Academic Faculty

by

Norbert Reményi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2012

CONTRIBUTIONS TO BAYESIAN WAVELET SHRINKAGE

Approved by:

Professor Brani Vidakovic, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor David Goldsman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Justin Romberg
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Xiaoming Huo
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: November 1, 2012

To my beloved family.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, Dr. Brani Vidakovic, for his support, patience, and trust. He was always kind and helpful to me, for which I am truly grateful. I also want to extend my gratitude to the members of my thesis committee, Dr. David Goldsman, Dr. Xiaoming Huo, Dr. Yajun Mei, and Dr. Justin Romberg, for the valuable suggestions during both my proposal and my defense.

I would also like to thank Dr. Sándor Kemény, my undergraduate advisor, and Dr. Leigh Murray, my master's thesis advisor, for their help and support. They both were a great inspiration to me to pursue studies in statistics.

I thank my fellow graduate students and friends, especially Namin Kim, Claudio Santiago, Helder Inacio, Martin Kistenmacher, Mallory Soldner, Seonghye Jeon, Carlos Valencia and Dan Steffy for their support, and for being good friends over the years. The time we spent together became a meaningful part of my life.

Finally, and most importantly, I would like to thank my family for their unconditional love and endless support throughout my entire life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis provides contributions to research in Bayesian modeling and shrinkage in the wavelet domain. Wavelets are a powerful tool to describe phenomena rapidly changing in time, and wavelet-based modeling has become a standard technique in many areas of statistics, and more broadly, in sciences and engineering. Bayesian modeling and estimation in the wavelet domain have found useful applications in nonparametric regression, image denoising, and many other areas. In this thesis, we build on the existing techniques and propose new methods for applications in nonparametric regression, image denoising, and partially linear models.

The thesis consists of an overview chapter and four main topics. In Chapter 1, we provide an overview of recent developments and the current status of Bayesian wavelet shrinkage research. The chapter contains an extensive literature review consisting of almost 100 references. The main focus of the overview chapter is on nonparametric regression, where the observations come from an unknown function contaminated with Gaussian noise. We present many methods which employ model-based and adaptive shrinkage of the wavelet coefficients through Bayes rules. These includes new developments such as dependence models, complex wavelets, and Markov chain Monte Carlo (MCMC) strategies. Some applications of Bayesian wavelet shrinkage, such as curve classification, are discussed.

In Chapter 2, we propose the Gibbs Sampling Wavelet Smoother ($GSWS$), an adaptive wavelet denoising methodology. We use the traditional mixture prior on the wavelet coefficients, but also formulate a fully Bayesian hierarchical model in the wavelet domain accounting for the uncertainty of the prior parameters by placing hyperpriors on them. Since a closed-form solution to the Bayes estimator does not

exist, the procedure is computational, in which the posterior mean is computed via MCMC simulations. We show how to efficiently develop a Gibbs sampling algorithm for the proposed model. The developed procedure is fully Bayesian, is adaptive to the underlying signal, and provides good denoising performance compared to state-of-the-art methods. Application of the method is illustrated on a real data set arising from the analysis of metabolic pathways, where an iterative shrinkage procedure is developed to preserve the mass balance of the metabolites in the system. We also show how the methodology can be extended to complex wavelet bases.

In Chapter 3, we propose a wavelet-based denoising methodology based on a Bayesian hierarchical model using a double Weibull prior. The interesting feature is that in contrast to the mixture priors traditionally used by some state-of-the-art methods, the wavelet coefficients are modeled by a single density. Two estimators are developed, one based on the posterior mean and the other based on the larger posterior mode; and we show how to calculate these estimators efficiently. The methodology provides good denoising performance, comparable even to state-of-the-art methods that use a mixture prior and an empirical Bayes setting of hyperparameters; this is demonstrated by simulations on standard test functions. An application to a real-word data set is also considered.

In Chapter 4, we propose a wavelet shrinkage method based on a neighborhood of wavelet coefficients, which includes two neighboring coefficients and a parental coefficient. The methodology is called Λ-neighborhood wavelet shrinkage, motivated by the shape of the considered neighborhood. We propose a Bayesian hierarchical model using a contaminated exponential prior on the total mean energy in the Λ-neighborhood. The hyperparameters in the model are estimated by the empirical Bayes method, and the posterior mean, median, and Bayes factor are obtained and used in the estimation of the total mean energy. Shrinkage of the neighboring coefficients is based on the

ratio of the estimated and observed energy. The proposed methodology is comparable and often superior to several established wavelet denoising methods that utilize neighboring information, which is demonstrated by extensive simulations. An application to a real-world data set from inductance plethysmography is considered, and an extension to image denoising is discussed.

In Chapter 5, we propose a wavelet-based methodology for estimation and variable selection in partially linear models. The inference is conducted in the wavelet domain, which provides a sparse and localized decomposition appropriate for nonparametric components with various degrees of smoothness. A hierarchical Bayes model is formulated on the parameters of this representation, where the estimation and variable selection is performed by a Gibbs sampling procedure. For both the parametric and nonparametric part of the model we are using point-mass-at-zero contamination priors with a double exponential spread distribution. In this sense we extend the model of Chapter 2 to partially linear models. Only a few papers in the area of partially linear wavelet models exist, and we show that the proposed methodology is often superior to the existing methods with respect to the task of estimating model parameters. Moreover, the method is able to perform Bayesian variable selection by a stochastic search for the parametric part of the model.

The thesis is concluded by an Appendix and References.

CHAPTER I

INTRODUCTION - BAYESIAN WAVELET SHRINKAGE STRATEGIES

In this chapter we 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.

1.1 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, 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 three-step procedure: (i) data are transformed into a set of wavelet coefficients; (ii) a shrinkage of the coefficients is performed; and (iii) 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 transforms 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 Walter and Shen

(2001).

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 transform, 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 Daubechies (1992), Ogden (1997), Vidakovic (1999), Walter and Shen (2001), among others. An introductory article is Vidakovic and Müller (1999).

1.1.1 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 transforms 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 (Mallat, 1989). The

wavelet decomposition of the vector $\boldsymbol{y}$ can be written as

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

Note that in (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

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

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 (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 (1) contains 2^j elements, and can be written as

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

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:

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

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

1.2 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 Vidakovic and Ruggeri (1999). The Bayes rules can be constructed to mimic the thresholding rules: to slightly shrink the large coefficients and heavily shrink the small coefficients. 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 Abramovich et al. (2000), in Vidakovic (1998b, 1999) and in Ruggeri and Vidakovic (2005). An edited volume on Bayesian modeling in the wavelet domain appeared 12 years ago (Müller and Vidakovic, 1999).

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 transforms $\boldsymbol{W}$ 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 simulations to produce the estimate. Bayes rules with respect to absolute or 0-1 loss functions are capable of producing bona fide thresholding rules.

1.2.1 An Illustrative Example

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

The BAMS originates from the observation that a realistic Bayes model should produce prior predictive distributions of the observations which "agree" with the observations. Other authors were previously interested in the empirical distribution of the wavelet coefficients; see, for example, Leporini and Pesquet (1998, 2001), Mallat (1989), Ruggeri (1999), Simoncelli (1999), and Vidakovic (1998b). 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, Mallat (1989) suggested to fit empirical distributions of wavelet coefficients by the exponential power model

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

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

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

Vidakovic (1998b) 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

$$\pi(\theta) = \epsilon\delta(0) + (1-\epsilon)\xi(\theta) \tag{3}$$

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 $0 \leq \epsilon \leq 1$ 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

$$\delta_\pi(d) = \frac{(1-\epsilon)\ m_\xi(d)\ \delta_\xi(d)}{(1-\epsilon)\ m_\xi(d) + \epsilon\ DE\left(0, \frac{1}{\sqrt{2\mu}}\right)}, \tag{4}$$

where

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

and

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

are the prior predictive distribution and the Bayes rule for the spread part of the prior, ξ. Rule (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. (1998) 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

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

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

$$\int_{-\infty}^{u} \pi^\star(\theta|d)d\theta = \frac{1}{2}. \tag{5}$$

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

$$\max \int_{-\infty}^{0^-} \pi^\star(\theta|d)d\theta = \frac{1}{2}, \tag{6}$$

and in case $d < 0$,

$$\min \int_{-\infty}^{0} \pi^\star(\theta|d)d\theta = \frac{1}{2}. \tag{7}$$

Because $\pi^\star(\theta|d)$ is a probability density, the integral in equation (5) is non-decreasing in u. Therefore, by using results (6) and (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 (5) becomes

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

Next, assume $d < 0$. Then the solution satisfies $u \leq 0$ and equation (5) becomes:

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

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

For $d > 0$,

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

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

$$\left.\left.+\frac{2/\tau}{2\mu-1/\tau^2}\right](\sqrt{2\mu}-1/\tau)\right\}$$

For $d < 0$,

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

$$\delta_M(d)=0$$

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

$$\left.\left.-e^{-(\sqrt{2\mu}-1/\tau)d}\right]\right\}$$

For $d = 0$,

$$\delta_M(d)=0. \tag{8}$$

The rule $\delta_M(d)$ based on algorithm (8) is the BAMS-MED rule. As evident from Figure 1, the BAMS-MED rule is a thresholding rule.

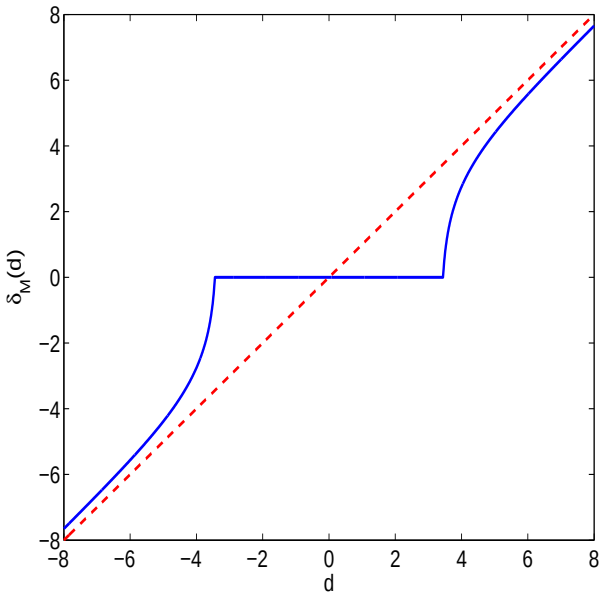

Figure 1: BAMS-MED rule (8) for $\epsilon = 0.9$, $\mu = 1$ and $\tau = 2$.

1.3 Bayesian Wavelet Regression

1.3.1 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 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 the whitening property of wavelet transforms (Flandrin, 1992), 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. (1997). Their approach is based on the stochastic search variable selection (SSVS) proposed by George and McCulloch (1997), with the assumption that σ is known.

Chipman et al. (1997) 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

$$\gamma_j \sim Ber(p_j).$$

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,

$$\delta(d) = \left[P(\gamma_j = 1|d)\frac{(c_j\tau_j)^2}{\sigma^2 + (c_j\tau_j)^2} + P(\gamma_j = 0|d)\frac{\tau_j^2}{\sigma^2 + \tau_j^2} \right] d, \tag{9}$$

where

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

and

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

For other early examples of the Bayesian approach to wavelet regression see papers, for example, by Abramovich et al. (1998), Clyde et al. (1998), Clyde and George (1998) and Vidakovic (1998a).

A more recent paper by Johnstone and Silverman (2005b) 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

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

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

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

where the posterior probability w_{post} is

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

and

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

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 (1992). It can be shown that the posterior mean is

$$\mathbb{E}(\mu|z) = w_{post}(z)\left[z - \frac{a\left[e^{-az}\Phi(z-a) - e^{az}\tilde{\Phi}(z+a)\right]}{e^{-az}\Phi(z-a) + e^{az}\tilde{\Phi}(z+a)}\right]. \tag{10}$$

A schematic picture of the posterior mean (10) is presented in Figure 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.

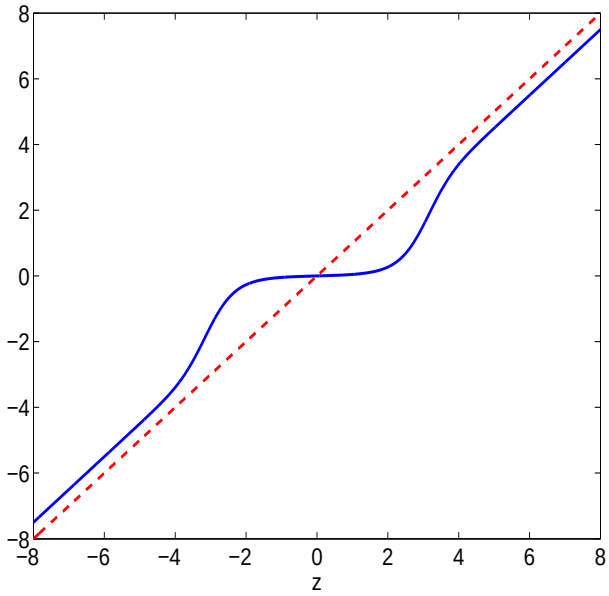

Figure 2: Posterior mean rule (10) for $w = 0.1$ and $a = 0.5$.

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 Johnstone and Silverman (2005b), and for more examples using the method, see Johnstone and Silverman (2005a).

Several more recent papers have considered term-by-term Bayesian wavelet shrinkage. Angelini and Sapatinas (2004) 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 (2004) 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. (2008) consider thresholding rules induced by a variation of the Bayesian 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.

1.3.2 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. (1997, 1998, 1999), Cai (1999, 2002), Cai and Silverman (2001) where wavelet coefficients are thresholded based on block sums of squares.

Abramovich et al. (2002) 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}:

$$\begin{aligned} b_{jK}|\gamma_{jK} &\sim N(0, \gamma_{jK} V_j), \\ \gamma_{jK} &\sim Ber(\pi_j). \end{aligned}$$

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

$$\mathbb{E}(b_{jK}|\hat{b}_{jK}) = \frac{1}{1 + O_{jK}} A_j \hat{b}_{jK}, \tag{11}$$

where

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

Rule (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 (1994). 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 Abramovich et al. (2002).

A paper by De Candiitis and Vidakovic (2004) 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. (1997). 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 Abramovich et al. (2002). 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 normal-inverse gamma 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

$$\mathbb{E}(\boldsymbol{\theta}_{jH}|\boldsymbol{d}_{jH}) = A_{jH}(\boldsymbol{d}_{jH})\boldsymbol{m}^{\star}_{jH} + (1 - A_{jH}(\boldsymbol{d}_{jH}))\boldsymbol{m}^{\star\star}_{jH}, \tag{12}$$

where

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

and

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

The posterior mean (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 De Candiitis and Vidakovic (2004).

Huerta (2005) 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

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

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 Huerta (2005).

Wang and Wood (2006) 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 $z = \|\hat{c}_B\|^2 = \sum_{i \in B} \hat{c}_i^2$, 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

$$\begin{aligned} f(z|\sigma^2, \theta) &= p\chi^2_m(\rho|0, \sigma^2) + (1 - p)\frac{(\lambda + 1)(2\sigma^2)^{\lambda+1}}{\Gamma(\frac{1}{2}m)z^{\lambda+2}}\gamma\left(\eta, \frac{z}{2\sigma^2}\right), \\ \mathbb{E}(\rho|z, \sigma^2, \theta) &= (1 - \pi)\left\{m\sigma^2 + z - \frac{m\sigma^2 + 2z}{z/(2\sigma^2)}\,\mathcal{C}_{\eta,1}\left(\frac{z}{2\sigma^2}\right) + \frac{4\sigma^4}{z}\mathcal{C}_{\eta,2}\left(\frac{z}{2\sigma^2}\right)\right\}, \end{aligned}$$

where

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

Hyperparameter σ^2 is estimated analogously to the median absolute deviation estimator suggested by Donoho and Johnstone (1994), 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

$$c_B = \hat{c}_B \{\mathcal{B}_{\sigma^2,\theta}(z)/z\}^{\frac{1}{2}}, \tag{13}$$

where $\mathcal{B}_{\sigma^2,\theta}(z)$ 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 Wang and Wood (2006).

There is a wide range of other articles considering Bayesian modeling of neighboring wavelet coefficients. To name a few, Romberg et al. (2001) use a Bayesian hidden Markov tree (HMT) to model the structure of wavelet coefficients in images. Jansen and A. (2001) 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 (2002) use parent-child neighboring relation and Laplacian bivariate prior to derive MAP estimators for wavelet coefficients. Pižurica et al. (2002) use a Markov random field (MRF) prior model to incorporate inter- and intrascale dependencies of wavelet coefficients. Portilla et al. (2003) models neighborhoods of image wavelet coefficients at adjacent positions and scales using scale mixture of Gaussians.

A recent non-Bayesian development was proposed by Fryzlewicz (2007) in a form of fast, hard-thresholding algorithm based on coupling parents and children in the wavelet coefficient tree.

1.3.3 Complex Wavelet Shrinkage

Wavelet shrinkage methods using complex-valued wavelets provide additional insights to shrinkage process. Lina and Mayrand (1995) 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 4 possible solutions to the defining equations, but only 2 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 (1993) 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 transform 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 transform 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, Lina and Macgibbon (1997), Lina (1997), and Lina et al. (1999) 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 (2004), 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

$$\begin{aligned}
\text{cov}\{\text{Re}(\boldsymbol{\varepsilon}), \text{Im}(\boldsymbol{\varepsilon})\} &= -\sigma^2 \text{Im}(WW^T)/2, \\
\text{cov}\{\text{Re}(\boldsymbol{\varepsilon}), \text{Re}(\boldsymbol{\varepsilon})\} &= \sigma^2\{I_n + \text{Re}(WW^T)\}/2, \\
\text{cov}\{\text{Im}(\boldsymbol{\varepsilon}), \text{Im}(\boldsymbol{\varepsilon})\} &= \sigma^2\{I_n - \text{Re}(WW^T)\}/2. \qquad (14)
\end{aligned}$$

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) for each dyadic level j. Noise variance σ^2 is estimated by the usual median absolute deviation by Donoho and Johnstone (1994).

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. (1998). 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

$$\mathbb{E}(\theta_{jk}) = \tilde{p}_{jk} \mu_{jk}, \qquad (15)$$

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 (2004) 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.

1.3.4 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 Markov chain Monte Carlo (MCMC) simulations.

We build on the results of Barber and Nason (2004) 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}:

$$\begin{aligned} d_{jk}|\theta_{jk}, \sigma^2 &\sim N_2(\theta_{jk}, \sigma^2 \Sigma_j) \\ \theta_{jk}|\epsilon_j, C_j &\sim (1-\epsilon_j)\delta_0 + \epsilon_j EP_2(\mu, C_j, \beta), \end{aligned} \tag{16}$$

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 Gomez et al. (1998). 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, Barber and Nason (2004) 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 (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:

$$\pi(\theta|C) = \frac{1}{8\pi|C|^{1/2}} \exp\left\{-\frac{1}{2}\left(\theta' C^{-1}\theta\right)^{1/2}\right\}. \tag{17}$$

The prior in (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 (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 Barber and Nason (2004) and mentioned above in (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 mixture of multivariate normal distributions, which is an essential step for efficient Monte Carlo simulation. From Gomez et al. (2008), the bivariate exponential power distribution with $\mu = 0$ and $\beta = 1/2$ can be represented as

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

which is a scale mixture of bivariate normal distributions with mixing distribution gamma. Using the specified hyperpriors and the mixture representation, the model in (16) extends to

$$\begin{aligned}
d_{jk}|\theta_{jk}, \sigma^2 &\sim N_2(\theta_{jk}, \sigma^2 \Sigma_j) \\
\sigma^2 &\sim IG(a, b) \\
\theta_{jk}|z_{jk}, v_{jk}, C_j &\sim (1 - z_{jk})\delta_0 + z_{jk} N_2(0, v_{jk} C_j) \\
z_{jk}|\epsilon_j &\sim Ber(\epsilon_j) \\
\epsilon_j &\sim U(0, 1) \\
v_{jk} &\sim Ga(3/2, 8) \\
C_j &\sim IW(A_j, w). \qquad (18)
\end{aligned}$$

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 mixture of normals, the hierarchical model in (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 $\hat{\theta}_{jk} = \sum_i \theta_{jk}^{(i)}/N$ 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 Nason (1996) and Abramovich et al. (1998, 2002). For more information about the data, refer to these papers.

The top part of Figure 3 shows a section of plethysmograph recording lasting approximately 80 s ($n = 4096$ observations), while the bottom part shows the reconstruction of the signal with the *CGSWS* method. In the reconstruction process we applied $N = 5000$ iterations of the Gibbs sampler of which the first 2000 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. (2002) 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. (2002), and better compared to the results obtained by other established methods analyzed in their paper.

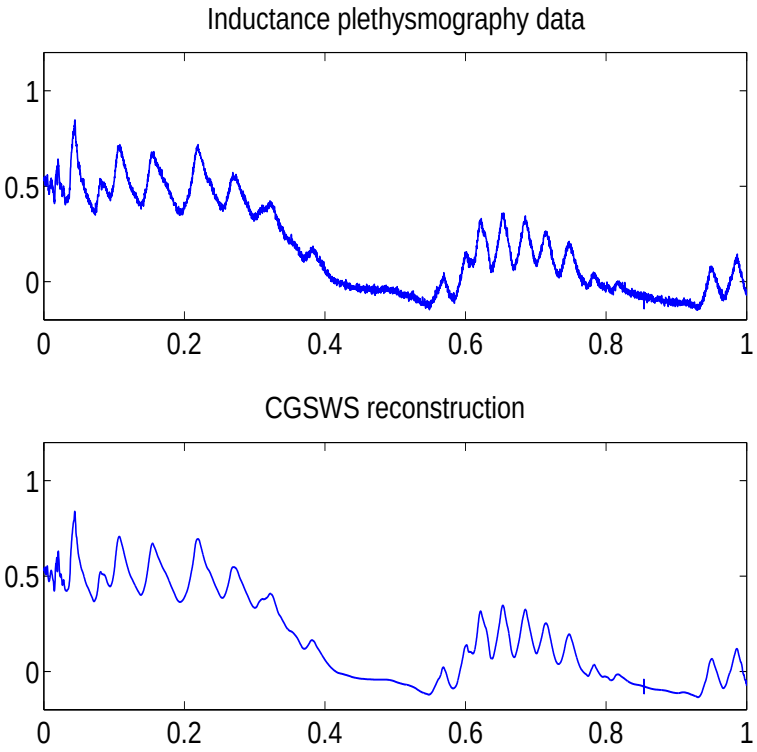

Figure 3: Reconstruction of the inductance plethysmography data (IPD) by *CGSWS*.

1.3.5 Bayesian Wavelet Shrinkage in Curve Classification

We consider the paper by Wang et al. (2007) 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. (2007) 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 $\boldsymbol{Y_i} = \boldsymbol{X\beta_i} + \boldsymbol{\varepsilon_i}$, where $\boldsymbol{X}$ is the discrete wavelet transform 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:

$$
\begin{aligned}
\text{Random function } \boldsymbol{Y_i} &\sim N(\boldsymbol{X\beta_i}, \sigma^2 I), \\
\boldsymbol{\beta_i}, \sigma^2 | \boldsymbol{\eta_i}, \boldsymbol{g} &\sim NIG(0, \mathrm{diag}(\boldsymbol{\eta_i})\mathrm{diag}(\boldsymbol{g}), a_\sigma, b_\sigma), \\
g_j &\sim IG(u_j, v_j), \\
\eta_{ijk} &\sim Ber(\rho_j). \\
\text{Binary outcome } z_i &\sim Ber(p_i), \\
T_i &\sim N(\boldsymbol{\beta_i^t \theta}, \tau^2), \quad \text{where } T_i = \mathrm{logit}(p_i), \\
\boldsymbol{\theta}, \tau^2 | \boldsymbol{\gamma}, \boldsymbol{h} &\sim NIG(0, \mathrm{diag}(\boldsymbol{\gamma})\mathrm{diag}(\boldsymbol{h}), a_\tau, b_\tau), \\
h_j &\sim IG(c_j, d_j), \\
\gamma_{jk} &\sim Ber(\pi_j)
\end{aligned}
\tag{19}
$$

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

The first part in (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 De Candiitis and Vidakovic (2004). 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 (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 (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 Markov chain Monte Carlo methods. Wang et al. (2007) derive the full conditional distributions for the parameters, which allow for implementation of a Gibbs sampling algorithm. The model in (19) is also extended to multicategory classification by the authors.

Application to Leaf Data Wang et al. (2007) analyzed a data set from Keogh et al. (2011) 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 Wang et al. (2007) can be seen in Figure 4.

The classification was carried out by randomly selecting 140 curves from the training and 10 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

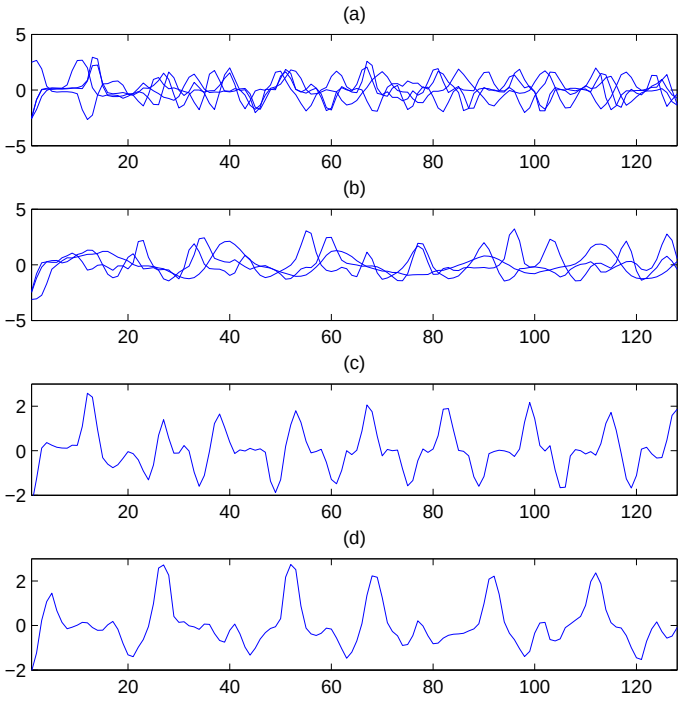

Figure 4: Adopted from Wang et al. (2007): "Pseudo-time series curves from leaf images. (a) and (b) Every other curve in two species in the data set, 33 of *Circinatum* and 42 of *Garryana*. (c) and (d) Example of single curve from two species, *Circinatum* and *Garryana*."

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 Wang et al. (2007).

1.3.6 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 Abramovich et al. (2000), Antoniadis (2007) and Nason (2008), for example. An excellent critical overview and simulation study comparing different wavelet shrinkage methods can be found in Antoniadis et al. (2001). Articles focusing only on Bayesian wavelet-based modeling include Vidakovic (1998b), Müller and Vidakovic (1999) and Ruggeri and Vidakovic (2005).

Some recent results about theoretical properties and optimality of Bayesian wavelet estimators can be found in Abramovich et al. (2004, 2007), Bochkina and Sapatinas (2006, 2009), Johnstone and Silverman (2005b), Pensky (2006) and Pensky and Sapatinas (2007).

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 (2001) and Moulin and Liu (1999), who use generalized Gaussian and complexity priors to model wavelet coefficients. Other articles in this group include Boubchir and Fadili (2006) using approximate α-stable prior, Chang et al. (2000) using generalized Gaussian distribution

(GCD) as a prior, Fadili and Boubchir (2005) using Bessel K forms (BKF) densities, and Leporini and Pesquet (2001) using Besov norm priors for modeling wavelet coefficients. Achim and Kuruoğlu (2005) 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 Cai (2002), Cai and Zhou (2009), Chicken (2003, 2005, 2007) and Efromovich (2004), to name a few. More general theoretical results about block empirical Bayes estimation appear in Zhang (2005).

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 (1999, 2000) and Johnstone and Silverman (1998, 2005b).

The usual assumptions for wavelet regression are equispaced sampling points with a sample size being a power of two, 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, Johnstone and Silverman (1997) consider wavelet thresholding with stationary correlated noise, and Kovac and Silverman (2000) 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 (1999). Ambler and Silverman (2004) 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 (2010) 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 Wang and Wood (2006). Ray and Mallick (2003) 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. (2000) 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. (2002) and Semadeni et al. (2004) derive Bayesian credible intervals for Bayesian wavelet regression estimates based on cumulants and saddlepoint approximation, respectively.

Olhede and Walden (2004) 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. (2006). A Bayesian interpretation of the FDR procedure and application to wavelet thresholding can be found in Tadesse et al. (2005).

Application of the Bayesian maximum a posteriori multiple testing (testimation) procedure to wavelet thresholding can be found in Abramovich et al. (2010).

CHAPTER II

ADAPTIVE WAVELET SHRINKAGE BY GIBBS SAMPLING

In this chapter we propose the Gibbs Sampling Wavelet Smoother ($GSWS$), an adaptive wavelet denoising methodology. The method is based on a fully Bayesian hierarchical model using a mixture prior on the wavelet coefficients. The heart of the procedure is computational, where the posterior mean is computed through Markov chain Monte Carlo (MCMC) simulations. We show that $GSWS$ has good performance, as demonstrated by simulations on well-known test functions and by comparisons with other commonly used denoising methods. The method is illustrated on a real data set arising from the analysis of metabolic pathways, and we also show how the methodology can be extended to complex wavelet bases.

2.1 Introduction

In the present chapter we consider a novel Bayesian model as a solution to the classical nonparametric regression problem

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{20}$$

where x_i, $i = 1, \ldots, n$, are equispaced sampling points, and the errors ε_i are i.i.d. normal random variables, with zero mean and variance σ^2. Our interest is to estimate the function f using the observations y_i. After applying a linear and orthogonal wavelet transform, the equation in (20) 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 ε, respectively. Note that ε_i and ε_{jk} are equal in distribution due to the orthogonality of wavelet transforms. Due to the whitening property of the wavelet transforms (Flandrin, 1992), many existing methods assume independence of the coefficients, and omit the double indices jk to work with a generic wavelet coefficient model

$$d = \theta + \varepsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{21}$$

When indices are needed for the clarity of exposition, they will be used.

To estimate θ in model (21), Bayesian shrinkage rules have been proposed in the literature by many authors. By a shrinkage rule, we mean that the observed wavelet coefficients d are replaced with their shrunken version $\hat{\theta} = \delta(d)$. Then f is estimated as the inverse wavelet transform of $\hat{\theta}$. Empirical distributions of detail wavelet coefficients for signals encountered in practical applications are (at each resolution level) centered around and peaked at zero (Mallat, 1989). A range of models, for which the unconditional distributions of wavelet coefficients mimic these properties, have been considered in the literature. The traditional Bayesian models consider the prior distribution on the wavelet coefficient θ as

$$\pi(\theta) = \epsilon\delta_0 + (1 - \epsilon)\xi(\theta), \tag{22}$$

where δ_0 is a point mass at zero, ξ is a symmetric and unimodal distribution, and ϵ is a fixed parameter in [0,1], usually level dependent, that controls the amount of shrinkage for values of d close to 0. This type of model was considered by Abramovich et al. (1998), Vidakovic (1998a), Vidakovic and Ruggeri (2001), and Johnstone and Silverman (2005b), among others.

In this chapter, we consider ξ to be the double exponential distribution. As commonly done in Bayesian modeling, we formulate a hierarchical model which accounts for the uncertainty of the prior parameters by placing hyperpriors on them.

Since a closed-form solution to the Bayes estimator does not exist, we apply MCMC methodology and compute an approximate estimator (posterior mean) from the output of simulational runs. Although the simplicity of a closed-form solution is lost, the procedure is fully Bayesian, adaptive to the underlying signal, and provides good performance in comparison to some existing state-of-the-art methods.

The chapter is organized as follows. Section 2.2 formalizes the model and presents some results related to it. Section 2.3 explains in detail the Gibbs sampling scheme developed for the hierarchical model. Section 2.4 discusses the selection of hyperparameters and contains simulations and comparisons to existing methods. In Section 2.5 we apply the method to a real data set related to Dynamic Flux Estimation (DFE). Section 2.6 briefly extends the methodology to complex wavelet bases, and in Section 2.7 we discuss some more extensions of the model. Conclusions and discussion are provided in Section 2.8.

2.2 Hierarchical Model

In this chapter we consider the following hierarchical Bayesian model in the wavelet domain,

$$\begin{aligned} d_{jk}|\theta_{jk},\sigma^2 &\sim \mathcal{N}(\theta_{jk},\sigma^2) \\ \sigma^2 &\sim \mathcal{IG}(a_1,b_1) \\ \theta_{jk}|\epsilon_j,\tau &\sim (1-\epsilon_j)\delta_0+\epsilon_j\mathcal{DE}(\tau) \\ \epsilon_j &\sim \mathcal{U}(0,1) \\ \tau &\sim \mathcal{G}a(a_2,b_2) \end{aligned} \tag{23}$$

where j pertains to the resolution level of d_{jk} and $\mathcal{N}$, $\mathcal{IG}$, $\mathcal{DE}$, $\mathcal{U}$, $\mathcal{G}a$ stand for the normal, inverse gamma, double exponential, uniform, and gamma distributions, respectively. Note that the model in (23) uses the well-established mixture prior with a point mass at zero, which accounts for the sparsity of the signal part in

the wavelet domain. Wavelet coefficients with large energies are captured by the spread part of the mixture prior, for which we propose the double exponential or Laplace distribution with variance $2/\tau^2$. The double exponential distribution is a popular choice for the spread part. It models wavelet coefficients with large energies and was used by several authors, for example, Vidakovic and Ruggeri (2001) and Johnstone and Silverman (2005b). The mixture prior on θ is specified levelwise, for each dyadic level j; however, the scale parameter τ is global. This serves the purpose of parsimony and contributes to ease of estimation. Specifying τ levelwise is possible, but this induces "double shrinkage" (Clyde et al., 1998) together with parameter ϵ_j, which did not improve the performance of the estimator. Parameter σ^2 represents the common noise variance for each resolution level.

For easier implementation purposes, we introduce the latent variable z_{jk} into the model. We rewrite the hierarchical model in (23) using variable z_{jk} as the following:

$$\begin{aligned} d_{jk}|\theta_{jk},\sigma^2 &\sim \mathcal{N}(\theta_{jk},\sigma^2) \\ \sigma^2 &\sim \mathcal{IG}(a_1,b_1) \\ \theta_{jk}|z_{jk},\tau &\sim (1-z_{jk})\delta_0 + z_{jk}\mathcal{DE}(\tau) \\ z_{jk}|\epsilon_j &\sim \mathcal{B}er(\epsilon_j) \\ \epsilon_j &\sim \mathcal{U}(0,1) \\ \tau &\sim \mathcal{G}a(a_2,b_2) \end{aligned} \tag{24}$$

where $\mathcal{B}er$ stands for the Bernoulli distribution. Here z_{jk} is a latent variable indicating whether our parameter θ_{jk} is coming from a point mass at zero ($z_{jk}=0$) or from a double exponential part ($z_{jk}=1$), with prior probability of $1-\epsilon_j$ or ϵ_j, respectively. The uniform $\mathcal{U}(0,1)$ prior on ϵ_j is equivalent to a beta $\mathcal{B}e(1,1)$ distribution, which is a conjugate prior for the Bernoulli distribution. Integrating out z from model (24) gives back the original model in (23).

The hierarchical model in (24) is not conjugate; however, with additional derivations and computational techniques it is possible to develop a fast Gibbs sampling algorithm for updating its parameters. In this context, estimators for θ_{jk} can be obtained from the simulation runs of the Gibbs sampling algorithm. We used the sample average as the standard MCMC estimator for the posterior mean.

Next we discuss results related to the model in (23) which are instrumental in developing the Gibbs sampler. In the following, d corresponds to an arbitrary d_{jk}, and the mean θ stands for the corresponding θ_{jk}. If we consider a $\mathcal{N}(\theta, \sigma^2)$ likelihood $f(d|\theta, \sigma^2)$ and elicit a $\mathcal{DE}(\tau)$ prior $p_1(\theta|\tau)$ on the θ, the marginal distribution becomes

$$m(d|\sigma^2, \tau) = \frac{\tau}{2} e^{\frac{\sigma^2\tau^2}{2}} \left\{ e^{-d\tau} \Phi\left(\frac{d}{\sigma} - \tau\sigma\right) + e^{d\tau} \Phi\left(-\frac{d}{\sigma} - \tau\sigma\right) \right\}, \tag{25}$$

and the posterior distribution becomes

$$\begin{aligned} &h(\theta|d, \sigma^2, \tau) = \\ &= \begin{cases} \dfrac{e^{-d\tau}}{e^{-d\tau}\Phi\left(\frac{d}{\sigma} - \tau\sigma\right) + e^{d\tau}\Phi\left(-\frac{d}{\sigma} - \tau\sigma\right)} \dfrac{1}{\sigma}\phi\left(\dfrac{\theta - (d - \sigma^2\tau)}{\sigma}\right), & \theta \geq 0 \\ \dfrac{e^{d\tau}}{e^{-d\tau}\Phi\left(\frac{d}{\sigma} - \tau\sigma\right) + e^{d\tau}\Phi\left(-\frac{d}{\sigma} - \tau\sigma\right)} \dfrac{1}{\sigma}\phi\left(\dfrac{\theta - (d + \sigma^2\tau)}{\sigma}\right), & \theta < 0 \end{cases} \end{aligned} \tag{26}$$

where ϕ and Φ denote the pdf and cdf of the standard normal distribution, respectively. For derivations of these results, see Appendix. From the representation in (26) we can see that the posterior distribution is a mixture of truncated normals, which will be utilized in our Gibbs sampling algorithm. Now if we consider the mixture prior $p(\theta|\tau) = (1 - \epsilon_j)\delta_0 + \epsilon_j p_1(\theta|\tau)$ on θ in (23), we get the posterior distribution as

$$\begin{aligned} \pi(\theta|d, \sigma^2, \tau) &= \frac{f(d|\theta, \sigma^2)p(\theta|\tau)}{\int_{-\infty}^{\infty} f(d|\theta, \sigma^2)p(\theta|\tau)d\theta} \\ &= \frac{(1 - \epsilon_j)f(d|\theta, \sigma^2)\delta_0 + \epsilon_j f(d|\theta, \sigma^2)p(\theta|\tau)}{(1 - \epsilon_j)f(d|0, \sigma^2) + \epsilon_j m(d|\sigma^2, \tau)} \\ &= \frac{(1 - \epsilon_j)f(d|0, \sigma^2)\delta_0 + \epsilon_j m(d|\sigma^2, \tau)h(\theta|d, \sigma^2, \tau)}{(1 - \epsilon_j)f(d|0, \sigma^2) + \epsilon_j m(d|\sigma^2, \tau)} \\ &= (1 - p_j)\delta_0 + p_j h(\theta|d, \sigma^2, \tau), \end{aligned} \tag{27}$$

where $f(d|0,\sigma^2)$ is the normal distribution with mean $\theta = 0$ and variance σ^2, and

$$p_j = \frac{\epsilon_j m(d|\sigma^2,\tau)}{(1-\epsilon_j)f(d|0,\sigma^2) + \epsilon_j m(d|\sigma^2,\tau)} \tag{28}$$

is the mixing weight for the posterior distribution. Thus, the posterior distribution of θ is a mixture of a point mass at zero and a mixture of truncated normal distributions $h(\theta|d,\sigma^2,\tau)$ with mixing weight p_j.

2.3 Gibbs Sampling Scheme

To obtain posterior inferences on the wavelet coefficients θ, we adopt a standard Gibbs sampling procedure. In this section we provide details of how to develop a Gibbs sampler for the model in (24). Gibbs sampling is an iterative algorithm that simulates from a joint posterior distribution through iterative simulation over the full conditional distributions. For more details on Gibbs sampling, see Casella and George (1992) or Robert and Casella (1999). For the model in (24), the full conditionals for all parameters can be determined exactly. We build on results given by (25)–(28). Derivations of some results in this section are deferred to Appendix.

Next, we will describe the full conditional distributions and updating schemes for parameters σ^2, z_{jk}, ϵ_j, θ_{jk}, and τ, which are necessary for the Gibbs sampler. Specification of the hyperparameters a_1, b_1, a_2, and b_2 will be done in Section 2.4.1.

2.3.1 Updating σ^2

Using a conjugate $\mathcal{IG}(a_1,b_1)$ prior on σ^2 results in an inverse gamma full conditional. Therefore, update σ^2 as

$$\sigma^{2(i)} \sim \mathcal{IG}\left(a_1 + n/2, \left[1/b_1 + 1/2\sum_{j,k}\left(d_{jk} - \theta_{jk}^{(i-1)}\right)^2\right]^{-1}\right), \tag{29}$$

where $n = 2^J - 2^{J_0}$ denotes the sample size of detail wavelet coefficients, and i denotes the i^{th} simulation run. $J-1$ and J_0 refer to the finest and coarsest levels in the wavelet

decomposition, respectively.

