Statistica Sinica **19** (2009), 1419-1438

# BAYESIAN FUNCTION ESTIMATION USING CONTINUOUS WAVELET DICTIONARIES

Jen-Hwa Chu, Merlise A. Clyde and Feng Liang

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

*Abstract:* We present a Bayesian approach for nonparametric function 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 may lead to sparser representations. The price for this flexibility is the computational challenge of searching over an infinite number of potential dictionary elements. We develop a novel reversible jump Markov chain Monte Carlo algorithm which utilizes local features in the proposal distributions to improve computational efficiency, and which leads to better mixing of the Markov chain. Performance comparison in terms of sparsity and mean squared 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:* Bayesian inference, nonparametric regression, overcomplete dictionaries, reversible jump Markov chain Monte Carlo, stochastic expansions, wavelets.

## 1. Introduction

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

$$Y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2), \tag{1.1}$$

measured with independent and identically distributed (i.i.d.) 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 (1989) and Donoho and Johnstone (1998)). 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. Given a wavelet function $\psi(x)$, if

$\psi_{jk}(x) \equiv 2^{j/2}\psi(2^j x - k)$, $j, k \in \mathbb{Z}$, then the $\psi_{jk}$'s form an orthonormal basis for $L_2$ functions and any $L_2$ function may be represented as

$$f(x) = \sum_{j,k} \theta_{jk}\psi_{jk}(x) \tag{1.2}$$

with coefficients $\{\theta_{jk}\}_{j\in\mathbb{Z},k\in\mathbb{Z}}$ (Vidakovic (1999, Sec. 3.5.2)). For equally-spaced locations $x_1, \ldots, x_n$, the coefficients $\theta_{jk}$ may be computed efficiently via the so-called Cascade algorithm (Mallat (1989)).

As the structure of the function $f$ is unknown in practice, 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, Donoho and Saunders (1998), Coifman, Meyer and Wickerhauser (1992), Lewicki and Sejnowski (1998), Donoho and Elad (2003), Wolfe, Godsill and Ng (2004) and Donoho, Elad and Temlyakov (2006)). Examples of overcomplete dictionaries include translation-invariant wavelet transforms (Dutilleux (1989) and Nason and Silverman (1995)), frames (Wolfe et al. (2004) and 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 greedy algorithms designed to search for one "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 (Wolfe et al., 2004). Because of the stochastic nature of MCMC algorithms, several "optimal" representations may be identified. In this paper, we propose a nonparametric Bayesian approach for function estimation using continuous wavelet dictionaries (CWD). As opposed to orthonormal wavelet basis functions that are subject to dyadic constraints on their locations and scales, as at (1.2), 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 novel reversible jump MCMC (RJ-MCMC) algorithm that utilizes "local" information to construct efficient proposal distributions in order to make inferences about the unknown function $f$.

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 for CWD. In Section 4 we describe posterior inference by means of a RJ-MCMC sampling scheme and discuss various estimates of $f$, including point estimates and a new method for simultaneous credible bands. In Section 5 we present results from simulation studies that show that our method leads to better performance in terms of sparsity and mean squared error. We then illustrate our method using non-equally spaced data. Concluding remarks are given in Section 6.

## 2. The Model

Suppose $\phi$ and $\psi$ 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)). In the continuous wavelet dictionary setting, we expand the mean function of (1.1) as

$$f(x) = f_0(x) + \sum_{k=0}^{K} \beta_{\lambda_k} \psi_{\lambda_k}(x), \tag{2.1}$$

where $f_0$ is a fixed scaling function representing coarse-scale features given by

$$f_0(x) = \sum_{m=1}^{M} \eta_m \phi_{\lambda_m}(x), \tag{2.2}$$

with $\phi_{\lambda_m}(t) \equiv a_m^{1/2}\phi(a_m(t - b_m))$ for some finite set of indices $\lambda_m = (a_m, b_m) \in (0, a_0) \times [0, 1]$, $i = 1, \ldots, M$. The second term in equation (2.1) describes the fine-scale "details" of the function with $\psi_{\lambda_k}(t) \equiv \sqrt{a_k}\psi(a_k(t - b_k))$ for $\lambda_k = (a_k, b_k) \in [a_0, \infty) \times [0, 1]$ constructed by scaling $(a_k)$ and shifting locations $(b_k)$ of the mother wavelet $\psi(\cdot)$. In this representation, in addition to the scaling and location parameters in $\lambda_k$, the number of wavelet elements $K$ from the CWD in the expansion is also an unknown parameter. We use the convention that when $k = 0$, $\beta_{\lambda_0} \equiv 0$, so that if $K = 0$ we obtain the null or zero function.

Motivated by Bayesian approaches, Abramovich, Sapatinas and Silverman (2000) studied properties of stochastic expansions in overcomplete wavelet dictionaries and conditions such that the functions would belong to a particular Besov space. Their stochastic expansions motivate the choice of prior distributions given in the next section.

