

BAYESIAN FUNCTION ESTIMATION USING CONTINUOUS WAVELET DICTIONARIES

Jen-Hwa Chu¹, Merlise A. Clyde² and Feng Liang^{2,3}

¹*Harvard Medical School*, ²*Duke University* and ³*University of Illinois at Urbana-Champaign*

Abstract: We present a Bayesian approach for nonparametric curve estimation based on a continuous wavelet dictionary, where the unknown function is modeled by a random sum of wavelet functions at arbitrary locations and scales. By avoiding the dyadic constraints for orthonormal wavelet bases, the continuous overcomplete wavelet dictionary has greater flexibility to adapt to the structure of the data, and leads to sparse representations. The price for this flexibility is the computational challenge of searching over an infinite number of potential dictionary elements. We develop a reversible jump Markov Chain Monte Carlo algorithm which utilizes local features in the proposal distributions and leads to better mixing of the Markov chain. Performance comparison in terms of sparsity and mean square error is carried out on standard wavelet test functions. Results on a non-equally spaced example show that our method compares favorably to methods using interpolation or imputation.

Key words and phrases: overcomplete dictionaries; Bayesian inference; wavelets; nonparametric regression; reversible jump Markov chain Monte Carlo; stochastic expansions;

1. Introduction

Suppose we have observed data $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ at points $x_1, \dots, x_n \in [0, 1]$ of some unknown function $f(x)$

$$Y_i = f(x_i) + \epsilon_i \quad \epsilon_i \stackrel{\text{iid}}{\sim} \mathbf{N}(0, \sigma^2) \quad (1.1)$$

measured with independent and identically distributed (iid) Gaussian noise. A standard approach in nonparametric function estimation is to expand f with respect to an orthonormal basis, such as Fourier, Hermite, Legendre or wavelet, and then to estimate the corresponding coefficients of the basis elements. Wavelets, as a popular choice of orthonormal bases, are widely used in nonparametric function

estimation and signal processing (Mallat, 1989a; Donoho and Johnstone, 1998). Given a wavelet function $\psi(x)$, let $\psi_{jk}(x) \equiv 2^{j/2}\psi(2^jx - k)$, $j, k \in \mathbb{Z}$, then the ψ_{jk} 's form an orthonormal basis for L_2 functions and any L_2 function can be represented as

$$f(x) = \sum_{j,k} \theta_{jk} \psi_{jk}(x).$$

For equally-spaced sample locations x_1, \dots, x_n , the coefficients θ_{jk} may be computed efficiently via the so-called Cascade algorithm (Mallat, 1989b,a). Several standard wavelet-based methods may be used to estimate these coefficients via thresholding or shrinkage, after which the estimated coefficients are transformed back to the data domain to provide an estimate of f .

Each wavelet is ideally suited to represent certain signal characteristics, so that just a few basis elements are needed to describe these features leading to a sparse representation of the signal. In practice, as the structure of the function f is unknown, it is desirable to have a representation with adaptive sparsity. Recently overcomplete (or redundant) representations have drawn considerable attention in the signal processing community due to their flexibility, adaptation and robustness (Chen et al., 2001; Coifman et al., 1992; Mallat, 1998; Wickerhauser, 1994; Lewicki and Sejnowski, 1998; Donoho and Elad, 2003; Wolfe et al., 2004; Donoho et al., 2006). Examples of overcomplete dictionaries include translation-invariant wavelet transforms (Duttilleux, 1989; Nason and Silverman, 1995), frames (Gröchenig, 2001; Wolfe et al., 2004; Kovačević and Chebira, 2007) and wavelet packets (Coifman and Meyer, 1990).

Due to the redundancy of overcomplete dictionaries, there is no unique solution to the representation problem. Efficient algorithms, such as matching pursuit (Mallat and Zhang, 1993), the best orthogonal basis (Coifman and Wickerhauser, 1992), and basis pursuit (Chen et al., 1998), are designed to search for the “best” representation. Bayesian methods offer another effective way to make inference using overcomplete representations, where regularization and shrinkage are introduced via prior distributions and efficient searching is guided via Markov Chain Monte Carlo (MCMC) algorithms. In this paper, we propose a nonparametric Bayesian approach in function estimation using continuous wavelet dictionaries (CWD). As opposed to orthonormal wavelet basis functions which are subject

to dyadic constraints on their locations and scales, the wavelet components in a CWD have arbitrary locations and scales. An additional advantage of a CWD is that it can be applied to non-equally spaced data without interpolation or imputation of the missing data. We develop a reversible jump MCMC algorithm for inference of the unknown function f and provide strategies to achieve better convergence.

The remainder of this paper is arranged as follows. In Section 2 we introduce the concept of stochastic expansions using a CWD. In Section 3 we discuss prior specifications. In Section 4 we describe posterior inference by means of a reversible jump Markov Chain Monte Carlo sampling scheme and discuss various estimates of f including point estimates and simultaneous credible bands. In Section 5 we present results from simulation studies and from a real example, which show that our new method leads to better performance in terms of sparsity and mean squared error. Concluding remarks are given in Section 6.

2. The Model

Suppose ϕ and ψ are the compact-supported scaling and wavelet functions that correspond to an r -regular multi-resolution analysis for some integer $r > 0$ (See Daubechies, 1992). Under the continuous wavelet dictionary (CWD) setting, we model the response variable Y as $N(f(x), \sigma^2)$ with

$$f(x) = f_0(x) + \sum_{k=1}^K \beta_{\lambda_k} \psi_{\lambda_k}(x), \quad (2.1)$$

where $\lambda_k = (a_k, b_k) \in [a_0, \infty) \times [0, 1]$, $k = 1, \dots, K$ represents the scaling and location of the wavelet functions used in the stochastic expansion and f_0 is a fixed scaling function given by

$$f_0(x) = \sum_{i=1}^M \eta_i \phi_{\lambda_i}(x), \quad (2.2)$$

where $\phi_{\lambda_i}(t) \equiv a_i^{1/2} \phi(a_i(t - b_i))$ for some finite set of indices $\lambda_i = (a_i, b_i) \in (0, a_0) \times [0, 1]$, $i = 1, \dots, M$. In the remaining of this paper, we take $f_0 = \bar{\mathbf{Y}}$, the sample mean. Such an empirical approach is similar to modelling f_0 as a constant function with a uniform prior on the corresponding coefficient.