2.3.2 Updating z_{jk}

In model (24) we saw that latent variable z_{jk} has a Bernoulli prior with parameter ϵ_j. Its full conditional distribution remains Bernoulli with parameter p_j as in (28). Therefore, the latent variable z_{jk} is updated as follows:

$$z_{jk}^{(i)} = \begin{cases} 0, & \text{wp.} \quad \dfrac{\left(1-\epsilon_j^{(i-1)}\right) f\left(d_{jk}\middle|0,\sigma^{2(i)}\right)}{\left(1-\epsilon_j^{(i-1)}\right) f\left(d_{jk}\middle|0,\sigma^{2(i)}\right) + \epsilon_j^{(i-1)} m\left(d_{jk}\middle|\sigma^{2(i)},\tau^{(i-1)}\right)} \\ 1, & \text{wp.} \quad \dfrac{\epsilon_j^{(i-1)} m\left(d_{jk}\middle|\sigma^{2(i)},\tau^{(i-1)}\right)}{\left(1-\epsilon_j^{(i-1)}\right) f\left(d_{jk}\middle|0,\sigma^{2(i)}\right) + \epsilon_j^{(i-1)} m\left(d_{jk}\middle|\sigma^{2(i)},\tau^{(i-1)}\right)} \end{cases} \tag{30}$$

2.3.3 Updating ϵ_j

Parameter ϵ_j is given a conjugate $\mathcal{B}e(1,1)$ prior. This results in a full conditional distributed as beta. Therefore, we update ϵ_j as

$$\epsilon_j^{(i)} \sim \mathcal{B}e\left(1+\sum_k z_{jk}^{(i)}, 1+\sum_k \left(1-z_{jk}^{(i)}\right)\right). \tag{31}$$

Note that other choices from the $\mathcal{B}e(\alpha,\beta)$ family are possible for the prior of ϵ_j. However, we used the noninformative choice $\alpha = 1$ and $\beta = 1$ to facilitate data-driven estimation of ϵ_j.

2.3.4 Updating θ_{jk}

We approach updating θ_{jk} with a method different than what is commonly done. A standard approach for handling the double exponential prior in MCMC computations of hierarchical models is to represent the double exponential distribution as a scale mixture of normals (Andrews and Mallows, 1974). This approach is used, for example, in Bayesian LASSO variable selection, where the double exponential prior is used on the regression parameters (Park and Casella, 2008; Yuan and Lin, 2005). However, this approach introduces an additional parameter corresponding to each θ_{jk}, which

needs to be updated. This adds $2^J - 2^{J_0}$ new parameters. A faster and more direct method to update θ_{jk} is possible by using results in (26) and (27). From the definition of the latent variable z_{jk} we can easily see that $\theta_{jk} = 0$ if $z_{jk} = 0$, because for such z_{jk}, θ_{jk} is distributed as δ_0 (point mass at zero). In the case $z_{jk} = 1$, θ_{jk} follows a mixture of truncated normal distributions. Therefore, the update for θ_{jk} is as follows:

$$\theta_{jk}^{(i)} \sim \begin{cases} \delta_0(\theta_{jk}), & \text{if} \quad z_{jk}^{(i)} = 0 \\ h\left(\theta_{jk} \middle| d_{jk}, \sigma^{2(i)}, \tau^{(i-1)}\right), & \text{if} \quad z_{jk}^{(i)} = 1 \end{cases} \tag{32}$$

where $\delta_0(\theta)$ is a point mass distribution at zero, and $h(\theta|d, \sigma^2, \tau)$ is a mixture of truncated normal distributions with the density provided in (26). Simulating random variables from $h(\theta|d, \sigma^2, \tau)$ is nonstandard, and regular built-in methods fail, because we need to simulate random variables from tails of the normal distribution having extremely low probability. The implementation of the updating algorithm is based on a fast algorithm proposed by Robert (1995).

2.3.5 Updating τ

The Gibbs updating scheme is completed with the discussion of how to update τ. In the hierarchical model (24), we impose a gamma prior on the scale parameter of the double exponential distribution. This turns out to be a conjugate problem; therefore, we update τ by

$$\tau^{(i)} \sim \mathcal{G}a\left(a_2 + \sum_{j,k} z_{jk}^{(i)}, \left[1/b_2 + \sum_{j,k}\left(z_{jk}^{(i)}|\theta_{jk}^{(i)}|\right)\right]^{-1}\right). \tag{33}$$

Note that the gamma distribution above is parameterized by its scale parameter. Now the derivation of the updating algorithm is complete. The implementation of the described Gibbs sampler requires simulation routines for standard distribution such as the gamma, Bernoulli, beta, and also a specialized routine to simulate from

the truncated normal. The procedure was implemented in MATLAB© and is available from the author.

In the following section, we apply the proposed Gibbs sampling algorithm to denoise simulated test functions.

2.4 Simulations

In this section, we apply the proposed Gibbs sampling algorithm and simulate posterior realizations for the model in (24). We call our method the Gibbs Sampling Wavelet Smoother ($GSWS$). Within each replication of our simulations we performed 10,000 Gibbs sampling iterations, of which the first 5,000 were used for burn-in. We used the sample average $\hat{\theta}_{jk} = \sum_i \theta_{jk}^{(i)}/N$ as the usual estimator for the posterior mean. In our set-up, $N = 5,000$.

First we discuss the selection of the hyperparameters, then present and compare the shrinkage performance results with other established methods on a standard battery of test functions (Donoho and Johnstone, 1994).

2.4.1 Selection of Hyperparameters

In any Bayesian modeling task, the selection of hyperparameters is critical for good performance of the model. It is also desirable to have a default way of selecting the hyperparameters which makes the shrinkage procedure automatic.

In order to apply the $GSWS$ method, we only need to specify hyperparameters a_1, b_1, a_2, and b_2 in the hyperprior distributions. The advantage of the fully Bayesian approach is that once the hyperpriors are set, the estimation of parameters σ^2, ϵ_j, θ_{jk}, and τ is automatic via the Gibbs sampling algorithm. The selection is governed by the data and hyperprior distributions on the parameters. Another advantage is that the method is robust to the choice of hyperparameters since they influence the model at a higher level of hierarchy.

The most critical parameter with respect to the performance of the shrinkage

is ϵ_j, which regulates the strength of shrinkage of a wavelet coefficient to zero. In model (24), we placed a noninformative uniform prior on this parameter; therefore, the estimation will be governed mostly by the data, which provides adaptiveness to the proposed method. In Abramovich et al. (1998), parameter ϵ_j is estimated by a theoretically justified but somewhat cumbersome method, and in Vidakovic and Ruggeri (2001), the estimation of this parameter depends on another hyperparameter γ, which is elicited based on empirical evidence. As the results will show, our method provides somewhat better performance because of the automatic adaptiveness to the underlying test signals.

An efficient way to elicit the hyperparameters of the model is through the empirical Bayes method performing maximization of the marginal likelihood. However, the likelihood function is nonconcave is most cases; therefore, clever optimization algorithms and carefully set starting values are crucial for the good performance of these methods. This method of estimating hyperparameters was used by Clyde and George (1999) and Johnstone and Silverman (2005b), among others. The empirical Bayes approach usually provides good average mean squared error (AMSE) performance, comparable to the proposed method of this chapter. Note that in order to use the empirical Bayes paradigm efficiently, one needs to have marginal distributions in a closed form.

Specification of the hyperparameters a_1, b_1, a_2, and b_2 in model (23) is given by the following:

- We set $a_1 = 2$ and $a_2 = 1$.
- Next we set $b_1 = 1/\hat{\sigma}^2$, so that the mean of the inverse gamma prior is $\hat{\sigma}^2$. We use $\hat{\sigma}^2 = MAD/0.6745$, which is the usual robust estimator of the noise variation in the wavelet shrinkage literature (Donoho and Johnstone, 1994). Here MAD stands for the median absolute deviation of the wavelet coefficients d_{jk} at the finest level of detail, and the constant 0.6745 calibrates the estimator

to be comparable with the sample standard deviation.

- Finally, we set $1/b_2 = \hat{\tau} = \sqrt{\max\{(\sigma_d^2 - \hat{\sigma}^2), 0\}}$, which sets the mean of the gamma prior on τ equal to an estimator of τ. This estimator is adopted from Vidakovic and Ruggeri (2001).

Note that this specification of the hyperparameters is appropriate in the wavelet shrinkage context, but the results are robust to changes of this specification.

2.4.2 Simulations and Comparisons with Various Methods

In this section, we discuss the performance of the proposed *GSWS* estimator and compare it to established and state-of-the-art wavelet-based estimators. In the simulations, four standard test functions (`Blocks, Bumps, Doppler, Heavisine`) were considered (Donoho and Johnstone, 1994). The functions were rescaled so that the added noise with $\sigma = 1$ produced a preassigned signal-to-noise ratio (SNR). The test functions were simulated at $n = 256,\ 512$, and 1024 points equally spaced in the unit interval. Four common SNR's were selected, SNR $= 3, 5, 7$, and 10. Wavelet bases used were: Symmlet 8 for `Heavisine` and `Doppler`, Daubechies 6 for `Bumps`, and Haar for `Blocks`. These pairings of bases and signals are standard in the wavelet literature. The coarsest decomposition level was $J_0 = 3$, which matches the suggested $J_0 = \lfloor \log_2(\log(n)) + 1 \rfloor$ from Antoniadis et al. (2001).

Reconstruction of the theoretical signal was assessed by the average mean squared error (AMSE), calculated as

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

where M is the number of simulation runs, $f(t_i),\ i = 1, \ldots, n$, are known values of the test functions considered, and $\hat{f}_k$ is the estimator from the kth simulation run. In each of these simulation runs we perform 10,000 Gibbs iterations to get the estimators $\hat{\theta}_{jk}$.

Performance of the proposed *GSWS* estimator will be compared to: the *EbayesThresh (EBAYES)* method of Johnstone and Silverman (2005b), the *BAMS* method of Vidakovic and Ruggeri (2001), the *Singlemean (SMEAN)* method of Clyde and George (1999) implemented by Antoniadis et al. (2001), the $\Gamma-$*minimax (GAMMA)* estimator of Angelini and Vidakovic (2004), the classical *VisuShrink (VISU)* of Donoho and Johnstone (1994), *Hybrid-SureShrink (SURE)* of Donoho and Johnstone (1995), the scale invariant term-by-term Bayesian *ABE* method of Figueiredo and Nowak (2001), the term-by-term False Discovery Rate *(FDR)* method of Abramovich and Benjamini (1995), and finally the *NeighCoeff (NC)* method of Cai and Silverman (2001).

Results are summarized in Tables 1 and 2, where boldface numbers indicate the smallest AMSE result for each test scenario. From the results, we can see that the proposed estimator is comparable to the established and state-of-the-art methods, and is superior for some combinations of signals, SNRs, and sample sizes. Out of the 48 test scenarios considered, the *GSWS* method gives the lowest AMSE in 16 cases, which is the best among the methods considered. As evident from Tables 1 and 2, for test functions `Blocks` and `Bumps`, the proposed *GSWS* method provides excellent results; and it also performs consistently well for `Doppler` and `Heavisine`. Figure 5 presents the boxplots of the MSE from $M = 1,000$ simulations for the above 10 methods based on $n = 512$ points and SNR $= 5$.

2.5 Application to Dynamic Flux Estimation

In this section, we apply the proposed *GSWS* method to the real-life problem of analysis of metabolic pathways. Dynamic Flux Estimation (DFE) is a methodological framework for estimating parameters for models of metabolic systems circumventing the costly integration of differential equations. The DFE method consists of two distinct phases: (a) a model- or assumption-free phase which includes data

Table 1: AMSE of the proposed *GSWS* estimator compared to other methods for test signals `Blocks` and `Doppler`.

Signal	N	Method	SNR=3	SNR=5	SNR=7	SNR=10	Signal	N	Method	SNR=3	SNR=5	SNR=7	SNR=10
Blocks	256	GSWS	**0.3135**	**0.2663**	0.2242	0.2012	Doppler	256	GSWS	0.3545	0.3691	0.3668	0.3889
		EBAYES	0.3285	0.2750	0.2278	0.2031			EBAYES	0.3572	0.3776	0.3775	0.3967
		BAMS	0.3343	0.2835	0.2412	0.2080			BAMS	0.3378	0.3821	0.3887	0.4114
		SMEAN	0.3247	0.2688	**0.2197**	0.1919			SMEAN	0.3651	0.3832	0.3876	0.4132
		GAMMA	0.3215	0.2775	0.2296	0.1995			GAMMA	0.3630	0.4067	0.4186	0.4587
		VISU	0.4606	0.3744	0.2467	**0.1854**			VISU	0.4960	0.5563	0.5456	0.5904
		SURE	0.5542	0.4904	0.4083	0.4077			SURE	0.5105	0.6642	0.6681	0.5509
		ABE	0.3356	0.2945	0.2591	0.2383			ABE	0.3520	0.3863	0.4039	0.4256
		FDR	0.4041	0.3225	0.2527	0.2477			FDR	0.4594	0.4796	0.4770	0.5101
		NC	0.6225	0.5716	0.4894	0.4122			NC	**0.3142**	**0.3321**	**0.3528**	**0.3764**
	512	GSWS	0.2060	**0.1841**	**0.1613**	**0.1431**		512	GSWS	0.1928	0.2234	0.2317	0.2367
		EBAYES	0.2122	0.1886	0.1670	0.1478			EBAYES	0.1962	0.2155	0.2211	0.2280
		BAMS	0.2101	0.1943	0.1763	0.1567			BAMS	0.1954	0.2131	0.2264	0.2391
		SMEAN	0.2136	0.1863	0.1640	0.1456			SMEAN	0.1989	0.2183	0.2227	0.2318
		GAMMA	**0.1988**	0.1915	0.1760	0.1596			GAMMA	0.1949	0.2146	0.2250	0.2415
		VISU	0.2769	0.2344	0.1945	0.1693			VISU	0.2578	0.2779	0.2862	0.2992
		SURE	0.3517	0.3653	0.3530	0.2939			SURE	0.2743	0.3797	0.4132	0.4680
		ABE	0.2221	0.2072	0.1967	0.1864			ABE	0.2108	0.2240	0.2325	0.2419
		FDR	0.2442	0.2095	0.1872	0.1712			FDR	0.2370	0.2523	0.2582	0.2675
		NC	0.4103	0.4031	0.3679	0.3199			NC	**0.1684**	**0.1784**	**0.1846**	**0.2046**
	1024	GSWS	0.1513	**0.1161**	**0.0990**	**0.0861**		1024	GSWS	0.1157	0.1397	0.1553	0.1650
		EBAYES	0.1510	0.1207	0.1038	0.0899			EBAYES	0.1168	0.1363	0.1473	0.1554
		BAMS	0.1583	0.1311	0.1107	0.0942			BAMS	0.1180	0.1350	0.1482	0.1590
		SMEAN	0.1489	0.1176	0.1015	0.0881			SMEAN	0.1183	0.1382	0.1503	0.1618
		GAMMA	**0.1486**	0.1241	0.1090	0.0950			GAMMA	0.1177	0.1400	0.1587	0.1690
		VISU	0.2161	0.1510	0.1231	0.1014			VISU	0.1552	0.1855	0.2085	0.2106
		SURE	0.3108	0.2926	0.2274	0.2128			SURE	0.1655	0.1964	0.2363	0.2712
		ABE	0.1695	0.1558	0.1472	0.1393			ABE	0.1554	0.1709	0.1786	0.1838
		FDR	0.1770	0.1358	0.1209	0.1077			FDR	0.1479	0.1738	0.1851	0.1879
		NC	0.3253	0.3088	0.2680	0.2250			NC	**0.0945**	**0.1160**	**0.1241**	**0.1302**

Table 2: AMSE of the proposed *GSWS* estimator compared to other methods for test signals `Bumps` and `Heavisine`.

Signal	N	Method	SNR=3	SNR=5	SNR=7	SNR=10	Signal	N	Method	SNR=3	SNR=5	SNR=7	SNR=10
Bumps	256	GSWS	**0.5255**	**0.5551**	**0.5880**	**0.6632**	Heavisine	256	GSWS	0.1366	0.1793	0.2069	0.2315
		EBAYES	0.5521	0.5874	0.6247	0.7100			EBAYES	0.1262	0.1799	0.2174	0.2497
		BAMS	0.6419	0.6996	0.7554	0.8607			BAMS	0.1462	0.1754	**0.1985**	**0.2245**
		SMEAN	0.5615	0.6073	0.6523	0.7492			SMEAN	0.1271	0.1827	0.2230	0.2589
		GAMMA	0.7348	0.8357	0.9217	1.0626			GAMMA	0.1211	**0.1661**	0.1992	0.2291
		VISU	1.0861	1.0817	1.1575	1.2612			VISU	0.1421	0.2563	0.3236	0.3497
		SURE	0.8599	0.7450	0.8533	0.9906			SURE	**0.1174**	0.2030	0.2454	0.3150
		ABE	0.6556	0.7147	0.7713	0.8520			ABE	0.1682	0.2214	0.2497	0.2704
		FDR	0.8181	0.8012	0.8473	0.9223			FDR	0.1426	0.2662	0.3284	0.3427
		NC	0.7965	0.7650	0.7722	0.7612			NC	0.1198	0.2146	0.2713	0.3105
	512	GSWS	**0.4067**	**0.4374**	**0.4610**	0.4696		512	GSWS	0.0940	0.1183	0.1413	0.1626
		EBAYES	0.4110	0.4417	0.4680	0.4830			EBAYES	0.0842	0.1205	0.1502	0.1742
		BAMS	0.4834	0.5132	0.5573	0.6093			BAMS	0.0957	0.1185	**0.1374**	**0.1584**
		SMEAN	0.4130	0.4511	0.4853	0.5074			SMEAN	0.0850	0.1241	0.1565	0.1824
		GAMMA	0.5182	0.5887	0.6604	0.7282			GAMMA	0.0838	**0.1163**	0.1395	0.1599
		VISU	0.7354	0.7630	0.8146	0.8789			VISU	0.0996	0.1583	0.2028	0.2548
		SURE	0.7052	0.5953	0.6497	0.7071			SURE	**0.0826**	0.1300	0.1751	0.2453
		ABE	0.4601	0.4983	0.5235	0.5396			ABE	0.1315	0.1614	0.1845	0.2065
		FDR	0.5496	0.5704	0.5985	0.5918			FDR	0.1037	0.1677	0.2057	0.2476
		NC	0.5828	0.5273	0.4779	**0.4385**			NC	0.0898	0.1438	0.1759	0.2067
	1024	GSWS	0.2787	0.3005	0.3018	0.3083		1024	GSWS	0.0586	**0.0668**	0.0817	0.0977
		EBAYES	**0.2713**	**0.2921**	0.2956	0.3042			EBAYES	0.0536	0.0693	0.0866	0.1038
		BAMS	0.2969	0.3263	0.3404	0.3508			BAMS	0.0607	0.0707	**0.0815**	**0.0958**
		SMEAN	0.2763	0.2980	0.3034	0.3129			SMEAN	0.0535	0.0704	0.0894	0.1083
		GAMMA	0.3282	0.3747	0.3953	0.3976			GAMMA	**0.0529**	0.0680	0.0840	0.0992
		VISU	0.4496	0.4808	0.4884	0.4532			VISU	0.0683	0.0937	0.1223	0.1619
		SURE	0.3840	0.4676	0.4907	0.4523			SURE	0.0534	0.0747	0.0955	0.1355
		ABE	0.3004	0.3193	0.3240	0.3320			ABE	0.1075	0.1233	0.1360	0.1455
		FDR	0.3566	0.3681	0.3590	0.3627			FDR	0.0734	0.0962	0.1237	0.1498
		NC	0.3217	0.3008	**0.2878**	**0.2923**			NC	0.0667	0.0894	0.0989	0.1127

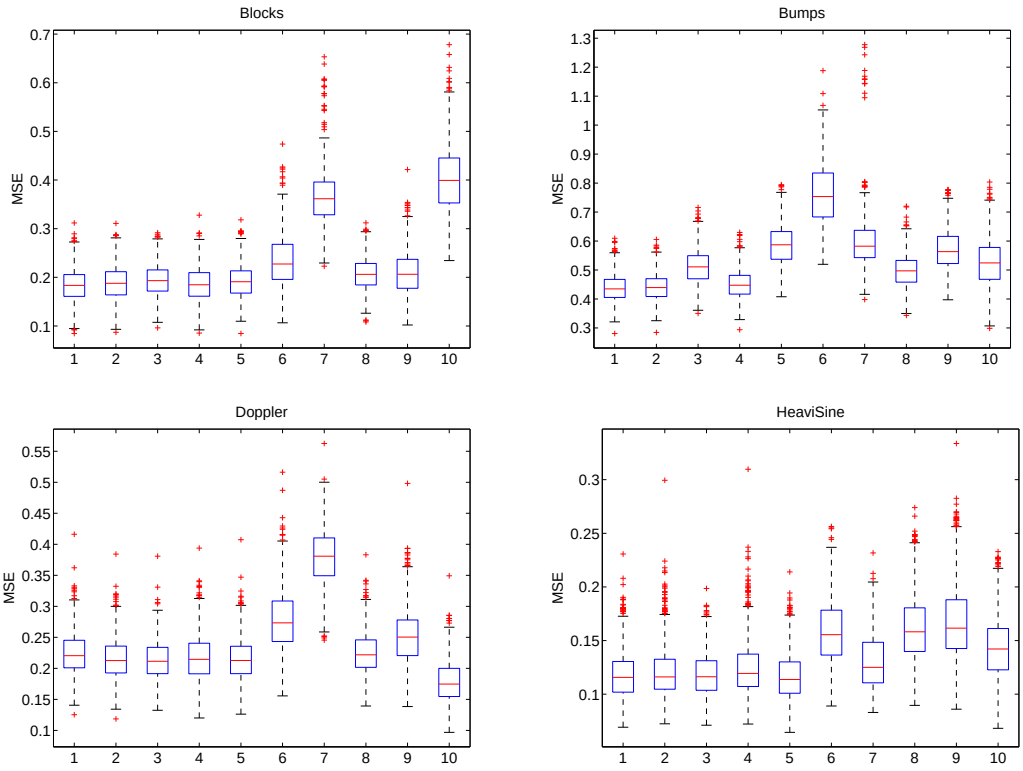

Figure 5: Boxplots of MSE for various denoising procedures based on $n = 512$ points and SNR $= 5$. (1) *GSWS*, (2) *EbayesThresh*, (3) *BAMS*, (4) *Singlemean*, (5) Γ-*minimax*, (6) *VisuShrink*, (7) *Hybrid-SureShrink*, (8) *ABE*, (9) *FDR*, (10) *NeighCoeff*

preprocessing, smoothing, slope estimation, and also uses slope estimators to obtain dynamic time series profiles of all fluxes in the system, and (b) a model-based phase consisting of mathematical characterization of process representations.

In the model-free phase of DFE, we need to smooth and balance the data in the sense that there is no gain or loss of material over time. The balance is checked against the stoichiometry of the system. Existing methods to do this include combined nonlinear programming and a moving-average algorithm to remove noise, and finite-difference approximations, cubic splines, etc. to fulfill smoothing and slope estimation. However, this unconstrained smoothing can be expected to lead to unbalanced time courses, which was actually observed. For more information about Dynamic Flux Estimation, the reader is referred to Goel et al. (2008).

In this section, we apply the *GSWS* method to smooth time series profiles of metabolites while at the same time satisfying the mass balance criteria. We satisfy this requirement by using the smoother iteratively with a constraint on the mass balance, because unconstrained smoothing leads to unbalanced time courses. More precisely, we smooth the time series $f_i(t),\ i = 1, \ldots, M$, under the condition such that their sum remains constant in time, that is, $\sum_i f_i(t) = C$. We propose an iterative smoothing method, in which after each smoothing step the smoothed functions $\hat{f}_i(t)$ are rescaled to balance the total mass of the metabolites. Rescaling simply means that we multiply each $\hat{f}_i(t)$ by the constant $C/\sum_i \hat{f}_i(t)$. This operation is done point-by-point and necessarily "unsmooths" the $\hat{f}_i$s. Then we repeat the process of smoothing and rescaling until convergence is reached. Convergence is achieved when the signal-to-noise ratios of all the $\hat{f}_i(t)$'s become larger than a preassigned threshold T. The procedure will be called the *Constrained Iterative GSWS* (*CI-GSWS*).

A step-by-step description of the *CI-GSWS* procedure is the following:

- STEP 1. Perform a discrete wavelet transform (DWT) of each metabolic time series $f_i(t)$.

- STEP 2. Using the *GSWS* method, shrink the detail wavelet coefficients d^i_{jk}, $j = J_0, \ldots, \log_2(n)$, for each time series i whose SNR is smaller than T. This results in $\hat{\theta}^i_{jk}$.

- STEP 3. Obtain $\hat{f}_i(t)$ by performing an inverse discrete wavelet transform (IDWT) on $\hat{\theta}^i_{jk}$.

- STEP 4. Rescale each smoothed function $\hat{f}_i(t)$ to recover the mass balance equation, $\sum_i \hat{f}_i(t) = C$. If the SNR of each rescaled $\hat{f}_i(t)$ is larger than T, STOP. Otherwise, set each $f_i(t) = \hat{f}_i(t)$, and return to Step 1.

The data set presented here was provided by the collaborators of Dr. Eberhard Voit's Laboratory for Biological Systems Analysis at the Georgia Institute of Technology, the group of Dr. Helena Santos of the Institute for Biotechnology (ITQB) at the New University of Lisbon (Portugal). The metabolomics data of the glycolysis in the bacterium *Lactococcus lactis* is in the form of time series concentration profiles of intermediate metabolites and end products, and were produced using non-invasive *in vivo* Nuclear Magnetic Resonance (NMR) techniques. *Lactococcus lactis* is an industrially important organism that plays an essential role in the manufacture of a wide range of fermented dairy products, such as cheeses and buttermilk. The schematic picture of this system is presented in Figure 6. For more information about the experiment, the reader is referred to Voit et al. (2006).

The data set contains 7 time series of different metabolites with 95 equally spaced observations. The observations were collected at the rate of 2 Hz, and this frequency was limited by the experimental set-up. The names of the measured metabolites were glucose, lactate, fructose-1,6-bisphosphate (FBP), ethanol, 2,3-butanediol, 3-phosphoglycerate (3-PGA) and uridine diphosphate glucose (UDP-glucose). Since the number of measurements was not a power of 2, each time series was extended to 128 observations by repeating the last 33 observations in a "mirror" fashion.

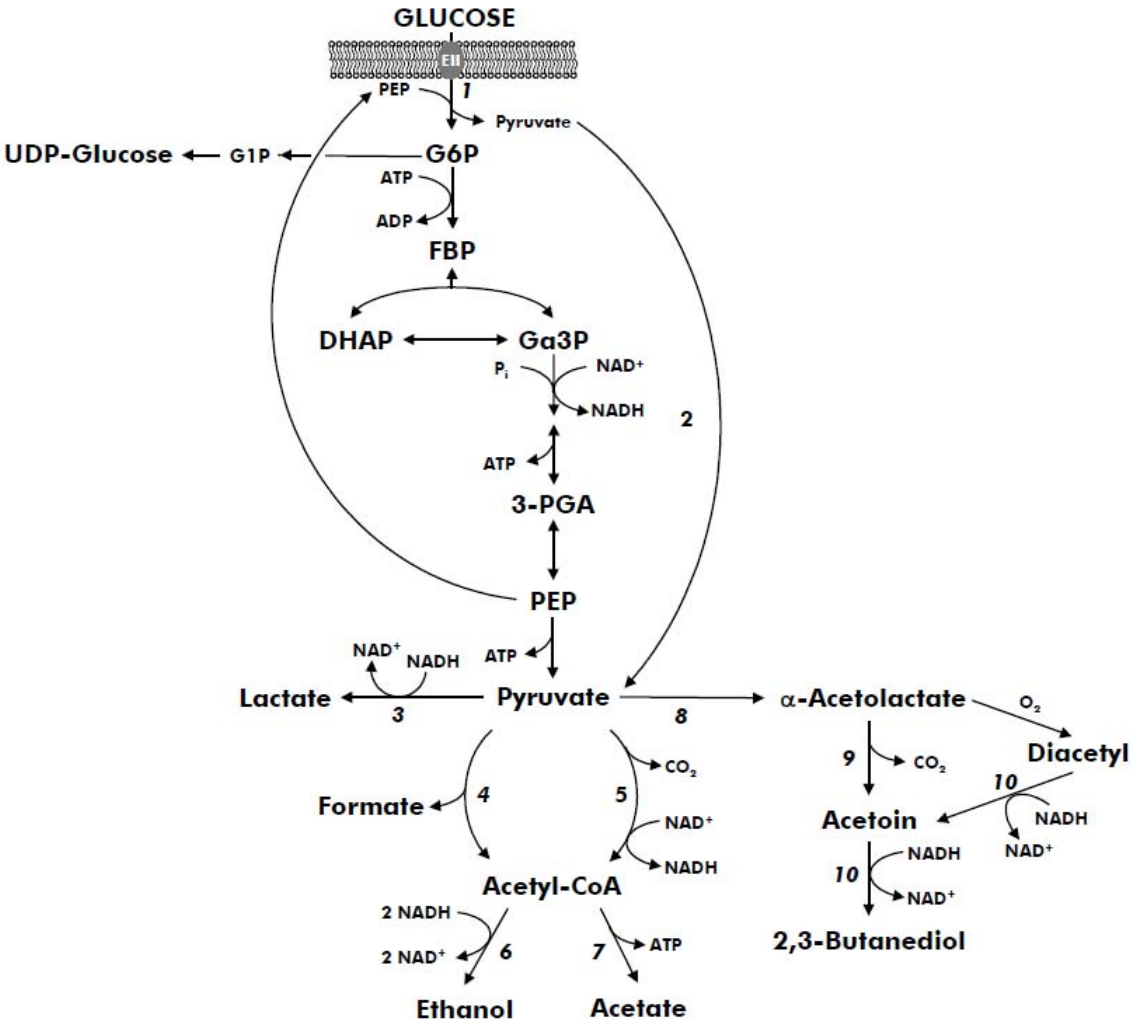

Figure 6: Simplified representation of glycolysis and lactate production in *Lactococcus lactis*.

Due to its superior smoothing performance and the small number of observations in the data set, we used the translation invariant wavelet transform (Coifman and Donoho, 1995; Nason and Silverman, 1995) in Steps 1 and 3. The translation invariant wavelet transform produces less artifacts in the reconstructed signal compared to the traditional (orthogonal) wavelet transform. It suppresses the artifacts by averaging the results of denoising over all circulant shifts of the signal. For our data set it produced smoother results, which can describe a biological phenomena more accurately; therefore, we chose to use it in the denoising procedure.

In Step 4, $T = 50$ was used; therefore, smoothing was only performed on the individual time series in each of the iterations if the SNR was less than 50. Convergence was reached after 3 iterations. Plots of the original and smoothed time series are shown in Figures 7 and 8.

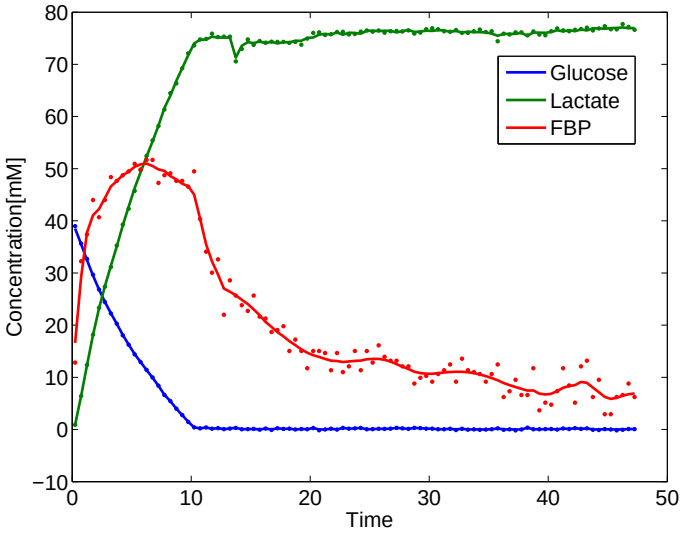

Figure 7: Plots of smoothed time series of concentration of metabolites for Glucose, Lactate, and FBP.

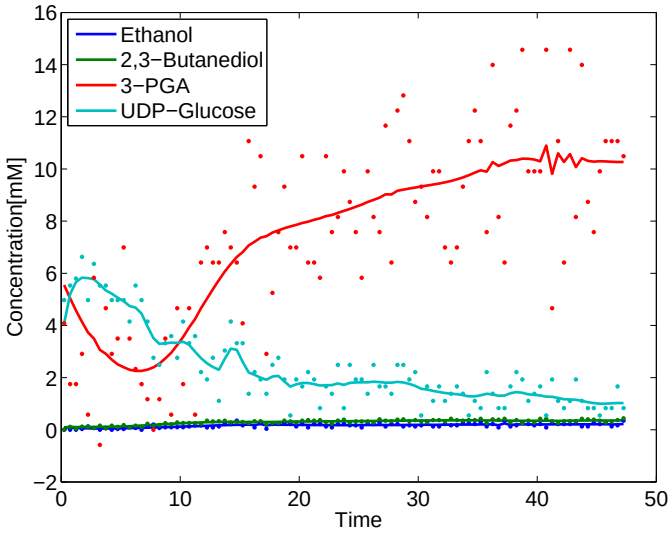

Figure 8: Plots of smoothed time series of concentration of metabolites for Ethanol, 2,3-Butanediol, 3-PGA, and UDP-Glucose.

2.6 Extension to complex wavelets

In Chapter 1, we discussed that wavelet shrinkage methods using complex-valued wavelets provide better denoising performance and additional insights into the shrinkage process. It was explained how the proposed model (23) can be extended to complex wavelets. The model and estimation procedure was briefly outlined in Chapter 1, but here we provide additional details and simulation results.

2.6.1 Model

After applying the complex wavelet transform to a real-valued signal, the observed wavelet coefficients d_{jk} at resolution j and location k become complex numbers. By representing the complex-valued wavelet coefficients as a bivariate real-valued random variables (Barber and Nason, 2004), we need to extend the model in (23) to the bivariate case. In Chapter 1 we introduced the following bivariate Bayesian hierarchical

model on the observed wavelet coefficients d_{jk}:

$$\begin{aligned} d_{jk}|\theta_{jk},\sigma^2 &\sim \mathcal{N}_2(\theta_{jk},\sigma^2\Sigma_j) \\ \sigma^2 &\sim \mathcal{IG}(a,b) \\ \theta_{jk}|z_{jk},v_{jk},C_j &\sim (1-z_{jk})\delta_0 + z_{jk}\,\mathcal{N}_2(0,v_{jk}C_j) \\ z_{jk}|\epsilon_j &\sim \mathcal{B}er(\epsilon_j) \\ \epsilon_j &\sim \mathcal{U}(0,1) \\ v_{jk} &\sim \mathcal{G}a(3/2,8) \\ C_j &\sim \mathcal{IW}(A_j,w), \end{aligned} \tag{34}$$

which is the extension of model (24) to the bivariate case. As it was argued, this model implicitly specifies a mixture of point mass at zero and a bivariate double exponential distribution as a prior on θ_{jk}. We used the term implicitly, because the bivariate double exponential distribution was represented as a scale mixture of bivariate normal distributions. Note again, that variable z_{jk} is indicating whether θ_{jk} comes from a point mass at zero ($z_{jk}=0$) or from a bivariate normal distribution ($z_{jk}=1$). Since the model is bivariate, we specify an inverse Wishart prior on the matrix C_j modeling the covariance structure of the spread prior of θ_{jk}. Similarly to model (24), we specify a "noninformative" uniform prior on mixing weight ϵ_j. Model (34) is tractable in this form because the full conditional distributions of all parameters become explicit, hence a Gibbs sampling algorithm can be developed to update the model parameters.

2.6.2 Gibbs sampling scheme

In this section we provide the details of the Gibbs sampler for model (34). This includes full conditional distributions and updating schemes for parameters σ^2, z_{jk}, ϵ_j, θ_{jk}, v_{jk} and C_j. Specification of hyperparameters a, b, A_j and w will be explained in Section 2.6.3.1. Derivations of results in this section are deferred to Appendix.

2.6.2.1 Updating σ^2

Using a conjugate $\mathcal{IG}(a, b)$ prior on σ^2 results in a full conditional which is inverse gamma. Therefore, update σ^2 as

$$\sigma^{2(i)} \sim \mathcal{IG}\left(a + n, \left[1/b + 1/2 \sum_{j,k} \left(d_{jk} - \theta_{jk}^{(i-1)}\right)' \Sigma_j^{-1} \left(d_{jk} - \theta_{jk}^{(i-1)}\right)\right]^{-1}\right), \quad (35)$$

where $n = 2^J - 2^{J_0}$ denotes the sample size, and i denotes the i^{th} simulation run.

2.6.2.2 Updating z_{jk} and ϵ_j

As for the real-valued case, the full conditional of z_{jk} is Bernoulli and is updated as follows:

$$z_{jk}^{(i)} = \begin{cases} 0, \quad \text{wp.} & \dfrac{\left(1 - \epsilon_j^{(i-1)}\right) f\left(d_{jk}|0, \sigma^{2(i)}\right)}{\left(1 - \epsilon_j^{(i-1)}\right) f\left(d_{jk}|0, \sigma^{2(i)}\right) + \epsilon_j^{(i-1)} m\left(d_{jk}|\sigma^{2(i)}, v_{jk}^{(i-1)}, C_j^{(i-1)}\right)} \\ 1, \quad \text{wp.} & \dfrac{\epsilon_j^{(i-1)} m\left(d_{jk}|\sigma^{2(i)}\right)}{\left(1 - \epsilon_j^{(i-1)}\right) f\left(d_{jk}|0, \sigma^{2(i)}\right) + \epsilon_j^{(i-1)} m\left(d_{jk}|\sigma^{2(i)}, v_{jk}^{(i-1)}, C_j^{(i-1)}\right)} \end{cases}, \quad (36)$$

where

$$\begin{aligned} f(d_{jk}|0, \sigma^2) &= \frac{1}{2\pi|\sigma^2\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2} d_{jk}' \Sigma_j^{-1} d_{jk}\right\}, \\ m(d_{jk}|\sigma^2, v_{jk}, C_j) &= \frac{1}{2\pi|\sigma^2\Sigma_j + v_{jk}C_j|^{1/2}} \exp\left\{-\frac{1}{2} d_{jk}' \left(\sigma^2\Sigma_j + v_{jk}C_j\right)^{-1} d_{jk}\right\}. \end{aligned}$$

Similarly, the full conditional distribution of parameter ϵ_j is beta, and the updating scheme is

$$\epsilon_j^{(i)} \sim \mathcal{B}e\left(1 + \sum_k z_{jk}^{(i)}, 1 + \sum_k \left(1 - z_{jk}^{(i)}\right)\right). \quad (37)$$

2.6.2.3 Updating θ_{jk}

From the conjugate setup of model (34) and using the latent variable z_{jk}, it follows that the full conditional distribution of θ_{jk} is either a point mass at zero ($z_{jk} = 0$),

or a bivariate normal distribution ($z_{jk} = 1$). Therefore, we will update θ_{jk} as follows:

$$\theta_{jk}^{(i)} \sim \begin{cases} \delta_0(\theta_{jk}), & \text{if} \quad z_{jk}^{(i)} = 0 \\ f\left(\theta_{jk} \middle| d_{jk}, \sigma^{2(i)}, v_{jk}^{(i-1)}, C_j^{(i-1)}\right), & \text{if} \quad z_{jk}^{(i)} = 1 \end{cases}, \tag{38}$$

where

$$\begin{aligned} f(\theta_{jk}|d_{jk}, \sigma^2, v_{jk}, C_j) &= \frac{1}{2\pi|\tilde{\Sigma}_{jk}|^{1/2}} \exp\left\{-\frac{1}{2}\tilde{\mu}'_{jk}\tilde{\Sigma}_{jk}^{-1}\tilde{\mu}_{jk}\right\}, \\ \tilde{\mu}_{jk} &= \tilde{\Sigma}_{jk}\frac{\Sigma_j^{-1}}{\sigma^2}d_{jk}, \\ \tilde{\Sigma}_{jk} &= \left(\Sigma_j^{-1}/\sigma^2 + C_j^{-1}/v_{jk}\right)^{-1}. \end{aligned}$$

2.6.2.4 Updating v_{jk}

In model (34) we placed a gamma prior on v_{jk} for the scale mixture of normals representation. The full conditional distribution of v_{jk} depends on the value of z_{jk}, and the updating scheme is:

$$v_{jk}^{(i)} \sim \begin{cases} \mathcal{G}a(3/2, 8), & \text{if} \quad z_{jk}^{(i)} = 0 \\ \mathcal{GIG}\left(1/4, \theta_{jk}^{(i)'}\left\{C_j^{(i-1)}\right\}^{-1}\theta_{jk}^{(i)}, 1/2\right), & \text{if} \quad z_{jk}^{(i)} = 1 \end{cases}. \tag{39}$$

Here $\mathcal{GIG}(a, b, p)$ denotes the generalized inverse Gaussian distribution (Johnson et al., 1994, p.284) with probability density function

$$f(x|a, b, p) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}x^{p-1}e^{-(ax+b/x)/2}, \quad x > 0; a, b > 0,$$

where K_p denotes the modified Bessel function of the third kind. Simulation of $\mathcal{GIG}$ random variates is available through a MATLAB implementation "randraw" based on Dagpunar (1989).

2.6.2.5 Updating C_j

Placing a conjugate inverse Wishart prior on covariance matrix C_j results in a full conditional distribution which is inverse Wishart. Therefore, update C_j as:

$$C_j^{(i)} \sim \mathcal{IW}\left(A_j + \sum_k z_{jk}^{(i)}\frac{\theta_{jk}^{(i)}\theta_{jk}^{(i)'}}{v_{jk}^{(i)}}, w + \sum_k z_{jk}^{(i)}\right). \tag{40}$$

The implementation of the described Gibbs sampler requires simulation routines for standard distribution such as inverse gamma, Bernoulli, beta, normal and also a specialized routine to simulate from generalized inverse Gaussian. In the following section we apply this extended Gibbs sampling algorithm to denoise simulated test functions.