## 3. Prior Distributions

The unknown parameters in the model (2.1) are the error variance $\sigma^2$, the number of wavelet elements $K$ and, given $K$, the corresponding coefficients and location-scale indices $\{\beta_{\lambda_k}, \lambda_k\}_{k=1}^{K}$ for each wavelet component. Following

Clyde and George (2000), we adopt a non-informative reference prior for $\sigma^2$, $p(\sigma^2) \propto 1/\sigma^2$. Although it is improper, it is easy to show that the corresponding posterior distribution is proper for $n \geq 2$. As in Abramovich et al. (2000), we may view $\{\beta_k, a_k, b_k\}_{k=1}^K$ as a realization of a compound Poisson process, leading to a Poisson distribution for $K$ and, conditional on $K$, the triplets $\{\beta_{\lambda_k}, a_k, b_k\}$ being independent and identically distributed for $k = 1, \ldots, K$.

### 3.1. Prior distribution for scaling parameters $a$

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

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

The lower bound $a_0$ corresponds to the coarsest-scale component allowable in the function. We have used $a_0 = 1$, which corresponds to the coarsest level wavelet ($j = 0$) in a regular discrete wavelet transform; for the more general setting in Abramovich et al. (2000), where the Poisson mean for $K$ may be infinite, $a_0$ needs to be larger than twice the support of the wavelet function $\psi$.

The upper bound $a_1$ corresponds to the smallest finest-scale component. Theoretically, $a_1$ may be infinity to span the whole space as in Abramovich et al. (2000). In practice, however, extremely large $a$ may lead to narrowly supported wavelet functions that have little or no effect on the likelihood. To see this, suppose we have 1,024 equally-spaced data points and use a mother wavelet with support of length 1. If we set $a_1 > 1,024$ , 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 could not be estimated effectively from the data and might lead to poor out-of-sample predictive properties. For this reason, we set an upper bound for $a$ so that the wavelet functions have large enough support. These bounds depend on the spacing in the data and the support of the wavelet $\psi$, and are discussed in more detail in the examples in Section 5.

Finally, the hyperparameter $\zeta$ controls the intensity or relative number of fine-scale wavelet components in the function. If $\zeta$ is large, *a priori* we have relatively few fine-scale (spiky) components in the function, while if $\zeta$ is small, fine-scale components predominate. We set $\zeta = 1.5$ for all examples in Section 5 and found that that posterior results were not very sensitive to this choice.

### 3.2. Prior distribution for location parameters $b$

Abramovich et al. (2000) place a uniform distribution on the location parameters. We construct a prior distribution composed of a mixture of a discrete

uniform on the observed data location and a continuous uniform distribution on $[0, 1]$,

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

where $\delta_{x_i}(b)$ is a point mass at the data point $x_i$. We set $\pi = 1/2$, although one could place a prior distribution on $\pi$. 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 leads to a more efficient RJ-MCMC algorithm by a novel proposal distribution for dictionary elements constructed using information from residuals (discussed in detail in the next section). Notice that when $\pi = 1$ and $p(b)$ has support on data points only, we return to the non-decimated discrete wavelet setting, and when $\pi = 0$, we have the continuous uniform distribution from Abramovich et al. (2000).

### 3.3. Prior for coefficients $\boldsymbol{\beta_\lambda}$

Given the location and scale of a wavelet function, the prior distribution of the corresponding wavelet coefficient $\beta_\lambda$ is independent normal (Abramovich et al. (2000)):

$$\beta_\lambda \mid a \sim \mathsf{N}(0, ca^{-\delta}), \tag{3.3}$$

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 $\delta$ 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 $\delta$ is large, we shrink the fine scale (spiky) wavelets more, resulting in a smoother function, and vice versa. For all examples in Section 5 we set $\delta = 2$.

In the random expansion (2.1), if the number of elements $K$ is infinite almost surely (a.s.), normality of $\beta_\lambda$ is one of the conditions for $f$ to be well-defined and to belong to a certain function space (see discussion in Section 3.5 and Abramovich et al. (2000)). However, as we shall see, our prior distribution on $K$ implies that $K$ is finite (a.s.), thus the normality of $\beta_\lambda$ is not necessary. We may replace the normal distribution with a heavier-tailed prior for $\beta$, e.g., Laplace with a scale parameter that depends on $a$ the same way as in (3.3),

$$p(\beta \mid a) = \frac{1}{2\tau} \exp\left\{ - \frac{|\beta|}{\tau} \right\}, \qquad \tau^2 = \frac{ca^{-\delta}}{2}, \tag{3.4}$$

or other scale mixtures of normals. These heavy-tailed priors have been shown to have theoretical advantages over normal distributions, and may lead to greater sparsity and improved performance (Clyde and George (2000) and Johnstone and Silverman (2004)).