While f_0 describes the coarse-scale features of the unknown function f , the second term in equation (2.1) describes the fine-scale features. In this representation, in addition to the scaling and location parameters in λ , the number of wavelet elements K from the CWD is also an unknown parameter. Note that in a regular discrete wavelet transformation where a and b have dyadic constraints and the data are on an equally spaced grid, we can obtain the coefficients β_λ through filters without evaluating the wavelet functions ψ_λ directly. In a CWD, the dictionary elements do not have a tree-like structure needed for the cascade algorithm and in addition our data may not be equally spaced, therefore it is necessary to evaluate each of the wavelet functions $\psi_\lambda(x)$. Here we use the Daubechies-Lagarias local pyramid algorithm (Vidakovic, 1999, Sec. 3.5.4), which enables us to evaluate ψ_λ at an arbitrary point with preassigned precision.

In practice, wavelets are often used to represent functions from certain Besov spaces. Naturally one would ask under what kind of conditions, the random function f will still be in the same Besov space almost surely (a.s.). If the number of wavelet element K is finite (a.s.), for example, if K has a Poisson prior with a finite intensity measure, equation (2.1) will have a finite number of elements (a.s.) and therefore f will belong to the same Besov space (a.s.) as the mother wavelet function ψ does, for any reasonable choice of the probability distribution for β_λ . However, extra conditions are needed for the random function f to be well-defined if K is not finite (a.s.), which is the case discussed by Abramovich et al. (2000). Though the focus of Abramovich et al. (2000) is not on Bayesian analysis, the stochastic expansion proposed in their paper suggests a prior choice for our model which is given in the next section.

3. The Prior

The unknown parameters in model (2.1) are the error variance σ^2 , the number of wavelet elements K , and the corresponding location-scale index and coefficient for each wavelet component (β_λ, λ) . We set a non-informative reference prior for σ^2 , $p(\sigma^2) \propto 1/\sigma^2$. Though it is improper, it is easy to show that the corresponding posterior distribution is proper for $n \geq 2$. Prior distributions on the remaining parameters are specified as follows.

3.1 Prior for $\lambda = (a, b)$

Following Abramovich et al. (2000), the prior for the scale parameter a takes

the form

$$p(a) \propto a^{-\zeta}, \quad a_0 \leq a \leq a_1 \text{ and } \zeta > 0. \quad (3.1)$$

The hyperparameter ζ controls the intensity or relative number of fine-scale wavelet components in the function. If ζ is large, *a priori* we will have relatively few fine-scale (spiky) components in the function, while if ζ is small fine-scale components will predominate. We set $\zeta = 1.5$ for all examples in Section 5. and found that that posterior results are not very sensitive to this choice in the examples that we have studied. The lower bound a_0 corresponds to the coarsest-scale component allowable in the function; a_0 needs to be larger than twice the support of the wavelet function ψ . The upper bound a_1 corresponds to the smallest finest-scale component. Theoretically, a can go up to infinity to span the whole space as in Abramovich et al. (2000). However in practice, allowing a to go up to infinity is not desirable, as when a increases we obtain spiky wavelet functions with very small support which have little or no effect on the likelihood. For example, suppose we have 1024 equally-spaced data points and use a mother wavelet with support of length 1. If we set $a_1 > 1024$, the support of a wavelet function could fall entirely between two data points and have no effect on the likelihood. As a result, the corresponding coefficient can not be estimated effectively, leading to potential over-fitting of the data and poor out of sample predictive properties. Therefore we set an upper bound for a so that the wavelet functions will have large enough support. These bounds will depend on the data and the support of the wavelet ψ .

The prior for the location parameter b takes the form

$$p(b) = \gamma \sum_{i=1}^n \frac{1}{n} \delta_{x_i}(b) + (1 - \gamma), \quad 0 < \gamma < 1, \quad (3.2)$$

which is a mixture of point masses on all the data points and a uniform distribution on $[0, 1]$. We take $\gamma = 1/2$, although one could place a prior distribution on γ . This prior is a compromise of flexibility, which allows b to be at arbitrary positions, and efficiency, which focuses on the data points where the information is abundant. This mixture prior also enables us to search the dictionary elements more efficiently by using the information from residuals, which we discuss in detail in the next section. Notice that when $\gamma = 1$ and $p(b)$ has support on data

points only, we return to the non-decimated discrete wavelet setting, and when $\gamma = 0$, we have the continuous distribution from Abramovich et al. (2000).

3.2 Prior for K

Abramovich et al. (2000) viewed $(a_k, b_k)_{k=1}^K$ as a realization from a compound Poisson process on $[a_0, a_1 = \infty] \times [0, 1]$, which results in a Poisson prior distribution on K with mean

$$\mu = \mathbb{E}(K) = c_1 \int_{a_0}^{a_1} \int_0^1 a^{-\zeta} db da,$$

where c_1 is some constant. By placing an additional Gamma distribution on μ we obtain a negative binomial prior on K

$$p(K|r, q) = \binom{r + K - 1}{K} q^r (1 - q)^K$$

as the negative binomial distribution may be obtained as a Gamma mixture of Poissons. This provides additional flexibility over the Poisson distribution, which has only one parameter that controls both the mean and the variance. The hyperparameters r and q are chosen by specifying the probability of the null model $p(K = 0)$ and a quantile of K (for example, the 95 percentile of $p(K)$). These two equations can be easily solved to obtain the values of r and q .

Both the Poisson and negative binomial priors can be regarded as a limiting case for the mixture prior from Clyde et al. (1998) when the model space moves from being finite to being infinite dimensional. Recall that in the orthonormal wavelet model with N wavelet basis functions, the mixture prior implies a Binomial distribution (N, π) on the number of non-zero coefficients. When N goes to infinity as in the continuous wavelet dictionary model, if we let π go to zero such that $N\pi$ converges to μ , we obtain the Poisson model with mean μ for the number of non-zero coefficients. If we have a Gamma distribution on μ , we obtain the negative binomial model above.

3.3 Prior for β_λ

Given the location and scale of a wavelet function, the prior distribution of the corresponding wavelet coefficient β_λ is independent normal as in Abramovich

et al. (2000):

$$p(\beta_\lambda | a) = \mathbf{N}(0, ca^{-\delta}), \quad (3.3)$$

where c is a fixed hyperparameter independent of a . One possible choice for c is to set $c = n$, the sample size, as in the unit-information prior (Kass and Wasserman, 1995). The hyperparameter δ controls the magnitude of the coefficients for the fine scale wavelet components relative to the coarse scale wavelet components, giving us more flexibility to adapt to the smoothness of the functions being modeled. For example, if δ is large, we will shrink the fine scale (spiky) wavelets more, resulting in a smoother function, and vice versa. For all examples in Section 5 we set $\delta = 2$.