2.6.3 Simulations

Similarly as before, in this section we apply the proposed Gibbs sampling algorithm for the model in (34). The complex extension of *GSWS* will be called Complex Gibbs Sampling Wavelet Smoother (*CGSWS*). The setup remains the same, within each replication of simulations we performed 10,000 Gibbs sampling iterations, of which the first 5,000 was burn-in. We used the sample average $\hat{\theta}_{jk} = \sum_i \theta_{jk}^{(i)}/N$ as the usual estimator for the posterior mean. In our set-up $N = 5,000$. After discussing the selection of hyperparameters, we compare the denoising performance of the method to the methods proposed by Barber and Nason (2004).

2.6.3.1 Selection of Hyperparameters

To implement model (34) we need to specify hyperparameters a, b, A_j and w. For simplicity we set $a = 2$ and $b = 1/\hat{\sigma}^2$, where

$$\hat{\sigma}^2 = \big(\text{MAD}(d_{jk}^{re}/0.6745)\big)^2 + \big(\text{MAD}(d_{jk}^{im}/0.6745)\big)^2, \quad j = \log_2(n) - 1.$$

This ensures that the mean of the inverse gamma prior on σ^2 is the standard robust estimator of the noise variation (Donoho and Johnstone, 1994). Here MAD stands for the median absolute deviation of the wavelet coefficients, which we calculate at the finest level of detail from both the real and imaginary parts of wavelet coefficients (Barber and Nason, 2004).

Hyperparameters A_j and w play an important role in the prior of the covariance matrix C_j. Since in the Gibbs updates $\sum_k z_{jk}^{(i)}$ and therefore $\sum_k z_{jk}^{(i)}\theta_{jk}^{(i)}\theta_{jk}^{(i)'}/v_{jk}^{(i)}$ can

possibly be zero, a noninformative Jeffreys prior on C_j is not computationally feasible. Also note that the mean of the inverse Wishart prior is $A_j/(w-p-1)$, where p is the dimension of A_j, being 2 in our case. Therefore, we set $A_j = (w-2-1)\hat{C}_j$, which forces the mean of the prior to be a pre-specified estimate of C_j. In the case of the mixture bivariate double exponential prior, the covariance of the signal part is $\mathrm{Cov}(\theta_{jk}) = \epsilon_j^2\, 12\, C_j$, where $12\, C_j$ is the covariance of a bivariate double exponential random variable (Gomez et al., 1998). Since the model assumes independence of signal and error parts, we have that $\mathrm{Cov}(d_{jk}) = \epsilon_j^2\, 12\, C_j + \sigma^2\Sigma_j$, where $\mathrm{Cov}(d_{jk})$ is the covariance of the observations d_{jk} at j^{th} dyadic level. We choose $\epsilon_j = 1/\sqrt{12}$ as a reasonable estimate, which additionally simplifies the equation in hand. Therefore, a reasonable estimator for C_j is

$$\hat{C}_j = \mathrm{Cov}(d_j) - \hat{\sigma}^2\Sigma_j, \quad J_0 \le j \le \log_2 n - 1, \tag{41}$$

where $\mathrm{Cov}(d_j)$ is the sample covariance estimator using observations d_{jk} at j^{th} dyadic level. Note that Σ_j is known, and $\hat{\sigma}^2$ is the usual robust estimator of the variance of wavelet coefficients introduced before. Note that when $\hat{C}_j$ is not positive definite, we regularize it by adding a multiple of the identity matrix. Finally, we set $w = 10$. Note, that $w = 4$ is the least informative choice in our case, however, we found that a slightly higher w worked better in practice.

2.6.3.2 Comparison with Barber and Nason (2004)

Here we perform a simulation study using the *CGSWS* method and compare its performance to two of the complex wavelet-based denoising methods introduced by Barber and Nason (2004). The first one (*CMWS-Hard*) is a phase preserving estimator based on hard thresholding of a "thresholding statistic" $d'_{jk}\Sigma_j^{-1}d_{jk}$. The second one (*CEB-Posterior mean*) is a bivariate posterior mean estimator based on an empirical Bayes procedure. The simulation setup is the same as before for the real case, except that we used the symmetric complex-valued Daubechies wavelet base with 3

vanishing moments for all the test functions.

Reconstruction of the theoretical signal was measured by the average mean squared error (AMSE) over 100 simulation runs. The results are summarized in Table 3, where boldface numbers indicate the smallest AMSE result for each test scenario. The results convey that the proposed *CGSWS* method outperforms both estimators in more than 50% of the cases, and in most cases it is very close in performance to the superior method. The improvement is most pronounced at small sample sizes ($n = 256$) and for the test function `Heavisine`. This result verifies the adaptiveness of the method and the advantage of using a heavy-tailed prior as prior distribution on the location of wavelet coefficients.

Table 3: AMSE of *CGSWS* method compared to estimators *CMWS-Hard* and *CEB-Posterior mean*.

Signal	N	Method	SNR=3	SNR=5	SNR=7	SNR=10	Signal	N	Method	SNR=3	SNR=5	SNR=7	SNR=10
Blocks	256	CGSWS	**0.4293**	**0.4533**	**0.4610**	**0.4499**	Doppler	256	CGSWS	**0.3093**	**0.3119**	**0.3251**	**0.3619**
		CMWS-H	0.4929	0.5476	0.5490	0.5021			CMWS-H	0.3332	0.3351	0.3644	0.4000
		CEB-PM	0.4343	0.4675	0.4715	0.4547			CEB-PM	0.3137	0.3158	0.3351	0.3723
	512	CGSWS	**0.2954**	**0.3180**	0.3138	0.3051		512	CGSWS	0.1854	0.2073	0.2052	**0.2095**
		CMWS-H	0.3481	0.3627	0.3457	0.3166			CMWS-H	0.2048	0.2217	0.2192	0.2289
		CEB-PM	0.3028	0.3202	**0.3126**	**0.2995**			CEB-PM	**0.1845**	**0.2007**	**0.2035**	0.2132
	1024	CGSWS	0.1991	0.2013	0.1991	0.1924		1024	CGSWS	**0.1034**	**0.1209**	0.1310	0.1467
		CMWS-H	0.2372	0.2230	0.2098	0.1944			CMWS-H	0.1160	0.1329	0.1432	0.1601
		CEB-PM	**0.1980**	**0.1988**	**0.1947**	**0.1879**			CEB-PM	0.1087	0.1225	**0.1302**	**0.1419**
Bumps	256	CGSWS	**0.4631**	**0.4825**	**0.4946**	**0.5181**	Heavisine	256	CGSWS	**0.1198**	**0.1640**	**0.1900**	**0.2030**
		CMWS-H	0.5972	0.5946	0.5853	0.5809			CMWS-H	0.1547	0.2075	0.2144	0.2198
		CEB-PM	0.4855	0.4996	0.5120	0.5390			CEB-PM	0.1338	0.1838	0.2098	0.2188
	512	CGSWS	**0.3273**	**0.3274**	**0.3235**	**0.3203**		512	CGSWS	**0.0799**	**0.1050**	**0.1258**	0.1429
		CMWS-H	0.3983	0.3760	0.3538	0.3317			CMWS-H	0.0959	0.1202	0.1357	**0.1371**
		CEB-PM	0.3295	0.3315	0.3287	0.3228			CEB-PM	0.0881	0.1167	0.1340	0.1427
	1024	CGSWS	0.1965	**0.1970**	**0.2009**	**0.2090**		1024	CGSWS	**0.0487**	**0.0650**	**0.0747**	0.0843
		CMWS-H	0.2137	0.2151	0.2134	0.2223			CMWS-H	0.0557	0.0746	0.0793	**0.0791**
		CEB-PM	**0.1919**	0.1986	0.2034	0.2122			CEB-PM	0.0564	0.0730	0.0794	0.0835

2.7 Extensions

It is worth to mention here that the introduced model can be flexible in the choice of prior distributions on the wavelet coefficients. Another reasonable choice is to use Student's t-distribution instead of the double exponential as the spread part in the hierarchical model (23). In general, any distribution which is a scale mixture or normals can be used in the model in a similar way. The model in case of Student's

t-distribution becomes

$$\begin{aligned} d_{jk}|\theta_{jk},\sigma^2 &\sim \mathcal{N}(\theta_{jk},\sigma^2) \\ \sigma^2 &\sim \mathcal{IG}(a_1,b_1) \\ \theta_{jk}|\epsilon_j,\tau &\sim (1-\epsilon_j)\delta_0+\epsilon_j\mathcal{T}(v,\tau^2) \\ \epsilon_j &\sim \mathcal{U}(0,1) \\ \tau^2 &\sim \mathcal{IG}(a_2,b_2) \end{aligned} \tag{42}$$

where $\mathcal{T}(v,\tau^2)$ denotes the 3-parameter Student's t-distribution with mean zero, degrees of freedom v, and scale parameter τ^2. Introducing the additional degrees of freedom parameter v in the prior might give us more flexibility in modeling, but it complicates the MCMC scheme. For sampling purposes, we can represent the Student's t-distribution as a scale mixture of normal distributions (Andrews and Mallows, 1974). After introducing a latent variable z, representing the Student's t-distribution as a scale mixture of normals and inducing a prior on the degrees of freedom v, the model in (42) becomes

$$\begin{aligned} d_{jk}|\theta_{jk},\sigma^2 &\sim \mathcal{N}(\theta_{jk},\sigma^2) \\ \sigma^2 &\sim \mathcal{IG}(a_1,b_1) \\ \theta_{jk}|z_{jk},\tau,\lambda_{jk} &\sim (1-z_{jk})\delta_0+z_{jk}\,\mathcal{N}(0,\tau^2/\lambda_{jk}) \\ z_{jk}|\epsilon_j &\sim \mathcal{B}er(\epsilon_j) \\ \epsilon_j &\sim \mathcal{U}(0,1) \\ \tau^2 &\sim \mathcal{IG}(a_2,b_2) \\ \lambda_{jk}|v &\sim \mathcal{G}a(v/2,2/v) \\ v &\sim \mathcal{LTE}xp(1,1), \end{aligned} \tag{43}$$

where $\mathcal{LTE}xp(\lambda,a)$ denotes the left-truncated exponential distribution with rate λ and truncation point a. Therefore, in the above model, we induced a left-truncated

exponential prior on the degrees of freedom (v), which restricts the degrees of freedom parameter to the interval $[1, \infty)$. In practice this means that the spread prior can have tails between a Cauchy ($v = 1$) and a normal ($v \to \infty$) distribution. With the exponential prior, we favor heavy-tailed distributions (v is small), which is appropriate for modeling wavelet coefficients. Since the full conditional distribution for parameter v is not available in a known form, we will use a Metropolis-within-Gibbs algorithm for inferential purposes. The algorithm for updating the parameters is the following:

Step 1.

$$\sigma^{2(i)} \sim \mathcal{IG}\left(a_1 + n/2, \left[1/b_1 + 1/2 \sum_{j,k} \left(d_{jk} - \theta_{jk}^{(i-1)}\right)^2\right]^{-1}\right)$$

Step 2.

$$z_{jk}^{(i)} = \begin{cases} 0, & \text{wp.} \quad \dfrac{\left(1-\epsilon_j^{(i-1)}\right) f\left(d_{jk}\middle|0, \sigma^{2(i)}\right)}{\left(1-\epsilon_j^{(i-1)}\right) f\left(d_{jk}\middle|0, \sigma^{2(i)}\right) + \epsilon_j^{(i-1)} m\left(d_{jk}\middle|\sigma^{2(i)}, \tau^{2(i-1)}, \lambda_{jk}^{(i-1)}\right)} \\ 1, & \text{wp.} \quad \dfrac{\epsilon_j^{(i-1)} m\left(d_{jk}\middle|\sigma^{2(i)}, \tau^{2(i-1)}, \lambda_{jk}^{(i-1)}\right)}{\left(1-\epsilon_j^{(i-1)}\right) f\left(d_{jk}\middle|0, \sigma^{2(i)}\right) + \epsilon_j^{(i-1)} m\left(d_{jk}\middle|\sigma^{2(i)}, \tau^{2(i-1)}, \lambda_{jk}^{(i-1)}\right)} \end{cases}$$

Step 3.

$$\epsilon_j^{(i)} \sim \mathcal{B}e\left(1 + \sum_k z_{jk}^{(i)}, 1 + \sum_k \left(1 - z_{jk}^{(i)}\right)\right)$$

Step 4.

$$\theta_{jk}^{(i)} \sim \begin{cases} \delta_0(\theta_{jk}), & \text{if} \quad z_{jk}^{(i)} = 0 \\ f\left(\theta_{jk}\middle|d_{jk}, \sigma^{2(i)}, \tau^{2(i-1)}, \lambda_{jk}^{(i-1)}\right), & \text{if} \quad z_{jk}^{(i)} = 1 \end{cases}$$

Step 5.

$$\tau^{2(i)} \sim \mathcal{IG}\left(a_2 + \sum_{j,k} z_{jk}^{(i)}/2, \left[1/b_2 + 1/2 \sum_{j,k} \left(z_{jk}^{(i)} \theta_{jk}^{2\ (i)} \lambda_{jk}^{(i-1)}\right)\right]^{-1}\right)$$

Step 6.

$$\lambda_{jk}^{(i)} \sim \begin{cases} \mathcal{G}a\left(v^{(i-1)}/2, 2/v^{(i-1)}\right), & \text{if} \quad z_{jk}^{(i)} = 0 \\ \mathcal{G}a\left(\left(v^{(i-1)}+1\right)/2, \left[1/2\left(v^{(i-1)} + \theta_{jk}^{2\ (i)}/\tau^{2(i)}\right)\right]^{-1}\right), & \text{if} \quad z_{jk}^{(i)} = 1 \end{cases}$$

Step 7. (Metropolis in Gibbs)

Let $v^{old} = v^{(i-1)}$ and sample $v^{new} \sim \mathcal{LTN}(v^{old}, \psi, 1)$

Let $u \sim \mathcal{U}(0,1)$

Let $r = \dfrac{\prod_{j,k}\left\{f_{Ga}\left(\lambda_{jk}^{(i)} | v^{new}/2, 2/v^{new}\right)\right\} f_{Exp}(v^{new}-1|1)\Phi_{(v\geq 1)}(v^{old}|v^{new},\psi)}{\prod_{j,k}\left\{f_{Ga}\left(\lambda_{jk}^{(i)} | v^{old}/2, 2/v^{old}\right)\right\} f_{Exp}(v^{old}-1|1)\Phi_{(v\geq 1)}(v^{new}|v^{old},\psi)}$

Set $v^{(i)} = v^{new}$ if $u \leq \min(r,1)$, otherwise $v^{(i)} = v^{old}$

(44)

In the above sampling scheme $f(d_{jk}|0,\sigma^2)$, $m(d_{jk}|\sigma^2,\tau^2,\lambda_{jk})$ and $f(\theta_{jk}|d_{jk},\sigma^2,\tau^2,\lambda_{jk})$ are normal distributions, because we represented the Student's t prior as scale mixture of normals. The distributions are given as

$$\begin{aligned} f(d_{jk}|0,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-d_{jk}^2/2\sigma^2} \\ m(d_{jk}|\sigma^2,\tau^2,\lambda_{jk}) &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2/\lambda_{jk})}} e^{-d_{jk}^2/2(\sigma^2+\tau^2/\lambda_{jk})} \\ f(\theta_{jk}|d_{jk},\sigma^2,\tau^2,\lambda_{jk}) &= \frac{1}{\sqrt{2\pi\tilde{\sigma}_{jk}^2}} e^{-(\theta_{jk}-\tilde{\mu}_{jk})^2/2\tilde{\sigma}_{jk}^2} \\ \text{where} \quad \tilde{\mu}_{jk} &= \frac{\tau^2/\lambda_{jk}}{\sigma^2+\tau^2/\lambda_{jk}} d_{jk}, \\ \text{and} \quad \tilde{\sigma}_{jk}^2 &= \frac{\tau^2/\lambda_{jk}}{\sigma^2+\tau^2/\lambda_{jk}} \sigma^2. \end{aligned}$$

In the Metropolis step of the sampling scheme $\mathcal{LTN}(\mu, \sigma^2, a)$ denotes the left-truncated normal distribution with truncation point a. Furthermore, f_{Ga}, f_{Exp} and $\Phi_{(v\geq 1)}$ denote the probability density functions of gamma, exponential and left-truncated normal distributions, respectively. Note that we sampled v from a truncated normal proposal distribution. Parameter ψ is the scale parameter of the truncated normal proposal distribution, which was tuned in the burn-in period to achieve acceptance rate of $25 - 50\%$ for parameter v (Müller, 1993). Simulations of the above algorithm provides results in AMSE sense similar to the *GSWS* method, which uses the double exponential as the spread prior. Detailed explanations of some of the results can be found in Appendix.

2.8 Conclusions

In this chapter we proposed Gibbs Sampling Wavelet Smoother (*GSWS*), a wavelet-based method for nonparametric regression. A fully Bayesian approach was taken, in which a hierarchical model was formulated accounting for the uncertainty of the prior parameters by placing hyperpriors on them. A mixture prior was specified on the wavelet coefficients with a double exponential spread distribution accounting for the large wavelet coefficients. Since all the full conditional distributions were available in an explicit distributional form, an efficient Gibbs sampling procedure was developed to estimate the parameters of the model. The *GSWS* provided excellent denoising performance, which was demonstrated by simulations on well-known test functions and by comparisons with other standardly used methods. The method was illustrated on a real-world data set from analysis of metabolic pathways, for which we applied the denoising procedure iteratively to preserve the mass balance of the system. We also extended the method to complex wavelet bases, which involved a bivariate hierarchical model. We showed how the Gibbs sampling procedure was applied in this case and compared its denoising performance to a state-of-the-art denoising method

that uses complex wavelet bases. Finally we demonstrated that the model is flexible and different distributions, for example Student's t, can be used as a spread part of a mixture prior on the wavelet coefficients.

CHAPTER III

WAVELET SHRINKAGE WITH DOUBLE WEIBULL PRIOR

In this chapter we propose a denoising methodology in the wavelet domain based on a Bayesian hierarchical model using double Weibull prior. We propose two estimators, one based on posterior mean ($DWWS$) and the other based on larger posterior mode ($DWWS$-LPM), and show how to calculate them efficiently. Traditionally, mixture priors have been used for modeling sparse wavelet coefficients. The interesting feature of this chapter is the use of non-mixture prior. We show that the methodology provides good denoising performance, comparable even to state-of-the-art methods that use mixture priors and empirical Bayes setting of hyperparameters, which is demonstrated by extensive simulations on standardly used test functions. An application to real-word data set is also considered.

3.1 Introduction

In the present chapter we consider a novel Bayesian model in the wavelet domain as a solution to the classical nonparametric regression problem

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{45}$$

where x_i, $i = 1, \ldots, n$, are equispaced sampling points, and the errors ε_i are i.i.d. normal random variables, with zero mean and variance σ^2. The interest is to estimate the function f from the observations y_i. After applying a linear and orthogonal wavelet transform, the equation in (45) 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 ε respectively. Note that ε_i and ε_{jk} are equal in distribution due to orthogonality of wavelet transforms. Due to the whitening property of the wavelet transforms (Flandrin, 1992) many of the existing methods assume independence of the coefficients, and omit the double indices jk to work with a generic wavelet coefficient model

$$d = \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{46}$$

The indices will be used when needed for clarity of the exposition.

To estimate θ in model (46) Bayesian shrinkage rules have been proposed in the literature by many authors. By a shrinkage rule the observed wavelet coefficients d are replaced with their shrunk version $\hat{\theta} = \delta(d)$. Then f is estimated as the inverse wavelet transform of $\hat{\theta}$. Empirical distributions of detail wavelet coefficients for signals encountered in practical applications are (at each resolution level) centered around and peaked at zero (Mallat, 1989). A range of models, for which unconditional distribution of wavelet coefficients mimic this observation, have been considered in the literature. The traditional Bayesian models consider prior distribution on the wavelet coefficient θ as

$$\pi(\theta) = \epsilon\delta_0 + (1 - \epsilon)\xi(\theta), \tag{47}$$

where δ_0 is a point mass at zero, ξ is symmetric about 0, unimodal distribution, and ϵ is a fixed parameter in [0,1], usually level dependent, that controls the amount of shrinkage for values of d close to 0. This type of model was considered by Abramovich et al. (1998), Vidakovic (1998a), Clyde and George (1999, 2000), Vidakovic and Ruggeri (2001) and Johnstone and Silverman (2005b), among others.

The above models provide good denoising performance because of their adaptivity provided by the point mass at zero. However, parameter ϵ, which controls the extent of shrinkage, needs to be specified. One of the contributions of this chapter is

simplification of the traditional mixture prior. We demonstrate that, in the wavelet context, a single prior can match the performance of more complex contamination priors from (47).

The chapter is organized as follows. Section 3.2 introduces the model and discusses the advantage of using the double Weibull prior. Section 3.3 explains the computation of two Bayes' estimators for our model, the posterior mean and the larger posterior mode. Section 3.4 contains simulations and comparisons to selected existing methods. Section 3.5 includes application of the method to inductance plethysmography data. Some remarks and discussion are provided in Sections 3.6 and 3.7.

3.2 Model

In our chapter we consider the following Bayesian model

$$\begin{aligned} d|\theta &\sim \mathcal{N}(\theta, \sigma^2) \\ \theta &\sim \mathcal{DW}(b, c), \end{aligned} \tag{48}$$

where $\mathcal{N}(\theta, \sigma^2)$ denotes the normal distribution and $\mathcal{DW}(b, c)$ denotes the double Weibull distribution with probability density function

$$\pi(\theta|b, c) = \frac{c}{2b}|\theta|^{c-1} \exp\left\{-\frac{|\theta|^c}{b}\right\},$$

where b and c are the scale and shape parameters, respectively. The standard Weibull distribution is popular for analyzing lifetime data. However, its symmetric relative, the double Weibull distribution, introduced by Balakrishnan and Kocherlakota (1985), is not extensively used in the literature, and have not been used in the wavelet shrinkage context previously. Balakrishnan and Kocherlakota (1985) considered a 3-parameter version of this distribution with location parameter a, but in our case $a = 0$ since the prior on the wavelet coefficient θ is always centered at zero, due to the definition of detail wavelet coefficients.

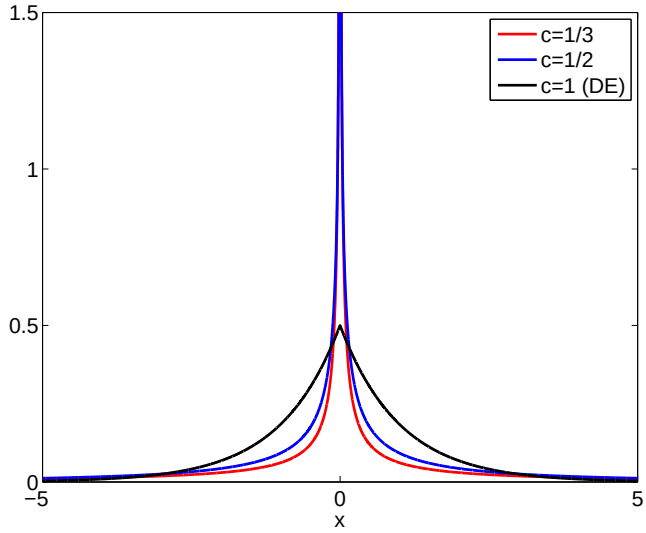

Figure 9: Double Weibull distribution for different values of c.

The double Weibull is a flexible family, which includes the double exponential distribution as its special case ($c = 1$). Figure 9 shows the double Weibull density for $b = 1$ and $c = 1/3, 1/2, 1$. In case of $c < 1$, the double Weibull density approaches infinity as $|\theta|$ approaches zero. This property of the prior will be crucial for the performance of the induced Bayes estimators. The singularity at zero mimics the effect of a point mass at zero in the mixture priors mentioned above. A prior with similar property was considered implicitly by Cutillo et al. (2008), and explicitly by Carvalho et al. (2010). Carvalho et al. (2010) consider the "Horseshoe" prior in form of a scale mixture of normal densities and use it in a context of sparse estimation. The Horseshoe prior, however, does not exist in a closed form.

The shrinkage estimator for the wavelet coefficient corresponding to the signal part θ, derived from (48) is fully specified by eliciting the hyperparameters σ^2, b and c. In this chapter, we consider two such estimators and evaluate their performance. The first is the posterior mean, which is a traditional choice in Bayesian estimation problems and the second is the "larger posterior mode", denoted as LPM in the sequel. The shrinkage procedure based on the posterior mean will be referred as

Double Weibull Wavelet Shrinker ($DWWS$), while the one based on the LPM will have the acronym *DWWS-LPM*. The existence of the LPM is an intrinsic characteristic of the considered Bayesian model (likelihood-prior). For more information on the LPM approach the reader is referred to Cutillo et al. (2008).

3.3 The Bayes Estimator

In this section we provide details of how to find the posterior mean and LPM as the proposed shrinkage estimators.

3.3.1 Posterior Mean

It is well known that the posterior mean, as an estimator of θ, has the following form

$$\delta(d) = \frac{\int \theta g(d,\theta) d\theta}{m(d)} = \frac{\int \theta f(d|\theta)\pi(\theta) d\theta}{\int f(d|\theta)\pi(\theta) d\theta}, \tag{49}$$

where g is the joint distribution, f is the likelihood, π is the prior, and m is the marginal distribution. From the marginal distribution

$$m(d) \propto \int e^{-\frac{(d-\theta)^2}{2\sigma^2}} |\theta|^{c-1} e^{-\frac{|\theta|^c}{b}} d\theta, \tag{50}$$

it can be seen that the integral does not exist in a closed form for fixed $c < 1$. However, the integral in (50) is finite, the posterior distribution is proper, and the posterior mean exists, as well. This is true because we are convolving the normal with the double Weibull distribution, which is integrable and all of its moments exist (Balakrishnan and Kocherlakota, 1985).

It is possible to evaluate this integral as a convolution using the characteristic functions of the likelihood and the prior, but the characteristic function of the double Weibull distribution does not have a simple form and involves special functions (Nadarajah, 2008). Therefore the posterior mean will be computed by numerical integration using adaptive Gauss-Kronrod quadrature, for which we utilized the function

script `quadgk(fun,a,b)` in MATLAB©. It is apparent in equation (50) that the integral has a singularity at $\theta = 0$ for $c < 1$. One can significantly increase the speed and accuracy of integration by removing this singularity, which can be done with a change of variable. After a change of variable $y = \{\text{sign}(\theta)\theta\}^c$ the posterior mean becomes

$$\delta(d) = \frac{\displaystyle\int_0^\infty y^{1/c} e^{-y/b} e^{-\frac{\left(d-y^{1/c}\right)^2}{2\sigma^2}} dy - \int_0^\infty y^{1/c} e^{-y/b} e^{-\frac{\left(d+y^{1/c}\right)^2}{2\sigma^2}} dy}{\displaystyle\int_0^\infty e^{-y/b} e^{-\frac{\left(d-y^{1/c}\right)^2}{2\sigma^2}} dy + \int_0^\infty e^{-y/b} e^{-\frac{\left(d+y^{1/c}\right)^2}{2\sigma^2}} dy}. \tag{51}$$

If c has a form of $1/n$, n odd, the posterior mean simplifies to

$$\delta(d) = \frac{\displaystyle\int_{-\infty}^\infty y^{1/c} e^{-|y|/b} e^{-\frac{\left(d-y^{1/c}\right)^2}{2\sigma^2}} dy}{\displaystyle\int_{-\infty}^\infty e^{-|y|/b} e^{-\frac{\left(d-y^{1/c}\right)^2}{2\sigma^2}} dy}. \tag{52}$$

Note that for any $c \in (0, 1)$ the posterior mean can be efficiently computed using 51. Figure 10 shows the posterior mean for $c = 1/3$, $b = 0.4$ and $\sigma^2 = 1$.

Figure 11 shows the marginal distribution $m(d)$, computed numerically for $c = 1/3$, $b = 1$ and $\sigma^2 = 1$. The marginal distribution is compared to a normal distribution with mean zero and standard deviation 2.6, which arises from matching the interquartile range of the two distributions. It is a desirable property in Bayesian wavelet shrinkage to produce a marginal that matches the observed empirical distribution of wavelet coefficients. We can see from Figure 11 that the marginal distribution corresponding to model (48) exhibits heavier tails, and it is more peaked than the normal density. This is in agreement with the observations of Mallat (1989) concerning the shape of empirical distributions of wavelet coefficients.

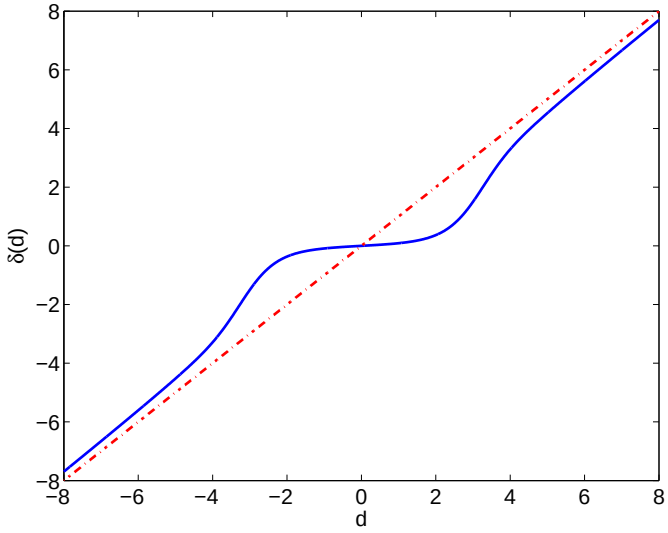

Figure 10: Posterior mean for $c = 1/3$.

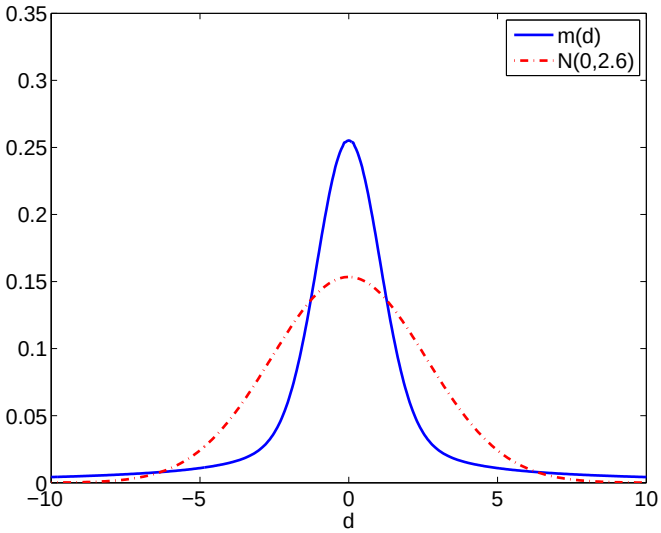

Figure 11: Marginal distribution of the wavelet coefficients.

Setting of the hyperparameters for the *DWWS* rule (49) is discussed in Section 3.4.1.

3.3.2 Larger Posterior Mode (LPM)

The LPM estimator was first introduced in the wavelet shrinkage context by Cutillo et al. (2008), and it is based on the Bayesian MAP (Maximum a Posteriori) principle. The LPM rule relates to the mode of the posterior distribution larger in absolute value. The MAP estimator of the wavelet coefficient θ is a rule maximizing the posterior $\pi(\theta|d)$, which is proportional to the joint distribution of d and θ, $g(d,\theta)$. Hence, the MAP estimator for θ also maximizes $g(d,\theta)$. For the model in (48) the joint distribution is

$$g(d,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(d-\theta)^2}{2\sigma^2}} \frac{c}{2b} |\theta|^{c-1} e^{-\frac{|\theta|^c}{b}}.$$

This leads to the posterior proportional to

$$\pi(\theta|d) \propto g(d,\theta) \propto e^{-\frac{(d-\theta)^2}{2\sigma^2}} |\theta|^{c-1} e^{-\frac{|\theta|^c}{b}}.$$

Figure 12 shows the posterior distribution for $c = 1/3$, $b = 1$, $\sigma^2 = 1$ and $d = -3, -2, -1, 1, 2, 3$. Note that the shape of posterior depends on the absolute magnitude of the observed wavelet coefficient d. If $|d|$ is small, the posterior mode is unique and equals to 0. For large values of $|d|$ there are two posterior modes and the one larger in magnitude is chosen.

The logarithm of the posterior is proportional to

$$l = \log \pi(\theta|d) \propto -\frac{(d-\theta)^2}{2\sigma^2} + (c-1)\log|\theta| - \frac{|\theta|^c}{b},$$

and has extrema at the solutions of the equation

$$\frac{d-\theta}{\sigma^2} + (c-1)\text{sign}(\theta)\frac{1}{|\theta|} - \frac{c}{b}\text{sign}(\theta)|\theta|^{c-1} = 0,$$

which is equivalent to the equation

$$-\frac{1}{\sigma^2}\theta^2 + \frac{d}{\sigma^2}\theta - \frac{c}{b}|\theta|^c + c - 1 = 0. \tag{53}$$

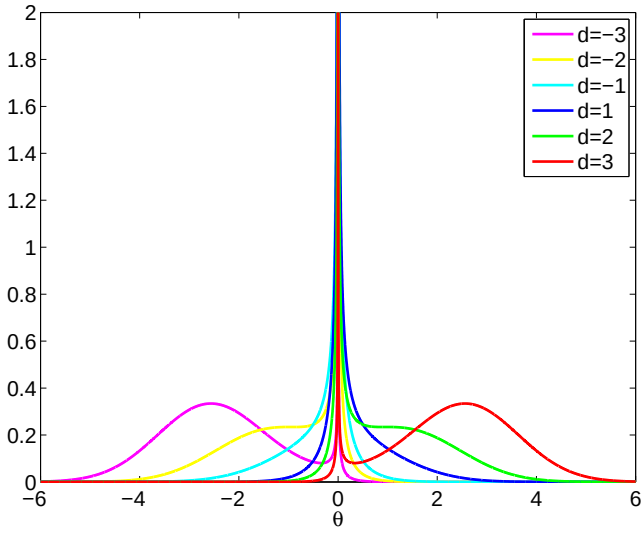

Figure 12: Posterior distribution of the wavelet coefficients.

For fixed $c < 1$ and by substituting $y = |\theta|^c$, equation (53) can be modified so that the solution is equivalent to a solution of a polynomial equation of order $2/c$. We will use the following numerical algorithm to find the LPM estimator from equation (53):

(1) Find the roots of the equation $-\frac{1}{\sigma^2}y^{2/c} + \text{sign}(d)\frac{d}{\sigma^2}y^{1/c} - \frac{c}{b}y + c - 1 = 0$. Denote the roots by $y^\star$ and the real roots by $y_r^\star$.

(2) If all the roots are complex ($y_r^\star$ is empty), $\delta_{LPM}(d) = 0$.

(3) If reals roots exist, $\delta_{LPM}(d) = \text{sign}(d)[\max(y_r^\star)]^{1/c}$.

Therefore, the LPM estimator for the model in (48) is

$$\delta_{LPM}(d) = \text{sign}(d)[\max(y_r^\star)]^{1/c},$$

where $\max(y_r^\star)$ is the maximum real root of the equation $-\frac{1}{\sigma^2}y^{2/c} + \text{sign}(d)\frac{d}{\sigma^2}y^{1/c} - \frac{c}{b}y + c - 1 = 0$. If no real root of this equation exist, $\delta_{LPM}(d) = 0$. In general, the roots can be computed by a nonlinear equation solver for any real $c \in (0, 1)$, but for a rational $c = m/n$ the roots can be found by a polynomial root solver, which was utilized in the implementation. Figure 13 shows the LPM rule for $c = 1/3$, $b = 0.4$ and $\sigma^2 = 1$. It is apparent from the figure that the rule is thresholding.

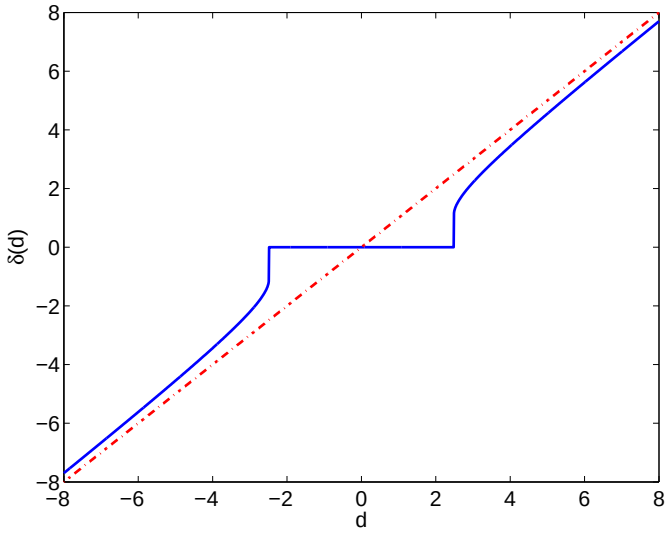

Figure 13: LPM rule for $c = 1/3$.

3.4 Simulations

In this section we apply the proposed shrinkage estimators and compare their performance to several existing and established wavelet denoising methods. For the *DWWS* and *DWWS-LPM* estimators we first discuss the selection of hyperparameters, then we present and compare the simulation results.

3.4.1 Selection of Hyperparameters

In any Bayesian modeling task the selection of hyperparameters is critical for good performance of the model. It is also desirable to have a default selection of the hyperparameters which makes the shrinkage procedure automatic. In the model (48) we need to specify parameters σ^2, b and c.

Parameter σ^2. Parameter σ^2 represents the variance of the random error ε. In the wavelet shrinkage literature σ^2 is frequently estimated by a robust estimator of the variance of wavelet coefficients at the finest level of detail (Donoho and Johnstone, 1994). We will adopt this practice and use the robust MAD estimator to estimate σ as $\hat{\sigma} = MAD/0.6745$. Here MAD stands for the median absolute deviation from the

median of the wavelet coefficients at finest level of detail, and the constant 0.6745 calibrates the estimator to be comparable with sample standard deviation.

Parameter b**.** Scale parameter b accounts for the spread of the double Weibull prior distribution. We propose a moment matching parameter specification, which was used for example by Cutillo et al. (2008) and Vidakovic and Ruggeri (2001). We propose to estimate b_j levelwise for all dyadic levels $J_0 \leq j \leq \log_2 n - 1$. Because of the linearity of wavelet transform, the i.i.d. normal noise with variance σ^2 transforms stochastically unchanged to each dyadic level. In the case of the double Weibull prior, the variance of the signal part is $\sqrt[c]{b_j^2}\Gamma(1 + 2/c)$. Since the model assumes independence of signal and error parts, we have $\sigma_{d_j}^2 = \sqrt[c]{b_j^2}\Gamma(1 + 2/c) + \sigma^2$, where $\sigma_{d_j}^2$ is the variance of the observations d_{jk} at j^{th} dyadic level. Therefore a reasonable estimator for b_j is

$$\hat{b}_j = \left\{ \frac{(\sigma_{d_j}^2 - \hat{\sigma}^2)_+}{\Gamma(1 + 2/c)} \right\}^{c/2}, \quad J_0 \leq j \leq \log_2 n - 1, \tag{54}$$

where $a_+ = \max(a, 0)$. In case $\hat{\sigma}^2 > \sigma_{d_j}^2$, we set $\hat{b}_j = 0$. Having $\hat{b}_j = 0$ is equivalent to a degenerate/point-mass-at-zero prior distribution on the wavelet coefficients. Therefore, if $\hat{b}_j = 0$, we set all the wavelet coefficients at level j to zero, similarly to Vidakovic and Ruggeri (2001).

Parameter c**.** Parameter c accounts for the shape of the prior distribution on the wavelet coefficients. When smaller than 1, parameter c controls the "strength of infinity" at zero. In this sense the role of c is similar to that of the point mass in the mixture prior models, and should be elicited depending on the signal regularity. In addition to this, parameter c also controls the tails of the prior distribution. We used $c = 1/3$ in our simulations, which empirically was the superior universal choice. Of course, c can be adaptively set depending on the input signal under consideration, as we will do in Section 3.5.

Figure 14 shows the exact risks of the posterior mean estimator for different values

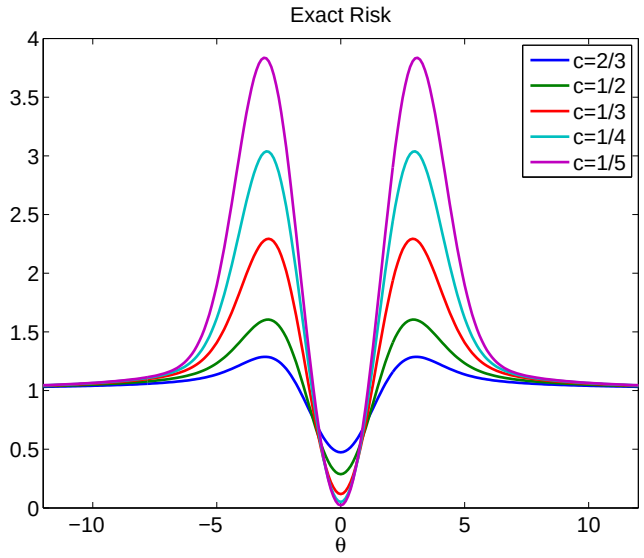

Figure 14: Exact risk plot for posterior mean, for $\sigma^2 = 1$ and $\sigma^2_{d_j} = 100$.

of c. We set $\sigma^2 = 1$, $\sigma^2_{d_j} = 100$ and specified b by equation (54) depending on different c's. From the plot we can see that the choice $c = 1/3$ is a good compromise in terms of risk. For $|\theta|$ close to zero $c = 1/3$ provides smaller risk than $c = 1/2$ or $c = 2/3$, and for larger $|\theta|$ the choice $c = 1/3$ has smaller risk than $c = 1/4$ or $c = 1/5$. Note that the pattern and shape of the plot depends on the quantity $\sigma^2_{d_j} - \sigma^2$, but $c = 1/3$ was an empirically superior choice.