Normality of β_λ was one of the conditions for f , as defined in (2.1), to be well-defined and to belong to the Besov space of the mother wavelet when K is infinite almost surely (Abramovich et al., 2000). However, since our prior distribution on K implies that K is finite (a.s.), the normality of β is not necessary. We may replace the normal distribution with a heavier-tailed prior for β , e.g. Laplace with a scale parameter that depends on a the same way as in (3.3)

$$p(\beta | a) = \frac{1}{2\sigma} \exp \frac{-|\beta|}{\sigma}, \quad \sigma^2 = ca^{-\delta}/2 \quad (3.4)$$

or other scale mixtures of normals. The heavy-tailed priors have been shown to have theoretical advantages over normal distributions, and may lead to greater sparsity and further reduction of the mean squared error (Johnstone and Silverman, 2004).

4. Posterior Inference

The task of searching over a continuous model space with infinitely many models can be extremely challenging. Since the dimensionality of the parameters varies, we propose a reversible jump Markov Chain Monte Carlo (RJ-MCMC) algorithm (Green, 1995) to explore the posterior distribution of models and model specific parameters. Our RJ-MCMC algorithm includes three types of moves: a birth step where we add a wavelet element, a death step where we delete a wavelet, and an update step where we move a wavelet element but leave the dimension K unchanged. For RJ-MCMC algorithms a good proposal distribution is necessary to speed up convergence. For example, proposing a “birth” of a new wavelet

dictionary element from the prior on $(\beta, a, b|K+1)$ may simplify the calculation of the Metropolis Hastings ratio, but often results in slow convergence since it does not necessarily lead to proposal values where the likelihood is high. Similarly, picking a component at random to remove may lead to frequent attempts to remove important wavelets. We provide highlights of the algorithm with more details available in the appendix.

4.1 RJ-MCMC Moves

Because of the local nature of wavelets, information in the residuals may aid in placing new wavelets. We choose a mixture proposal for the location parameter b of the new wavelet functions which is a mixture of point masses on the data points with weights that depend on the current residuals and uniform on $[0, 1]$. In particular, the proposal for the birth step is

$$q(b_{K+1}) = \gamma \sum_{i=1}^n \delta_{x_i}(b_{K+1})v_i + (1 - \gamma), \quad 0 < \gamma < 1, \quad (4.1)$$

where

$$v_i = \frac{|Y_i - \hat{f}(x_i)|}{\sum_{j=1}^n |Y_j - \hat{f}(x_j)|}$$

is proportional to the magnitude of the residual. Since the prior for b is also a mixture of point masses and uniform, it has density on the same measure as the proposal, which is a necessary condition for the transition kernel to be reversible.

The proposal for the death step is inversely proportional to the wavelet coefficient:

$$q(b_k | K) = \frac{1/|\beta_k|}{\sum_{i=1}^K (1/|\beta_i|)}, \quad (4.2)$$

so that small magnitude coefficients are more likely to be removed.

Finally, the proposal for the update step is

$$q(\tilde{b}_k | b_k) = \delta_{b_k}(\tilde{b}_k)u_k + \mathbf{N}(\tilde{b}_k; b_k, \sigma_b^2)(1 - u_k), \quad (4.3)$$

where

$$u_k = \begin{cases} 1 & \text{if } b_k \text{ is a data point} \\ 0 & \text{otherwise,} \end{cases}$$

which is a point mass at b_k if b_k is a data point and a random walk otherwise.

These proposal distributions can improve convergence in practice since a successful birth is more likely where the residual is large, and killing a wavelet of which the coefficient is small will likely not change the likelihood dramatically.

After T MCMC iterations post burn-in, each collection of the parameters $\{\boldsymbol{\beta}, \mathbf{a}, \mathbf{b}, K\}$ represents a sample from the posterior distribution, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_K)$ and \mathbf{a} and \mathbf{b} are defined similarly. At each iteration we plug $\{\boldsymbol{\beta}, \mathbf{a}, \mathbf{b}, K\}$ back into equation (2.1), obtaining posterior samples $f^{(t)}(x_i), t = 1, \dots, T$ from $p(f | \mathbf{Y})$ which provide a full description of the posterior distribution of f given the data \mathbf{Y} .

4.2 Point Estimates for f

A natural point estimate for $f(x)$ is the posterior mean, which is approximated by the ergodic average of MCMC samples,

$$\hat{f}_{\text{AVE}}(x) = \frac{1}{T} \sum_{t=1}^T \hat{f}^{(t)}(x) = \frac{1}{T} \sum_{t=1}^T \mathbb{E}(\mathbf{Y} | x, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, \boldsymbol{\beta}^{(t)}, K^{(t)}), \quad (4.4)$$

where T is the number of MCMC iterations after burn-in and $\hat{f}^{(t)}$ represents the estimate from the t -th MCMC iteration.

While the posterior mean of f is an average over many sparse models, the average itself is not necessarily sparse. When the goal is selection of a single model, we choose to report the model which is closest to the posterior mean in terms of mean squared error:

$$f^* = \arg \min_{t \in \{1, \dots, T\}} \sum_{i=1}^n \{\hat{f}(x_i) - \hat{f}^{(t)}(x_i)\}^2. \quad (4.5)$$

If β has a normal prior, we can reduce the Monte Carlo variation in estimating the mean under model selection by replacing $\beta^{(t)}$ by its posterior mean when we calculate $\hat{f}^{(t)}$,

$$\hat{\beta}^{(t)} = \mathbb{E}(\beta | \mathbf{Y}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, K^{(t)}).$$

4.3 Simultaneous Credible Bands for f

We can also construct a credible region which contains $f(x)$ simultaneously at all x with at least $1 - \alpha$ posterior probability. Specifically, a credible band corresponds to a pair of functions $l(x)$ and $u(x)$ which define an envelope along

x ,

$$C = \{f : l(x) \leq f(x) \leq u(x), \text{ for all } x\},$$

such that

$$P(f \in C \mid \mathbf{Y}) \geq 1 - \alpha.$$

In practice, the posterior probability is approximated by the empirical distribution based on MCMC samples and the condition “for all x ” is approximated by “for a fine grid (x^1, \dots, x^m) on the range of x ”, where the x^j s could be observed data points, but not necessarily.

Several ways have been proposed to construct such credible regions. For example, the Bayesian credible band in Crainiceanu et al. (2007) takes the following symmetric form

$$\hat{f}_{\text{AVE}}(x^j) \pm M_\alpha \cdot \text{sd}[f(x^j)],$$

where $\text{sd}[f(x^j)]$ denotes the posterior standard deviation of $f(x^j)$ estimated from MCMC samples and M_α denotes the $(1 - \alpha)100$ quantile of

$$\max_{1 \leq j \leq m} \frac{|\hat{f}^{(t)}(x^j) - \hat{f}_{\text{AVE}}(x^j)|}{\text{sd}[f(x^j)]}. \quad (4.6)$$

By the definition of the credible band, one could just select $100(1 - \alpha)\%$ of the MCMC samples of f , and their minimal and maximal values at each grid point x^j form a credible band. In Besag et al. (1995), the set of the MCMC samples are selected as follows: at each grid point x^j , the corresponding T values $\{f^{(t)}(x^j)\}_{t=1}^T$ are ranked; a MCMC sample $f^{(t)}$ with an extremely high or low rank is less preferable, as the corresponding credible band at x^j will be unnecessarily wide; therefore, Besag et al. (1995) propose to select $100(1 - \alpha)\%$ of the MCMC samples by minimizing their worst rank (too high or too low) across all the grid points. In our empirical study in Section 5, we found both methods to be conservative. Here we propose to construct credible bands based on an L_2 ball of errors, which is motivated by earlier work of Cox (1993) and Baraud (2004).

First we start with a ball defined as

$$\{f : \|f - \hat{f}_{\text{AVE}}\|_\Sigma \leq D_\alpha\}, \quad (4.7)$$

where $\|\cdot\|_\Sigma$ is the L_2 norm normalized by the estimated covariance matrix Σ of

the $f(x^i)$'s,

$$\|a\|_{\Sigma} = a'\Sigma^{-1}a,$$

and D_{α} is the $100(1 - \alpha)\%$ quantile of all such scaled L_2 distances from MCMC samples. This ball gives the $1 - \alpha$ probability bound in the estimation error in scaled L_2 loss. For better visualization, the credible region takes the form of a hyper-rectangle containing the ball defined in (4.7). The $(1 - \alpha)$ credible band is given as follows:

1. For the t -th MCMC iteration, calculate the scaled L_2 distance D_t to the ergodic average estimate from (4.4):

$$D^{(t)} = \|\hat{f}^{(t)} - \hat{f}_{\text{AVE}}\|_{\Sigma}.$$

2. Calculate D_{α} , the $100(1 - \alpha)\%$ quantile of $D^{(t)}$.
3. Let T_{α}^D be the collection of indices of MCMC samples of f of which the distance to \hat{f}_{AVE} is below D_{α} :

$$T_{\alpha}^D = \{t : 1 \leq t \leq T, \quad D^{(t)} \leq D_{\alpha}\}.$$

4. Then our simultaneous credible region C is defined as the minimum hyper-rectangle that contains all the posterior samples in T_{α}^D , namely,

$$l(x^i) = \min_{t \in T_{\alpha}^D} f^{(t)}(x^i), \quad u(x^i) = \max_{t \in T_{\alpha}^D} f^{(t)}(x^i),$$

$$C = \{f : l(x^i) \leq f(x^i) \leq u(x^i), \text{ for all } i\}.$$

It is straight forward to show that the posterior coverage of C is bigger than or equal to $100(1 - \alpha)\%$.

5. Examples

We illustrate our Bayesian CWD method and compare it to other approaches in the literature in a series of simulation studies and a real example. Throughout we use the `1a8` wavelet (this is the default in `R`) with $a_0 = 8$ except were noted. The hyperparameters were set to $\delta = 2$ and $\zeta = 1.5$ as discussed previously. The prior for the number of coefficients K is negative binomial with $r = 1$

and $q = 0.01$, which corresponds to 0.01 probability of the null model and 95% percentile at $K = 298$. The prior distribution on K is relatively flat and covers a wide range of possible models (see Figure 1(b)).

5.1 Simulation Studies

As the stochastic representation allows extremely flexible representations, an initial concern is that the method may lead to over-fitting of the data. To test this, we apply the CWD method to the null function $f(x) = 0$ observed with noise with $n = 1024$. We set $a_1 = 500$ (approximately $n/2$) so that each wavelet supports an adequate number of data points. The results from the posterior simulation are shown in Figure 1. We can see from the posterior histogram that the null model ($K=0$) is the one with the highest posterior probability. We compare the results with the empirical Bayes (EBayes) method (Johnstone and Silverman, 2005), an overcomplete method using translational invariant wavelets. EBayes shrinks and thresholds nJ wavelet coefficients, where J is the number of levels in the multi-resolution expansion. In addition to the wavelet coefficients there are n scaling coefficients that do not undergo any shrinkage or thresholding. In this example we use EBayes with the Laplace prior and $J = 4$ (the default). Even though EBayes thresholds all but one of the wavelet coefficients to zero, the estimate still appears “bumpy” due to the included 1024 scaling coefficients.

We carry out a simulation study on four standard test functions from Donoho and Johnstone (1994): `bumps`, `blocks`, `doppler`, and `heavysine`. For each test function, 100 replications are generated with a fixed signal-to-noise ratio of 7. In each replicate, the data are simulated at 1024 equally spaced points in $[0, 1]$.

The default choice of wavelet in R (`1a8`) is used in all functions except for `blocks`, where we use the Haar wavelet with $a_0 = 2$. Unlike many other wavelet methods, we do not assume a boundary correction here, since some of the functions (e.g. `doppler`) are clearly not periodic. We set the upper bound $a_1 = 500$ as in the null model simulation.

Usual convergence diagnostic methods, such as Gelman and Rubin (1992), do not apply to assessing convergence of the joint posterior distribution since we are moving within an infinite model space and the parameters are not common to all models. Instead we look at K and the mean squared error, which have a coherent interpretation throughout the model space (Brooks and Giudici, 2000).