For $c = 1/3$, equation (53) becomes

$$-\frac{1}{\sigma^2}\theta^2 + \frac{d}{\sigma^2}\theta - \frac{1}{3b}|\theta|^{1/3} - 2/3 = 0,$$

and the algorithm to find the LPM estimator becomes equivalent to solving the equation

$$-\frac{1}{\sigma^2}y^6 + \text{sign}(d)\frac{d}{\sigma^2}y^3 - \frac{1}{3b}y - 2/3 = 0. \tag{55}$$

Note that it is possible to specify parameter c levelwise similar to specifying the weight parameter in the Bayesian mixture prior models. Therefore, if elicited levelwise, c could be set up to increase from the finest to the coarsest dyadic levels of wavelet coefficients. However, because of simplicity and the good performance provided, c

was held fixed through the dyadic levels and the levelwise elicitation of parameter b provided the adaptiveness of the shrinkage rule.

3.4.2 Simulations and Comparisons with Various Methods

In this section we discuss the performance of the proposed *DWWS* and *DWWS-LPM* estimators and compare them to some established wavelet-based methods for reconstructing noisy signals. Four standard test functions (`Blocks, Bumps, Doppler, Heavisine`) were considered (Donoho and Johnstone, 1994) in the simulation study. The functions were rescaled such that the added noise ($\sigma^2 = 1$) produced the preassigned signal-to-noise ratio (SNR). The test functions were simulated at $n = 512, 1024,$ and 2048 points equally spaced in the unit interval. Three common SNRs were selected, SNR $= 3, 5,$ and 7. The standard wavelet bases were used: Symmlet 8 for `Heavisine` and `Doppler`, Daubechies 6 for `Bumps` and Haar for `Blocks`. The coarsest decomposition level was $J_0 = 3$, which matches the suggested $J_0 = \lfloor \log_2(\log(n)) + 1 \rfloor$ by Antoniadis et al. (2001). Note, that for computing the *DWWS* estimator, MATLAB's built-in Gauss-Kronrod quadrature method was used, and the *DWWS-LPM* estimator is the solution of equation (55), for which MATLAB's built-in polynomial root-solver was used.

Reconstruction of the theoretical signal was evaluated by the average mean squared error (AMSE), calculated as

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

where M is the number of simulation runs and $f(t_i),\ i = 1, \ldots, n$ are known values of the test functions considered. We denote by $\hat{f}_k(t_i),\ i = 1, \ldots, n$ the estimator from the kth simulation run.

The proposed estimators were compared to the *EbayesThresh* method of Johnstone and Silverman (2005b) using the posterior mean, the *BAMS* method of Vidakovic and Ruggeri (2001), the *LPM* method from Model 1 of Cutillo et al. (2008), the classical

VisuShrink and *Hybrid-SureShrink* of Donoho and Johnstone (1994, 1995), the scale invariant term-by-term *ABE* method of Figueiredo and Nowak (2001), and finally the *NeighCoeff* method of Cai and Silverman (2001). Note that methods *EbayesThresh*, *BAMS*, *LPM* and *ABE* are Bayesian.

Results are summarized in Tables 4 and 5, where boldface numbers indicate the smallest AMSE result for each test scenario. The number of simulations performed was $M = 1000$. From the results we can see that the proposed estimators are comparable to the established shrinkage methods. In some scenarios involving `Heavisine` signal the *DWWS* is superior. Simulations further indicate that the *DWWS* estimator outperforms the *BAMS* estimator in 64% of the cases, and the *EbayesThresh* method in 28% of the cases. This is remarkable considering that these Bayesian methods are based on a more complicated mixture model with a point mass, and the latter one uses an empirical Bayes procedure to estimate the hyperparameters. It is also evident from Tables 4 and 5, that the *DWWS-LPM* estimator outperforms the *LPM* estimator in 67% of the cases. Note that for the model in Cutillo et al. (2008) the posterior distribution is not proper for all values of the hyperparameter k, hence the posterior mean does not exist. For the proposed model in (48) the posterior mean always exists and the resulting *DWWS* estimator uniformly outperforms the *DWWS-LPM* estimator. However, *DWWS-LPM* is computationally more robust and faster to compute. Also note, that the authors of *LPM* select hypermarameter k separately for each simulated test function, so the results are optimal. In our simulation study we kept hyperparameter c default for each test function. It can also be seen from the results that the *DWWS-LPM* estimator outperforms the *ABE* method in 81% of the cases. The difference in AMSE was the most pronounced for signals `Doppler` and `Heavisine`. The *ABE* is also using a single prior model and the MAP approach. Finally, the proposed methods outperform the non-Bayesian methods *VisuShrink*, *Hybrid-SureShrink* and *NeighCoeff* under most test scenarios.

Graphical summary of the results is presented in Figure 15 where the boxplots of the MSE are given for $n = 1024$ and SNR $= 5$.

Table 4: AMSE of the proposed *DWWS* and *DWWS-LPM* estimators compared to other methods for test signals `Blocks` and `Doppler`.

Signal	N	Method	SNR=3	SNR=5	SNR=7	Signal	N	Method	SNR=3	SNR=5	SNR=7
Blocks	512	DWWS	0.2174	0.1917	0.1790	Doppler	512	DWWS	0.2002	0.2244	0.2296
		DWWS-LPM	0.2223	0.1940	0.1826			DWWS-LPM	0.2061	0.2315	0.2389
		EBAYES	0.2122	**0.1886**	**0.1670**			EBAYES	0.1962	0.2155	0.2211
		BAMS	**0.2101**	0.1943	0.1763			BAMS	0.1954	0.2131	0.2264
		LPM	0.2217	0.1949	0.1756			LPM	0.2110	0.2258	0.2353
		VISU	0.2769	0.2344	0.1945			VISU	0.2578	0.2779	0.2862
		SURE	0.3517	0.3653	0.3530			SURE	0.2743	0.3797	0.4132
		ABE	0.2221	0.2072	0.1967			ABE	0.2108	0.2240	0.2325
		NC	0.4103	0.4031	0.3679			NC	**0.1684**	**0.1784**	**0.1846**
	1024	DWWS	0.1563	0.1289	0.1241		1024	DWWS	0.1141	0.1348	0.1469
		DWWS-LPM	0.1567	0.1329	0.1281			DWWS-LPM	0.1241	0.1456	0.1561
		EBAYES	**0.1510**	**0.1207**	**0.1038**			EBAYES	0.1168	0.1363	0.1473
		BAMS	0.1583	0.1311	0.1107			BAMS	0.1180	0.1350	0.1482
		LPM	0.1596	0.1284	0.1130			LPM	0.1349	0.1584	0.1681
		VISU	0.2161	0.1510	0.1231			VISU	0.1552	0.1855	0.2085
		SURE	0.3108	0.2926	0.2274			SURE	0.1655	0.1964	0.2363
		ABE	0.1695	0.1558	0.1472			ABE	0.1554	0.1709	0.1786
		NC	0.3253	0.3088	0.2680			NC	**0.0945**	**0.1160**	**0.1241**
	2048	DWWS	0.0919	0.0816	0.0795		2048	DWWS	0.0624	0.0771	0.0884
		DWWS-LPM	0.0944	0.0852	0.0835			DWWS-LPM	0.0685	0.0846	0.0953
		EBAYES	**0.0865**	**0.0730**	**0.0603**			EBAYES	0.0642	0.0773	0.0860
		BAMS	0.0921	0.0788	0.0665			BAMS	0.0687	0.0783	0.0868
		LPM	0.0914	0.0774	0.0643			LPM	0.0755	0.0887	0.0978
		VISU	0.1172	0.0919	0.0712			VISU	0.0835	0.1003	0.1121
		SURE	0.1740	0.1815	0.1629			SURE	0.0845	0.1184	0.1514
		ABE	0.1227	0.1161	0.1108			ABE	0.1158	0.1242	0.1297
		NC	0.1938	0.1798	0.1587			NC	**0.0511**	**0.0636**	**0.0714**

3.5 Application to Inductance Plethysmography Data

In this section we apply the proposed wavelet estimators to a real-world data set from anaesthesiology 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. This was analysed by several authors, for example Nason (1996) and Abramovich et al. (1998, 2002). For more information about the data set, we refer the reader to these two papers.

Figure 16 shows a section of plethysmograph recording lasting approximately 80 s ($n = 4096$ observations). Figure 17 shows the reconstructions of the signal with the *DWWS* and *DWWS-LPM* methods. In our reconstruction we set $c = 1/5$,

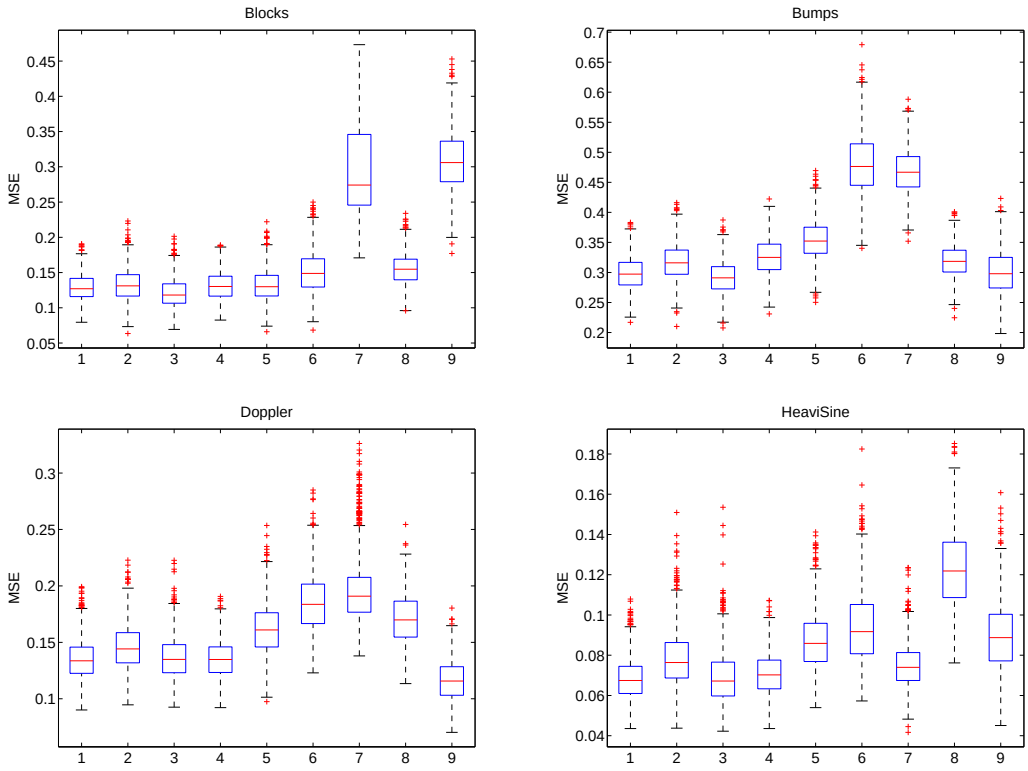

Figure 15: Boxplots of MSE for various shrinking procedures based on $n = 1024$ points and SNR $= 5$. (1) *DWWS*, (2) *DWWS-LPM*, (3) *EbayesThresh*, (4) *BAMS*, (5) *LPM*, (6) *VisuShrink*, (7) *Hybrid-SureShrink*, (8) *ABE*, (9) *NeighCoeff*

Table 5: AMSE of the proposed *DWWS* and *DWWS-LPM* estimators compared to other methods for test signals `Bumps` and `Heavisine`.

Signal	N	Method	SNR=3	SNR=5	SNR=7	Signal	N	Method	SNR=3	SNR=5	SNR=7
Bumps	512	DWWS	0.4659	0.4733	0.4875	Heavisine	512	DWWS	**0.0793**	0.1199	0.1534
		DWWS-LPM	0.4908	0.5128	0.5270			DWWS-LPM	0.0912	0.1337	0.1696
		EBAYES	**0.4110**	**0.4417**	**0.4680**			EBAYES	0.0842	0.1205	0.1502
		BAMS	0.4834	0.5132	0.5573			BAMS	0.0957	**0.1185**	**0.1374**
		LPM	0.4606	0.4885	0.5052			LPM	0.0932	0.1445	0.1800
		VISU	0.7354	0.7630	0.8146			VISU	0.0996	0.1583	0.2028
		SURE	0.7052	0.5953	0.6497			SURE	0.0826	0.1300	0.1751
		ABE	0.4601	0.4983	0.5235			ABE	0.1315	0.1614	0.1845
		NC	0.5828	0.5273	0.4779			NC	0.0898	0.1438	0.1759
	1024	DWWS	0.2855	0.2986	0.3004		1024	DWWS	**0.0504**	**0.0683**	0.0890
		DWWS-LPM	0.3057	0.3174	0.3156			DWWS-LPM	0.0583	0.0783	0.1008
		EBAYES	**0.2713**	**0.2921**	0.2956			EBAYES	0.0536	0.0693	0.0866
		BAMS	0.2969	0.3263	0.3404			BAMS	0.0607	0.0707	**0.0815**
		LPM	0.3168	0.3318	0.3308			LPM	0.0635	0.0867	0.1121
		VISU	0.4496	0.4808	0.4884			VISU	0.0683	0.0937	0.1223
		SURE	0.3840	0.4676	0.4907			SURE	0.0534	0.0747	0.0955
		ABE	0.3004	0.3193	0.3240			ABE	0.1075	0.1233	0.1360
		NC	0.3217	0.3008	**0.2878**			NC	0.0667	0.0894	0.0989
	2048	DWWS	0.1717	0.1871	0.1905		2048	DWWS	**0.0313**	0.0457	0.0560
		DWWS-LPM	0.1836	0.1965	0.2007			DWWS-LPM	0.0376	0.0534	0.0630
		EBAYES	**0.1668**	**0.1816**	**0.1866**			EBAYES	0.0339	**0.0456**	0.0543
		BAMS	0.1823	0.1978	0.2049			BAMS	0.0402	0.0471	**0.0531**
		LPM	0.2033	0.2110	0.2120			LPM	0.0395	0.0609	0.0760
		VISU	0.2766	0.2948	0.2863			VISU	0.0416	0.0653	0.0887
		SURE	0.2438	0.2907	0.3071			SURE	0.0344	0.0506	0.0709
		ABE	0.2039	0.2132	0.2167			ABE	0.0925	0.1037	0.1103
		NC	0.1824	0.1840	0.1877			NC	0.0435	0.0543	0.0599

which provided a smoother, visually more pleasing result, although this choice is not necessarily AMSE superior. Both methods remove the noise well, however, the *DWWS* estimator based on the posterior mean provides a slightly smoother result. Abramovich et al. (2002) report the height of the maximum peak while analysing this data set. In our case the height is 0.8410 for the *DWWS* method and 0.8421 for the *DWWS-LPM*. These are quite close to the result 0.8433, obtained by Abramovich et al. (2002), and better compared to some established methods reported in their paper.

3.6 Remarks

It is worth mentioning here that a slight modification of the double Weibull prior can lead to a Bayes rule which can be expressed as a closed form using special functions.

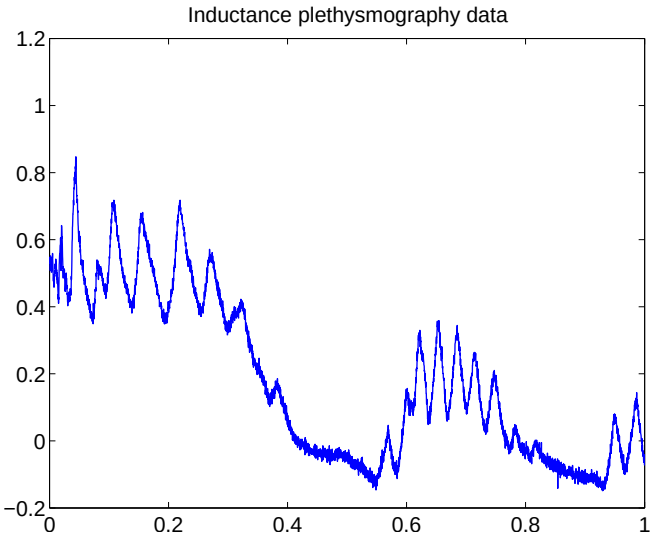

Figure 16: A section of inductance plethysmography data with $n = 4096$.

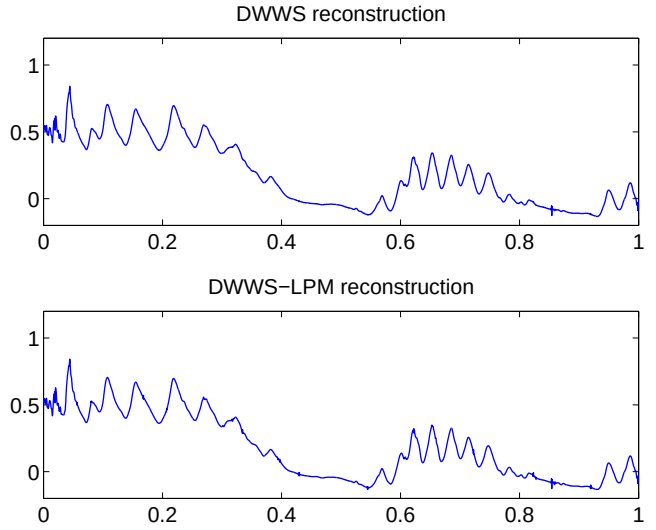

Figure 17: Reconstruction of inductance plethysmography data obtained by the *DWWS* and *DWWS-LPM* methods.

Consider the following prior distribution on the wavelet coefficient θ:

$$\pi(\theta|b,c) = \frac{1}{2\Gamma(c)b^c}|\theta|^{c-1}\exp\left\{-\frac{|\theta|}{b}\right\},$$

which is the one dimensional special case of the more general Kotz distribution (Nadarajah, 2003) with $p = 1$, $\mu = 0$, $\Sigma = 1$, $N = (c+1)/2$, $s = 1/2$ and $r = 1/b$. Using an integral identity (Gradshteyn and Ryzhik, 1980, p.337), the marginal distribution and the posterior mean can be expressed as

$$\begin{aligned} m(d) &= \frac{\sigma^c e^{-d^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}2b^c}\left\{e^{(\sigma/2b-d/2\sigma)^2}D_{-c-1}(\sigma/b-d/\sigma) - e^{(\sigma/2b+d/2\sigma)^2}D_{-c-1}(\sigma/b+d/\sigma)\right\}, \\ \delta(d) &= c\sigma\frac{e^{-d/2b}D_{-c-1}(\sigma/b-d/\sigma) - e^{d/2b}D_{-c-1}(\sigma/b+d/\sigma)}{e^{-d/2b}D_{-c}(\sigma/b-d/\sigma) - e^{d/2b}D_{-c}(\sigma/b+d/\sigma)}, \end{aligned}$$

where $D_v(x)$ is the parabolic cylinder function (Abramowitz and Stegun, 1964).

Because the marginal distribution is available in a closed form, the empirical Bayes procedure is a possibility for eliciting the hyperparameters of the prior. However, in practice, this estimator is computationally more expensive than *DWWS*, *DWWS-LPM*, and the performance in terms of AMSE is somewhat inferior.

3.7 Conclusions

In this chapter we have proposed a methodology for Bayesian wavelet denoising. A hierarchical model was specified in which the double Weibull distribution was utilized as the prior on the locations of wavelet coefficients. In contrast to mixture priors used by some state-of-the-art methods, the wavelet coefficients were modeled by a density with single expression. The flexibility of the double Weibull distribution was able to mimic the characteristics of mixture priors consisting of a point mass at zero and a heavy-tailed spread part. Two Bayesian estimators were proposed, one as the posterior mean (*DWWS*) and the other as the larger posterior mode (*DWWS-LPM*). We also showed how to compute them efficiently. Simulations on standard test functions and comparisons with numerous existing methods demonstrated that

the methodology provides good and comparable denoising performance, even compared to state-of-the-art methods that use mixture priors and empirical Bayes setting of hyperparameters. Once again, we emphasize that the aim was the simplicity of the model, and demonstration that a carefully selected single prior could match the performance of more complex mixture priors. An application to real-word data set (inductance plethysmography) was also considered. The methodology performed well in both denoising and preserving the important features of the real data.

Future improvements of the method are possible by specifying hyperparameter c based on dyadic levels and signal regularity. Another avenue for future improvement can be the approximation of integral in (49) to evaluate the posterior mean. However, if approximations are asymptotic, this would work satisfactorily only in the case of shrinkage of multiple related signals (Chang and Vidakovic, 2002).

In the spirit of reproducible research we made MATLAB scripts used in simulation for *DWWS* and *DWWS-LPM* available at http://gtwavelet.bme.gatech.edu/.

CHAPTER IV

Λ-NEIGHBORHOOD WAVELET SHRINKAGE

We propose a wavelet-based denoising methodology based on total energy of a neighboring pair of coefficients plus their "parental" coefficient. The model is based on a Bayesian hierarchical model using a contaminated exponential prior on the total mean energy in a neighborhood of wavelet coefficients. The hyperparameters in the model are estimated by the empirical Bayes method, and the posterior mean, median and Bayes factor are obtained and used in the estimation of the total mean energy. Shrinkage of the neighboring coefficients are based on the ratio of the estimated and observed energy. It is shown that the methodology is comparable and often superior to several existing and established wavelet denoising methods that utilize neighboring information, which is demonstrated by extensive simulations on a standard battery of test functions. An application to real-word data set from inductance plethysmography is also considered and an extension to image denoising is discussed.

4.1 Introduction

In the present chapter we consider a new Bayesian model as a solution to the classical nonparametric regression problem

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{56}$$

where x_i, $i = 1, \ldots, n$, are equispaced sampling points, and the random errors ε_i are i.i.d. normal, with zero mean and variance σ^2. The interest is to estimate the function f using the observations Y_i. After applying a linear and orthogonal wavelet transform, the equation in (56) becomes

$$d_{j,k} = \theta_{j,k} + \varepsilon_{j,k},$$

where $d_{j,k}$, $\theta_{j,k}$ and $\varepsilon_{j,k}$ are the wavelet coefficients (at resolution j and position k) corresponding to Y, f and ε, respectively. Due to the the whitening property of wavelet transforms (Flandrin, 1992) many existing methods assume independence of the coefficients, and omit the double indices j, k to work with a generic model

$$d = \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{57}$$

To estimate θ in model (57) Bayesian shrinkage rules have been proposed in the literature by many authors. The observed wavelet coefficients d are replaced with their shrunk version $\hat{\theta} = \delta(d)$ representing a Bayes estimator of θ. Most of the signals encountered in practical applications have (at each resolution level) empirical distributions of detail wavelet coefficients centered around and peaked at zero (Mallat, 1989). A range of models complying with Mallat's observation have been considered in the literature. The traditional Bayesian models consider prior distribution on the wavelet coefficient θ as

$$\pi(\theta) = \epsilon\delta_0 + (1 - \epsilon)\xi(\theta), \tag{58}$$

where δ_0 is a point mass at zero, ξ is a unimodal distribution symmetric about 0, and ϵ is a fixed parameter in [0,1], usually level dependent, that controls the amount of shrinkage for values of d close to 0. This type of model was considered by Abramovich et al. (1998), Vidakovic (1998b), Vidakovic and Ruggeri (2001), and Johnstone and Silverman (2005b), among others.

On the other hand, many authors argued that shrinkage performance can be improved by considering the neighborhoods of wavelet coefficients (blocks, parent-child relations, cones of influence, etc.). These authors report improvement over the coefficient-by-coefficient or diagonal methods. Examples include block thresholding methods by Hall et al. (1997, 1998, 1999), Cai (1999, 2002), Cai and Silverman (2001) where wavelet coefficients are thresholded based on block sums of squares. Abramovich et al. (2002) and De Candiitis and Vidakovic (2004) consider Bayesian

block shrinkage methods allowing for dependence between the wavelet coefficients. Wang and Wood (2006) considered a Bayesian block shrinkage approach based directly on the block sum of squares. Sendur and Selesnick (2002) and Fryzlewicz (2007), among others, use parent-child neighboring relation to improve on shrinkage performance. In Fryzlewicz (2007), the coupling of wavelet coefficients from different levels leads to a bivariate model in which the energy, under appropriate assumptions, is χ^2-distributed.

In this chapter the neighboring structure is enhanced by looking simultaneously at two neighboring coefficients at the same level j and their common parental coefficient from the level $j-1$ in the wavelet ordering given by the parametrization $2^{j/2}\phi(2^j x - k)$, $j, k \in \mathbb{Z}$. This leads to a joint "energy" distributed as noncentral χ^2, in which the Bayesian model accounts for the noncentrality parameter and leads to simple and fast shrinkage rules. The idea of considering neighboring and parental coefficients in the denoising procedure has been used in signal and particularly in image denoising algorithms to improve the performance and visual appearance of the procedures. One example is the tree-based wavelet thresholding estimator, see for example Autin (2008) and Autin et al. (2011). Another popular method that incorporates neighboring structure is the Hidden Markov Tree (HMT) model which is explored for example by Crouse et al. (1998) and Romberg et al. (2001).

The chapter is organized as follows. Section 4.2 introduces the model, then derives and discusses the properties of the shrinkage rule. Section 4.3 explains the elicitation of hyperparameters via the empirical Bayes method. Section 4.4 contains simulations and comparisons to existing methods in terms of average mean squared error and Section 4.5 contains an application of the method to real-world data. Extension to image denoising is discussed in Section 4.6, and finally, conclusions and discussion are provided in Section 4.7.

4.2 Model and Estimation

The idea of our method is to estimate wavelet coefficients $\theta_{j,k}$ by forming a χ^2_3 variable composed of two neighboring and their parental wavelet coefficient. Our model is based on the sum of the energy of this "family" or "clique". We will call it Λ-neighborhood motivated by the geometric shape of the neighborhood. This also motivates the name of the induced shrinkage methodology: Λ-neighborhood wavelet shrinkage ($LNWS$). The idea of using a χ^2_p variable as a "thresholding statistics" was considered in different contexts by several authors, for example by Downie and Silverman (1998), Barber and Nason (2004), Wang and Wood (2006) and Fryzlewicz (2007), among others. Wang and Wood (2006) considered forming a model based on blocks of m neighboring wavelet coefficients, while Fryzlewicz (2007) considered two bivariate thresholding methods, one using basis averaging and the other using parental coefficients. Here we build on these ideas and form a block wavelet shrinkage estimator supervised by their parent coefficient. We form the "thresholding statistics" as

$$x_{j,l} = \left(d_{j,k}, d_{j,k+1}, d_{j-1,\lceil k/2 \rceil}\right) \Sigma_j^{-1} \left(d_{j,k}, d_{j,k+1}, d_{j-1,\lceil k/2 \rceil}\right)', \tag{59}$$

where l is short for $\lceil k/2 \rceil$. Σ_j is the covariance matrix of the Λ-neighborhood, for which a schematic picture is provided in Figure 18. Note that the location index of the parental coefficient is $\lceil k/2 \rceil$ for locations k and $k+1$ of the children coefficients.

It is important to emphasize that shrinkage induced by statistic (59) will be applied only to a pair of coefficients in the same level and not on their parent. Since the discrete wavelet transform (DWT) is orthonormal and decorrelating, we model the wavelet coefficients as independent. Hence, we can take $\Sigma_j = \sigma^2 I_3$. Using this assumption, the "thresholding statistics" becomes

$$x_{j,l} = \left(d^2_{j,k} + d^2_{j,k+1} + d^2_{j-1,\lceil k/2 \rceil}\right) / \sigma^2. \tag{60}$$

In (60) the sum of the energy of the wavelet coefficients is normalized by the noise

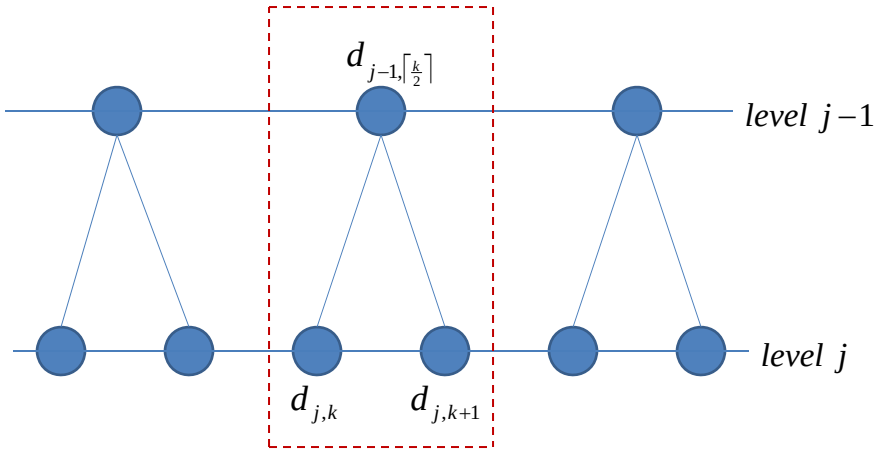

Figure 18: Λ-neighborhood of wavelet coefficients.

variance. We use the notion of "parent" in the same sense as it was used by Fryzlewicz (2007), where the parental coefficient $d_{j-1,\lceil k/2 \rceil}$ is located directly above $d_{j,k}$ and $d_{j,k+1}$ in the binary tree of wavelet coefficients (Figure 18). When the signal part is energetic, i.e., significant, the magnitudes of all three coefficients in (60) are expected to be connected because the coefficients are local in their representation of the signal. Speaking in the jargon of wavelet shrinkage, the whole Λ-neighborhood will fall into the energetic "cone of influence".

Let n be the size of signal f and J_0, $0 \leq J_0 < \log_2(n)$ be the coarsest level of detail in the wavelet decomposition of f. Given $d_{j,k}, j = J_0, \ldots, \log_2(n) - 1,\ k = 0, 1, \ldots, 2^j - 1$, our method forms the thresholding statistics $x_{j,l}, j = J_0, \ldots, \log_2(n) - 1,\ l = 0, 1, \ldots, 2^{j-1} - 1$. The number of $x_{j,l}$s is equal to half of the number of detail wavelet coefficients $d_{j,k}$, because we form nonoverlapping blocks of size 2 at each dyadic level j. With each of these blocks, we also consider the parental coefficient $d_{j-1,\lceil k/2 \rceil}$.

Because the wavelet coefficients $d_{j,k}$ are distributed as normal with mean $\theta_{j,k}$ and variance σ^2, $x_{j,l}$ will have noncentral χ^2 distribution with 3 degrees of freedom and noncentrality parameter $\lambda_{j,l} = \theta^2_{j,k} + \theta^2_{j,k+1} + \theta^2_{j-1,\lceil k/2 \rceil}$. Now, omitting the indices j, l in $x_{j,l}$ we propose the following Bayesian model on the thresholding statistic:

$$\begin{aligned} x|\lambda &\sim \chi^2_3(\lambda) \\ \lambda &\sim \epsilon_j \delta_0 + (1-\epsilon_j)\pi(\lambda), \end{aligned} \tag{61}$$

where $\chi^2_3(\lambda)$ denotes the noncentral chi-square distribution with 3 degrees of freedom, noncentrality parameter λ, and probability density function

$$f(x|\lambda) = \frac{1}{2} e^{-(x+\lambda)/2} \left(\frac{x}{\lambda}\right)^{1/4} I_{1/2}(\sqrt{\lambda x}).$$

In the above equation I_p denotes the modified Bessel function of the first kind (Abramowitz and Stegun, 1964; Gradshteyn and Ryzhik, 1980).

Note that the model in (61) is using the well established mixture prior with point mass at zero, which accounts for the sparsity of the wavelet coefficients, however, the prior is not on the coefficients but on their energies. In our prior formulation the large energies of the family of wavelet coefficients are captured by a spread distribution $\pi(\lambda)$, for which we propose the exponential distribution with mean $1/b$. For estimating the noncentrality parameter in a Bayesian fashion, Berger et al. (1998) proposed the noninformative prior distribution $\pi(\lambda) = \lambda^{-c}$ for the case of independent observations, and showed that choice of $c = 1/2$ has certain optimality properties. A closed-form Bayes rule for this prior was derived, but since both the prior and the associated marginal are improper, they can not be used for empirical Bayes (marginal maximum likelihood) hyperparameter estimation. Fourdrinier et al. (2000) consider $\pi(\lambda) = \lambda^{-c}$ and the related family of gamma priors $\pi(\lambda) \propto \exp\{-b\lambda\}\lambda^{-c}$ and prove that under Stein's loss functions $L_1(\lambda, x) = \frac{x}{\lambda} - \log\left(\frac{x}{\lambda}\right) - 1$ and $L_{-1}(\lambda, x) = \frac{\lambda}{x} - \log\left(\frac{\lambda}{x}\right) - 1$ the Bayes estimator corresponding to prior $\pi(\lambda) = \lambda^{-c}$ is admissible. Using gamma prior as a spread distribution in model (61) leads to computationally unstable marginals

and Bayes rules. In addition, the exponential distribution is maximizing the entropy among all distributions supported on $[0, \infty)$ with fixed mean. In that sense, the choice of exponential distribution is noninformative. Furthermore, in case $b \approx 0$ our prior approximates the noninformative prior $\pi(\lambda) = \lambda^{-c}$ with $c = 0$, which is optimal for large λ in case of a weighted loss (Berger et al., 1998).

For these reasons, we propose the exponential distribution as our spread prior. This formulation leads to a simple and computationally tractable Bayes rule. Therefore, the model in (61) becomes:

$$\begin{aligned} x|\lambda &\sim \chi^2_3(\lambda) \\ \lambda &\sim \epsilon_j \delta_0 + (1-\epsilon_j)\mathcal{E}(b). \end{aligned} \tag{62}$$

Note that in our model the weight of point mass ϵ_j is specified adaptively at each dyadic level j, but the scale parameter b for the spread distribution is specified globally. This serves the purpose of parsimony and ease of estimation. Specifying b_j levelwise did not improve the performance of the estimator. We also found that specifying a global $\mathcal{E}(b)$ will result in b small, to accommodate for Λ-neighborhoods with large energies. In such a case the prior will approximate a mixture prior with point mass and a noninformative spread distribution. At the same time, an increase of ϵ_j towards the finest scales will decrease the variance of the mixture prior in (62), accounting for the sparsity of wavelet decomposition.

For model (62) the marginal distribution becomes

$$\begin{aligned} m(x) &= \epsilon_j \sqrt{\frac{2}{\pi}} x^{1/2} e^{-x/2} + (1-\epsilon_j) \frac{b}{\sqrt{1+2b}} e^{-\frac{bx}{1+2b}} \mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] \\ &= \epsilon_j m_0(x) + (1-\epsilon_j) m_1(x), \end{aligned} \tag{63}$$

which is a mixture of a central chi-square distribution $m_0(x)$ with 3 degrees of freedom, and another distribution arising from $\mathcal{E}(b)$ part in the prior on λ in (62). The details are provided in Appendix.

Our goal for model in (62) is to estimate λ, the noncentrality parameter representing the mean energy of the Λ-neighborhood. An analytically tractable estimator is the posterior mean, which in this context becomes

$$\delta(x) = (1 - p_j) \frac{1 + 2b + x + \sqrt{\frac{2x(1+2b)}{\pi}} e^{-\frac{x}{2+4b}} \Big/ \text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}{(1+2b)^2}, \tag{64}$$

with

$$p_j = \epsilon_j \frac{m_0(x)}{m(x)}. \tag{65}$$

The derivation is deferred to Appendix. Other types of Bayes location estimators that are based on simple algorithmic forms, such as the posterior median, are considered later. The posterior mean $\delta(x)$ is fully specified by eliciting the hyperparameters ϵ_j and b. Figure 19 shows the shape of $\delta(x)$ for $b = 0.01$ and $\epsilon = 0.9$.

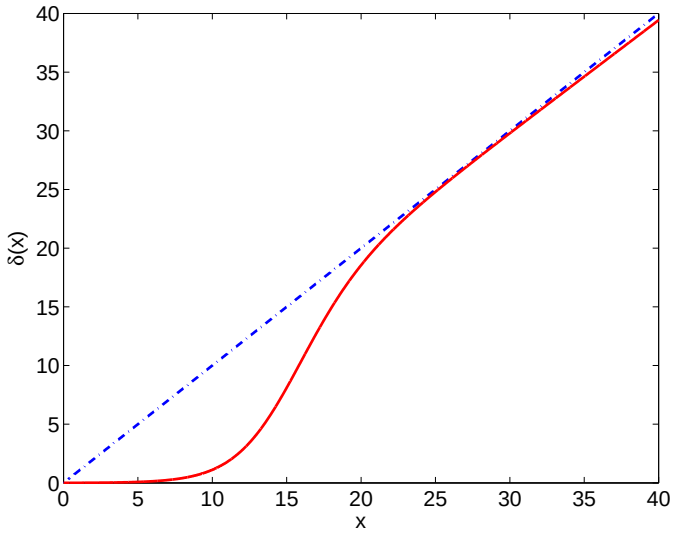

Figure 19: Bayes rule (64) for $b = 0.01$ and $\epsilon = 0.9$.

Note that the shape is desirable, as the rule heavily shrinks small and slightly shrinks large energy Λ-neighborhoods, which is a smooth approximation to hard thresholding.

It is interesting to examine the behavior of the Bayes rule (64) in terms of b when x gets large. For $b > 0$ the Bayes rule becomes non-robust for large values of x

and slightly levels off from the 45° line. This can be seen in Figure 19. We can analytically show this by examining the behavior of $\delta(x)$ when x gets larger. Because $\lim_{x\to\infty} p_j = 0$ (see Appendix) and Erf $\left[\sqrt{\frac{x}{2+4b}}\right]$ is bounded

$$\delta(x) - x \;\approx\; \frac{1}{1+2b} - x\left[1 - \frac{1}{(1+2b)^2}\right],$$

as x gets large. Therefore for any $b > 0$ the rule levels off from the 45° line in a linear fashion with slope $-(1 - 1/(1+2b)^2)$. As b gets larger, the rule levels off faster. Note, that this behavior of the Bayes rule is not unseen in the wavelet shrinkage literature. For instance, Angelini and Vidakovic (2004) consider a model, which incorporates prior belief on the boundedness of the energy of the signal, and the Bayes rule levels off from the 45° line for large wavelet coefficients.

In order to recover unknown signal f in (56) we need to estimate a pair of wavelet coefficients $\theta_{j,k}$ and $\theta_{j,k+1}$ from each Λ-neighborhood. Above, we derived an estimator $\delta(x)$ for the energy of the family of wavelet coefficients, which is

$$\hat{\lambda}_{j,l} = \hat{\theta}^2_{j,k} + \hat{\theta}^2_{j,k+1} + \hat{\theta}^2_{j-1,\lceil k/2\rceil} = \delta(x_{j,l}). \tag{66}$$

Shrinkage Rule. A natural way to estimate the wavelet coefficients $\theta_{j,k}$ and $\theta_{j,k+1}$ is to take

$$\begin{bmatrix} \hat{\theta}_{j,k} \\ \hat{\theta}_{j,k+1} \end{bmatrix} = \left\{\frac{\hat{\lambda}_{j,l}}{x_{j,l}}\right\}^{1/2} \begin{bmatrix} d_{j,k} \\ d_{j,k+1} \end{bmatrix}, \tag{67}$$

where $\hat{\lambda}_{j,l}$ is a Bayes estimator of $\lambda_{j,l}$, given for example by rule (64) as in (66). In practice this means that we shrink wavelet coefficients $d_{j,k}$ and $d_{j,k+1}$ in the block by the same factor, which is the square root of the ratio of the estimated and the observed energy of the block. The energy of the block contains the energy of the parental wavelet coefficient, which only contributes to the shrinkage procedure, but it is not shrunk at this stage. The shrinkage of $d_{j-1,\lceil k/2\rceil}$ is done when level $j-1$ is considered. Note that similar procedure for estimating the individual wavelet coefficients from the energy of the blocks was applied by Wang and Wood (2006).

Bayes rules under the squared error loss and regular models are never thresholding rules. A thresholding rule is preferable to smooth shrinkage rules in many applications, like model selection, data compression, dimension reduction, and related statistical tasks in which it is beneficial to replace by zero a majority of the processed wavelet coefficients. If thresholding rule is desirable, we can use the posterior median or the Bayes factor procedure to replace the posterior mean in (66) to get $\hat{\lambda}_{j,l}$ for shrinkage rule (67).

The posterior median of λ is

$$\hat{\lambda} = \delta_M(x) = u \ \mathbf{1}\left(p_j < \frac{1}{2}\right), \tag{68}$$

where u is the solution of the equation

$$1 - (1-\epsilon_j)\frac{1}{m(x)}\frac{b}{\sqrt{1+2b}}e^{-\frac{bx}{1+2b}}\left(\mathrm{Erf}\left[\frac{(1+2b)\sqrt{u}+\sqrt{x}}{\sqrt{2+4b}}\right] - \mathrm{Erf}\left[\frac{(1+2b)\sqrt{u}-\sqrt{x}}{\sqrt{2+4b}}\right]\right) = 0. \tag{69}$$

Equation (69) is transcendental, but can be efficiently solved with a built-in root finder algorithm available in standard computing packages. It can also be formulated and solved as an optimization problem. Derivation of the rule is deferred to Appendix.

The Bayes factor procedure to estimate λ is

$$\hat{\lambda} = \delta_{BF}(x) = x \ \mathbf{1}\left(p_j < \frac{1}{2}\right). \tag{70}$$

The derivation can be found in Appendix and for more details, see Vidakovic (1998a).