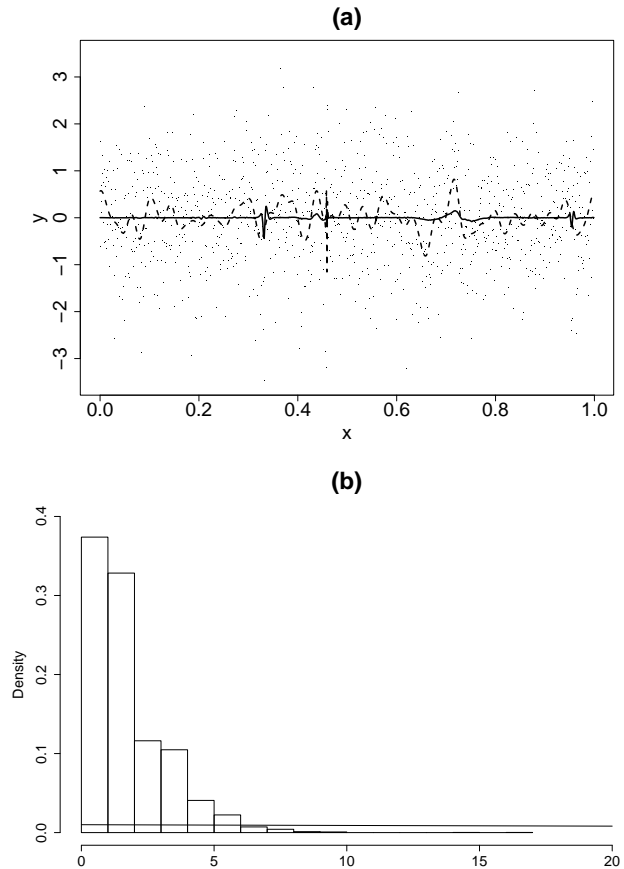


Figure 1: (a) The EBayes (Johnstone and Silverman, 2004) (dash line) and CWD (solid line) fits of the null function and (b) the posterior histogram for K overlaid with the $\text{NB}(1, 0.01)$ prior distribution.

The trace plots and the Gelman-Rubin shrink factor for K and mean squared error suggest that convergence usually occurs within 1 million MCMC iterations. The following results are based on 5 million iterations, which takes about 8-9 hours to run on a computer cluster.

We compare the CWD fits with EBayes using a Laplace prior based on mean-squared error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \{\hat{f}(x_i) - f(x_i)\}^2. \quad (5.1)$$

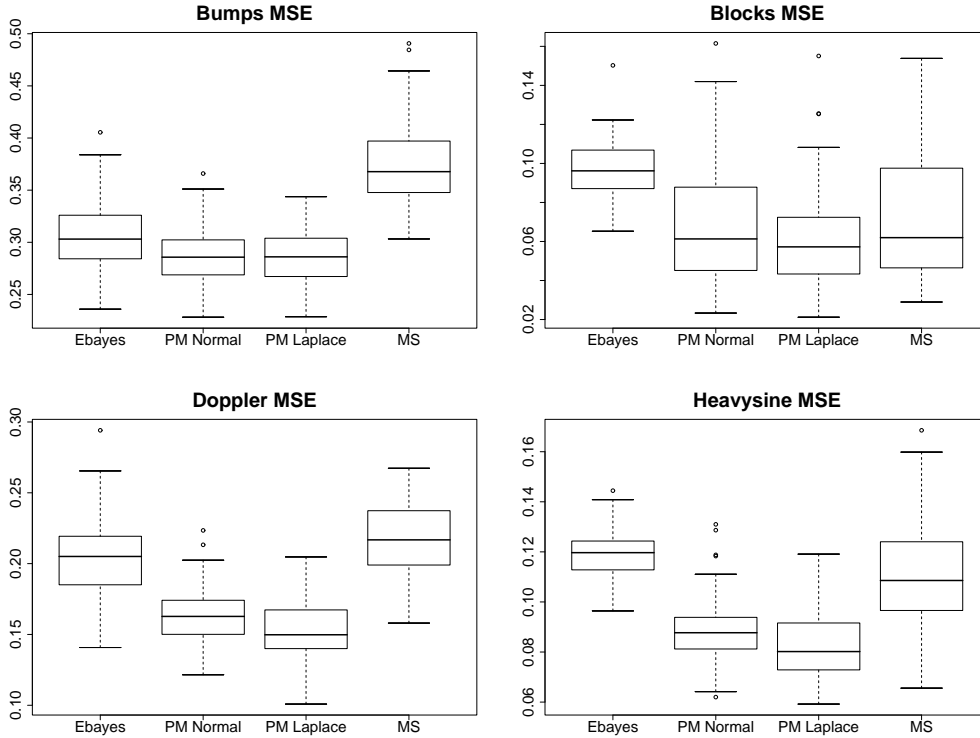


Figure 2: Box plots for mean squared error for the four test functions using the EBayes method of Johnstone and Silverman (2005) and the continuous wavelet dictionary (CWD) method with the posterior mean (PM) under the normal and Laplace prior distributions and with model selection (MS) under the normal prior distribution.

For CWD we calculate point estimates of f based on the posterior mean (PM) for both the normal and Laplace prior for β , along with the model selection (MS) estimate from (4.5) with the normal prior. Figure 2 shows that the PM estimate in (4.4) has smaller MSE than EBayes for all four functions. The heavy-tailed Laplace prior leads to an additional reduction in MSE except for **bumps**. Using model selection under squared error loss, we find that the EBayes estimate is better for **bumps** and **doppler**. However, if we compare the number of non-zero coefficients in \hat{f} (See Figure 3) then CWD method clearly gives a much sparser representation than EBayes. We note that the EBayes summaries for K do not include the 1024 coefficients from the scaling function, which are not shrunk or thresholded.

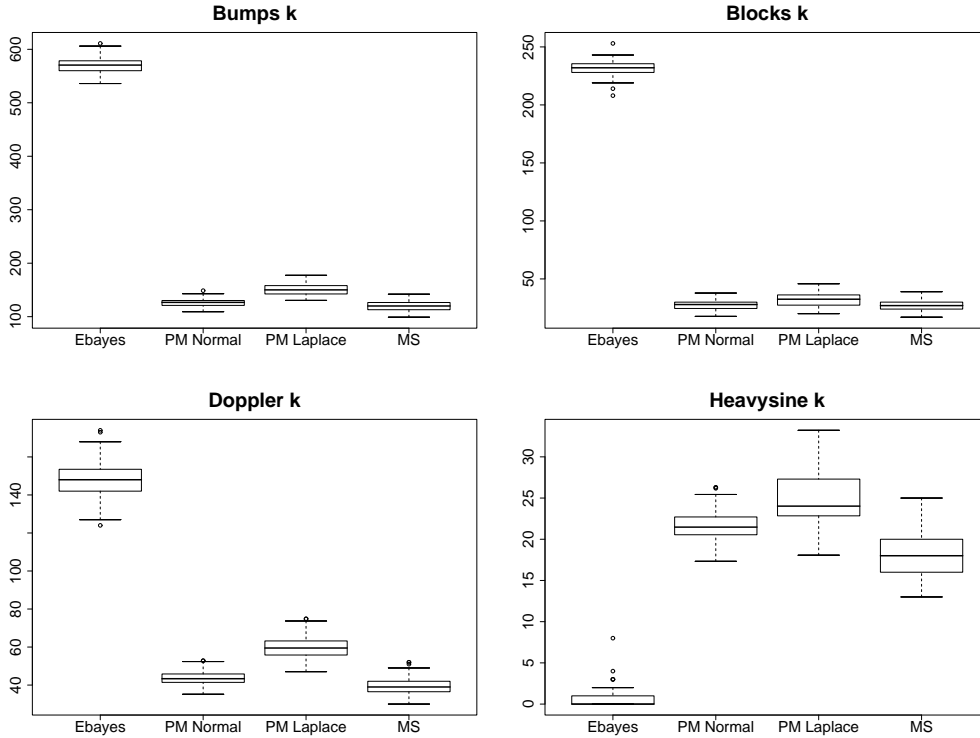


Figure 3: Box plots for the number or average number (for PM) of non-zero wavelet coefficients for the four test functions using the EBayes method Johnstone and Silverman (2005) and continuous wavelet dictionary (CWD) method with the posterior mean (PM) under the normal and Laplace priors and with model selection (MS) under the normal prior distribution.

5.2 Real Application

One of the advantages of the CWD based method is that it can be applied directly to non-equally spaced data sets. To illustrate this point, we apply our method to a well-studied data set, **ethanol** data, from Brinkman (1981). This data set consists of $n = 88$ measurements from an experiment where ethanol was burned in a single cylinder engine. The concentration of the total amount of nitric oxide and nitrogen dioxide in the engine exhaust, normalized by the work done by the engine is related to the “equivalence ratio”, a measure of the richness of the air ethanol mixture.

We apply our CWD method with 4, 8, and 10 vanishing moments of the least

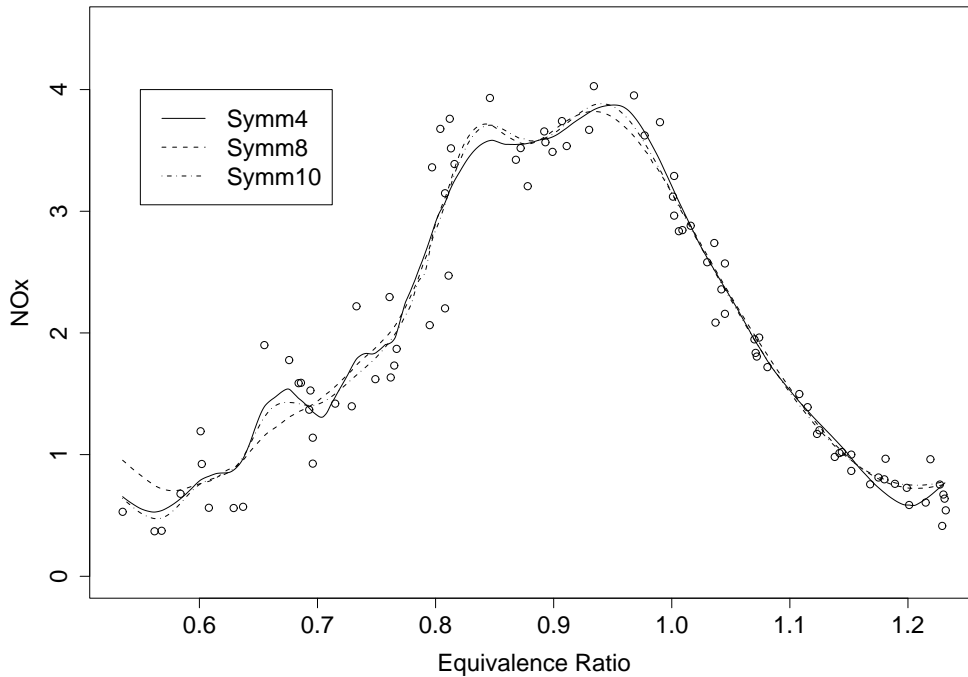


Figure 4: The posterior mean using the CWD with the normal prior for the ethanol data from Brinkman (1981).

asymmetric Daubechies' wavelets (`symm4`, `symm8` and `symm10`). We set the upper bound a_1 to 100 as there are fewer data points. We report the estimated posterior mean curves defined in (4.4), in Figure 4 and the 95% simultaneous credible band with `symm8` in Figure 5. Notice that with the Daubechies-Lagarias algorithm we can evaluate the function at arbitrary locations, not just the observed data points. The posterior samples $f^{(t)}$'s are evaluated on 512 equally-spaced grid points covering the range of the data.

This same data set has been studied by Nason (2002) using a linear interpolation method to address the problem of non-equally spaced observations. To compare with their result, we perform a leave-one-out cross validation study and

calculate the cross validation score

$$\text{CV-score} = \frac{1}{n} \sum_{i=1}^n \{ \hat{f}^{-i}(x_i) - Y_i \}. \quad (5.2)$$

where \hat{f}^{-i} is the estimated f from all the data except the i th point. With no attempts to optimize the hyperparameters, the CV score from CWD with `symm8` ranks 2nd out of the 60 combinations reported in Nason (2002), and the estimated functions look very similar to their best combination. We can see that over the left region where there are fewer data points, there seems to be greater uncertainty, as the credible band is wider and the estimates have greater disagreement. On the right hand side where all estimates seem to agree on the same downward slope, the credible region is much narrower as we have more information here. We can see that CWD has managed to capture the main feature of the data without over-fitting.