Figure 20 shows rules $\delta(x)$ (64), $\delta_{BF}(x)$ (70) and $\delta_M(x)$ (68) for $b = 0.01$ and $\epsilon = 0.9$.

4.3 Eliciting the hyperparameters

The described model and hence the Bayes estimators depend on hyperparameters that have to be specified. Purely subjective elicitation is only possible when considerable knowledge about the underlying signal is available. We followed the empirical

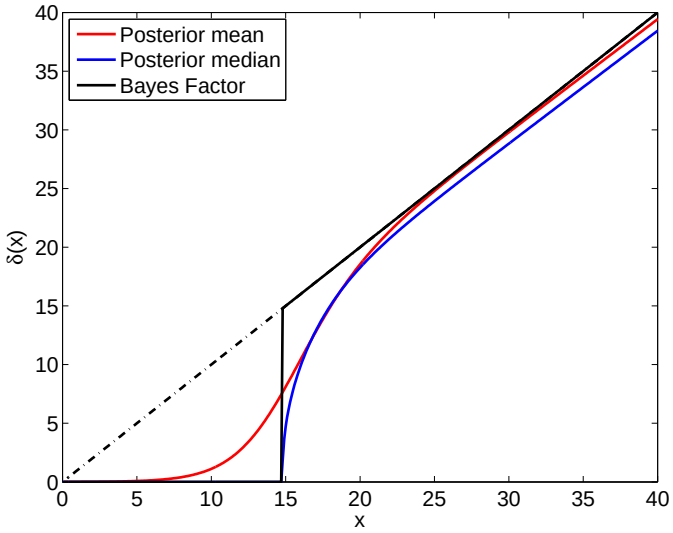

Figure 20: Comparison of Bayes rules (64), (70) and (68) for $b = 0.01$ and $\epsilon = 0.9$.

Bayes paradigm in this chapter: Johnstone and Silverman (2005b), Clyde and George (1999, 2000), and Abramovich et al. (2002) estimate the hyperparameters by marginal maximum likelihood (MLII) method in the wavelet denoising context. We also used this completely data-driven procedure to estimate the hyperparameters.

More specifically, we are interested in maximum likelihood estimates of hyperparameters ϵ_j and b. Parameter ϵ_j is specified at each resolution level j while parameter b is a global scale parameter in the exponential spread part of the prior (62). The marginal log-likelihood function ℓ is

$$\ell(\epsilon_j, b) = \sum_{j \geq J_0} \sum_{l=0}^{2^{j-1}-1} \log \left\{ \epsilon_j \sqrt{\frac{2}{\pi}} x_{j,l}^{1/2} e^{-x_{j,l}/2} + (1-\epsilon_j) \frac{b e^{-\frac{b x_{j,l}}{1+2b}} \mathrm{Erf}\left[\sqrt{\frac{x_{j,l}}{2+4b}}\right]}{\sqrt{1+2b}} \right\}. \tag{71}$$

Since a closed-form solution for the maximum of the log-likelihood is not available, we rely on numerical techniques to find estimates of ϵ_j and b. There are various approaches suggested in Clyde and George (2000), including direct maximization and the EM algorithm. We used the direct maximization approach for all the parameters,

which was also followed by Johnstone and Silverman (2005b) and Abramovich et al. (2002). We used MATLAB©'s built-in `fmincon` function with the "interior-point" option to minimize $-\ell$ with respect to ϵ_j and b. We carried out the optimization in one step for all levels j, since parameter b is global for all detail coefficients, which differs from the level-by-level optimization approach taken in the references above. The procedure is not sensitive to starting values of ϵ_j, but it is important to specify a proper starting value for b in order for the marginal log-likelihood ℓ to be finite. Starting values $\epsilon_j^0 = 0.5$ and $b^0 = 1/\bar{x}_{j,l}$ worked well.

Note, that in order to fully specify the thresholding statistics $x_{j,l}$ in (60), we also need to estimate parameter σ^2, which represents the variance of the random error ε. In the wavelet shrinkage literature σ^2 is frequently estimated by a robust estimator of the variance of wavelet coefficients at the finest level of detail (Donoho and Johnstone, 1994). We adopted this practice to estimate σ as $\hat{\sigma} = MAD/0.6745$. Here MAD stands for the median absolute deviation from the median of wavelet coefficients at the finest level of detail and constant 0.6745 calibrates the estimator to be comparable to sample standard deviation.

4.4 Simulations and Comparisons

In this section we discuss the performance of the proposed estimator (67) and compare it to established methods from the literature considering block-type wavelet denoising. In our simulations four standard test functions (`Blocks`, `Bumps`, `Doppler`, `Heavisine`) were considered (Donoho and Johnstone, 1994). The functions were rescaled such that the added noise produced preassigned signal-to-noise ratio (SNR), as standardly done. The test functions were simulated at $n = 256$, 512, and 1024 equally spaced points in the interval $[0, 1]$. Five commonly considered SNR's were selected, SNR $= 1, 3, 5, 7$ and 10. The standard wavelet bases were also used: Symmlet 8 for `Heavisine` and `Doppler`, Daubechies 6 for `Bumps` and Haar for `Blocks`. The coarsest decomposition

level was $J_0 = 3$ which matches $J_0 = \lfloor \log_2(\log(n)) + 1 \rfloor$ suggested by Antoniadis et al. (2001). Note here, that in practice we carried out the discrete wavelet decomposition to level $J_0 - 1$ so that the parental detail coefficients were available for level J_0.

Reconstruction of the theoretical signal was measured by the average mean squared error (AMSE), calculated as

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

where M is the number of simulation runs and $f(t_i),\ i = 1, \ldots, n$ are known values of the test functions considered. We denote by $\hat{f}_k(t_i),\ i = 1, \ldots, n$ the estimator from the k-th simulation run.

For comparison with other methods the posterior mean (64) is used to estimate $\lambda_{j,l}$, as in (66), because it gives better performance than the posterior median (68) and the Bayes factor procedure (70). It can be seen in Table 6 that the *LNWS* method based on the posterior mean gives better AMSE results in most test scenarios. However, for test function `Bumps` and for very high level of noise (SNR = 1) the posterior median results in better performance. Note that boldface numbers indicate the smallest AMSE result for each test scenario.

Table 6: AMSE comparison of the *LNWS* method based on posterior mean, median and Bayes factor.

Signal	N	Method	SNR=1	SNR=3	SNR=5	SNR=7	SNR=10	Signal	N	Method	SNR=1	SNR=3	SNR=5	SNR=7	SNR=10
Blocks	256	Mean	0.3186	0.3643	0.3240	0.3050	0.2962	Doppler	256	Mean	0.2378	**0.3367**	**0.3482**	**0.3604**	**0.3742**
		Med	**0.2888**	**0.3537**	**0.3060**	**0.2883**	**0.2864**			Med	**0.2162**	0.3687	0.3833	0.3954	0.4099
		BF	0.3531	0.3703	0.3121	0.2912	0.2881			BF	0.2519	0.3818	0.3906	0.4017	0.4135
	512	Mean	0.2006	**0.2261**	**0.2190**	**0.2072**	**0.1974**		512	Mean	0.1412	**0.1978**	**0.2239**	**0.2316**	**0.2400**
		Med	**0.1877**	0.2276	0.2324	0.2148	0.2004			Med	**0.1313**	0.1989	0.2449	0.2561	0.2634
		BF	0.2152	0.2343	0.2352	0.2162	0.2012			BF	0.1445	0.2041	0.2487	0.2581	0.2648
	1024	Mean	0.1228	**0.1509**	**0.1394**	**0.1314**	0.1245		1024	Mean	0.0738	**0.1080**	**0.1198**	**0.1251**	**0.1368**
		Med	**0.1174**	0.1627	0.1451	0.1341	**0.1242**			Med	**0.0688**	0.1139	0.1273	0.1288	0.1405
		BF	0.1270	0.1660	0.1464	0.1347	0.1244			BF	0.0742	0.1161	0.1287	0.1302	0.1419
Bumps	256	Mean	0.4264	**0.4896**	**0.5024**	**0.5052**	**0.5049**	Heavisine	256	Mean	0.0928	**0.1297**	**0.1930**	**0.2104**	**0.2190**
		Med	**0.3946**	0.4938	0.5221	0.5315	0.5119			Med	**0.0843**	0.1299	0.2039	0.2255	0.2359
		BF	0.5026	0.5196	0.5342	0.5381	0.5160			BF	0.0944	0.1318	0.2063	0.2272	0.2380
	512	Mean	0.3242	0.3542	0.3531	0.3562	0.3753		512	Mean	0.0426	**0.0808**	**0.1130**	**0.1355**	**0.1379**
		Med	**0.2939**	**0.3428**	**0.3456**	**0.3461**	**0.3702**			Med	**0.0410**	0.0835	0.1218	0.1444	0.1539
		BF	0.3538	0.3579	0.3530	0.3516	0.3756			BF	0.0427	0.0842	0.1225	0.1453	0.1547
	1024	Mean	0.1996	0.2204	0.2325	0.2465	0.2638		1024	Mean	0.0230	**0.0449**	**0.0577**	**0.0797**	**0.0953**
		Med	**0.1802**	**0.2049**	**0.2220**	**0.2397**	**0.2610**			Med	**0.0227**	0.0479	0.0592	0.0836	0.1065
		BF	0.2057	0.2126	0.2276	0.2440	0.2643			BF	0.0232	0.0483	0.0596	0.0843	0.1073

The mean-square performance of our method is compared to the *NCPmn-2* method

of Wang and Wood (2006), the *BITUP* method of Fryzlewicz (2007), the *Neigh-Coeff* method of Cai and Silverman (2001), the *BlockPostMean (PMN)* method of Abramovich et al. (2002), the *BBS* method of De Candiitis and Vidakovic (2004) and the *TSW* method of Autin et al. (2011). Results are summarized in Table 7, where boldface numbers indicate the best performance in each test scenario. From the results we can see that the proposed estimator is comparable to the established block shrinkage methods and superior for some combinations of signals, SNRs and sample sizes. As evident from Table 7, for test function `Bumps` the *LNWS* estimator has the lowest AMSE, except for the case SNR = 1, and this is also true in many cases of the `Heavisine` signal. For test signals `Blocks` and `Doppler` our method performs comparable to the existing methods. Note that *TSW* is the only other method which considers both neighboring and parental relations. It is apparent that the *LNWS* estimator performs better in most test scenarios than *TSW*, because it is based on a more sophisticated shrinkage rule instead of simple thresholding. Graphical summary of the results is presented in Figure 21 where the boxplots of the MSE are shown for $n = 1024$ and SNR = 3. The number of simulations performed was $M = 1000$.

4.5 Application to Inductance Plethysmography Data

In this section we apply the proposed *LNWS* method to a real-world data set from anaesthesiology generated 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 measurements are the output voltage of the inductance plethysmograph over time. The data set is popular in the wavelet denoising literature and was used as an example by several authors, for example Nason (1996), Abramovich et al. (1998, 2002) and Johnstone and Silverman (2005b). For more information about the data set, please refer to Nason (1996).

Figure 22 shows a section of plethysmograph recording lasting approximately 80

Table 7: AMSE of the proposed *LNWS* estimator compared to other methods.

Signal	N	Method	SNR=1	SNR=3	SNR=5	SNR=7	SNR=10	Signal	N	Method	SNR=1	SNR=3	SNR=5	SNR=7	SNR=10
Blocks	256	LNWS	0.3136	0.3643	0.3240	0.3050	0.2962	Doppler	256	LNWS	0.2378	0.3367	0.3482	0.3604	0.3742
		NCP	0.2798	0.3709	0.3407	0.3226	0.3146			NCP	0.2160	0.3218	**0.3090**	**0.3245**	**0.3433**
		BITUP	0.3768	0.4100	0.3287	0.3040	0.3036			BITUP	0.2724	0.4608	0.4773	0.4598	0.4435
		NC	0.3503	0.6227	0.5716	0.4895	0.4122			NC	0.2121	0.3168	0.3332	0.3529	0.3740
		PMN	0.3203	0.5197	0.5056	0.5021	0.4898			PMN	0.2685	0.3414	0.3538	0.4049	0.4618
		BBS	**0.2510**	0.3258	0.2792	0.2310	0.1956			BBS	**0.1951**	**0.3018**	0.3256	0.3631	0.4334
		TSW	0.3433	**0.2784**	**0.2140**	**0.1791**	**0.1734**			TSW	0.2689	0.4583	0.4862	0.4681	0.5065
	512	LNWS	0.2006	0.2261	0.2190	0.2072	0.1974		512	LNWS	0.1412	0.1978	0.2239	0.2316	0.2400
		NCP	0.1975	0.2573	0.2468	0.2328	0.2233			NCP	0.1254	0.1798	0.1909	0.1928	0.2103
		BITUP	0.2390	0.2338	0.2167	0.2055	0.2011			BITUP	0.1660	0.2374	0.2700	0.2860	0.2889
		NC	0.2629	0.4104	0.4032	0.3679	0.3199			NC	**0.1158**	0.1696	0.1803	0.1867	**0.2069**
		PMN	0.2303	0.3846	0.3567	0.3531	0.3389			PMN	0.1635	0.1983	0.1748	0.2060	0.2704
		BBS	**0.1745**	0.1992	0.1907	0.1756	0.1591			BBS	0.1173	**0.1576**	**0.1632**	**0.1822**	0.2318
		TSW	0.2250	**0.1898**	**0.1583**	**0.1176**	**0.1111**			TSW	0.1681	0.2450	0.2765	0.2700	0.2625
	1024	LNWS	0.1228	0.1509	0.1394	0.1314	0.1245		1024	LNWS	0.0738	0.1080	0.1198	0.1251	0.1368
		NCP	0.1287	0.1904	0.1714	0.1610	0.1537			NCP	0.0614	0.1032	0.1205	0.1258	0.1312
		BITUP	0.1496	0.1502	0.1375	0.1319	0.1265			BITUP	0.0898	0.1368	0.1447	0.1559	0.1728
		NC	0.1867	0.3253	0.3088	0.2680	0.2250			NC	0.0608	**0.0954**	**0.1162**	**0.1239**	**0.1301**
		PMN	0.1746	0.3576	0.2915	0.2800	0.2743			PMN	0.1094	0.1466	0.1379	0.1444	0.1838
		BBS	**0.1222**	0.1489	0.1240	0.1084	0.0939			BBS	**0.0605**	0.0980	0.1181	0.1314	0.1352
		TSW	0.1527	**0.1378**	**0.0810**	**0.0706**	**0.0703**			TSW	0.0943	0.1483	0.1735	0.1961	0.1938
Bumps	256	LNWS	0.4264	**0.4896**	**0.5024**	**0.5052**	**0.5049**	Heavisine	256	LNWS	0.0928	0.1297	0.1930	**0.2104**	**0.2190**
		NCP	0.4383	0.5036	0.5393	0.5627	0.5784			NCP	**0.0555**	0.1289	0.2113	0.2467	0.2832
		BITUP	0.5356	0.6875	0.6075	0.5883	0.5645			BITUP	0.1207	0.1651	0.2070	0.2127	0.2194
		NC	0.5667	0.8026	0.7694	0.7771	0.7669			NC	0.0627	**0.1217**	0.2139	0.2682	0.3069
		PMN	0.4706	0.5853	0.7215	0.8302	0.9178			PMN	0.1307	0.1844	0.2425	0.2917	0.3664
		BBS	**0.3927**	0.7232	0.8462	0.9516	1.1201			BBS	0.0932	0.1332	**0.1810**	0.2174	0.2479
		TSW	0.5685	0.8765	0.8910	0.9985	1.1621			TSW	0.0659	0.1447	0.2354	0.2724	0.3008
	512	LNWS	0.3242	**0.3542**	**0.3531**	**0.3562**	**0.3753**		512	LNWS	0.0426	**0.0808**	0.1130	**0.1355**	**0.1379**
		NCP	**0.3084**	0.3986	0.4005	0.4041	0.4105			NCP	**0.0303**	0.0836	0.1195	0.1596	0.1800
		BITUP	0.4100	0.3997	0.3863	0.3860	0.4006			BITUP	0.0572	0.1014	0.1395	0.1432	0.1402
		NC	0.4542	0.5851	0.5300	0.4803	0.4384			NC	0.0372	0.0903	0.1422	0.1719	0.1999
		PMN	0.3522	0.4750	0.5224	0.5399	0.5697			PMN	0.0825	0.1243	0.1585	0.1956	0.2435
		BBS	0.3127	0.5124	0.5847	0.6593	0.7429			BBS	0.0499	0.0820	**0.1114**	0.1369	0.1601
		TSW	0.4468	0.6001	0.6499	0.7260	0.7851			TSW	0.0359	0.0980	0.1501	0.1855	0.2240
	1024	LNWS	0.1996	**0.2204**	**0.2325**	**0.2465**	**0.2638**		1024	LNWS	0.0230	**0.0449**	**0.0577**	0.0797	0.0953
		NCP	**0.1886**	0.2480	0.2666	0.2741	0.2849			NCP	**0.0199**	0.0507	0.0636	0.0847	0.1033
		BITUP	0.2451	0.2291	0.2401	0.2519	0.2700			BITUP	0.0322	0.0543	0.0680	0.0866	0.0999
		NC	0.2759	0.3228	0.3022	0.2881	0.2892			NC	0.0203	0.0667	0.0878	0.0951	0.1070
		PMN	0.2404	0.3430	0.3521	0.3851	0.4208			PMN	0.0649	0.0915	0.1022	0.1189	0.1517
		BBS	0.2058	0.3203	0.3691	0.3941	0.4015			BBS	0.0288	0.0492	0.0629	**0.0768**	**0.0940**
		TSW	0.2845	0.3862	0.4464	0.4582	0.4351			TSW	0.0236	0.0632	0.0824	0.1174	0.1515

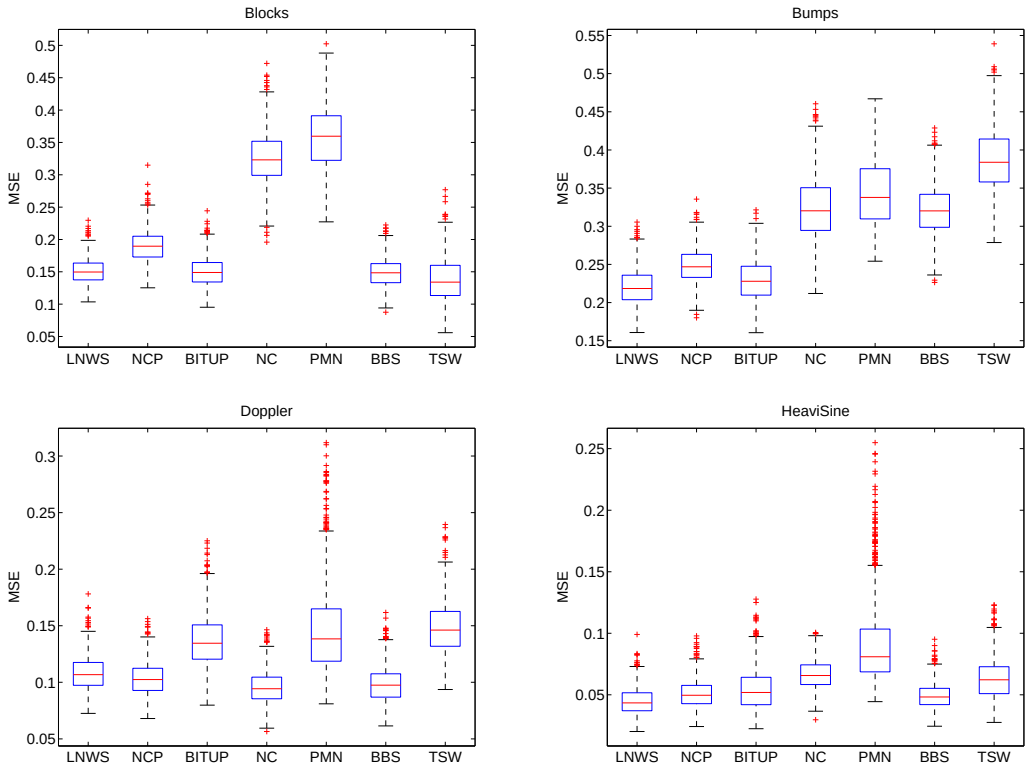

Figure 21: Boxplots of MSE for various block-shrinkage procedures based on $n = 1024$ points and SNR $= 3$.

s ($n = 4096$ observations). Figure 23 shows the reconstruction of the signal with the *LNWS* method using posterior mean. Using posterior median or Bayes factor as an estimator lead to essentially identical results. It is apparent that the proposed method removes the noise well while preserving the important features of the signal. Abramovich et al. (2002) report the heights of the first peak while analysing this data set. For the *LNWS* method the heights are 0.8433, 0.8431 and 0.8433 using the posterior mean, median and Bayes factor, respectively. These numbers are the same as the result obtained by Abramovich et al. (2002), and better compared to some established methods reported in their paper. The empirical Bayes method by Johnstone and Silverman (2005b) reports 0.842 as a result.

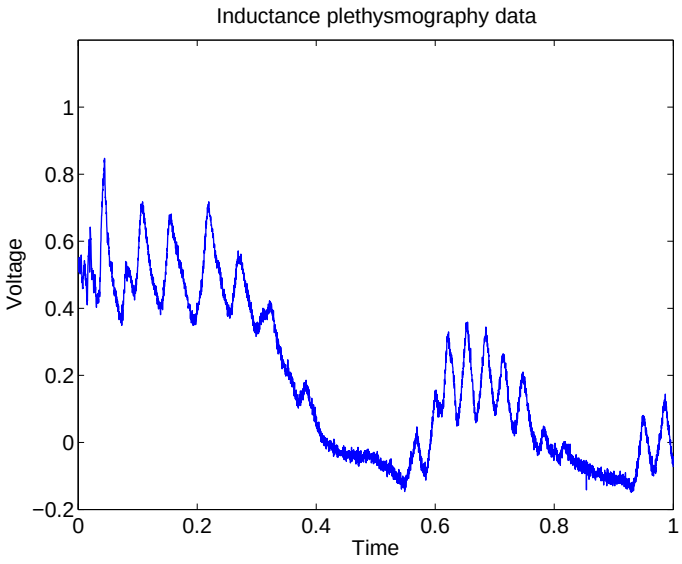

Figure 22: A section of inductance plethysmography data with $n = 4096$.

4.6 Extension to Image Denoising

In this section we show briefly how the proposed methodology can easily be extended to two-dimensional signals/images. For images, the neighboring structures comprise 1 parent and 4 children coefficients, as it was used by Crouse et al. (1998) and Romberg

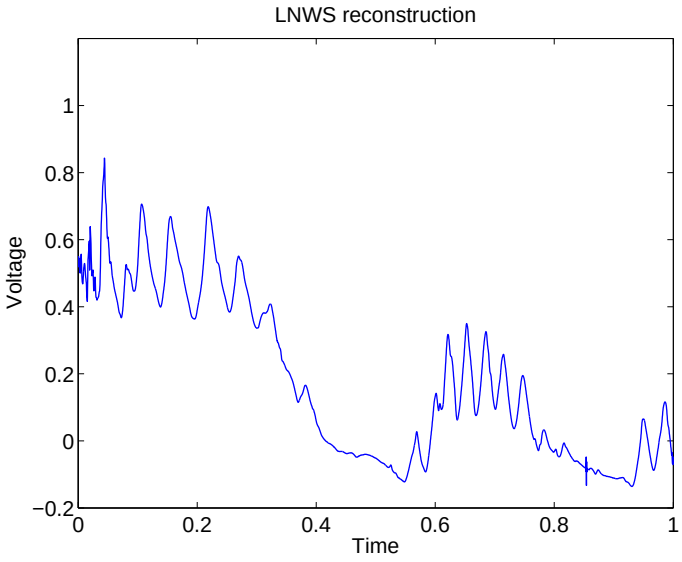

Figure 23: Reconstruction of inductance plethysmography data obtained by the *LNWS* method.

et al. (2001), among others. Therefore, the "thresholding statistics" (60) becomes

$$x_{j,l} = \left(d^2_{j,\{k_1,k_2\}} + d^2_{j,\{k_1+1,k_2\}} + d^2_{j,\{k_1,k_2+1\}} + d^2_{j,\{k_1+1,k_2+1\}} + d^2_{j-1,\{\lceil k_1/2\rceil,\lceil k_2/2\rceil\}}\right)/\sigma^2.$$

In the above $j = J_0, \ldots, \log_2(n)$ denotes the scale or subband in the 2D wavelet decomposition, and $\{k_1, k_2\} \in \{0, \ldots, 2^j - 1\} \times \{0, \ldots, 2^j - 1\}$ denotes the location of wavelet coefficients. The Bayesian model on the "thresholding statistics" becomes

$$\begin{aligned} x|\lambda &\sim \chi^2_5(\lambda) \\ \lambda &\sim \epsilon_j\delta_0 + (1-\epsilon_j)\pi(\lambda), \end{aligned} \tag{72}$$

where $\chi^2_5(\lambda)$ denotes the noncentral chi-square distribution with 5 degrees of freedom. For model (72) the marginal distribution becomes

$$\begin{aligned} m(x) &= \epsilon_j \frac{1}{\Gamma(5/2)2^{5/2}} x^{3/2}e^{-x/2} + \\ &\quad (1-\epsilon_j)\left\{b\sqrt{1+2b}e^{-\frac{bx}{1+2b}}\,\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] - b\sqrt{2x/\pi}e^{-x/2}\right\} \\ &= \epsilon_j m_0(x) + (1-\epsilon_j)m_1(x), \end{aligned} \tag{73}$$

while the posterior mean becomes

$$\delta(x) = (1-p_j)\frac{\sqrt{\frac{2x(1+2b)}{\pi}}e^{-\frac{x}{2+4b}} + (x-1-2b)\text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}{-(1+2b)^{3/2}\sqrt{2x/\pi}e^{-\frac{x}{2+4b}} + (1+2b)^2\text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}, \tag{74}$$

with

$$p_j = \epsilon_j \frac{m_0(x)}{m(x)}. \tag{75}$$

The derivations are analogous to the one-dimensional case and explained in the Appendix. Equations for the posterior median and the Bayes factor procedure are omitted here, but can be derived similarly, as before. Naturally, the shrinkage rule to estimate $\theta_{j,\{k_1,k_2\}}$, $\theta_{j,\{k_1+1,k_2\}}$, $\theta_{j,\{k_1,k_2+1\}}$ and $\theta_{j,\{k_1+1,k_2+1\}}$ modifies to

$$\begin{bmatrix} \hat{\theta}_{j,\{k_1,k_2\}} \\ \hat{\theta}_{j,\{k_1+1,k_2\}} \\ \hat{\theta}_{j,\{k_1,k_2+1\}} \\ \hat{\theta}_{j,\{k_1+1,k_2+1\}} \end{bmatrix} = \left\{\frac{\hat{\lambda}_{j,l}}{x_{j,l}}\right\}^{1/2} \begin{bmatrix} d_{j,\{k_1,k_2\}} \\ d_{j,\{k_1+1,k_2\}} \\ d_{j,\{k_1,k_2+1\}} \\ d_{j,\{k_1+1,k_2+1\}} \end{bmatrix}, \tag{76}$$

where $\hat{\lambda}_{j,l}$ is a Bayes estimator of $\lambda_{j,l}$; in this case the posterior mean given by rule (74). Again, we shrink wavelet coefficients $d_{j,\{k_1,k_2\}}$, $d_{j,\{k_1+1,k_2\}}$, $d_{j,\{k_1,k_2+1\}}$ and $d_{j,\{k_1+1,k_2+1\}}$ in the block by the same factor, which is the square root of the ratio of the estimated and the observed energy of the block. The energy of the block contains the energy of the parental wavelet coefficient $d_{j-1,\{\lceil k_1/2\rceil,\lceil k_2/2\rceil\}}$, which contributes to the shrinkage procedure. We refer to the procedure as *LNWS-2D* in the future.

As before, estimation of the hyperparameters ϵ_j and b are done by numerically maximizing the marginal log-likelihood function, and we estimate parameter σ^2 by a robust estimator of the variance of wavelet coefficients at the finest level of detail.

To demonstrate the method, we compared the AMSE performance of *LNWS-2D* to the Hidden Markov Tree (*HMT*) model of Romberg et al. (2001). In the simulations three standard test images (`Lena`, `Peppers`, `Barbara`) of size 512×512

were considered. We added i.i.d. normal noise to the test images. Three different noise levels were considered, $\sigma = 10$, $\sigma = 25$ and $\sigma = 50$. The coarsest decomposition level was set $J_0 = 3$. Reconstruction of the image was evaluated by the average mean squared error (AMSE), calculated as

$$\frac{1}{M}\sum_{k=1}^{M}\left(\frac{\|\hat{\boldsymbol{Y}}_k - \boldsymbol{Y}\|_2^2}{n^2}\right),$$

where M is the number of simulation runs and $\boldsymbol{Y}$ is the $n \times n$ known values of the test image considered. We denote by $\hat{\boldsymbol{Y}}_k$ the estimator from the kth simulation run. Note, that for method *HMT* the pixel values were normalized to [0,1] as suggested by the authors. We used the Daubechies 4 wavelet filter and the number of simulation runs was $M = 50$. Results are summarized in Table 8.

Table 8: AMSE of the *LNWS-2D* method compared to *HMT*.

Picture	Method	$\sigma = 10$	$\sigma = 25$	$\sigma = 50$
Lena	LNWS-2D	28.95	78.41	156.11
	HMT	27.25	76.66	157.09
Peppers	LNWS-2D	30.69	78.23	165.25
	HMT	30.37	86.88	174.69
Barbara	LNWS-2D	46.79	160.23	324.44
	HMT	47.37	148.64	321.12

It is evident that the proposed method *LNWS-2D* performs very similar compared to the established *HMT* procedure. Note, that the computational requirements of the two procedures are virtually the same. To visually illustrate the results, Figure 24 is provided. It shows a `Peppers` image with $\sigma = 25$ denoised by the two procedures.

4.7 Conclusions

In this chapter we proposed a wavelet shrinkage method based on a neighborhood of wavelet coefficients, which includes two neighboring and a parental coefficient. We called the methodology Λ-neighborhood wavelet shrinkage, motivated by the shape of the considered neighborhood. A Bayesian model was formulated on the total energy

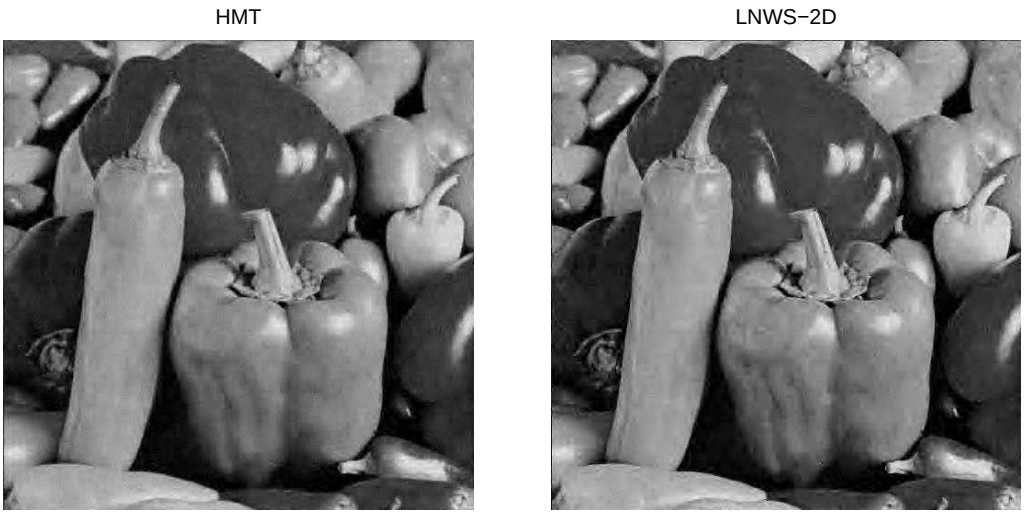

Figure 24: Denoised `Peppers` with $\sigma = 25$.

of the coefficients in the neighborhood, and different Bayes estimators of the mean energy were derived and explored. Shrinkage of the neighboring wavelet coefficients were based on the ratio of the estimated and observed energy. Extensive simulations on standard test functions showed that the method performs comparable and often superior to several existing block-shrinkage methods. Possible explanation for the noted improvement is that the Λ-neighborhood mimics a "cone of influence" in which the local energy spreads from a parent to children coefficients. An application to inductance plethysmography data set was also considered. The proposed method performed well in both denoising and preserving the important features of the real data. Finally, we showed how the method can be extended to image denoising.

In the model we used a global scale parameter b for the spread distribution. Possible future work may include models with scale parameter b_j set levelwise as well as the elicitation of these parameters. Other future improvement can be to explore the possibility and performance of using neighborhoods of different size, for example 4 neighboring and 2 parental wavelet coefficients. Size of the neighborhoods could be specified depending on the dyadic level and nature of the signal.

In the spirit of reproducible research we made MATLAB scripts used in simulation

for *LNWS* available at http://gtwavelet.bme.gatech.edu/.

CHAPTER V

FULLY BAYESIAN ESTIMATION AND VARIABLE SELECTION IN PARTIALLY LINEAR WAVELET MODELS

In this chapter we propose a wavelet-based methodology for estimation and variable selection in partially linear models. The inference is conducted in the wavelet domain, which provides a sparse and localized decomposition appropriate for nonparametric components with various degrees of smoothness. A hierarchical Bayes model is formulated on the parameters of this representation, where the estimation and variable selection is performed by a Gibbs sampling procedure. For both the parametric and nonparametric part of the model we are using point-mass-at-zero contamination priors with a double exponential spread distribution. In this sense we extend the model of Chapter 2 to partially linear models. Only a few papers in the area of partially linear wavelet models exist, and we show that the proposed methodology is often superior to the existing methods with respect to the task of estimating model parameters. Moreover, the method is able to perform Bayesian variable selection by a stochastic search for the parametric part of the model.

5.1 Introduction

In the present chapter we consider a novel Bayesian approach for solving the following regression problem

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{f}(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{77}$$

where t_i, $i = 1, \ldots, n$, are equispaced sampling points, $\boldsymbol{x}_i$, $i = 1, \ldots, n$, are known p-dimensional design points, $\boldsymbol{\beta}$ is an unknown p-dimensional parameter vector, $\boldsymbol{f}$ is

an unknown and potentially non-smooth function, and the random errors ε_i are i.i.d. normal, with zero mean and variance σ^2. The model can be written in matrix-vector form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\varepsilon}. \tag{78}$$

Our interest is to simultaneously estimate the unknown parameter vector $\boldsymbol{\beta}$ and nonparametric function $\boldsymbol{f}$ using the observations $\boldsymbol{Y}$. Another task is to identify important (non-zero) components of $\boldsymbol{\beta}$, that is to perform dimension reduction via variable selection on $\boldsymbol{\beta}$.

The model in (77) is called a partially linear model (PLM) in the literature. Engle et al. (1986) were among the first to use PLM to analyze electricity sales data. The model is semiparametric in nature because it combines parametric (linear) and nonparametric parts. In this chapter we consider a model with one nonparametric part in it. The monograph by Härdle et al. (2000) discusses the general PLM model extensively.

Several approaches are proposed in the literature to represent the nonparametric component $\boldsymbol{f}$ of the model in (78). These all build on existing nonparametric regression techniques, such as the kernel method, the local linear method (local polynomial or trigonometric polynomial techniques), or splines. In the most recent papers, wavelets are used (Chang and Qu, 2004; Fadili and Bullmore, 2005; Qu, 2006; Gannaz, 2007; Ding et al., 2011), which allows the nonparametric component to be parsimoniously represented by a limited number of coefficients. The wavelet representation can include a wide variety of nonparametric parts, including non-smooth signals, and reduces the bias in estimating the parametric component.

In this chapter we consider the latter approach and use the wavelet decomposition to represent $\boldsymbol{f}$. We use the Bayesian approach to formulate a hierarchical model in the wavelet domain and estimate its parameters. Only a few papers used wavelets in the partially linear model context, and besides Qu (2006), all used a penalized least

squares estimation procedure. Therefore, using a fully Bayesian approach can be of interest.

After applying a linear and orthogonal wavelet transform, the model in (77) becomes

$$d_{jk} = \boldsymbol{u}_{jk}^T \boldsymbol{\beta} + \theta_{jk} + \tilde{\varepsilon}_{jk}, \tag{79}$$

where d_{jk}, θ_{jk} and $\tilde{\varepsilon}_{jk}$ are the wavelet coefficients (at resolution j and location k) corresponding to $\boldsymbol{Y}$, $\boldsymbol{f}$ and $\boldsymbol{\varepsilon}$, and $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{X}$, where $\boldsymbol{W}$ is an orthogonal matrix implementing the wavelet transform. In a matrix-vector form,

$$\boldsymbol{W}\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{f} + \boldsymbol{W}\boldsymbol{\varepsilon},$$

which becomes

$$\boldsymbol{d} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\theta} + \tilde{\boldsymbol{\varepsilon}}. \tag{80}$$

Note that because of the orthogonality of $\boldsymbol{W}$, $\tilde{\boldsymbol{\varepsilon}} \sim \mathcal{N}(0, \sigma^2 I)$. Due to the whitening property of wavelet transforms (Flandrin, 1992), we can assume independence of the coefficients d_{jk}. To estimate β_i and θ_{jk} in model (79) in a Bayesian fashion, we build on results from the Bayesian linear models and wavelet regression literature.

To estimate θ_{jk} in a simple nonparametric regression model $Y_i = \boldsymbol{f}(t_i) + \varepsilon_i$, Bayesian shrinkage rules have been proposed in the literature by many authors. By a shrinkage rule, we mean that the observed wavelet coefficients d are replaced with their shrunken version $\hat{\theta} = \delta(d)$. The traditional Bayesian models consider a prior distribution on a generic wavelet coefficient θ as

$$\pi(\theta) = \epsilon\delta_0 + (1-\epsilon)\xi(\theta), \tag{81}$$

where δ_0 is a point mass at zero, ξ is a symmetric about 0 and unimodal distribution, and ϵ is a fixed parameter in [0,1], usually level dependent, that controls the amount of shrinkage for values of d close to 0. This type of model was considered by Abramovich

et al. (1998), Vidakovic (1998a), Vidakovic and Ruggeri (2001), and Johnstone and Silverman (2005b), among others. We also considered this type of model as a part of a fully Bayesian approach in Chapter 2.

The mixture prior approach was also utilized in Bayesian estimation and variable selection of linear models, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where a mixture prior is specified on parameters β_i. This type of model was considered for example by George and McCulloch (1993, 1997), and Yuan and Lin (2004, 2005). It is natural to combine these approaches; therefore, we build on these modeling ideas to formulate a fully Bayesian model in the partially linear model context.

The chapter is organized as follows. Section 5.2 formalizes the Bayesian model and presents some results related to it. In Section 5.3 we explain the estimation through a Gibbs sampling procedure developed for the hierarchical model. Section 5.4 discusses the selection of hyperparameters, contains simulations and comparisons to existing methods, and discusses how variable selection can be performed. Conclusions and discussion are provided in Section 5.5.

5.2 Hierarchical Model

In this section we propose a hierarchical model in which we use a mixture prior approach for both the parametric and the nonparametric components of the partially linear model. The model on the nonparametric part is the same as the model introduced in Chapter 2; therefore, the proposed model is an extension of that. As a consequence, a number of details and results related to the following model are the same, but for completeness, we present all the details here.

Let us consider the following hierarchical Bayesian model for a partially linear

model in the wavelet domain (80):

$$\begin{aligned}
\boldsymbol{d}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta},\sigma^2 &\sim \mathcal{N}(\boldsymbol{U}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}+\boldsymbol{\theta},\sigma^2 I)\\
\sigma^2 &\sim \mathcal{IG}(a_1,b_1)\\
\beta_i|\gamma_i,\tau_\beta &\sim (1-\gamma_i)\delta_0+\gamma_i\mathcal{DE}(\tau_\beta),\quad i=1,\ldots,p\\
\theta_{jk}|z_{jk},\tau_\theta &\sim (1-z_{jk})\delta_0+z_{jk}\mathcal{DE}(\tau_\theta)\\
\gamma_i|q &\sim \mathcal{B}er(q),\quad i=1,\ldots,p\\
z_{jk}|\epsilon_j &\sim \mathcal{B}er(\epsilon_j)\\
q &\sim \mathcal{U}(0,1)\\
\epsilon_j &\sim \mathcal{U}(0,1),
\end{aligned} \tag{82}$$

where j pertains to the resolution level of d_{jk} and $\mathcal{N}$, $\mathcal{IG}$, $\mathcal{DE}$, $\mathcal{B}er$, and $\mathcal{U}$ stand for the normal, inverse gamma, double exponential, Bernoulli, and uniform distributions, respectively. Index i refers to the regression coefficients in $\boldsymbol{\beta}$. Note that $\boldsymbol{\gamma}$ is an indicator vector of binary elements; therefore, subscript $\boldsymbol{\gamma}$ indicates that only those columns or elements of $\boldsymbol{U}$ and $\boldsymbol{\beta}$ with the corresponding $\boldsymbol{\gamma}$ element of 1 are included.