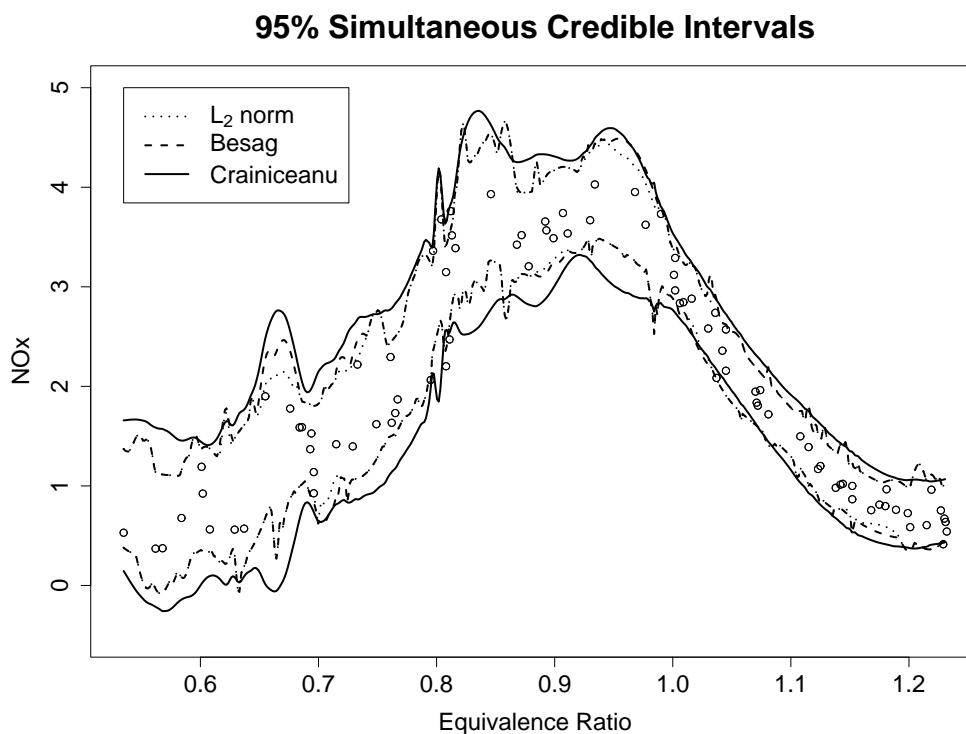


Figure 5: Simultaneous credible bands ($\alpha = 0.05$) for `symm8`

We also construct 95% simultaneous credible bands for $f(x)$ with `symm8` using methods by Besag et al. (1995) and Crainiceanu et al. (2007). From Figure 5 we can see that while all credible bands take a similar shape, the L_2 loss based credible region is narrower than Crainiceanu's, and the difference between our credible band and Besag's is negligible. We can also see on the left part of the graph where the data are sparser and there is more uncertainty, Crainiceanu et al. (2007) gives a wider interval than the other methods.

We can calculate the area of the credible region by numerical integration:

$$\text{Area} = \frac{x^n - x^1}{n} \sum_{i=1}^n (u(x^i) - l(x^i)), \quad (5.3)$$

where x^1, \dots, x^n are the grid locations where the functions are being evaluated. The L_2 based method has the smallest area of 0.6915, while Besag's method gives 0.6974 and Crainiceanu is the largest with 0.9159. In general with the same coverage rate a smaller credible region is more desirable. Therefore, we can say our method performs slightly better than the other methods in this particular case.

6. Conclusion

In this paper we have introduced a Bayesian method for function estimation based on a stochastic expansion in a continuous wavelet dictionary. Despite the richness of the potential representations and computational challenges of evaluating the wavelet functions and model search, RJ-MCMC algorithms are able to identify sparse representations in a reasonable time frame. The simulation study shows that the new method leads to greater sparsity and improved mean squared error performance over current wavelet-based methods. Because the models do not require the data to be equally spaced, this will permit wavelet methods to be used in a greater variety of applications. We have also introduced a new approach for constructing simultaneous credible bands in the overcomplete setting which appears to give narrower bands than other existing methods.

Acknowledgment The authors would like to thank Brani Vidakovic for suggesting the Daubechies-Lagarias algorithm for evaluating continuous wavelets. The authors acknowledge support of the National Science Foundation (NSF)

through grants DMS-0422400 and DMS 0406115. Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF.

Appendix: Reversible Jump MCMC Algorithm

We follow the general framework by Green (1995) and Denison et al. (1998) and include three types of movement in the MCMC algorithm:

1. Birth step. (add a wavelet)
2. Death Step. (delete a wavelet)
3. Update Step. (move a wavelet)

As in Green (1995), the birth and death probabilities are chosen to be

$$p_b(K) = c \min\{1, p(K+1)/p(K)\},$$

$$p_d(K) = c \min\{1, p(K)/p(K+1)\},$$

where $c < 0.5$ is some constant. For the birth step, we propose to add a wavelet coefficient β_{K+1} with scale a_{K+1} and b_{K+1} from some joint proposal $q(\beta, a, b)$. Let $\hat{f}(x)$ be the mean estimate for the current model and $\tilde{f}(x)$ be the mean estimate for the proposed model, then the likelihood ratio is

$$LR = \frac{\mathbf{N}(\mathbf{Y}; \tilde{f}(x), \sigma^2 I)}{\mathbf{N}(\mathbf{Y}; \hat{f}(x), \sigma^2 I)}. \quad (\text{A.1})$$

The acceptance ratio for the birth step is

$$LR \times \text{prior ratio} \times \text{proposal ratio} \times \text{Jacobian},$$

where the prior ratio is

$$\begin{aligned} & \frac{p(K+1)p(\beta_{1:K+1}, a_{1:K+1}, b_{1:K+1})}{p(K)p(\beta_{1:K}, a_{1:K}, b_{1:K})} \\ &= \frac{p(K+1)(K+1)! \prod_{k=1}^{K+1} p(\beta_k, a_k, b_k)}{p(K)K! \prod_{k=1}^K p(\beta_k, a_k, b_k)} \\ &= \frac{p(K+1)(K+1)p(\beta_{K+1}, a_{K+1}, b_{K+1})}{p(K)}, \end{aligned}$$

and the proposal ratio is

$$\frac{p_d(K+1)q'(\beta_{K+1}, a_{K+1}, b_{K+1} | K+1)}{p_b(K)q(\beta_{K+1}, a_{K+1}, b_{K+1} | K)},$$

where $q'(\beta, a, b)$ is the proposal for the death step. In particular, $q'(\beta_{K+1}, a_{K+1}, b_{K+1} | K+1)$ is the probability of proposing to delete $\{\beta_{K+1}, a_{K+1}, b_{K+1}\}$ given that the current model has $K+1$ wavelets.

The Jacobian here is 1. Using Green's birth and death probabilities, these will cancel with the prior ratio and the acceptance rate becomes:

$$AR = LR \times \frac{p(\beta_{K+1}, a_{K+1}, b_{K+1})(K+1)q'(\beta_{K+1}, a_{K+1}, b_{K+1} | K+1)}{q(\beta_{K+1}, a_{K+1}, b_{K+1} | K)}. \quad (\text{A.2})$$

Notice that with the normal prior for β as in (3.3), the full posterior for β_{K+1} is also normal

$$p(\beta_{K+1} | \beta_{1:K}, a_{1:(K+1)}, b_{1:(K+1)}, \mathbf{Y}) \sim \mathbf{N}(\hat{\beta}, \hat{\sigma}^2_\beta), \quad (\text{A.3})$$

where

$$\hat{\sigma}^2_\beta = \left(\frac{1}{ca^{-\delta}} + \frac{\psi_{a_{K+1}, b_{K+1}} \psi'_{a_{K+1}, b_{K+1}}}{\sigma^2} \right)^{-1},$$

and

$$\hat{\beta} = \frac{\hat{\sigma}^2_\beta}{\sigma^2} \psi_{a_{K+1}, b_{K+1}}'(\mathbf{Y} - \hat{f}),$$

This normal proposal can also apply to heavy-tailed priors where the posterior for β does not have a close-form.