Note that the model in (82) uses the well-established mixture prior on θ_{jk} with a point mass at zero, which accounts for the sparsity of the nonparametric part in the wavelet domain. Wavelet coefficients with large magnitudes are captured by the spread part of the mixture prior, for which we propose the double exponential or Laplace distribution with variance $2/\tau_\theta^2$. The double exponential distribution is a popular choice for the spread part. It models wavelet coefficients with large energies and was used by several authors, for example Vidakovic and Ruggeri (2001), and Johnstone and Silverman (2005b). The mixture prior on θ_{jk} is specified levelwise, for each dyadic level j; however, the scale parameter τ_θ is global. This serves the purpose of parsimony and contributes to the ease of estimation. Here z_{jk} is a latent variable indicating whether our parameter θ_{jk} is coming from a point mass at zero ($z_{jk}=0$) or from a double exponential part ($z_{jk}=1$), with prior probability of $1-\epsilon_j$

or ϵ_j, respectively. For the prior probability ϵ_j we assume a "noninformative" uniform prior. The uniform $\mathcal{U}(0,1)$ prior is equivalent to a beta $\mathcal{B}e(1,1)$ distribution, which is a conjugate prior for the Bernoulli distribution. Note that this specification of the nonparametric part of the model is the same as in Chapter 2.

In our model we naturally propose the same mixture prior to model the regression parameters $\beta_i,\ i = 1, \ldots, p$. Yuan and Lin (2004, 2005) used this prior in the Bayesian variable selection context for linear models. In case $\gamma_i = 0$ the model forces $\beta_i = 0$ and if $\gamma_i = 1$ then β_i is modeled with a double exponential prior accommodating large regression coefficients. For the elements of binary vector $\boldsymbol{\gamma}$ we use the Bernoulli prior with common parameter q. This prior assumes that each predictor enters the model independently with prior probability q. Although it does not take into account the possible correlation between the predictors, this type of prior works well in practice, and it was used by George and McCulloch (1993) and George and Foster (2000), to name a few. Unlike George and McCulloch (1993), who prespecified q, we introduce another level of hierarchy by assuming a uniform "noninformative" prior on q. Since it is not clear, in general, how to specify q, it makes sense to put a prior distribution on the parameter, instead of using $q = 1/2$, which is a common suggestion in practice. As opposed to the fully Bayesian approach, George and Foster (2000) used the empirical Bayes approach to estimate q.

Parameter σ^2 represents the common noise variance for each resolution level on which we specified a conjugate inverse gamma prior. Spread parameters τ_θ and τ_β will be given priors after a reformulated version of the model (82) is discussed.

The hierarchical model in (82) is not conjugate; however, with additional transformations, derivations and computational techniques, it is possible to develop a fast Gibbs sampling algorithm for updating of its parameters. Note that a standard approach for handling the double exponential prior in Markov chain Monte Carlo

(MCMC) computations of hierarchical models is to represent the double exponential distribution as a scale mixture of normal distributions (Andrews and Mallows, 1974). This approach is used for example in Bayesian LASSO variable selection, where the double exponential prior (without point mass) is used on the regression parameters (Park and Casella, 2008). Here we will only use the scale mixture approach for the double exponential prior on β_i. This introduces an additional parameter v_i corresponding to each β_i, which needs to be updated. Using the scale mixture representation, the model in (82) becomes

$$\begin{aligned}
\boldsymbol{d}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta},\sigma^2 &\sim \mathcal{N}(\boldsymbol{U}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}+\boldsymbol{\theta},\sigma^2 I)\\
\sigma^2 &\sim \mathcal{IG}(a_1,b_1)\\
\beta_i|\gamma_i,v_i,\eta^2 &\sim (1-\gamma_i)\delta_0+\gamma_i\mathcal{N}(0,v_i\eta^2),\quad i=1,\ldots,p\\
v_i &\sim \mathcal{E}xp(1),\quad i=1,\ldots,p\\
\theta_{jk}|z_{jk},\tau_\theta &\sim (1-z_{jk})\delta_0+z_{jk}\mathcal{DE}(\tau_\theta)\\
\gamma_i|q &\sim \mathcal{B}er(q),\quad i=1,\ldots,p\\
z_{jk}|\epsilon_j &\sim \mathcal{B}er(\epsilon_j)\\
q &\sim \mathcal{U}(0,1)\\
\epsilon_j &\sim \mathcal{U}(0,1)\\
\eta^2 &\sim \mathcal{IG}(a_2,b_2)\\
\tau_\theta &\sim \mathcal{G}a(a_3,b_3)
\end{aligned} \tag{83}$$

In the model above $\eta=\sqrt{2}/\tau_\beta$. If we integrate out v_is from (83), we get back the model in (82), which follows from the scale mixture representation of the double exponential distribution. For the spread parameters η^2 and τ_θ, inverse gamma and gamma priors are specified in the model, which turn out to be conjugate.

For parameters θ_{jk} it is possible to derive the full conditional distributions without resorting to the scale mixture representation. This improves the speed of the Gibbs

sampling algorithm. In order to do this, we first discuss some results related to model (83), which are instrumental in developing the Gibbs sampler.

First let $d^\star_{jk} = d_{jk} - (\boldsymbol{U}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})_{jk}$ from which it follows that $d^\star_{jk} \sim \mathcal{N}(\theta_{jk}, \sigma^2)$. In the following notation $d^\star$ refers to an arbitrary $d^\star_{jk}$ and the mean θ stands for the corresponding θ_{jk}. If we consider a $\mathcal{N}(\theta, \sigma^2)$ likelihood $f(d^\star|\theta, \sigma^2)$ and elicit a double exponential $\mathcal{DE}(\tau)$ prior $p_1(\theta|\tau)$ on the θ, the marginal distribution becomes

$$m(d^\star|\sigma^2, \tau) = \frac{\tau}{2} e^{\frac{\sigma^2\tau^2}{2}} \left\{ e^{-d^\star\tau} \Phi\left(\frac{d^\star}{\sigma} - \tau\sigma\right) + e^{d^\star\tau} \Phi\left(-\frac{d^\star}{\sigma} - \tau\sigma\right) \right\}, \quad (84)$$

and the posterior distribution of θ becomes

$$h(\theta|d^\star, \sigma^2, \tau) = \\ = \begin{cases} \dfrac{e^{-d^\star\tau}}{e^{-d^\star\tau}\Phi\left(\frac{d^\star}{\sigma} - \tau\sigma\right) + e^{d^\star\tau}\Phi(-d^\star/\sigma - \tau\sigma)} \dfrac{1}{\sigma}\phi\left(\dfrac{\theta - (d^\star - \sigma^2\tau)}{\sigma}\right), & \theta \geq 0 \\ \dfrac{e^{d^\star\tau}}{e^{-d^\star\tau}\Phi\left(\frac{d^\star}{\sigma} - \tau\sigma\right) + e^{d^\star\tau}\Phi\left(-\frac{d^\star}{\sigma} - \tau\sigma\right)} \dfrac{1}{\sigma}\phi\left(\dfrac{\theta - (d^\star + \sigma^2\tau)}{\sigma}\right), & \theta < 0 \end{cases}, \quad (85)$$

where ϕ and Φ respectively denote the pdf and cdf of the standard normal distribution. For derivations of these results, see Appendix. From the representation in (85) we can see that the posterior distribution is a mixture of truncated normals, which will be utilized in the Gibbs sampling algorithm. If we consider the mixture prior $p(\theta|\tau) = (1 - \epsilon_j)\delta_0 + \epsilon_j p_1(\theta|\tau)$ on θ in (82), we obtain the posterior distribution as

$$\begin{aligned} \pi(\theta|d^\star, \sigma^2, \tau) &= \frac{f(d^\star|\theta, \sigma^2)p(\theta|\tau)}{\int_{-\infty}^{\infty} f(d^\star|\theta, \sigma^2)p(\theta|\tau)d\theta} \\ &= \frac{(1 - \epsilon_j)f(d^\star|\theta, \sigma^2)\delta_0 + \epsilon_j f(d^\star|\theta, \sigma^2)p_1(\theta|\tau)}{(1 - \epsilon_j)f(d^\star|0, \sigma^2) + \epsilon_j m(d^\star|\sigma^2, \tau)} \\ &= \frac{(1 - \epsilon_j)f(d^\star|0, \sigma^2)\delta_0 + \epsilon_j m(d^\star|\sigma^2, \tau)h(\theta|d^\star, \sigma^2, \tau)}{(1 - \epsilon_j)f(d^\star|0, \sigma^2) + \epsilon_j m(d^\star|\sigma^2, \tau)} \\ &= (1 - p_j)\delta_0 + p_j h(\theta|d^\star, \sigma^2, \tau), \end{aligned} \quad (86)$$

where $f(d^\star|0, \sigma^2)$ is the normal distribution with mean $\theta = 0$ and variance σ^2, and

$$p_j = \frac{\epsilon_j m(d^\star|\sigma^2, \tau)}{(1 - \epsilon_j)f(d^\star|0, \sigma^2) + \epsilon_j m(d^\star|\sigma^2, \tau)} \quad (87)$$

is the mixing weight. Thus, the posterior distribution of θ is a mixture of point mass at zero and a mixture of truncated normal distributions $h(\theta|d^\star, \sigma^2, \tau)$ with mixing weight p_j.

5.3 Gibbs sampling scheme

To conduct posterior inference on the parameters θ_{jk} and β_i, we adopt a standard Gibbs sampling procedure. Gibbs sampling is an iterative algorithm that simulates from a joint posterior distribution through iterative simulation of the full conditional distributions. For more details on Gibbs sampling see Casella and George (1992) or Robert and Casella (1999). For the model in (83), full conditionals for all parameters can be determined exactly. We build on results given as (85), (86) and results derived by Yuan and Lin (2004). Derivations of the results in this section are deferred to Appendix.

Next we will find full conditional distributions and updating schemes for parameters γ_i, β_i, v_i, η^2, q, σ^2, z_{jk}, ϵ_j, θ_{jk}, and τ_θ, which are necessary to run the Gibbs sampler. Specification of the hyperparameters a_1, b_1, a_2, b_2, a_3 and b_3 will be done in Section 5.4.1.

5.3.1 Updating γ_i, β_i and v_i

In each Gibbs sampling iteration we first update the block (γ_i, β_i) by updating γ_i and β_i for $i = 1, \ldots, p$, and then we generate v_i for $i = 1, \ldots, p$.

5.3.1.1 Updating γ_i and β_i as a block

Here we follow the results of Yuan and Lin (2004) and we get

$$P(\gamma_i = 1|\boldsymbol{d}, \boldsymbol{\theta}, \sigma^2, \eta^2, \boldsymbol{\beta}^{[-i]}, v_i, \boldsymbol{\gamma}^{[-i]}) = \frac{1}{1 + \frac{f(\boldsymbol{d}|\boldsymbol{\theta}, \sigma^2, \eta^2, \boldsymbol{\beta}^{[-i]}, v, \boldsymbol{\gamma}^{[-i]}, \gamma_i = 0) P(\boldsymbol{\gamma}^{[-i]}, \gamma_i = 0)}{f(\boldsymbol{d}|\boldsymbol{\theta}, \sigma^2, \eta^2, \boldsymbol{\beta}^{[-i]}, v, \boldsymbol{\gamma}^{[-i]}, \gamma_i = 1) P(\boldsymbol{\gamma}^{[-i]}, \gamma_i = 1)}},$$

where

$$f(\boldsymbol{d}|\boldsymbol{\theta}, \sigma^2, \eta^2, \boldsymbol{\beta}^{[-i]}, v, \boldsymbol{\gamma}^{[-i]}, \gamma_i = 0) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\boldsymbol{Z}'\boldsymbol{Z}}{2\sigma^2}\right\},$$

and

$$f(\boldsymbol{d}|\boldsymbol{\theta},\sigma^2,\eta^2,\boldsymbol{\beta}^{[-i]},v,\boldsymbol{\gamma}^{[-i]},\gamma_i=1)=$$
$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\boldsymbol{Z}'\boldsymbol{Z}}{2\sigma^2}\right\}\sqrt{\frac{\sigma^2}{v_i\eta^2\boldsymbol{U}_i'\boldsymbol{U}_i+\sigma^2}}\exp\left\{\frac{v_i\eta^2(\boldsymbol{Z}'\boldsymbol{U}_i)^2}{2\sigma^2(v_i\eta^2\boldsymbol{U}_i'\boldsymbol{U}_i+\sigma^2)}\right\}.$$

Note that

$$\boldsymbol{Z}=\boldsymbol{d}-\boldsymbol{U}_{\boldsymbol{\gamma}^{[-i]},\gamma_i=0}\boldsymbol{\beta}_{\boldsymbol{\gamma}^{[-i]},\gamma_i=0}-\boldsymbol{\theta},$$

and

$$\frac{P(\boldsymbol{\gamma}^{[-i]},\gamma_i=0)}{P(\boldsymbol{\gamma}^{[-i]},\gamma_i=1)}=\frac{1-q^{(l-1)}}{q^{(l-1)}}.$$

Here the notation $\boldsymbol{\gamma}^{[-i]}$ and $\boldsymbol{\beta}^{[-i]}$ refers to vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ without the i^{th} element and $\boldsymbol{U}_i$ indicates the i^{th} column of matrix $\boldsymbol{U}$. Therefore, in the l^{th} iteration of the Gibbs sampling, update γ_i as a Bernoulli random variable with probabilities given

$$\gamma_i^{(l)}=\begin{cases}0, & \text{wp.} \quad 1-P\left(\gamma_i=1\middle|\boldsymbol{d},\boldsymbol{\theta}^{(l-1)},\sigma^{2(l-1)},\eta^{2(l-1)},\boldsymbol{\beta}^{[-i](l)},v_i^{(l-1)},\boldsymbol{\gamma}^{[-i](l)}\right)\\ 1, & \text{wp.} \quad P\left(\gamma_i=1\middle|\boldsymbol{d},\boldsymbol{\theta}^{(l-1)},\sigma^{2(l-1)},\eta^{2(l-1)},\boldsymbol{\beta}^{[-i](l)},v_i^{(l-1)},\boldsymbol{\gamma}^{[-i](l)}\right)\end{cases}. \tag{88}$$

Then it is straightforward to update β_i as

$$\beta_i^{(l)}\sim\begin{cases}\delta_0(\beta_i), & \text{if} \quad \gamma_i^{(l)}=0\\ \mathcal{N}\left(\frac{v_i^{(l-1)}\eta^{2(l-1)}(\boldsymbol{Z}'\boldsymbol{U}_i)^2}{v_i^{(l-1)}\eta^{2(l-1)}\boldsymbol{U}_i'\boldsymbol{U}_i+\sigma^{2(l-1)}},\frac{v_i^{(l-1)}\eta^{2(l-1)}\sigma^{2(l-1)}}{v_i^{(l-1)}\eta^{2(l-1)}\boldsymbol{U}_i'\boldsymbol{U}_i+\sigma^{2(l-1)}}\right), & \text{if} \quad \gamma_i^{(l)}=1\end{cases}. \tag{89}$$

Note that in the above equation $\boldsymbol{Z}=\boldsymbol{d}-\boldsymbol{U}_{\boldsymbol{\gamma}^{[-i]},\gamma_i=0}\boldsymbol{\beta}_{\boldsymbol{\gamma}^{[-i]},\gamma_i=0}-\boldsymbol{\theta}$ in which we substitute $\boldsymbol{\gamma}^{[-i](l)}$, $\boldsymbol{\beta}^{(l)}$ and $\boldsymbol{\theta}^{(l-1)}$. Also, $\delta_0(\beta_i)$ is a point mass distribution at zero, which is equivalent to $\beta_i=0$.

5.3.1.2 Updating v_i

For the scale mixture of normals representation of the double exponential distribution, we placed an exponential prior on v_i in model (83). We update v_i depending on the

value of the latent variable γ_i, whether β_i comes from a point mass or a normal prior. The updating scheme for v_i is

$$v_i^{(l)} \sim \begin{cases} \mathcal{E}xp(1), & \text{if} \quad \gamma_i^{(l)} = 0 \\ \mathcal{GIG}\left(2, \beta_i^{2(l)}/\eta^{2(l-1)}, 1/2\right), & \text{if} \quad \gamma_i^{(l)} = 1 \end{cases}, \tag{90}$$

where $\mathcal{GIG}(a, b, p)$ denotes the generalized inverse Gaussian distribution (Johnson et al., 1994, p.284) with probability density function

$$f(x|a, b, p) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} e^{-(ax+b/x)/2}, \quad x > 0; a, b > 0.$$

Here K_p denotes the modified Bessel function of the third kind. Simulation of $\mathcal{GIG}$ random variates is available through a MATLAB© implementation "randraw" based on Dagpunar (1989).

5.3.2 Updating η^2, q, ϵ_j and σ^2

Using a conjugate $\mathcal{IG}(a_2, b_2)$ prior on η^2 results in an inverse gamma full conditional distribution. Therefore, update η^2 as

$$\eta^{2(l)} \sim \mathcal{IG}\left(a_2 + 1/2 \sum_i {\gamma_i}^{(l)}, \left[1/b_2 + 1/2 \sum_i \left({\gamma_i}^{(l)} \beta_i^{2(l)}/v_i^{(l)}\right)\right]^{-1}\right). \tag{91}$$

Parameter q has a conjugate $\mathcal{B}e(1, 1)$ prior. This results in a full conditional distributed as beta,

$$q^{(l)} \sim \mathcal{B}e\left(1 + \sum_i {\gamma_i}^{(l)}, 1 + \sum_i \left(1 - {\gamma_i}^{(l)}\right)\right). \tag{92}$$

Similarly, parameter ϵ_j is given a conjugate $\mathcal{B}e(1, 1)$ prior, and the update is

$$\epsilon_j^{(l)} \sim \mathcal{B}e\left(1 + \sum_k z_{jk}^{(l)}, 1 + \sum_k \left(1 - z_{jk}^{(l)}\right)\right). \tag{93}$$

Note that other choices from the $\mathcal{B}e(\alpha, \beta)$ family are possible for the prior of ϵ_j and q, similarly. However, we used the noninformative choice $\alpha = 1$ and $\beta = 1$ to facilitate data-driven estimation of ϵ_j and q.

Using a conjugate $\mathcal{IG}(a_1, b_1)$ prior on σ^2 also results in an inverse gamma full conditional distribution. This leads to an update for σ^2 as

$$\sigma^{2(l)} \sim \mathcal{IG}\left(a_1 + n/2, [1/b_1 + \boldsymbol{Z}'\boldsymbol{Z}/2]^{-1}\right), \tag{94}$$

where $\boldsymbol{Z} = \boldsymbol{d} - \boldsymbol{U}_{\boldsymbol{\gamma}^{(l)}}\boldsymbol{\beta}^{(l)}_{\boldsymbol{\gamma}^{(l)}} - \boldsymbol{\theta}^{(l-1)}$ and $n = 2^J - 2^{J_0}$ denotes the sample size. $J-1$ and J_0 refer to the finest and coarsest levels in the wavelet decomposition, respectively.

5.3.3 Updating z_{jk}

We saw in model (83) that latent variable z_{jk} has a Bernoulli prior with parameter ϵ_j. Its full conditional distribution remains Bernoulli with parameter p_j as in (87). Thus, the latent variable z_{jk} is updated as follows:

$$z^{(l)}_{jk} = \begin{cases} 0, \quad \text{wp.} & \dfrac{\left(1-\epsilon_j^{(l-1)}\right) f\left(d^\star_{jk}|0,\sigma^{2(l)}\right)}{\left(1-\epsilon_j^{(l-1)}\right) f\left(d^\star_{jk}|0,\sigma^{2(l)}\right) + \epsilon_j^{(l-1)} m\left(d^\star_{jk}|\sigma^{2(l)},\tau_\theta^{(l-1)}\right)} \\ 1, \quad \text{wp.} & \dfrac{\epsilon_j^{(l-1)} m\left(d^\star_{jk}|\sigma^{2(l)},\tau_\theta^{(l-1)}\right)}{\left(1-\epsilon_j^{(l-1)}\right) f\left(d^\star_{jk}|0,\sigma^{2(l)}\right) + \epsilon_j^{(l-1)} m\left(d^\star_{jk}|\sigma^{2(l)},\tau_\theta^{(l-1)}\right)} \end{cases} \tag{95}$$

where $d^\star_{jk} = d_{jk} - \left(\boldsymbol{U}_{\boldsymbol{\gamma}^{(l)}}\boldsymbol{\beta}^{(l)}_{\boldsymbol{\gamma}^{(l)}}\right)_{jk}$.

5.3.4 Updating θ_{jk}

We approach updating θ_{jk} in a novel way. As we mentioned before, the common approach for handling the double exponential prior in hierarchical models is the scale mixture representation. This approach, however, introduces an additional parameter corresponding to each θ_{jk}, which needs to be updated. This adds $2^J - 2^{J_0}$ new parameters. A faster and more direct method to update θ_{jk} is possible by using results in (85) and (86). From the definition of latent variable z_{jk} we can easily see that $\theta_{jk} = 0$ if $z_{jk} = 0$, because for such z_{jk}, θ_{jk} is distributed as point mass at zero. In case $z_{jk} = 1$, θ_{jk} follows a mixture of truncated normal distributions a posteriori.

Therefore, the update for θ_{jk} is as follows:

$$\theta_{jk}^{(l)} \sim \begin{cases} \delta_0(\theta_{jk}), & \text{if} \quad z_{jk}^{(l)} = 0 \\ h\left(\theta_{jk} \middle| d_{jk}^{\star}, \sigma^{2(l)}, \tau_{\theta}^{(l-1)}\right), & \text{if} \quad z_{jk}^{(l)} = 1 \end{cases}, \tag{96}$$

where $d_{jk}^{\star} = d_{jk} - \left(\boldsymbol{U}_{\boldsymbol{\gamma}^{(l)}} \boldsymbol{\beta}_{\boldsymbol{\gamma}^{(l)}}^{(l)}\right)_{jk}$, $\delta_0(\theta)$ is a point mass distribution at zero, and $h(\theta|d^{\star}, \sigma^2, \tau_{\theta})$ is a mixture of truncated normal distributions with the density provided in (85). Simulating random variables from $h(\theta|d^{\star}, \sigma^2, \tau_{\theta})$ is nonstandard, and regular built-in methods fail, because we need to simulate random variables from tails of normal distributions having extremely low probability. The implementation of the updating algorithm is based on vectorizing a fast algorithm proposed by Robert (1995).

5.3.5 Updating τ_{θ}

The Gibbs updating scheme is completed with the discussion of how to update τ_{θ}. In the hierarchical model (83), we impose a gamma prior on the scale parameter of the double exponential distribution. This turns out to be a conjugate problem; therefore, we update τ_{θ} by

$$\tau_{\theta}^{(l)} \sim \mathcal{G}a\left(a_3 + \sum_{j,k} z_{jk}^{(l)}, \left[1/b_3 + \sum_{j,k}\left(z_{jk}^{(l)}|\theta_{jk}^{(l)}|\right)\right]^{-1}\right). \tag{97}$$

Note that the gamma distribution above is parameterized by its scale parameter.

Now the derivation of the updating algorithm is complete. Implementation of the described Gibbs sampler requires simulation routines for standard distributions such as the gamma, inverse gamma, Bernoulli, beta, exponential, normal, and also specialized routines to simulate from truncated normal, and generalized inverse Gaussian. The procedure was implemented in MATLAB and available from the author.

The Gibbs sampling procedure can be summarized as

(i) Choose initial values for parameters

(ii) Repeat steps (iii) - (xi) for $l = 1, \ldots, M$

(iii) Update the block (γ_i, β_i) for $i = 1, \ldots, p$

(iv) Update v_i for $i = 1, \ldots, p$

(v) Update η^2

(vi) Update q

(vii) Update σ^2

(viii) Update z_{jk} for $j = J_0, \ldots, \log_2(n) - 1,\ k = 0, \ldots, 2^j - 1$

(ix) Update ϵ_j for $j = J_0, \ldots, \log_2(n) - 1$

(x) Update θ_{jk} for $j = J_0, \ldots, \log_2(n) - 1,\ k = 0, \ldots, 2^j - 1$

(xi) Update τ_θ.

Note that the updating steps of vectors $\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\epsilon}$, and $\boldsymbol{\theta}$ are vectorized in the implementation, which considerably speeds up the computation.

5.4 Simulations

In this section, we apply the proposed Gibbs sampling algorithm and simulate posterior realizations for the model in (83). We will name our method *GS-WaPaLiM*, which is an acronym for Gibbs Sampling Wavelet-based Partially Linear Model (*GS-WaPaLiM*) method. Within each simulation step 20,000 Gibbs sampling iterations were performed, of which the first 5,000 were used for burn-in. We used the sample averages $\hat{\theta}_{jk} = \sum_l \theta_{jk}^{(l)}/L$ and $\hat{\beta}_i = \sum_l \beta_i^{(l)}/L$ as the usual estimator for the posterior mean. In our set-up, $L = 15,000$.

In what follows, we first discuss the selection of the hyperparameters, then compare the estimation performance with other methods on two simulated examples. Finally, variable selection will be demonstrated on an example.

5.4.1 Selection of Hyperparameters

In any Bayesian modeling task, the selection of hyperparameters is critical for good performance of the model. It is also desirable to have a default choice of the hyperparameters which makes the procedure automatic.

In order to apply the *GS-WaPaLiM* method, we only need to specify hyperparameters a_1, b_1, a_2, b_2, a_3, and b_3 in the hyperprior distributions. The advantage of the fully Bayesian approach is that once the hyperpriors are set, the estimation of parameters γ_i, β_i, v_i, η^2, q, σ^2, z_{jk}, ϵ_j, θ_{jk}, and τ_θ is automatic via the Gibbs sampling algorithm. The selection is governed by the data and hyperprior distributions on the parameters. Another advantage is that the method is relatively robust to the choice of hyperparameters since they influence the model at a higher level of hierarchy.

Critical parameters with respect to the performance of the shrinkage are ϵ_j and q, which control the strength of shrinkage of θ_{jk} and β_i to zero. In model (83), we placed a uniform prior on these parameters; therefore, the estimation will be governed mostly by the data, which provides a degree of adaptiveness. Parameter q represents the probability that a predictor enters the model a priori. When a priori information is available, it can be incorporated into the model, however, this is rarely the case. In the wavelet regression context, Abramovich et al. (1998) estimated parameter ϵ_j by a theoretically justified but somewhat involved method, and in Vidakovic and Ruggeri (2001), the estimation of this parameter depends on another hyperparameter γ, which is elicited based on empirical evidence. The proposed method provides a better alternative because of its automatic adaptiveness to the underlying nonparametric part of the model.

Another efficient way to elicit the hyperparameters of the model is through the empirical Bayes method performing maximization of the marginal likelihood. This approach was followed by Qu (2006) in the context of estimating partially linear

wavelet models. However, the likelihood function is nonconcave; therefore, clever optimization algorithm and carefully set starting values are crucial for the performance of this method. The same method of estimating hyperparameters was used for example by Clyde and George (1999) and Johnstone and Silverman (2005b) in the wavelet regression context, and by George and Foster (2000) in the linear regression context. Note that for the mixture priors specified on the parametric and nonparametric parts in model (83) the empirical Bayes approach might not be computationally tractable; therefore, the fully Bayesian approach provides a good alternative.

Default specification of hyperparameters a_1, b_1, a_2, b_2, a_3, and b_3 in model (83) is given by the following:

- We set $a_1 = 2$, $a_2 = 2$ and $a_3 = 1$.

- Then we compute naive estimators from the data

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$
$$\boldsymbol{Y_f} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS},$$

  where $\boldsymbol{Y_f}$ is an estimator of the nonparametric part of model (78), and $\hat{\boldsymbol{\beta}}_{OLS}$ is the ordinary least squares estimator for $\boldsymbol{\beta}$, although computed from the raw partially linear data.

- Then we set $b_1 = 1/\hat{\sigma}^2$, so that the mean of the inverse gamma prior becomes $\hat{\sigma}^2$. We use $\hat{\sigma}^2 = MAD/0.6745$, which is the usual robust estimator of the noise variation in the wavelet shrinkage literature (Donoho and Johnstone, 1994). Here MAD stands for the median absolute deviation of the wavelet coefficients $d_{jk}^{\boldsymbol{f}}$ at the finest level of detail and the constant 0.6745 calibrates the estimator to be comparable with the sample standard deviation. Note that coefficients $d_{jk}^{\boldsymbol{f}}$ correspond to $\boldsymbol{Y_f}$, therefore, $d_{jk}^{\boldsymbol{f}} = d_{jk} - (\boldsymbol{U}\hat{\boldsymbol{\beta}}_{OLS})_{jk}$.

- After this we set $b_3 = \hat{\tau}_\theta = \left(\sqrt{\max\{(\sigma_f^2 - \hat{\sigma}^2), 0\}}\right)^{-1}$, which sets the mean of the gamma prior on τ_θ equal to an estimator of τ_θ. This estimator is adopted from Vidakovic and Ruggeri (2001), where $\sigma_f^2 = \mathrm{Var}(\boldsymbol{Y_f})$.

- Finally we set $b_2 = 1/\hat{\eta}^2$, so that the mean of the inverse gamma prior is a pre-specified value, $\hat{\eta}^2$. Results in the estimation of β_is turned out to be somewhat sensitive to $\hat{\eta}^2$ for small sample size and small number of linear predictors. We used $\hat{\eta}^2 = \left(3 \max_i\{|\hat{\boldsymbol{\beta}}_{OLS_i}|\}\right)^2$, which specified a prior on β_i with large enough variance to work well in practice.

5.4.2 Simulations and Comparisons with Various Methods

In this section, we discuss the estimation performance of the proposed *GS-WaPaLiM* method and compare it to three methods from the partially linear wavelet model literature. The first one is the wavelet Backfitting algorithm (*BF*) proposed by Chang and Qu (2004), the second one is the *LEGEND* algorithm proposed by Gannaz (2007) and the last one is the double penalized PLM wavelet estimator (*DPPLM*) by Ding et al. (2011). A Bayesian wavelet-based algorithm for the same problem was proposed by Qu (2006). However, we found that the implementation of that algorithm is not robust to different simulated examples and initial values of the empirical Bayes procedure, therefore, we omitted it from our discussion.

The coarsest wavelet decomposition level was $J_0 = \lfloor \log_2(\log(n)) + 1 \rfloor$, as suggested from Antoniadis et al. (2001). Reconstruction of the theoretical signal was measured by the average mean squared error (AMSE), calculated as

$$\mathrm{AMSE} = \frac{1}{Mn} \sum_{m=1}^{M} \sum_{i=1}^{n} \left(\hat{Y}_i^{(m)} - Y_i\right)^2,$$

where M is the number of simulation runs, and $Y_i,\ i = 1, \ldots, n$ are known values of the simulated functions considered. We denote by $\hat{Y}_i^{(m)},\ i = 1, \ldots, n$ the estimator from the mth simulation run. Note again, that in each of these simulation runs we

perform 20,000 Gibbs sampling iterations in order to get the estimators $\hat{\theta}_{jk}$ and $\hat{\beta}_i$. Also note that $\hat{\boldsymbol{Y}} = \boldsymbol{W}'\hat{\boldsymbol{d}}$, where $\hat{\boldsymbol{d}} = \boldsymbol{U}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}}$. We also assess the performance in estimating the parametric part of the model by $\text{AMSE}_{\boldsymbol{\beta}}$, calculated as

$$\text{AMSE}_{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{p}\left(\hat{\beta}_i^{(m)} - \beta_i\right)^2.$$

In the following simulation study we also used a modification of the wavelet Backfitting algorithm proposed by Chang and Qu (2004). The original algorithm, denoted as *BF*, uses $\hat{\sigma}\sqrt{2\log(n)}$ as a soft threshold value in each iteration. In the modified algorithm we run the iterative algorithm a second time using the generalized cross-validation threshold as in Jansen et al. (1997). This simple modification significantly improves the performance of the original algorithm. The method will be denoted as *BFM* in the sequel.

The procedure based on Gannaz (2007), denoted as *LEGEND*, is a wavelet thresholding based estimation procedure solved by the proposed *LEGEND* algorithm. The formulation of the problem is similar to the one in Chang and Qu (2004) and Fadili and Bullmore (2005), penalizing only the wavelet coefficients of the nonparametric part, but the solution is faster by recognizing the connection with Huber's M-estimation of a standard linear model with outliers.

The algorithm by Ding et al. (2011) will be denoted as *DPPLM* in the simulations. The authors discuss several simulation results based on how the Lasso penalty parameter λ_2 was chosen and whether the adaptive Lasso algorithm was used or not in the estimation procedure. It was reported that the *GCV* criteria with adaptive Lasso provided the smallest AMSE results, therefore, that version of the algorithm is used in the present simulations. We will refer to the method as *DPPLM-GCV* in the future.

For comparison purposes we use two simulation examples, one from Qu (2006), and another one from Ding et al. (2011).

Example 1

The first example is based on an example in Qu (2006). The simulated data are generated from

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{f}(t_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \sim N(0,1)$ and $\boldsymbol{\beta} = (0.5, 1)'$ with $p = 2$. The nonparametric test functions are $\boldsymbol{f}(t) = c_j \boldsymbol{f}_j(t),\ j = 1, \ldots, 4$, where $\boldsymbol{f}_1(t) = \texttt{Blocks}$, $\boldsymbol{f}_2(t) = \texttt{Bumps}$, $\boldsymbol{f}_3(t) = \texttt{Doppler}$ and $\boldsymbol{f}_4(t) = \texttt{Heavisine}$. These are four standard test functions considered by Donoho and Johnstone (1994). We chose $c_1 = 3$, $c_2 = 7$, $c_3 = 18$ and $c_4 = 2$ to have reasonable signal-to-noise ratios (SNR). The test functions were simulated at $n = 64, 128, 256$ and 512 points, and the nonparametric components were equally spaced in the unit interval. The standard wavelet bases were used: Symmlet 8 for `Heavisine` and `Doppler`, Daubechies 6 for `Bumps` and Haar for `Blocks`. The two columns of the design matrix were generated as independent $N(0,1)$ random variables.

Results of the simulation are presented in Table 9. It can be seen that the proposed *GS-WaPaLiM* method gives better AMSE and $\text{AMSE}_{\boldsymbol{\beta}}$ results in most test scenarios. It is apparent that the modified version of the Backfitting algorithm (BFM) provides better results than the original backfitting algorithm (BF). Note that an additional uncertainty results from estimating the noise variance σ^2, which was assumed to be known in the simulations by Chang and Qu (2004). $LEGEND$ provides comparable results to the BF algorithm, since both are using the same least squares formulation penalizing only the wavelet coefficients of the nonparametric part of the model. The solution algorithm and estimation of the noise is different in these methods. Note that boldface numbers indicate the smallest AMSE result for each test scenario.

Example 2

The second example is based on a simulation example from Ding et al. (2011). The

Table 9: AMSE comparison of the *GS-WaPaLiM* method to other methods for Example 1.

Signal	N	Method	AMSE	$\text{AMSE}_{\boldsymbol{\beta}}$	Signal	N	Method	AMSE	$\text{AMSE}_{\boldsymbol{\beta}}$
Blocks	64	GS-WaPaLiM	**0.6104**	**0.1119**	Doppler	64	GS-WaPaLiM	1.0189	0.1552
		BF	8.4895	0.6351			BF	3.6740	0.2322
		BFM	1.0921	0.1773			BFM	1.1392	**0.1274**
		LEGEND	7.4006	0.5465			LEGEND	4.7976	0.3049
		DPPLM-GCV	0.9103	0.1497			DPPLM-GCV	1.0281	0.1295
	128	GS-WaPaLiM	**0.3904**	**0.0285**		128	GS-WaPaLiM	**0.4839**	**0.0365**
		BF	4.6501	0.1660			BF	2.4664	0.0777
		BFM	0.6247	0.0435			BFM	0.6709	0.0402
		LEGEND	3.5357	0.1372			LEGEND	2.8148	0.0827
		DPPLM-GCV	0.6120	0.0418			DPPLM-GCV	0.6570	0.0421
	256	GS-WaPaLiM	**0.2513**	**0.0108**		256	GS-WaPaLiM	**0.3785**	**0.0137**
		BF	2.3339	0.0427			BF	1.7548	0.0270
		BFM	0.4546	0.0161			BFM	0.4909	0.0152
		LEGEND	1.9722	0.0368			LEGEND	1.8112	0.0271
		DPPLM-GCV	0.4579	0.0163			DPPLM-GCV	0.4873	0.0150
	512	GS-WaPaLiM	**0.1737**	**0.0040**		512	GS-WaPaLiM	**0.2266**	**0.0044**
		BF	1.3563	0.0111			BF	0.9508	0.0078
		BFM	0.3377	0.0056			BFM	0.3081	0.0048
		LEGEND	1.2783	0.0105			LEGEND	1.0479	0.0079
		DPPLM-GCV	0.3345	0.0056			DPPLM-GCV	0.3071	0.0047
Bumps	64	GS-WaPaLiM	**0.7876**	**0.1732**	Heavisine	64	GS-WaPaLiM	**0.4273**	0.0657
		BF	8.0091	0.6035			BF	0.5362	0.0515
		BFM	1.3190	0.2130			BFM	0.4584	**0.0485**
		LEGEND	9.4988	0.5074			LEGEND	1.5180	0.1042
		DPPLM-GCV	1.1293	0.2015			DPPLM-GCV	0.4501	0.0508
	128	GS-WaPaLiM	**0.7320**	0.0782		128	GS-WaPaLiM	**0.2836**	0.0208
		BF	7.3514	0.2355			BF	0.4733	0.0228
		BFM	1.1271	0.0789			BFM	0.3490	**0.0202**
		LEGEND	7.8628	0.2150			LEGEND	0.9679	0.0317
		DPPLM-GCV	1.0199	**0.0764**			DPPLM-GCV	0.3543	0.0208
	256	GS-WaPaLiM	**0.5555**	**0.0198**		256	GS-WaPaLiM	**0.1993**	**0.0100**
		BF	3.9389	0.0501			BF	0.3631	0.0119
		BFM	0.7420	0.0226			BFM	0.2654	0.0108
		LEGEND	3.9856	0.0482			LEGEND	0.6548	0.0133
		DPPLM-GCV	0.7366	0.0225			DPPLM-GCV	0.2676	0.0107
	512	GS-WaPaLiM	**0.4339**	**0.0066**		512	GS-WaPaLiM	**0.1278**	**0.0039**
		BF	2.8462	0.0200			BF	0.2546	0.0044
		BFM	0.5914	0.0090			BFM	0.1669	0.0041
		LEGEND	2.8794	0.0206			LEGEND	0.4198	0.0049
		DPPLM-GCV	0.5903	0.0090			DPPLM-GCV	0.1677	0.0041

simulated data are generated from

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{f}(t_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \sim N(0,1)$ and $\boldsymbol{\beta} = (1.5, 2, 2.5, 3, 0, \ldots, 0)'$ with $p = 20$. The parametric part of the model is sparse, where only the first 4 regression variables are significant. The nonparametric test functions are $\boldsymbol{f}(t) = c_j \boldsymbol{f}_j(t),\ j = 1, 2$, where $\boldsymbol{f}_1(t) =$ `PiecePoly` given in Nason (1996) and $\boldsymbol{f}_2(t) =$ `Bumps`. We chose $c_1 = 9$ and $c_2 = 3$ to have reasonable signal-to-noise ratios (SNR). The test functions were simulated at $n = 128, 256$ and 512 points, and Daubechies 8 wavelet base were used in both cases of the test functions. Rows of the design matrix $\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T$ were independently generated from 20-dimensional multivariate normal distribution with zero mean vector, variance 1 and pairwise correlation coefficient between consecutive elements of the rows $\rho = 0.4$.

Results of the simulation are presented in Table 10. Note that boldface numbers indicate the smallest AMSE results for each test scenario. It can be seen that the proposed *GS-WaPaLiM* method gives better AMSE and AMSE$_{\boldsymbol{\beta}}$ results in all test scenarios. In this example the parametric part of the model is sparse, therefore, the double penalized wavelet estimator is superior to the wavelet backfitting and $LEGEND$ algorithms, especially in estimating β_is. Since the true $\boldsymbol{\beta}$ is a sparse vector, penalized estimation of the coefficients provides superior results as opposed to the BF, BFM and $LEGEND$ methods, which only penalize the wavelet coefficients corresponding to the nonparametric part in the estimation procedure. Similarly to Example 1, $LEGEND$ provides comparable results to the BF algorithm. The proposed *GS-WaPaLiM* method provides superior performance both in estimating the overall signal and the linear regression coefficients compared to the non-Bayesian methods considered.

Table 10: AMSE comparison of the *GS-WaPaLiM* method to other methods for Example 2.

Signal	N	Method	AMSE	$\text{AMSE}_{\boldsymbol{\beta}}$	Signal	N	Method	AMSE	$\text{AMSE}_{\boldsymbol{\beta}}$
PiecePoly	128	GS-WaPaLiM	**0.2786**	**0.0618**	Bumps	128	GS-WaPaLiM	**0.6700**	**0.1359**
		BF	0.4520	0.4067			BF	2.5257	1.3275
		BFM	0.4101	0.3985			BFM	1.1061	0.9452
		LEGEND	0.4379	0.4027			LEGEND	2.8525	1.4145
		DPPLM-GCV	0.3613	0.1790			DPPLM-GCV	0.8800	0.6331
	256	GS-WaPaLiM	**0.1799**	**0.0267**		256	GS-WaPaLiM	**0.4847**	**0.0398**
		BF	0.3366	0.1752			BF	1.8561	0.4884
		BFM	0.2585	0.1639			BFM	0.6999	0.3254
		LEGEND	0.2985	0.1688			LEGEND	1.9165	0.4954
		DPPLM-GCV	0.2289	0.0732			DPPLM-GCV	0.6378	0.2121
	512	GS-WaPaLiM	**0.1160**	**0.0137**		512	GS-WaPaLiM	**0.3900**	**0.0173**
		BF	0.2440	0.0831			BF	1.5642	0.2001
		BFM	0.1672	0.0772			BFM	0.5188	0.1287
		LEGEND	0.2133	0.0805			LEGEND	1.4924	0.1935
		DPPLM-GCV	0.1524	0.0325			DPPLM-GCV	0.5072	0.0819

5.4.3 Variable selection

A distinguishing feature of the proposed algorithm is that it can be used for variable selection. The method proposed by Ding et al. (2011) was developed for variable selection, but in the Bayesian framework, the method proposed by Qu (2006) is not able to perform this important task.

The proposed methodology can simply mimic the machinery of SSVS (stochastic search variable selection) by George and McCulloch (1993). Recall, that latent variable γ_i indicates whether predictor i should be included in the model or not. We can select the best subset of linear predictors by using Gibbs sampling to identify models with higher posterior probability $f(\boldsymbol{\gamma}|d)$. In the Gibbs sampling procedure we generate the sequence $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(l)}$ which converges to the posterior distribution $f(\boldsymbol{\gamma}|\boldsymbol{d})$. Simple calculation of the empirical frequency of $\boldsymbol{\gamma}$ or different strategies mentioned in George and McCulloch (1993) can be used to identify the best subsets of predictors.

To illustrate this, we show how variable selection works on Example 2 from the previous section, using `Bumps` for the nonparametric component and $n = 128$. Remember that $p = 20$, therefore, there are 2^{20} candidate models. Table 11 shows 10

models with the highest estimated posterior probability based on 20,000 runs (5,000 was burn-in) of the Gibbs sampling algorithm. We can see that the method identifies the true model with distinctively highest posterior probability, even for $n = 128$. In case $n = 256$, the estimated posterior probability of the true model is 0.8128.

Table 11: Subset models with highest estimated posterior probabilities.

Variables	Posterior probability
x_1, x_2, x_3, x_4	0.2885
$x_1, x_2, x_3, x_4, x_{19}$	0.1020
x_1, x_2, x_3, x_4, x_9	0.0646
x_1, x_2, x_3, x_4, x_6	0.0321
$x_1, x_2, x_3, x_4, x_{16}$	0.0304
$x_1, x_2, x_3, x_4, x_{15}$	0.0271
$x_1, x_2, x_3, x_4, x_9, x_{15}$	0.0236
$x_1, x_2, x_3, x_4, x_{20}$	0.0197
$x_1, x_2, x_3, x_4, x_{16}, x_{19}$	0.0167
$x_1, x_2, x_3, x_4, x_{15}, x_{19}$	0.0164

5.5 Conclusions

In this chapter we proposed a wavelet-based method for estimation and variable selection in partially linear models. Because wavelets provide efficient representation for wide ranges of functions, the inference was conducted in the wavelet domain. A fully Bayesian approach was taken, in which a mixture prior was specified on both the parametric and nonparametric components of the model, unifying modeling approaches from both the Bayesian linear models and the wavelet shrinkage literature. Estimation and variable selection was performed by a Gibbs sampling procedure. It was shown through simulated examples that the methodology provides superior performance compared to the penalized least squares approach, most common in the existing literature.

The developed algorithm is efficient; however, the computational time considerably increases when the number of covariates in the linear part of the model grows. Another limitation is the usual assumptions of wavelet regression, that is, we assumed

equally spaced sampling points without replicates for the nonparametric component, and the number of observations was assumed to be a power of two. This can be a limitation for analyzing real-world data sets, however, wavelet transforms extending these assumptions can be found in the literature, see for example Kovac and Silverman (2000).

APPENDIX A

DERIVATIONS OF SOME RESULTS

A.1 Derivations for Chapter 2

First we provide derivation of results (25) and (26). The joint distribution $f(d, \theta|\sigma^2, \tau)$ using prior $p_1(\theta|\tau)$ is

$$\begin{aligned}
f(d,\theta|\sigma^2,\tau) &= f(d|\theta,\sigma^2)p_1(\theta|\tau) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(d-\theta)^2}{2\sigma^2}}\frac{\tau}{2}e^{-\tau|\theta|} \\
&= \frac{\tau}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\{\theta^2-2\theta(-\text{sign}(\theta)\sigma^2\tau+d)+d^2\}} \\
&= \frac{\tau}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\{\theta-(d-\text{sign}(\theta)\sigma^2\tau)^2\}} - e^{-\frac{1}{2\sigma^2}\{-(d-\text{sign}(\theta)\sigma^2\tau)^2+d^2\}} \\
&= \frac{\tau e^{\frac{\sigma^2\tau^2}{2}}e^{-\text{sign}(\theta)d\tau}}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\{\theta-(d-\text{sign}(\theta)\sigma^2\tau)\}^2} \\
&= \begin{cases} \dfrac{\tau e^{\frac{\sigma^2\tau^2}{2}}e^{-d\tau}}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d-\sigma^2\tau)]^2}, & \theta \geq 0 \\ \\ \dfrac{\tau e^{\frac{\sigma^2\tau^2}{2}}e^{d\tau}}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d+\sigma^2\tau)]^2}, & \theta < 0 \end{cases}.
\end{aligned}$$

The marginal distribution becomes

$$\begin{aligned}
m(d|\sigma^2,\tau) &= \int_{-\infty}^{\infty} f(d,\theta|\sigma^2,\tau)d\theta \\
&= \frac{\tau}{2}e^{\frac{\sigma^2\tau^2}{2}}\left\{e^{d\tau}\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d+\sigma^2\tau)]^2}d\theta + \right. \\
&\quad \left. e^{-d\tau}\int_{0}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d-\sigma^2\tau)]^2}d\theta\right\} \\
&= \frac{\tau}{2}e^{\frac{\sigma^2\tau^2}{2}}\left\{e^{d\tau}\Phi\left(\frac{-d-\sigma^2\tau}{\sigma}\right) + e^{-d\tau}\Phi\left(\frac{d-\sigma^2\tau}{\sigma}\right)\right\}.
\end{aligned}$$

Combining the two equations above, we get the posterior as

$$h(\theta|d,\sigma^2,\tau) = \frac{f(d,\theta|\sigma^2,\tau)}{m(d|\sigma^2,\tau)} =$$

$$\begin{cases} \dfrac{e^{-d\tau}}{e^{-d\tau}\Phi\left(\frac{d}{\sigma}-\tau\sigma\right)+e^{d\tau}\Phi\left(-\frac{d}{\sigma}-\tau\sigma\right)}\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d-\sigma^2\tau)]^2}, & \theta \geq 0 \\ \\ \dfrac{e^{d\tau}}{e^{-d\tau}\Phi\left(\frac{d}{\sigma}-\tau\sigma\right)+e^{d\tau}\Phi\left(-\frac{d}{\sigma}-\tau\sigma\right)}\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d+\sigma^2\tau)]^2}, & \theta < 0 \end{cases}.$$

These results were also derived by Pericchi and Smith (1992) and used by Johnstone and Silverman (2005b).

Now we derive the results used for the Gibbs sampling algorithm of model (24). To derive the full conditional distribution for a parameter of interest we look at the joint distribution of all the parameters and collect the terms which contain the desired parameter. Let us denote $\boldsymbol{d} = \{d_{jk} : j = J_0, \ldots, \log_2(n)-1,\ k = 0, \ldots, 2^j-1\}$, $\boldsymbol{\theta} = \{\theta_{jk} : j = J_0, \ldots, \log_2(n)-1,\ k = 0, \ldots, 2^j-1\}$, $\boldsymbol{z} = \{z_{jk} : j = J_0, \ldots, \log_2(n)-1,\ k = 0, \ldots, 2^j-1\}$ and $\boldsymbol{\epsilon} = \{\epsilon_j : j = J_0, \ldots, \log_2(n)-1\}$. The joint distribution of the data and parameters for model in (24) becomes

$$\begin{aligned} f(\boldsymbol{d},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\epsilon},\sigma^2,\tau) \;=\; & \left[\prod_{j,k}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(d_{jk}-\theta_{jk})^2}\right]\frac{1}{\Gamma(a_1){b_1}^{a_1}}(\sigma^2)^{-a_1-1}e^{-\frac{1}{\sigma^2}\frac{1}{b_1}}\cdot \\ & \left[\prod_{j,k}\left\{(1-z_{jk})\delta_0 + z_{jk}\frac{\tau}{2}e^{-\tau|\theta_{jk}|}\right\}\right]\cdot \\ & \left[\prod_{j,k}\epsilon_j^{z_{jk}}(1-\epsilon_j)^{(1-z_{jk})}\right]\cdot \\ & \left[\prod_{j}\mathbf{1}\{0\leq\epsilon_j\leq 1\}\right]\frac{1}{\Gamma(a_2){b_2}^{a_2}}\tau^{a_2-1}e^{-\tau/b_2}. \end{aligned}$$

From the joint distribution, the full conditional distribution of σ^2 is

$$\begin{aligned} p(\sigma^2|\boldsymbol{\theta}, \boldsymbol{d}) &\propto (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{j,k}(d_{jk}-\theta_{jk})^2} (\sigma^2)^{-a_1-1} e^{-\frac{1}{\sigma^2}\frac{1}{b_1}} \\ &= (\sigma^2)^{-a_1-n/2-1} \exp\left\{-\frac{1}{\sigma^2}\left(1/b_1 + 1/2\sum_{j,k}(d_{jk}-\theta_{jk})^2\right)\right\} \\ &= \mathcal{IG}\left(a_1 + n/2, \left[1/b_1 + 1/2\sum_{j,k}(d_{jk}-\theta_{jk})^2\right]^{-1}\right). \end{aligned}$$

The conditional distribution of z_{jk} remains Bernoulli with posterior probability derived by

$$P(z_{jk}=1|d_{jk},\sigma^2,\tau,\epsilon_j) = \frac{P(z_{jk}=1|\epsilon_j) f(d_{jk}|\sigma^2,\tau,z_{jk}=1)}{\sum_{i\in\{0,1\}} P(z_{jk}=i|\epsilon_j) f(d_{jk}|\sigma^2,\tau,z_{jk}=i)} =$$
$$\frac{P(z_{jk}=1|\epsilon_j)\int_{-\infty}^{\infty} f(d_{jk}|\theta_{jk},\sigma^2) p(\theta_{jk}|\tau,z_{jk}=1) d\theta_{jk}}{\sum_{i\in\{0,1\}} P(z_{jk}=i|\epsilon_j)\int_{-\infty}^{\infty} f(d_{jk}|\theta_{jk},\sigma^2) p(\theta_{jk}|\tau,z_{jk}=i) d\theta_{jk}} =$$
$$\frac{\epsilon_j m\left(d_{jk}|\sigma^2,\tau\right)}{\left(1-\epsilon_j\right) f\left(d_{jk}|0,\sigma^2\right) + \epsilon_j m\left(d_{jk}|\sigma^2,\tau\right)}.$$

Here $p(\theta_{jk}|\tau, z_{jk}=i),\ i\in\{0,1\}$ denote the two parts of the mixture prior in model (24), depending on the value of latent variable z_{jk}. Similar result was used by Yuan and Lin (2005).

The full conditional distribution of ϵ_j is

$$\begin{aligned} p(\epsilon_j|\boldsymbol{z}) &\propto \left[\prod_k \epsilon_j^{z_{jk}}(1-\epsilon_j)^{(1-z_{jk})}\right] \mathbf{1}\{0\le\epsilon_j\le 1\} \\ &= \epsilon_j^{\sum_k z_{jk}}(1-\epsilon_j)^{\sum_k(1-z_{jk})} \\ &= \mathcal{B}e\left(1+\sum_k z_{jk}, 1+\sum_k(1-z_{jk})\right). \end{aligned}$$

Similarly, the full conditional distribution of θ_{jk} is

$$\begin{aligned} p(\theta_{jk}|d_{jk}, z_{jk}, \sigma^2, \tau) &\propto \exp\left\{-\frac{1}{2\sigma^2}(d_{jk}-\theta_{jk})^2\right\}\left\{(1-z_{jk})\delta_0 + z_{jk}\frac{\tau}{2}e^{-\tau|\theta_{jk}|}\right\} \\ &= \begin{cases} \delta_0(\theta_{jk}), & \text{if} \quad z_{jk}=0 \\ h(\theta_{jk}|d_{jk}, \sigma^2, \tau), & \text{if} \quad z_{jk}=1 \end{cases}, \end{aligned}$$

where the distribution $h(\theta_{jk}|d_{jk}, \sigma^2, \tau)$ comes from the result in (26) and was derived above.

Finally, the full conditional distribution of τ is

$$\begin{aligned} p(\tau|\boldsymbol{\theta}, \boldsymbol{z}) &\propto \left[\prod_{j,k}\left\{(1-z_{jk})\delta_0 + z_{jk}\frac{\tau}{2}e^{-\tau|\theta_{jk}|}\right\}\right]\frac{1}{\Gamma(a_2){b_2}^{a_2}}\tau^{a_2-1}e^{-\tau/b_2} \\ &\propto \tau^{a_2+\sum_{j,k} z_{jk}-1}\exp\left\{-\tau\left(\sum_{j,k}(z_{jk}|\theta_{jk}|) + 1/b_2\right)\right\} \\ &= \mathcal{G}a\left(a_2 + \sum_{j,k} z_{jk}, \left[1/b_2 + \sum_{j,k}(z_{jk}|\theta_{jk}|)\right]^{-1}\right). \end{aligned}$$

Next we present some results used for the Gibbs sampling algorithm of the bivariate model (34). The notation is the same as before, but $\boldsymbol{d}$ and $\boldsymbol{\theta}$ represent vectors with bivariate components d_{jk} and θ_{jk}, respectively. Let us denote $\boldsymbol{v} = \{v_{jk} : j = J_0, \ldots, \log_2(n)-1,\ k = 0, \ldots, 2^j - 1\}$ and $\boldsymbol{C} = \{C_j : j = J_0, \ldots, \log_2(n)-1\}$ the vector containing matrices C_j for resolution levels j. The joint distribution of the

data and parameters is

$$
\begin{aligned}
f(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\epsilon}, \boldsymbol{v}, \sigma^2, \boldsymbol{C}) \;=\; & \left[\prod_{j,k} \frac{1}{\sqrt{2\pi}|\sigma^2\Sigma_j|^{1/2}} \cdot \right. \\
& \left. \exp\left\{-\frac{1}{2\sigma^2}(d_{jk}-\theta_{jk})'\Sigma_j^{-1}(d_{jk}-\theta_{jk})\right\}\right] \cdot \\
& \frac{1}{\Gamma(a)b^a}(\sigma^2)^{-a-1}e^{-\frac{1}{\sigma^2}\frac{1}{b}} \left[\prod_{j,k}\{(1-z_{jk})\delta_0 + \right. \\
& \left. z_{jk}\frac{1}{\sqrt{2\pi}|v_{jk}C_j|^{1/2}}\exp\left\{-\frac{1}{2v_{jk}}\theta_{jk}'C_j^{-1}\theta_{jk}\right\}\right\}\right] \cdot \\
& \left[\prod_{j,k}\epsilon_j^{z_{jk}}(1-\epsilon_j)^{(1-z_{jk})}\right]\left[\prod_j \mathbf{1}\{0\le\epsilon_j\le 1\}\right] \cdot \\
& \left[\prod_{j,k}\frac{1}{\Gamma(3/2)8^{3/2}}v_{jk}^{3/2-1}e^{-v_{jk}/8}\right] \cdot \\
& \left[\prod_j |C_j|^{-(w+d+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(A_jC_j^{-1}\right)\right\}\right].
\end{aligned}
$$

From the joint distribution, the full conditional distribution of σ^2 is

$$
\begin{aligned}
p(\sigma^2|\boldsymbol{\theta}, \boldsymbol{d}) \;&\propto\; \left(\frac{1}{\sigma^2}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{j,k}(d_{jk}-\theta_{jk})'\Sigma_j^{-1}(d_{jk}-\theta_{jk})\right\}\left(\sigma^2\right)^{-a-1}e^{-\frac{1}{\sigma^2}\frac{1}{b}} \\
&=\; (\sigma^2)^{-a-n-1}\exp\left\{-\frac{1}{\sigma^2}\left(1/b + 1/2\sum_{j,k}(d_{jk}-\theta_{jk})'\Sigma_j^{-1}(d_{jk}-\theta_{jk})\right)\right\} \\
&=\; \mathcal{IG}\left(a+n, \left[1/b + 1/2\sum_{j,k}(d_{jk}-\theta_{jk})'\Sigma_j^{-1}(d_{jk}-\theta_{jk})\right]^{-1}\right).
\end{aligned}
$$

Similarly as before, the conditional distribution of z_{jk} is Bernoulli with success probability

$$
P(z_{jk}=1|d_{jk}, \sigma^2, \epsilon_j, v_{jk}, C_j) = \frac{\epsilon_j m\left(d_{jk}|\sigma^2, v_{jk}, C_j\right)}{(1-\epsilon_j)\, f\left(d_{jk}|0, \sigma^2\right) + \epsilon_j m\left(d_{jk}|\sigma^2, v_{jk}, C_j\right)},
$$

where

$$\begin{aligned} f(d_{jk}|0,\sigma^2) &= \frac{1}{2\pi|\sigma^2\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}d'_{jk}\Sigma_j^{-1}d_{jk}\right\}, \\ m(d_{jk}|\sigma^2,v_{jk},C_j) &= \frac{1}{2\pi|\sigma^2\Sigma_j+v_{jk}C_j|^{1/2}}\exp\left\{-\frac{1}{2}d'_{jk}\left(\sigma^2\Sigma_j+v_{jk}C_j\right)^{-1}d_{jk}\right\}. \end{aligned}$$

The marginal distribution $m(d_{jk}|\sigma^2,v_{jk},C_j)$ is a bivariate normal distribution with zero mean and covariance matrix $\sigma^2\Sigma_j+v_{jk}C_j$. The derivation is a standard result of having a multivariate normal prior on the mean of the multivariate normal distribution. The result in a general form can be found for example in Lindley and Smith (1972).

The full conditional distribution of ϵ_j remains the same as for the real-valued model. The full conditional distribution of θ_{jk} is

$$\begin{aligned} p(\theta_{jk}|d_{jk},z_{jk},\sigma^2,v_{jk},C_j) &\propto \exp\left\{-\frac{1}{2\sigma^2}(d_{jk}-\theta_{jk})'\Sigma_j^{-1}(d_{jk}-\theta_{jk})\right\}\cdot \\ &\quad [(1-z_{jk})\delta_0+ \\ &\quad z_{jk}\frac{1}{\sqrt{2\pi}|v_{jk}C_j|^{1/2}}\exp\left\{-\frac{1}{2v_{jk}}\theta'_{jk}C_j^{-1}\theta_{jk}\right\}] \\ &= \begin{cases} \delta_0(\theta_{jk}), & \text{if} \quad z_{jk}=0 \\ f\left(\theta_{jk}|d_{jk},\sigma^2,v_{jk},C_j\right), & \text{if} \quad z_{jk}=1 \end{cases}, \end{aligned}$$

where

$$\begin{aligned} f(\theta_{jk}|d_{jk},\sigma^2,v_{jk},C_j) &= \frac{1}{2\pi|\tilde{\Sigma}_{jk}|^{1/2}}\exp\left\{-\frac{1}{2}\tilde{\mu}'_{jk}\tilde{\Sigma}_{jk}^{-1}\tilde{\mu}_{jk}\right\}, \\ \tilde{\mu}_{jk} &= \tilde{\Sigma}_{jk}\frac{\Sigma_j^{-1}}{\sigma^2}d_{jk}, \\ \tilde{\Sigma}_{jk} &= \left(\Sigma_j^{-1}/\sigma^2+C_j^{-1}/v_{jk}\right)^{-1}. \end{aligned}$$

Derivation of $f(\theta_{jk}|d_{jk},\sigma^2,v_{jk},C_j)$ is also a standard result contained for example in Lindley and Smith (1972) and was used in the wavelet shrinkage context by Barber and Nason (2004).

The full conditional distribution of v_{jk} is proportional to

$$\begin{aligned} p(v_{jk}|\theta_{jk}, z_{jk}, C_j) &\propto \left[(1-z_{jk})\delta_0 + z_{jk}\frac{1}{\sqrt{2\pi}|v_{jk}C_j|^{1/2}}\exp\left\{-\frac{1}{2v_{jk}}\theta'_{jk}C_j^{-1}\theta_{jk}\right\}\right]\cdot \\ & \quad v_{jk}^{3/2-1}\exp\left\{-\frac{v_{jk}}{8}\right\}. \end{aligned}$$

In case $z_{jk}=0$, this becomes

$$\begin{aligned} p(v_{jk}|\theta_{jk}, z_{jk}=0, C_j) &\propto v_{jk}^{3/2-1}\exp\left\{-\frac{v_{jk}}{8}\right\} \\ &= \mathcal{G}a(3/2, 8), \end{aligned}$$

and when $z_{jk}=1$, it becomes

$$\begin{aligned} p(v_{jk}|\theta_{jk}, z_{jk}=1, C_j) &\propto \frac{1}{v_{jk}}\exp\left\{-\frac{1}{2v_{jk}}\theta'_{jk}C_j^{-1}\theta_{jk}\right\} v_{jk}^{3/2-1}\exp\left\{-\frac{v_{jk}}{8}\right\} \\ &= v_{jk}^{1/2-1}\exp\left\{-\frac{1}{2}\left(\frac{1}{4}v_{jk} + \theta'_{jk}C_j^{-1}\theta_{jk}\frac{1}{v_{jk}}\right)\right\} \\ &= \mathcal{GIG}\left(1/4, \theta'_{jk}C_j^{-1}\theta_{jk}, 1/2\right). \end{aligned}$$

Here $\mathcal{GIG}(a,b,p)$ denotes the generalized inverse Gaussian distribution (Johnson et al., 1994, p.284) with probability density function

$$f(x|a,b,p) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}x^{p-1}e^{-(ax+b/x)/2}, \quad x>0; a,b>0,$$

where K_p denotes the modified Bessel function of the third kind.

Finally, the full conditional distribution of C_j is given as

$$\begin{aligned} p(C_j|\boldsymbol{\theta_j}, \boldsymbol{z_j}, \boldsymbol{v_j}) &\propto \prod_k \left[(1-z_{jk})\delta_0 + z_{jk}\frac{1}{\sqrt{2\pi}|v_{jk}C_j|^{1/2}}\exp\left\{-\frac{1}{2v_{jk}}\theta'_{jk}C_j^{-1}\theta_{jk}\right\}\right]\cdot \\ & \quad |C_j|^{-(w+d+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(A_jC_j^{-1}\right)\right\} \\ &= \prod_k \left[(1-z_{jk})\delta_0 + z_{jk}\frac{1}{\sqrt{2\pi}|v_{jk}C_j|^{1/2}}\cdot\right. \\ & \quad \left.\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\frac{\theta_{jk}\theta'_{jk}}{v_{jk}}C_j^{-1}\right)\right\}\right]|C_j|^{-(w+d+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(A_jC_j^{-1}\right)\right\} \\ &\propto |C_j|^{-\left(\sum_k z_{jk}+w+d+1\right)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left[A_j + \sum_k z_{jk}\frac{\theta_{jk}\theta'_{jk}}{v_{jk}}\right]C_j^{-1}\right)\right\} \\ &= \mathcal{IW}\left(A_j + \sum_k z_{jk}\frac{\theta_{jk}\theta'_{jk}}{v_{jk}}, w + \sum_k z_{jk}\right), \end{aligned}$$

where $\mathcal{IW}$ denotes the inverse Wishart distribution.

Finally, we briefly present the results used for the Gibbs sampling algorithm (44) of model (43). Some of the results are the same or similar to the ones presented before, therefore, we will not discuss them in detail.

The full conditional distributions of σ^2 and ϵ_j are the same as for model (24), and they were explained before. The full conditional distribution of z_{jk} can be derived in the same way as it was done for model (24), but in the case of model (43) the marginal distribution m becomes a normal distribution with pdf

$$m(d_{jk}|\sigma^2, \tau^2, \lambda_{jk}) = \frac{1}{\sqrt{2\pi(\sigma^2 + \tau^2/\lambda_{jk})}} e^{-d_{jk}^2/2(\sigma^2+\tau^2/\lambda_{jk})},$$

since we represented the Student's t prior as scale mixture of normals. This is a standard result for a marginal distribution arising from a model involving a $\mathcal{N}(\theta_{jk}, \sigma^2)$ likelihood and a $\mathcal{N}(0, \tau^2/\lambda_{jk})$ prior on θ_{jk}. See, for example Carlin and Louis (2000).

The full conditional distribution of θ_{jk} can be found by similar considerations as before. In case $z_{jk} = 1$, it becomes

$$\begin{aligned} f(\theta_{jk}|d_{jk}, \sigma^2, \tau^2, \lambda_{jk}) &= \frac{1}{\sqrt{2\pi\tilde{\sigma}_{jk}^2}} e^{-(\theta_{jk}-\tilde{\mu}_{jk})^2/2\tilde{\sigma}_{jk}^2} \\ \text{where} \quad \tilde{\mu}_{jk} &= \frac{\tau^2/\lambda_{jk}}{\sigma^2 + \tau^2/\lambda_{jk}} d_{jk}, \\ \text{and} \quad \tilde{\sigma}_{jk}^2 &= \frac{\tau^2/\lambda_{jk}}{\sigma^2 + \tau^2/\lambda_{jk}} \sigma^2, \end{aligned}$$

which is also a standard result for a posterior distribution arising from a $\mathcal{N}(\theta_{jk}, \sigma^2)$ likelihood and a $\mathcal{N}(0, \tau^2/\lambda_{jk})$ prior on θ_{jk}. See, for example Carlin and Louis (2000).

The full conditional distribution of τ is

$$\begin{aligned} p(\tau|\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\lambda}) &\propto \left[\prod_{j,k}\left\{(1-z_{jk})\delta_0 + z_{jk}\frac{1}{\sqrt{2\pi\tau^2/\lambda_{jk}}}\exp\left\{-\frac{1}{2\tau^2/\lambda_{jk}}\theta_{jk}^2\right\}\right\}\right]\cdot \\ &\quad \left(\tau^2\right)^{-a_2-1} e^{-\frac{1}{\tau^2}\frac{1}{b_2}} \\ &\propto \tau^{a_2+\sum_{j,k} z_{jk}/2-1}\exp\left\{-\tau\left(1/2\sum_{j,k}\left(z_{jk}\theta_{jk}^2\lambda_{jk}\right)+1/b_2\right)\right\} \\ &= \mathcal{IG}\left(a_2+\sum_{j,k} z_{jk}/2, \left[1/b_2+1/2\sum_{j,k}\left(z_{jk}\theta_{jk}^2\lambda_{jk}\right)\right]^{-1}\right). \end{aligned}$$

The full conditional distribution of λ_{jk} is proportional to

$$\begin{aligned} p(\lambda_{jk}|\theta_{jk}, z_{jk}, \tau, v) &\propto \left[(1-z_{jk})\delta_0 + z_{jk}\frac{1}{\sqrt{2\pi\tau^2/\lambda_{jk}}}\exp\left\{-\frac{1}{2\tau^2/\lambda_{jk}}\theta_{jk}^2\right\}\right]\cdot \\ &\quad \lambda_{jk}^{v/2-1}\exp\left\{-\frac{\lambda_{jk}}{2/v}\right\}. \end{aligned}$$

In case $z_{jk}=0$, this becomes

$$\begin{aligned} p(\lambda_{jk}|\theta_{jk}, z_{jk}=0, \tau, v) &\propto \lambda_{jk}^{v/2-1}\exp\left\{-\frac{\lambda_{jk}}{2/v}\right\} \\ &= \mathcal{G}a(v/2, 2/v), \end{aligned}$$

and when $z_{jk}=1$, it becomes

$$\begin{aligned} p(\lambda_{jk}|\theta_{jk}, z_{jk}=1, \tau, v) &\propto \lambda_{jk}^{1/2}\exp\left\{-\frac{1}{2\tau^2/\lambda_{jk}}\theta_{jk}^2\right\}\lambda_{jk}^{v/2-1}\exp\left\{-\frac{\lambda_{jk}}{2/v}\right\} \\ &= \lambda_{jk}^{(v+1)/2-1}\exp\left\{-\lambda_{jk}\left(v+\theta_{jk}^2/\tau^2\right)/2\right) \\ &= \mathcal{G}a\left((v+1)/2, \left[\left(v+\theta_{jk}^2/\tau^2\right)/2\right]^{-1}\right). \end{aligned}$$

To update parameter v we use a Metropolis step in the Gibbs sampling algorithm, because the full conditional distribution of v is not available in an explicit distributional form:

$$p(v|\boldsymbol{\lambda}) \propto \prod_{j,k}\left[\lambda_{jk}^{v/2-1}\exp\left\{-\frac{\lambda_{jk}}{2/v}\right\}\right]\exp\{-(v-1)\}I(v\geq 1).$$

In general, the Metropolis algorithm to update parameter v can be described as follows (Robert, 1994). Given density $p(v)$ known up to a normalizing factor and a conditional density (proposal distribution) $q(v|v')$, the algorithm updates $v^{(i-1)}$ to $v^{(i)}$ by

- Generate $\xi \sim q\left(\xi|v^{(i-1)}\right)$
- Define $r = \frac{p(\xi)q\left(v^{(i-1)}|\xi\right)}{p\left(v^{(i-1)}\right)q\left(\xi|v^{(i-1)}\right)}$
- Take $v^{(i)} = \begin{cases} \xi, & \text{with probability} \quad \min(1,r) \\ v^{(i-1)}, & \text{otherwise} \end{cases}$.

Choosing the proposal distribution q as a left-truncated normal with truncation point 1 and scale parameter ψ, the algorithm in Step 7 of (44) easily follows.

A.2 Derivations for Chapter 4

In this part of the Appendix we present derivation of the marginal distributions (63), shrinkage rules (64), (68), (70) and show that $\lim_{x\to\infty} p_j = 0$. At the end we briefly explain how to extend the results to the image denoising case and get marginal distribution (73) and shrinkage rule (74).

Marginal distribution For the derivation of the marginal distribution (63) and the posterior mean (64) we use Lemma A from Fourdrinier et al. (2000), which states that

$$\int_0^\infty \lambda^a \frac{1}{2}\left(\frac{x}{\lambda}\right)^{(p-2)/4} e^{-(x+\lambda)/2} I_{(p-2)/2}(\sqrt{\lambda x}) e^{-b\lambda}\lambda^{-c} d\lambda = e^{-x/2}x^{(p-2)/2}\frac{\Gamma(1+a-c)\,{}_1F_1\left(1+a-c; p/2; \frac{x}{2(2b+1)}\right)}{(1+2b)^{1+a-c}2^{p/2-1-a+c}\Gamma(p/2)}. \tag{98}$$

Substituting $p = 3$, $c = 0$ and $a = 0$ into result (98) and multiplying with b we get

$$\begin{aligned}\int_0^\infty f(x|\lambda)be^{-b\lambda}d\lambda &= \int_0^\infty \frac{1}{2}e^{-(x+\lambda)/2}\left(\frac{x}{\lambda}\right)^{1/4} I_{1/2}(\sqrt{\lambda x})be^{-b\lambda}d\lambda \\ &= \sqrt{\frac{2}{\pi}}\frac{b}{1+2b}\sqrt{x}e^{-x/2}{}_1F_1\left(1;3/2;\frac{x}{2(2b+1)}\right) \\ &= \frac{b}{\sqrt{1+2b}}e^{-\frac{bx}{1+2b}}\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]. \end{aligned} \tag{99}$$

The last equality follows from the identity (`http://functions.wolfram.com/07.20.03.0051.01`)

$${}_1F_1\left(1;3/2;z\right) = \frac{e^z\sqrt{\pi}}{2\sqrt{z}}\mathrm{Erf}\left[\sqrt{z}\right].$$

Using result (99) the marginal distribution becomes

$$\begin{aligned} m(x) &= \int_0^\infty \frac{1}{2}e^{-(x+\lambda)/2}\left(\frac{x}{\lambda}\right)^{1/4} I_{1/2}(\sqrt{\lambda x})\left\{\epsilon_j\delta_0(\lambda) + (1-\epsilon_j)be^{-b\lambda}\right\}d\lambda \\ &= \epsilon_j\sqrt{\frac{2}{\pi}}x^{1/2}e^{-x/2} + (1-\epsilon_j)\frac{b}{\sqrt{1+2b}}e^{-\frac{bx}{1+2b}}\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] \\ &= \epsilon_j m_0(x) + (1-\epsilon_j)m_1(x), \end{aligned}$$

where $m_0(x)$ is the pdf of the central χ^2_p distribution with $p = 3$. This is true because the noncentral $\chi^2_p(\lambda)$ distribution with $\lambda = 0$ reduces to a central χ^2_p distribution, which can be easily seen from the infinite sum representation of the noncentral chi-square pdf:

$$\chi^2_p(x;\lambda) = \frac{1}{2}\left(\frac{x}{\lambda}\right)^{(p-2)/4} e^{-(x+\lambda)/2}I_{(p-2)/2}(\sqrt{\lambda x}) = \sum_{i=0}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^i}{i!}\chi^2_{p+2i}(x).$$

Posterior mean To derive Bayes shrinkage rule (64), first substitute $p = 3$, $c = 0$ and $a = 1$ into (98) and then multiply with b. We get that

$$\begin{aligned}
&\int_0^\infty \lambda f(x|\lambda) b e^{-b\lambda} d\lambda = \\
&\int_0^\infty \lambda \frac{1}{2} e^{-(x+\lambda)/2} \left(\frac{x}{\lambda}\right)^{1/4} I_{1/2}(\sqrt{\lambda x}) b e^{-b\lambda} d\lambda = \\
&\frac{2^{3/2}}{\sqrt{\pi}} \frac{b}{(1+2b)^2} \sqrt{x} e^{-x/2} {}_1F_1\left(2; 3/2; \frac{x}{2(2b+1)}\right) = \\
&\frac{b e^{-x/2} \left(2\sqrt{x+2bx} + e^{\frac{x}{2+4b}} \sqrt{2\pi}(1+2b+x) \mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]\right)}{(1+2b)^{5/2}\sqrt{2\pi}},
\end{aligned} \tag{100}$$

where last equality follows from the identity (`http://functions.wolfram.com/07.20.03.0069.01`)

$${}_1F_1\left(2; 3/2; z\right) = \frac{e^z \sqrt{\pi}(2z+1)\mathrm{Erf}\left[\sqrt{z}\right] + 2\sqrt{z}}{4\sqrt{z}}.$$

For the model in (62) the posterior mean can be derived as

$$\begin{aligned}
\delta(x) &= \frac{\int_0^\infty \lambda f(x|\lambda) \left\{\epsilon_j \delta_0 + (1-\epsilon_j) b e^{-b\lambda}\right\} d\lambda}{\int_0^\infty f(x|\lambda) \left\{\epsilon_j \delta_0 + (1-\epsilon_j) b e^{-b\lambda}\right\} d\lambda} \\
&= \frac{(1-\epsilon_j) \int_0^\infty \lambda f(x|\lambda) b e^{-b\lambda} d\lambda}{\epsilon_j m_0(x) + (1-\epsilon_j) m_1(x)} \\
&= \left(1 - \frac{\epsilon_j m_0(x)}{\epsilon_j m_0(x) + (1-\epsilon_j) m_1(x)}\right) \frac{\int_0^\infty \lambda f(x|\lambda) b e^{-b\lambda} d\lambda}{m_1(x)} \\
&= (1-p_j)\, \delta_E(x).
\end{aligned}$$

Here $\delta_E(x)$ is the posterior mean induced by an exponential prior on a noncentral chi-square likelihood, which, by (99) and (100), can be expressed as

$$\begin{aligned}
\delta_E(x) &= \frac{\int_0^\infty \lambda f(x|\lambda) b e^{-b\lambda} d\lambda}{\int_0^\infty f(x|\lambda) b e^{-b\lambda} d\lambda} \\
&= \frac{b e^{-x/2} \left(2\sqrt{x+2bx} + e^{\frac{x}{2+4b}} \sqrt{2\pi}(1+2b+x) \mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]\right)}{\frac{b}{\sqrt{1+2b}} e^{-\frac{bx}{1+2b}} \mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] (1+2b)^{5/2}\sqrt{2\pi}} \\
&= \frac{1+2b+x+\sqrt{\frac{2x(1+2b)}{\pi}} e^{-\frac{x}{2+4b}} \Big/ \mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}{(1+2b)^2}.
\end{aligned}$$

Therefore, shrinkage rule (64) becomes

$$\delta(x) = (1-p_j)\frac{1+2b+x+\sqrt{\frac{2x(1+2b)}{\pi}}e^{-\frac{x}{2+4b}}\Big/\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}{(1+2b)^2},$$

where

$$p_j = \epsilon_j \frac{m_0(x)}{m(x)}.$$

Posterior median Since the posterior distribution is absolutely continuous, the median is the solution u to equation

$$\varphi(u) = \int_0^u f(\lambda|x)d\lambda = \int_0^u \frac{f(x|\lambda)\left\{\epsilon_j\delta_0(\lambda) + (1-\epsilon_j)be^{-b\lambda}\right\}}{m(x)}d\lambda = \frac{1}{2}.$$

Using identities (`http://functions.wolfram.com/03.02.03.0004.01`) and (`http://functions.wolfram.com/01.03.21.0266.01`) it follows that

$$I_{1/2}(z) = \frac{e^z - e^{-z}}{\sqrt{2\pi z}},$$

and

$$\int e^{dz+cz^2}dz = \frac{\sqrt{\pi}e^{-\frac{d^2}{4c}}\mathrm{Erfi}\left[\frac{d+2cz}{2\sqrt{c}}\right]}{2\sqrt{c}}.$$

By a change of variable $y = \sqrt{\lambda}$ and the fact that $\mathrm{Erfi}[z] = -i\mathrm{Erf}[iz] = i\mathrm{Erf}[-iz]$, we get

$$\begin{aligned}
&\int_0^u e^{-\lambda/2}\left(\frac{x}{\lambda}\right)^{1/4} I_{1/2}(\sqrt{\lambda x})e^{-b\lambda}d\lambda = \\
&\frac{1}{\sqrt{2\pi}}\left\{\int_0^u \lambda^{-1/2}e^{-\lambda(b+1/2)+\sqrt{\lambda}\sqrt{x}}d\lambda - \int_0^u \lambda^{-1/2}e^{-\lambda(b+1/2)-\sqrt{\lambda}\sqrt{x}}d\lambda\right\} = \\
&\frac{\sqrt{2}}{\sqrt{\pi}}\left\{\int_0^{\sqrt{u}} e^{-y^2(b+1/2)+y\sqrt{x}}dy - \int_0^{\sqrt{u}} e^{-y^2(b+1/2)-y\sqrt{x}}dy\right\} = \\
&\frac{1}{\sqrt{b+1/2}}e^{\frac{x}{2+4b}}\left\{\mathrm{Erf}\left[\frac{-\sqrt{x}-(1+2b)\sqrt{u}}{\sqrt{2+4b}}\right] - \mathrm{Erf}\left[\frac{\sqrt{x}-(1+2b)\sqrt{u}}{\sqrt{2+4b}}\right] + \right. \\
&\left. 2\mathrm{Erf}\left[\frac{\sqrt{x}}{\sqrt{2+4b}}\right]\right\}.
\end{aligned}$$

Using the above result, the integral $\varphi(u)$ becomes

$$\varphi(u) =$$
$$\frac{1}{m(x)} \int_0^u \frac{1}{2} e^{-(x+\lambda)/2} \left(\frac{x}{\lambda}\right)^{1/4} I_{1/2}(\sqrt{\lambda x}) \left\{\epsilon_j \delta_0(\lambda) + (1-\epsilon_j) b e^{-b\lambda}\right\} d\lambda =$$
$$\frac{1}{m(x)} \left(\epsilon_j \sqrt{\frac{2}{\pi}} x^{1/2} e^{-x/2} + (1-\epsilon_j) \frac{b}{\sqrt{1+2b}} e^{-\frac{bx}{1+2b}} \left\{ \text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] + \right.\right.$$
$$\left.\left. \frac{1}{2}\text{Erf}\left[\frac{(1+2b)\sqrt{u} - \sqrt{x}}{\sqrt{2+4b}}\right] - \frac{1}{2}\text{Erf}\left[\frac{(1+2b)\sqrt{u} + \sqrt{x}}{\sqrt{2+4b}}\right] \right\} \right).$$

Because $\varphi(0) = p_j$, the algorithm to find the posterior median becomes

$$\delta_M(x) = u \; \mathbf{1}\left(p_j < \frac{1}{2}\right),$$

where u is the solution of the equation

$$1 - (1-\epsilon_j) \frac{1}{m(x)} \frac{b}{\sqrt{1+2b}} e^{-\frac{bx}{1+2b}} \left(\text{Erf}\left[\frac{(1+2b)\sqrt{u} + \sqrt{x}}{\sqrt{2+4b}}\right] - \right.$$
$$\left. \text{Erf}\left[\frac{(1+2b)\sqrt{u} - \sqrt{x}}{\sqrt{2+4b}}\right] \right) = 0,$$

in case $p_j < \frac{1}{2}$.