Similarly, the acceptance rate for a death step is:

$$AR = LR \times \frac{q(\beta_k, a_k, b_k | K-1)}{p(\beta_k, a_k, b_k)Kq'(\beta_k, a_k, b_k | K)}, \quad (\text{A.4})$$

where $q'(\beta_k, a_k, b_k | K)$ is given in (4.2).

In an update step, we randomly pick an index k from $\text{Unif}(1 : K)$, and propose a scale a_k and location b_k from a random-walk proposal and propose the wavelet coefficient β_k from (A.3) so that

$$q(\tilde{\beta}_k, \tilde{a}_k, \tilde{b}_k) = p(\beta_k | \beta_{-k}, a_{1:K}, b_{1:K}, \mathbf{Y}) \mathbf{N}([\tilde{a}_k, \tilde{b}_k]; [a_k, b_k], [\sigma_a^2, \sigma_b^2]^T I_2). \quad (\text{A.5})$$

The second part cancels the reverse proposal so that the acceptance rate

$$AR = LR \times \frac{p(\tilde{\beta}_k, \tilde{a}_k, \tilde{b}_k)}{p(\beta_k, a_k, b_k)} \times \frac{p(\beta_k | \beta_{-k}, a_{1:K}, b_{1:K}, \mathbf{Y})}{p(\tilde{\beta}_k | \beta_{-k}, a_{1:K}, b_{1:K}, \mathbf{Y})}. \quad (\text{A.6})$$

The reversible jump algorithm goes as follows:

1. Initially, select K_0 wavelet coefficients and scale and location parameters $\{\beta, a, b\}_0$.
2. Find the mean estimates $f(x|\{\beta, a, b\}_0)$.
3. Generate a uniform (0,1) random number u ,
 - (a) If $u < p_b(K)$, perform the birth step.
 - (b) If $p_b(K) < u < p_b(K) + p_d(K)$, perform the death step.
 - (c) If $u > p_b(K) + p_d(K)$, perform the update step.
4. Update σ^2 by Gibb Sampling:

$$\sigma_{\text{new}}^2 \sim IG(n/2, 2/SSE)$$

where $SSE = \sum_{i=1}^n (Y_i - \hat{f}(x_i))^2$.

5. Repeat steps 2-4.

References

- F. Abramovich, T. Sapatinas, and B. W. Silverman. Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields*, 117: 133–144, 2000.
- Y. Baraud. Confidence balls in Gaussian regression. *Ann. Stat.*, 32:528–551, 2004.
- J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, 1995.

- N. Brinkman. Ethanol – a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions*, 90:1410–1424, 1981.
- S. Brooks and P. Giudici. Markov chain monte carlo convergence assessment via two-way analysis of variance. *Journal of Computational and Graphical Statistics*, 9(2):266–285, 2000.
- S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.
- S. S. Chen, D. L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- M. Clyde, G. Parmigiani, and B. Vidakovic. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85:391–401, 1998.
- R. R. Coifman and Y. Meyer. Orthonormal wave packet bases. Preprint, 1990.
- R. R. Coifman and M. Wickerhauser. Entropy-based algorithms for best-basis selection. *IEEE Transactions on Information Theory*, 38:713–718, 1992.
- R. R. Coifman, Y. Meyer, and M. Wickerhauser. Adapted waveform analysis, wavelet packets and applications. *ICIAM 1991, Proceedings of the Second International Conference on Industrial and Applied Mathematics*, pages 41–50, 1992.
- D. Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Stat.*, 21:903–923, 1993.
- C. M. Crainiceanu, D. Ruppert, R. J. Carroll, A. Joshi, and B. Goodner. Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computation and Graphical Statistics*, 16(2):265–288, 2007.
- I. Daubechies. *Ten Lectures on Wavelets*. Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia, 1992.
- D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Automatic Bayesian curve fitting. *J. R. Statist. Soc. B*, 60:333–350, 1998.

- D. Donoho and I. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26:879–921, 1998.
- D. Donoho, M. Elad, and M. Temlyakov. Stable recovery of sparse overcomplete representation in the presence of noise. *Information Theory, IEEE Transactions on*, 52:6–18, 2006.
- D. L. Donoho and M. Elad. Maximal sparsity representation via l_1 minimization. *Proc. Nat. Aca. Sci.*, 100:2197–2202, Mar. 2003.
- D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
- P. Dutilleul. An implementation of the “algorithme á trous” to compute the wavelet transform. In J.-M. Combes, A. Grossman, and P. Tchamitchain, editors, *Wavelets: Time frequency methods and phase space*, Inverse problems and theoretical imaging, pages 298–304. Springer-Verlag, Berlin, 1989.
- A. Gelman and D. P. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.
- P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.
- K. Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhäuser, Boston, 2001.
- I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Stat.*, 32:1594–1649, 2004.
- I. M. Johnstone and B. W. Silverman. Empirical Bayes selection of wavelet thresholds. *Ann. Stat.*, 33:1700–1752, 2005.
- R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Ass.*, 90(431): 928–934, 1995.
- J. Kovačević and A. Chebira. Life beyond bases: The advent of frames (part i). *IEEE signal Rrocessing Magazine*, 24(4):86–104, 2007.

- M. Lewicki and T. Sejnowski. Learning overcomplete representations. *Neuron Computation*, 1998.
- S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE trans. on Pratt. Anal. Mach. Intell.*, 11:674–693, 1989a.
- S. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 315:69–87, 1989b.
- S. Mallat. *A wavelet tour of signal processing*. Academic Press, second edition, 1998.
- S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Information Theory*, 41:3397–3415, 1993.
- G. Nason. Choice of wavelet smoothness, primary solution and threshold in wavelet shrinkage. *Statistics and Computing*, 12:219–227, 2002.
- G. P. Nason and B. W. Silverman. The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pages 281–300. Springer-Verlag, New York, 1995.
- B. Vidakovic. *Statistical Modeling by Wavelets*. Computational & Graphical Statistics. John Wiley & Sons, New York, NY, 1999. ISBN 0-471-29365-2.
- M. Wickerhauser. *Adapted wavelet analysis from theory to software*. Wellesley, 1994.
- P. J. Wolfe, S. J. Godsill, and W.-J. Ng. Bayesian variable selection and regularization for time-frequency surface estimation. *J. R. Statist. Soc. B*, 66:575–589, 2004.

Channing Laboratory, Brigham and Women’s Hospital, Harvard Medical School

E-mail: jen-hwa.chu@channing.harvard.edu

Department of Statistical Science, Duke University

E-mail: clyde@stat.duke.edu

Department of Statistics, University of Illinois at Urbana-Champaign

E-mail: liangf@uiuc.edu