Bayes factor Testing the hypothesis $H_0 : \lambda = 0$, versus $H_1 : \lambda \neq 0$ in the Bayesian framework is possible with the Bayes factor procedure, which results in a thresholding rule. In general, the Bayes factor procedure with a prior that has a point mass component (Vidakovic, 1998a) is

$$\hat{\lambda} = x \; \mathbf{1}\left(P(H_0|x) < \frac{1}{2}\right),$$

where

$$P(H_0|x) = \left(1 + \frac{1-\epsilon_j}{\epsilon_j} \frac{1}{B}\right)^{-1}$$

is the posterior probability of the H_0 hypothesis and

$$B = \frac{m_0(x)}{m_1(x)}.$$

Therefore, the Bayes factor procedure becomes

$$\delta_{BF}(x) = x \ \mathbf{1}\left(p_j < \frac{1}{2}\right),$$

where

$$p_j = \epsilon_j \frac{m_0(x)}{m(x)}.$$

Limit of p_j Lastly, we show that $\lim_{x\to\infty} p_j = 0$. Assume $0 < \epsilon_j < 1$. Since

$$p_j = \frac{\epsilon_j m_0(x)}{\epsilon_j m_0(x) + (1-\epsilon_j) m_1(x)},$$

it follows that

$$\frac{1}{p_j} = 1 + \frac{1-\epsilon_j}{\epsilon_j} \frac{m_1(x)}{m_0(x)}.$$

Now, for $b > 0$

$$\begin{aligned} \frac{m_1(x)}{m_0(x)} &= \frac{b e^{-\frac{bx}{1+2b}} \text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}{\sqrt{1+2b}\sqrt{\frac{2}{\pi}} x^{1/2} e^{-x/2}} \\ &= K \ \text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] x^{-1/2} e^{\frac{x}{2(1+2b)}}, \end{aligned}$$

where K is a constant. Since Erf is a bounded function,

$$\lim_{x\to\infty} \frac{m_1(x)}{m_0(x)} = \infty,$$

therefore

$$\lim_{x\to\infty} p_j = 0.$$

Extension to image denoising To derive marginal distribution (73) substitute $p = 5$, $c = 0$ and $a = 0$ into result (98) and multiply with b. We get that

$$\begin{aligned} \int_0^\infty f(x|\lambda) b e^{-b\lambda} d\lambda &= \int_0^\infty \frac{1}{2} e^{-(x+\lambda)/2} \left(\frac{x}{\lambda}\right)^{3/4} I_{3/2}(\sqrt{\lambda x}) b e^{-b\lambda} d\lambda \\ &= \left\{ b\sqrt{1+2b} e^{-\frac{bx}{1+2b}} \text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] - b\sqrt{2x/\pi} e^{-x/2} \right\}. \end{aligned} \tag{101}$$

Above we used the identity (http://functions.wolfram.com/07.20.03.0053.01)

$$ {}_1F_1\left(1;5/2;z\right) = \frac{3e^z\sqrt{\pi}\mathrm{Erf}\left[\sqrt{z}\right]}{4z^{3/2}} - \frac{3}{2z}. $$

Using result (101) the marginal distribution becomes

$$ \begin{aligned} m(x) &= \int_0^\infty \frac{1}{2}e^{-(x+\lambda)/2}\left(\frac{x}{\lambda}\right)^{3/4} I_{3/2}(\sqrt{\lambda x})\left\{\epsilon_j\delta_0(\lambda) + (1-\epsilon_j)be^{-b\lambda}\right\}d\lambda \\ &= \epsilon_j\frac{1}{\Gamma(5/2)2^{5/2}}x^{3/2}e^{-x/2} + \\ &\quad (1-\epsilon_j)\left\{b\sqrt{1+2b}e^{-\frac{bx}{1+2b}}\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right] - b\sqrt{2x/\pi}e^{-x/2}\right\} \\ &= \epsilon_j m_0(x) + (1-\epsilon_j)m_1(x), \end{aligned} $$

where $m_0(x)$ is the pdf of the central χ^2_p distribution with $p = 5$, and $m_1(x)$ is result (101).

To derive the posterior mean in (74), substitute $p = 5$, $c = 0$ and $a = 1$ into (98) and then multiply with b. We get that

$$ \begin{aligned} &\int_0^\infty \lambda f(x|\lambda)be^{-b\lambda}d\lambda = \\ &\int_0^\infty \lambda\frac{1}{2}e^{-(x+\lambda)/2}\left(\frac{x}{\lambda}\right)^{3/4} I_{3/2}(\sqrt{\lambda x})be^{-b\lambda}d\lambda = \\ &\frac{b}{(1+2b)^{3/2}}\left(e^{-x/2}\frac{\sqrt{2x}}{\sqrt{\pi}}\sqrt{1+2b} + e^{-\frac{bx}{1+2b}}(x-1-2b)\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]\right), \end{aligned} \tag{102} $$

where we used the identity (http://functions.wolfram.com/07.20.03.0070.01)

$$ {}_1F_1\left(2;5/2;z\right) = \frac{3e^z\sqrt{\pi}(2z-1)\mathrm{Erf}\left[\sqrt{z}\right] + 6\sqrt{z}}{8z^{3/2}}. $$

Similarly as before, using results (101) and (102) we get that

$$ \delta_E(x) = \frac{\sqrt{\frac{2x(1+2b)}{\pi}}e^{-\frac{x}{2+4b}} + (x-1-2b)\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}{-(1+2b)^{3/2}\sqrt{2x/\pi}e^{-\frac{x}{2+4b}} + (1+2b)^2\mathrm{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}, $$

from which the posterior mean in (74) simply follows as before:

$$\delta(x) = (1-p_j)\frac{\sqrt{\frac{2x(1+2b)}{\pi}}e^{-\frac{x}{2+4b}} + (x-1-2b)\text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]}{-(1+2b)^{3/2}\sqrt{2x/\pi}e^{-\frac{x}{2+4b}} + (1+2b)^2\text{Erf}\left[\sqrt{\frac{x}{2+4b}}\right]},$$

where

$$p_j = \epsilon_j \frac{m_0(x)}{m(x)}.$$

A.3 Derivations for Chapter 5

Some of the following results are equivalent to the results of Section A.1 of the Appendix, since the model in Chapter 5 builds on the model of Chapter 2. However, for completeness, we present all the results here. First we provide derivation of results (84) and (85). The joint distribution $f(d^\star, \theta|\sigma^2)$ using prior $p_1(\theta|\tau)$ is

$$\begin{aligned}
f(d^\star, \theta|\sigma^2, \tau) &= f(d^\star|\theta, \sigma^2)p_1(\theta|\tau) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(d^\star-\theta)^2}{2\sigma^2}}\frac{\tau}{2}e^{-\tau|\theta|} \\
&= \frac{\tau}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\{\theta^2-2\theta(-\text{sign}(\theta)\sigma^2\tau+d^\star)+d^{\star 2}\}} \\
&= \frac{\tau}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\{\theta-(d^\star-\text{sign}(\theta)\sigma^2\tau)^2\}} - e^{-\frac{1}{2\sigma^2}\{-(d^\star-\text{sign}(\theta)\sigma^2\tau)^2+d^{\star 2}\}} \\
&= \frac{\tau e^{\frac{\sigma^2\tau^2}{2}}e^{-\text{sign}(\theta)d^\star\tau}}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\{\theta-(d^\star-\text{sign}(\theta)\sigma^2\tau)\}^2} \\
&= \begin{cases} \dfrac{\tau e^{\frac{\sigma^2\tau^2}{2}}e^{-d^\star\tau}}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d^\star-\sigma^2\tau)]^2}, & \theta \geq 0 \\ \\ \dfrac{\tau e^{\frac{\sigma^2\tau^2}{2}}e^{d^\star\tau}}{2\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d^\star+\sigma^2\tau)]^2}, & \theta < 0 \end{cases}.
\end{aligned}$$

The marginal distribution becomes

$$\begin{aligned}
m(d^\star|\sigma^2, \tau) &= \int_{-\infty}^{\infty} f(d^\star, \theta|\sigma^2, \tau)d\theta \\
&= \frac{\tau}{2}e^{\frac{\sigma^2\tau^2}{2}}\left\{e^{d^\star\tau}\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d^\star+\sigma^2\tau)]^2}d\theta + \right. \\
&\quad \left. e^{-d^\star\tau}\int_{0}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d^\star-\sigma^2\tau)]^2}d\theta\right\} \\
&= \frac{\tau}{2}e^{\frac{\sigma^2\tau^2}{2}}\left\{e^{d^\star\tau}\Phi\left(\frac{-d^\star-\sigma^2\tau}{\sigma}\right) + e^{-d^\star\tau}\Phi\left(\frac{d^\star-\sigma^2\tau}{\sigma}\right)\right\}.
\end{aligned}$$

Combining the two equations above, we get the posterior as

$$h(\theta|d^\star,\sigma^2,\tau) = \frac{f(d^\star,\theta|\sigma^2,\tau)}{m(d^\star|\sigma^2,\tau)} =$$

$$\begin{cases} \dfrac{e^{-d^\star\tau}}{e^{-d^\star\tau}\Phi\left(\frac{d^\star}{\sigma}-\tau\sigma\right)+e^{d^\star\tau}\Phi\left(-\frac{d^\star}{\sigma}-\tau\sigma\right)}\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d^\star-\sigma^2\tau)]^2}, & \theta\geq 0 \\ \\ \dfrac{e^{d^\star\tau}}{e^{-d^\star\tau}\Phi\left(\frac{d^\star}{\sigma}-\tau\sigma\right)+e^{d^\star\tau}\Phi\left(-\frac{d^\star}{\sigma}-\tau\sigma\right)}\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[\theta-(d^\star+\sigma^2\tau)]^2}, & \theta<0 \end{cases}.$$

These results were also derived by Pericchi and Smith (1992) and used by Johnstone and Silverman (2005b).

Now we derive the results used for the Gibbs sampling algorithm of model (83). To derive the full conditional distribution for a parameter of interest we look at the joint distribution of all the parameters and collect the terms which contain the desired parameter. Let us denote $\boldsymbol{d} = \{d_{jk} : j = J_0,\ldots,\log_2(n)-1,\ k = 0,\ldots,2^j-1\}$, $\boldsymbol{\beta} = \{\beta_i : i = 1,\ldots,p\}$, $\boldsymbol{\theta} = \{\theta_{jk} : j = J_0,\ldots,\log_2(n)-1,\ k = 0,\ldots,2^j-1\}$, $\boldsymbol{\gamma} = \{\gamma_i : i = 1,\ldots,p\}$, $\boldsymbol{z} = \{z_{jk} : j = J_0,\ldots,\log_2(n)-1,\ k = 0,\ldots,2^j-1\}$, $\boldsymbol{\epsilon} = \{\epsilon_j : j = J_0,\ldots,\log_2(n)-1\}$ and $\boldsymbol{v} = \{v_i : i = 1,\ldots,p\}$. The joint distribution of the data and parameters for model in (83) becomes

$$f(\boldsymbol{d},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{z},q,\boldsymbol{\epsilon},\boldsymbol{v},\sigma^2,\tau_\theta,\eta^2) = \left[\prod_{j,k}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(d_{jk}-(\boldsymbol{U}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})_{jk}-\theta_{jk})^2}\right]\cdot$$
$$\frac{1}{\Gamma(a_1){b_1}^{a_1}}(\sigma^2)^{-a_1-1}e^{-\frac{1}{\sigma^2}\frac{1}{b_1}}\left[\prod_i\left\{(1-\gamma_i)\delta_0+\gamma_i\frac{1}{\sqrt{2\pi v_i\eta^2}}e^{-\frac{1}{2v_i\eta^2}\beta_i^2}\right\}\right]\cdot$$
$$\left[\prod_{j,k}\left\{(1-z_{jk})\delta_0+z_{jk}\frac{\tau_\theta}{2}e^{-\tau_\theta|\theta_{jk}|}\right\}\right]\left[\prod_i q^{\gamma_i}(1-q)^{(1-\gamma_i)}\right]\cdot$$
$$\left[\prod_{j,k}\epsilon_j^{z_{jk}}(1-\epsilon_j)^{(1-z_{jk})}\right]\mathbf{1}\{0\leq\ q\leq 1\}\left[\prod_j\mathbf{1}\{0\leq\epsilon_j\leq 1\}\right]\left[\prod_i e^{-v_i}\right]\cdot$$
$$\frac{1}{\Gamma(a_2){b_2}^{a_2}}(\eta^2)^{-a_2-1}e^{-\frac{1}{\eta^2}\frac{1}{b_2}}\frac{1}{\Gamma(a_3){b_3}^{a_3}}\tau_\theta^{a_3-1}e^{-\tau_\theta/b_3}.$$

The full conditional distribution of parameters β_i and γ_i simply follows from Yuan and Lin (2004) with using $\boldsymbol{Z} = \boldsymbol{d} - \boldsymbol{U}_{\boldsymbol{\gamma}^{[-i]},\gamma_i=0}\boldsymbol{\beta}_{\boldsymbol{\gamma}^{[-i]},\gamma_i=0} - \boldsymbol{\theta}$.

The full conditional distribution of v_i is

$$\begin{aligned} p(v_i|\beta_i,\gamma_i,\eta^2) &\propto \left\{(1-\gamma_i)\delta_0 + \gamma_i \frac{1}{\sqrt{2\pi v_i\eta^2}} e^{-\frac{1}{2v_i\eta^2}\beta_i^2}\right\} e^{-v_i} \\ &= \begin{cases} \mathcal{E}xp(1), & \text{if} \quad \gamma_i = 0 \\ \mathcal{GIG}\left(2, \beta_i^2/\eta^2, 1/2\right), & \text{if} \quad \gamma_i = 1 \end{cases}, \end{aligned}$$

where $\mathcal{GIG}(a,b,p)$ denotes the generalized inverse Gaussian distribution (Johnson et al., 1994, p.284) with probability density function

$$f(x|a,b,p) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} e^{-(ax+b/x)/2}, \quad x>0; a,b>0,$$

where K_p denotes the modified Bessel function of the third kind.

The full conditional distribution of η^2 is

$$\begin{aligned} p(\eta^2|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{v}) &\propto \left[\prod_i \left\{(1-\gamma_i)\delta_0 + \gamma_i \frac{1}{\sqrt{2\pi v_i\eta^2}} e^{-\frac{1}{2v_i\eta^2}\beta_i^2}\right\}\right] \cdot \\ &\quad \frac{1}{\Gamma(a_2){b_2}^{a_2}} (\eta^2)^{-a_2-1} e^{-\frac{1}{\eta^2}\frac{1}{b_2}} \\ &\propto (\eta^2)^{-a_2-1/2\sum_i \gamma_i - 1} \exp\left\{-\frac{1}{\eta^2}\left(1/b_2 + 1/2\sum_i \left(\gamma_i\beta_i^2/v_i\right)\right)\right\} \\ &= \mathcal{IG}\left(a_2 + 1/2\sum_i \gamma_i, \left[1/b_2 + 1/2\sum_i \left(\gamma_i\beta_i^2/v_i\right)\right]^{-1}\right). \end{aligned}$$

The full conditional distribution of q can be derived as

$$\begin{aligned} p(q|\boldsymbol{\gamma}) &= \left[\prod_i q^{\gamma_i}(1-q)^{(1-\gamma_i)}\right] \mathbf{1}\{0 \le q \le 1\} \\ &\propto q^{\sum_i \gamma_i}(1-q)^{p-\sum_i \gamma_i} \mathbf{1}\{0 \le q \le 1\} \\ &= \mathcal{B}e\left(1 + \sum_i \gamma_i, 1 + \sum_i (1-\gamma_i)\right). \end{aligned}$$

The full conditional distribution of σ^2 is

$$p(\sigma^2|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{d}) \propto (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{j,k}(d_{jk}-(\boldsymbol{U_\gamma\beta_\gamma})_{jk}-\theta_{jk})^2} (\sigma^2)^{-a_1-1} e^{-\frac{1}{\sigma^2}\frac{1}{b_1}} =$$
$$(\sigma^2)^{-a_1-n/2-1} \exp\left\{-\frac{1}{\sigma^2}\left(1/b_1 + 1/2\sum_{j,k}\left(d_{jk}-(\boldsymbol{U_\gamma\beta_\gamma})_{jk}-\theta_{jk}\right)^2\right)\right\} =$$
$$\mathcal{IG}\left(a_1+n/2, \left[1/b_1 + 1/2\sum_{j,k}\left(d_{jk}-(\boldsymbol{U_\gamma\beta_\gamma})_{jk}-\theta_{jk}\right)^2\right]^{-1}\right).$$

In the following, we denote $d^\star_{jk} = d_{jk} - (\boldsymbol{U_\gamma\beta_\gamma})_{jk}$. The conditional distribution of z_{jk} remains Bernoulli with posterior probability derived by

$$P(z_{jk}=1|d^\star_{jk},\sigma^2,\tau,\epsilon_j) = \frac{P(z_{jk}=1|\epsilon_j) f(d^\star_{jk}|\sigma^2,\tau,z_{jk}=1)}{\sum_{i\in\{0,1\}} P(z_{jk}=i|\epsilon_j) f(d^\star_{jk}|\sigma^2,\tau,z_{jk}=i)} =$$
$$\frac{P(z_{jk}=1|\epsilon_j)\int_{-\infty}^{\infty} f(d^\star_{jk}|\theta_{jk},\sigma^2)p(\theta_{jk}|\tau,z_{jk}=1)d\theta_{jk}}{\sum_{i\in\{0,1\}} P(z_{jk}=i|\epsilon_j)\int_{-\infty}^{\infty} f(d^\star_{jk}|\theta_{jk},\sigma^2)p(\theta_{jk}|\tau,z_{jk}=i)d\theta_{jk}} =$$
$$\frac{\epsilon_j m\left(d^\star_{jk}|\sigma^2,\tau\right)}{(1-\epsilon_j)\, f\left(d^\star_{jk}|0,\sigma^2\right) + \epsilon_j m\left(d^\star_{jk}|\sigma^2,\tau\right)}.$$

Here $p(\theta_{jk}|\tau,z_{jk}=i),\ i\in\{0,1\}$ denote the two parts of the mixture prior in model (83), depending on the value of latent variable z_{jk}. Similar result for the full conditional of γ_i was used by Yuan and Lin (2005).

The full conditional distribution of ϵ_j is

$$\begin{aligned} p(\epsilon_j|\boldsymbol{z}) &\propto \left[\prod_k \epsilon_j^{z_{jk}}(1-\epsilon_j)^{(1-z_{jk})}\right]\mathbf{1}\{0\le\epsilon_j\le 1\} \\ &= \epsilon_j^{\sum_k z_{jk}}(1-\epsilon_j)^{\sum_k(1-z_{jk})} \\ &= \mathcal{B}e\left(1+\sum_k z_{jk}, 1+\sum_k(1-z_{jk})\right). \end{aligned}$$

Similarly, the full conditional distribution of θ_{jk} is

$$\begin{aligned} p(\theta_{jk}|d_{jk}, \boldsymbol{\beta}, \boldsymbol{\gamma}, z_{jk}, \sigma^2, \tau_\theta) &\propto \exp\left\{-\frac{1}{2\sigma^2}(d_{jk} - (\boldsymbol{U}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})_{jk} - \theta_{jk})^2\right\} \cdot \\ &\quad \left\{(1 - z_{jk})\delta_0 + z_{jk}\frac{\tau_\theta}{2}e^{-\tau_\theta|\theta_{jk}|}\right\} \\ &= \begin{cases} \delta_0(\theta_{jk}), & \text{if} \quad z_{jk} = 0 \\ h(\theta_{jk}|d^\star_{jk}, \sigma^2, \tau_\theta), & \text{if} \quad z_{jk} = 1 \end{cases}, \end{aligned}$$

where the distribution $h(\theta_{jk}|d^\star_{jk}, \sigma^2, \tau_\theta)$ comes from the result in (85) and was derived above.

Finally, the full conditional distribution of τ_θ is

$$\begin{aligned} p(\tau_\theta|\boldsymbol{\theta}, \boldsymbol{z}) &\propto \left[\prod_{j,k}\left\{(1 - z_{jk})\delta_0 + z_{jk}\frac{\tau_\theta}{2}\exp\left(-\tau_\theta|\theta_{jk}|\right)\right\}\right] \cdot \\ &\quad \frac{1}{\Gamma(a_3){b_3}^{a_3}}\tau_\theta^{a_3-1}\exp\left(-\tau/b_3\right) \\ &\propto \tau_\theta^{a_3+\sum_{j,k} z_{jk}-1}\exp\left\{-\tau_\theta\left(\sum_{j,k}\left(z_{jk}|\theta_{jk}|\right) + 1/b_3\right)\right\} \\ &= \mathcal{G}a\left(a_3 + \sum_{j,k} z_{jk}, \left[1/b_3 + \sum_{j,k}\left(z_{jk}|\theta_{jk}|\right)\right]^{-1}\right). \end{aligned}$$

Bibliography

F. Abramovich, U. Amato, C. Angelini, On optimality of Bayesian wavelet estimators, Scandinavian Journal of Statistics 31 (2004) 217–234.

F. Abramovich, C. Angelini, D. Canditiis, Pointwise optimality of Bayesian wavelet estimators, Annals of the Institute of Statistical Mathematics 59 (2007) 425–434.

F. Abramovich, T.C. Bailey, T. Sapatinas, Wavelet analysis and its statistical applications, The Statistician 49 (2000) 1–29.

F. Abramovich, Y. Benjamini, Thresholding of wavelet coefficients as multiple hypotheses testing procedure, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, volume 103 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1995, pp. 5–14.

F. Abramovich, Y. Benjamini, D.L. Donoho, I.M. Johnstone, Adapting to unknown sparsity by controlling the false discovery rate, The Annals of Statistics 34 (2006) 584–653.

F. Abramovich, P. Besbeas, T. Sapatinas, Empirical Bayes approach to block wavelet function estimation, Computational Statistics & Data Analysis 39 (2002) 435–451.

F. Abramovich, V. Grinshtein, A. Petsa, T. Sapatinas, On Bayesian testimation and its application to wavelet thresholding, Biometrika 97 (2010) 181–198.

F. Abramovich, T. Sapatinas, B.W. Silverman, Wavelet thresholding via a Bayesian approach, Journal of the Royal Statistical Society, Series B 60 (1998) 725–749.

M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, Dover, New York, Fifth edition, 1964.

A. Achim, E.E. Kuruoğlu, Image denoising using bivariate α-stable distributions in the complex wavelet domain, IEEE Signal Processing Letters 12 (2005) 17–20.

G.K. Ambler, B.W. Silverman, Perfect simulation for Bayesian wavelet thresholding with correlated coefficients, Technical Report 04:01, Department of Mathematics, University of Bristol, 2004.

D.F. Andrews, C.L. Mallows, Scale mixtures of normal distributions, Journal of the Royal Statistical Society, Series B 36 (1974) 99–102.

C. Angelini, T. Sapatinas, Empirical Bayes approach to wavelet regression using ϵ-contaminated priors, Journal of Statistical Computation and Simulation 74 (2004) 741–764.

C. Angelini, B. Vidakovic, Γ-minimax wavelet shrinkage: A robust incorporation of information about energy of a signal in denoising applications, Statistica Sinica 14 (2004) 103–125.

A. Antoniadis, Wavelet methods in statistics: Some recent developments and their applications, Statistics Surveys 1 (2007) 16–55.

A. Antoniadis, J. Bigot, T. Sapatinas, Wavelet estimators in nonparametric regression: A comparative simulation study, Journal of Statistical Software 6 (2001) 1–83.

F. Autin, On the performance of a new thresholding procedure using tree structure, Electronic Journal of Statistics 2 (2008) 412–431.

F. Autin, J.M. Freyermuth, R. von Sachs, Ideal denoising within a family of tree-structured wavelet estimators, Journal of Statistical Software 5 (2011) 829–855.

N. Balakrishnan, S. Kocherlakota, On the Double Weibull distribution: Order statistics and estimation, Sankhyā: The Indian Journal of Statistics, Series B 47 (1985) 161–178.

S. Barber, G.P. Nason, Real nonparametric regression using complex wavelets, Journal of the Royal Statistical Society, Series B 66 (2004) 927–939.

S. Barber, G.P. Nason, B.W. Silverman, Posterior probability intervals for wavelet thresholding, Journal of the Royal Statistical Society, Series B 64 (2002) 189–205.

J.O. Berger, A. Philippe, C.P. Robert, Estimation of quadratic functions: Noninformative priors for non-centrality parameters, Statistica Sinica 8 (1998) 359–375.

N. Bochkina, T. Sapatinas, On pointwise optimality of Bayes factor wavelet regression estimators, Sankhyā: The Indian Journal of Statistics, Series B 68 (2006) 513–541.

N. Bochkina, T. Sapatinas, Minimax rates of convergence and optimality of Bayes factor wavelet regression estimators under pointwise risks, Statistica Sinica 19 (2009) 1389–1406.

L. Boubchir, J.M. Fadili, A closed-form nonparametric Bayesian estimator in the wavelet domain of images using an approximate α-stable prior, Pattern Recognition Letters 27 (2006) 1370–1382.

C.J.F. ter Braak, Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising, Computational Statistics & Data Analysis 51 (2006) 1232–1242.

T.T. Cai, Adaptive wavelet estimation: a block thresholding and oracle inequality approach, Annals of Statistics 27 (1999) 898–924.

T.T. Cai, On block thresholding in wavelet regression: adaptivity, block size, and threshold level, Statistica Sinica 12 (2002) 1241–1273.

T.T. Cai, B.W. Silverman, Incorporating information on neighboring coefficients into wavelet estimation, Sankhyā: The Indian Journal of Statistics, Series B 63 (2001) 127–148.

T.T. Cai, H.H. Zhou, A data-driven block thresholding approach to wavelet estimation, Annals of Statistics 37 (2009) 569–595.

B. Carlin, T. Louis, Bayes and Empirical Bayes Methods for Data Analysis, Second Edition, Texts in statistical science, Taylor & Francis, 2000.

C.M. Carvalho, N.G. Polson, J.G. Scott, The horseshoe estimator for sparse signals, Biometrika 97 (2010) 465–480.

G. Casella, E.I. George, Explaining the Gibbs sampler, The American Statistician 46 (1992) 167–174.

S.G. Chang, B. Yu, M. Vetterli, Adaptive wavelet thresholding for image denoising and compression, IEEE Transactions on Image Processing 9 (2000) 1532–1546.

W. Chang, B. Vidakovic, Wavelet estimation of a base-line signal from repeated noisy measurements by vertical block shrinkage, Computational Statistics and Data Analysis 40 (2002) 317–328.

X.W. Chang, L. Qu, Wavelet estimation of partially linear models, Computational Statistics & Data Analysis 47 (2004) 31–48.

E. Chicken, Block thresholding and wavelet estimation for nonequispaced samples, Journal of Statistical Planning and Inference 116 (2003) 113–129.

E. Chicken, Block-dependent thresholding in wavelet regression, Journal of Nonparametric Statistics 17 (2005) 467–491.

E. Chicken, Nonparametric regression with sample design following a random process, Communications in Statistics - Theory and Methods 36 (2007) 1915–1934.

H.A. Chipman, E.D. Kolaczyk, R.E. McCulloch, Adaptive Bayesian wavelet shrinkage, Journal of The American Statistical Association 92 (1997) 1413–1421.

M. Clyde, G. Parmigiani, B. Vidakovic, Multiple shrinkage and subset selection in wavelets, Biometrika 85 (1998) 391–401.

M.A. Clyde, E.I. George, Robust empirical Bayes estimation in wavelets. ISDS discussion paper, Technical Report 98-21, Duke University, 1998.

M.A. Clyde, E.I. George, Empirical Bayes estimation in wavelet nonparametric regression, in: P. Müller, B. Vidakovic (Eds.), Bayesian Inference in Wavelet Based Models, volume 141 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1999, pp. 309–322.

M.A. Clyde, E.I. George, Flexible empirical Bayes estimation for wavelets, Journal of the Royal Statistical Society, Series B 62 (2000) 681–698.

R.R. Coifman, D.L. Donoho, Translation-invariant de-noising, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, volume 103 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1995, pp. 125–150.

M.S. Crouse, R.D. Nowak, R.G. Baraniuk, Wavelet-based statistical signal processing using hidden markov models, IEEE Transactions On Signal Processing 46 (1998) 886–902.

L. Cutillo, Y.Y. Jung, F. Ruggeri, B. Vidakovic, Larger posterior mode wavelet thresholding and applications, Journal of Statistical Planning and Inference 138 (2008) 3758–3773.

J. Dagpunar, An easily implemented generalized inverse Gaussian generator, Communications in Statistics - Simulation and Computation 18 (1989) 703–710.

I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, First edition, 1992.

D. De Candiitis, B. Vidakovic, Wavelet Bayesian block shrinkage via mixtures of normal-inverse gamma priors, Journal of Computational and Graphical Statistics 13 (2004) 383–398.

H. Ding, G. Claeskens, M. Jansen, Variable selection in partially linear wavelet models, Statistical Modelling 11 (2011) 409–427.

D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (1994) 425–455.

D.L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, Journal of the American Statistical Association 90 (1995) 1200–1224.

T.R. Downie, B.W. Silverman, The discrete multiple wavelet transform and thresholding methods, IEEE Transactions in Signal Processing 46 (1998) 2558–2561.

S. Efromovich, Analysis of blockwise shrinkage wavelet estimates via lower bounds for no-signal setting, Annals of the Institute of Statistical Mathematics 56 (2004) 205–223.

R.F. Engle, C.W.J. Granger, J. Rice, A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, Journal of the American Statistical Association 81 (1986) 310–320.

J. Fadili, E. Bullmore, Penalized partially linear models using sparse representations with an application to fmri time series, IEEE Transactions on Signal Processing 53 (2005) 3436–3448.

J.M. Fadili, L. Boubchir, Analytical form for a Bayesian wavelet estimator of images using the Bessel K form densities, IEEE Transactions on Image Processing 14 (2005) 231–240.

M.A.T. Figueiredo, R.D. Nowak, Wavelet-based image estimation: An empirical Bayes approach using Jeffreys' noninformative prior, IEEE Transactions on Image Processing 10 (2001) 1322–1331.

P. Flandrin, Wavelet analysis and synthesis of fractional Brownian motion, IEEE Transactions on Information Theory 38 (1992) 910–917.

D. Fourdrinier, A. Philippe, C.P. Robert, Estimation of a noncentrality parameter under Stein-type-like losses, Journal of Statistical Planning and Inference 87 (2000) 43–54.

P. Fryzlewicz, Bivariate hard thresholding in wavelet function estimation, Statistica Sinica 17 (2007) 1457–1481.

I. Gannaz, Robust estimation and wavelet thresholding in partially linear models, Statistics and Computing 17 (2007) 293–310.

E.I. George, D.P. Foster, Calibration and empirical Bayes variable selection, Biometrika 87 (2000) 731–747.

E.I. George, R.E. McCulloch, Variable selection via Gibbs sampling, Journal of the American Statistical Association 88 (1993) 881–889.

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

G. Goel, I.C. Chou, E.O. Voit, System estimation from metabolic time-series data, Bioinformatics 24 (2008) 2505–2511.

E. Gomez, M. Gomez-Villegas, J. Marin, A multivariate generalization of the power exponential family of distributions, Communications in Statistics - Theory and Methods 27 (1998) 589–600.

E. Gomez, M.A. Gomez-Villegas, J.M. Marin, Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications, Communications in Statistics - Theory and Methods 37 (2008) 972–985.

I.S. Gradshteyn, I.M. Ryzhik, Table of Integrals, Series, and Products, Academic Press, New York, 1980.

P. Hall, G. Kerkyacharian, D. Picard, Block threshold rules for curve estimation using kernel and wavelet methods, Annals of Statistics 26 (1998) 922–942.

P. Hall, G. Kerkyacharian, D. Picard, On the minimax optimality of block thresholded wavelet estimators, Statista Sinica 9 (1999) 33–50.

P. Hall, S. Penev, G. Kerkyacharian, D. Picard, Numerical performance of block thresholded wavelet estimators, Statistics and Computing 7 (1997) 115–124.

W. Härdle, H. Liang, J. Gao, Partially Linear Models, Physica-Verlag, 2000.

G. Huerta, Multivariate Bayes wavelet shrinkage and applications, Journal of Applied Statistics 32 (2005) 529–542.

M. Jansen, M. Malfait, A. Bultheel, Generalized cross validation for wavelet thresholding, Signal Processing 56 (1997) 33–44.

M.J. Jansen, B. A., Empirical Bayes approach to improve wavelet thresholding for image noise reduction, Journal of the American Statistical Association 96 (2001) 629–639.

N.L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, Volume 1, Wiley-Interscience, Second edition, 1994.

I.M. Johnstone, B.W. Silverman, Wavelet threshold estimators for data with correlated noise, Journal of the Royal Statistical Society, Series B 59 (1997) 319–351.

I.M. Johnstone, B.W. Silverman, Empirical Bayes approaches to mixture problems and wavelet regression, Technical Report, Department of Statistics, Stanford University, 1998.

I.M. Johnstone, B.W. Silverman, Ebayesthresh: R programs for empirical Bayes thresholding, Journal of Statistical Software 12 (2005a) 1–38.

I.M. Johnstone, B.W. Silverman, Empirical Bayes selection of wavelet thresholds, The Annals of Statistics 33 (2005b) 1700–1752.

E. Keogh, Q. Zhu, B. Hu, Y. Hao, X. Xi, L. Wei, C.A. Ratanamahatana, The UCR Time Series Classification/Clustering Homepage, 2011.

R. Kohn, J.S. Marron, P. Yau, Wavelet estimation using Bayesian basis selection and basis averaging, Statistica Sinica 10 (2000) 109–128.

A. Kovac, B.W. Silverman, Extending the scope of wavelet regression methods by coefficient-dependent thresholding, Journal of the American Statistical Association 95 (2000) 172–183.

W. Lawton, Applications of complex valued wavelet transforms to subband decomposition, IEEE Transactions on Signal Processing 41 (1993) 3566–3568.

D. Leporini, J.C. Pesquet, Wavelet thresholding for a wide class of noise distributions, in: EUSIPCO98, Rhodes, Greece, pp. 993–996.

D. Leporini, J.C. Pesquet, Bayesian wavelet denoising: Besov priors and non-Gaussian noises, Signal Processing 81 (2001) 55–67.

J.M. Lina, Image processing with complex Daubechies wavelets, Journal of Mathematical Imaging and Vision 7 (1997) 211–223.

J.M. Lina, B. Macgibbon, Non-linear shrinkage estimation with complex Daubechies wavelets, in: Proceedings of SPIE, Wavelet Applications in Signal and Image Processing V, 3169, pp. 67–79.

J.M. Lina, M. Mayrand, Complex Daubechies wavelets, Applied and Computational Harmonic Analysis 2 (1995) 219–229.

J.M. Lina, P. Turcotte, B. Goulard, Complex dyadic multiresolution analyses, in: P.W. Hawkes (Ed.), Advances in Imaging and Electron Physics, volume 109 of *Advances in Imaging and Electron Physics*, Elsevier, 1999, pp. 163 – 197.

D.V. Lindley, A.F.M. Smith, Bayes estimates for the linear model, Journal of the Royal Statistical Society, Series B 34 (1972) 1–41.

S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (1989) 674–693.

P. Moulin, J. Liu, Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors, IEEE Transactions on Information Theory 45 (1999) 909–919.

P. Müller, A generic approach to posterior integration and Gibbs sampling, Technical Report, ISDS, Duke University, 1993.

P. Müller, B. Vidakovic (Eds.), Bayesian Inference in Wavelet Based Models, volume 141 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1999.

S. Nadarajah, The Kotz-type distribution with applications, Statistics 37 (2003) 341–358.

S. Nadarajah, Discussion letter to the editor, Coastal Engineering 55 (2008) 189–190.

G.P. Nason, Wavelet shrinkage using cross-validation, Journal of the Royal Statistical Society, Series B 58 (1996) 463–479.

G.P. Nason, Wavelet Methods in Statistics with R, Springer, First edition, 2008.

G.P. Nason, B.W. Silverman, The stationary wavelet transform and some statistical applications, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, volume 103 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1995, pp. 281–300.

R.T. Ogden, Essential Wavelets for Statistical Applications and Data Analysis, Birkhäuser, Boston, Cambridge, MA, USA, 1997.

S.C. Olhede, A.T. Walden, "Analytic" wavelet thresholding, Biometrika 91 (2004) 955–973.

T. Park, G. Casella, The Bayesian Lasso, Journal of the American Statistical Association 103 (2008) 681–686.

M. Pensky, Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise, Annals of Statistics 34 (2006) 769–807.

M. Pensky, T. Sapatinas, Frequentist optimality of Bayes factor thresholding estimators in wavelet regression models, Statistica Sinica 17 (2007) 599–633.

L. Pericchi, A. Smith, Exact and approximate posterior moments for a normal location parameter, Journal of the Royal Statistical Society, Series B 54 (1992) 793–804.

A. Pižurica, W. Philips, I. Lemahieu, M. Acheroy, A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising, IEEE Transactions on Image Processing 11 (2002) 545–557.

J. Portilla, V. Strela, M.J. Wainwright, E.P. Simoncelli, Image denoising using scale mixtures of Gaussians in the wavelet domain, IEEE Transactions on Image Processing 12 (2003) 1338–1351.

L. Qu, Bayesian wavelet estimation of partially linear models, Journal of Statistical Computation and Simulation 76 (2006) 605–617.

S. Ray, B.K. Mallick, A Bayesian transformation model for wavelet shrinkage, IEEE Transactions on Image Processing 12 (2003) 1512–1521.

C. Robert, The Bayesian Choice: A Decision-Theoretic Motivation, Springer texts in statistics, Springer-Verlag, 1994.

C.P. Robert, Simulation of truncated normal variables, Statistics and Computing 5 (1995) 121–125.

C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer-Verlag, New York, First edition, 1999.

J.K. Romberg, H. Choi, R.G. Baraniuk, Bayesian tree-structured image modeling using wavelet-domain hidden Markov models, IEEE Transactions on Image Processing 10 (2001) 1056–1068.

F. Ruggeri, Robust Bayesian and Bayesian decision theoretic wavelet shrinkage, in: P. Müller, B. Vidakovic (Eds.), Bayesian Inference in Wavelet Based Models, volume 141 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1999, pp. 139–154.

F. Ruggeri, B. Vidakovic, Bayesian modeling in the wavelet domain, in: D.K. Dey, C.R. Rao (Eds.), Bayesian Thinking: Modeling and Computation, volume 25 of *Handbook of Statistics*, North-Holland, Amsterdam, 2005, pp. 315–338.

C. Semadeni, A. Davison, D. Hinkley, Posterior probability intervals in Bayesian wavelet estimation, Biometrika 91 (2004) 497–505.

L. Sendur, I.W. Selesnick, Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency, IEEE Transactions on Signal Processing 50 (2002) 2744–2756.

E.P. Simoncelli, Bayesian denoising of visual images in the wavelet domain, in: P. Müller, B. Vidakovic (Eds.), Bayesian Inference in Wavelet Based Models, volume 141 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1999, pp. 291–308.

M.G. Tadesse, J.G. Ibrahim, M. Vannucci, R. Gentleman, Wavelet thresholding with Bayesian false discovery rate control, Biometrics 61 (2005) 25–35.

M. Vannucci, F. Corradi, Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective, Journal of the Royal Statistical Society, Series B 61 (1999) 971–986.

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

B. Vidakovic, Wavelet-based nonparametric Bayes methods, in: D.K. Dey, P. Müller, P. Sinha (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics, volume 133 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1998b, pp. 133–155.

B. Vidakovic, Statistical Modeling by Wavelets, Wiley series in probability and mathematical statistics: Applied probability and statistics, John Wiley & Sons, Inc., New York, 1999.

B. Vidakovic, P. Müller, An introduction to wavelets, in: P. Müller, B. Vidakovic (Eds.), Bayesian Inference in Wavelet Based Models, volume 141 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1999, pp. 1–18.

B. Vidakovic, F. Ruggeri, Expansion estimation by Bayes rules, Journal of Statistical Planning and Inference 79 (1999) 223–235.

B. Vidakovic, F. Ruggeri, BAMS method: Theory and simulations, Sankhyā: The Indian Journal of Statistics, Series B 63 (2001) 234–249.

E.O. Voit, J. Almeida, S. Marino, R. Lall, G. Goel, A.R. Neves, H. Santos, Regulation of glycolysis in Lactococcus lactis: an unfinished systems biological case study, IEE Proceedings - Systems Biology 153 (2006) 286–298.

G.G. Walter, X. Shen, Wavelets and other orthogonal systems, Studies in advanced mathematics, Chapman & Hall/CRC, Boca Raton, Second edition, 2001.

X. Wang, S. Ray, B.K. Mallick, Bayesian curve classification using wavelets, Journal of the American Statistical Association 102 (2007) 962–973.

X. Wang, A.T.A. Wood, Empirical Bayes block shrinkage of wavelet coefficients via the noncentral χ^2 distribution, Biometrika 93 (2006) 705–722.

X. Wang, A.T.A. Wood, Wavelet estimation of an unknown function observed with correlated noise, Communications in Statistics - Simulation and Computation 39 (2010) 287–304.

M. Yuan, Y. Lin, Efficient empirical Bayes variable selection and estimation in linear models, Technical Report 1092, Department of Statistics, University of Wisconsin, Madison, http://www.stat.wisc.edu/public/ftp/yilin/tr1092.pdf, 2004.

M. Yuan, Y. Lin, Efficient empirical Bayes variable selection and estimation in linear models, Journal of the American Statistical Association 100 (2005) 1215–1225.

C.H. Zhang, General empirical Bayes wavelet methods and exactly adaptive minimax estimation, Annals of Statistics 33 (2005) 54–100.

VITA

Norbert Reményi is a Ph.D. candidate in Industrial and Systems Engineering (ISyE) at Georgia Institute of Technology with a specialization in statistics. He joined the Ph.D. program in 2007 and started working with Dr. Brani Vidakovic in the Summer of 2009. His research focus is on Bayesian modeling in the wavelet domain with applications to nonparametric regression, image denoising and estimation of partially linear models. He has been working with the Centers for Disease Control and Prevention (CDC) since July 2012 as an ORISE research fellow. Before coming to Georgia Tech, Norbert earned a master's degree in Engineering Management at Budapest University of Technology and Economics (BME) as well as a master's degree in Experimental Statistics at New Mexico State University (NMSU).