### 3.4. Prior distribution for $K$

The coefficients and indices $\{\beta_k, a_k, b_k\}_{k=1}^K$ in the stochastic expansion can be viewed as i.i.d. samples generated from a compound Poisson process. Our prior specification on $(\beta, a, b)$ induces an intensity measure $\gamma a^{-\zeta} p(\beta \mid a) p(b) d\beta\, da\, db$ on $\mathbb{R} \times [a_0, a_1] \times [0,1]$, where $\gamma$ is a non-negative scalar. This leads to a Poisson prior distribution for $K$ with mean given by

$$\mathbb{E}(K) = \gamma \iiint_{\mathbb{R}\times[a_0,a_1]\times[0,1]} a^{-\zeta} p(\beta \mid a) p(b)\, d\beta\, db\, da = \gamma\, c(a_0, a_1, \zeta),$$

where

$$c(a_0, a_1, \zeta) = \frac{1 - (a_0/a_1)^{\zeta-1}}{(\zeta - 1) a_0^{\zeta-1}}, \text{if } \zeta \neq 1; \quad \log \frac{a_1}{a_0}, \text{ otherwise,}$$

is always finite as long as $a_1 < \infty$. Because the mean and variance of the Poisson distribution are equal, the distribution for $K$ may be too concentrated about the mean *a priori*. By assigning $\gamma$ a Gamma prior distribution, $\mathsf{G}(\alpha_\gamma, \beta_\gamma)$, with rate $\beta_\gamma$, we obtain a negative binomial prior distribution $\mathsf{NB}(s, q)$ for $K$:

$$p(k|s,q) = \binom{s+k-1}{k} q^s (1-q)^K \quad k = 0, 1, \ldots, \tag{3.5}$$

where $s = \alpha_\gamma$ and $q = \beta_\gamma/(\beta_\gamma + c(a_0, a_1, \zeta))$. This provides additional flexibility and over-dispersion compared to the Poisson distribution. The hyperparameters $s$ and $q$ may be determined by the gamma hyperparameters $(\alpha_\gamma, \beta_\gamma)$ and $c(a_0, a_1, \zeta)$, or specified directly. We have chosen the hyperparameters $s$ and $q$ 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 solved to obtain the values of $s$ and $q$ (and thus $\alpha_\gamma$ and $\beta_\gamma$).

### 3.5. Function spaces

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 be in the same Besov space (a.s.) as the mother wavelet. If the number of wavelet elements $K$ is finite (a.s.), for example if $K$ has a Poisson or Negative Binomial prior distribution with a finite mean, (2.1) will have a finite number of elements (a.s.) and $f$ will belong to the same Besov space (a.s.) as

does the mother wavelet function $\psi$ for any reasonable choice of the probability distribution for $\beta_\lambda$. In our case, as long as $a_1 < \infty$ this is always ensured. However, if $a_1 = \infty$ and $\zeta \leq 1$, extra conditions that involve $\zeta$ and prior distributions on $\beta$ are needed for the random function $f$ to be well-defined (see Abramovich et al. (2000) for more details). As a practical matter, since we can only deal with finite representations from a computational standpoint, restricting attention to the Poisson or Negative Binomial with finite mean presents no problems.

## 4. Posterior Inference

For the purpose of this paper we first center the data $\mathbf{Y}$ and take $f_0(x) = 0$, so that $f(x)$ represents departures from the overall mean. In principle, for an $f_0(x)$ given by (2.2), it is straightforward to update the coefficients $\eta_m$ via a Gibbs step using a multivariate normal distribution (assuming normal or reference prior distributions).

In a standard discrete wavelet transformation, where $a$ and $b$ have dyadic constraints and the data are on an equally spaced grid, the empirical wavelet coefficients, which are sufficient statistics, may be obtained through filters without evaluating the wavelet functions $\psi_\lambda$ directly. In a CWD, the dictionary elements do not have the tree-like structure needed for the cascade algorithm, therefore it is necessary to evaluate each of the wavelet functions $\psi_\lambda(x)$. The wavelet functions in Daubechies' family (except the Haar wavelet) and many other families have no explicit representations. The Daubechies-Lagarias local pyramid algorithm (Vidakovic (1999, Sec. 3.5.4)), however, does enable one to evaluate $\psi_\lambda(\cdot)$ at an arbitrary point with preassigned precision using the wavelet filter coefficients. This involves a product of $m$ matrices of size $(2N-1) \times (2N-1)$, where $N$ is the number of vanishing moments of $\psi(\cdot)$, and $m$ is chosen to obtain the desired degree of accuracy. We have found that taking $m = 5$ for the `la8` wavelet ($N = 4$) gives a desired level of accuracy, while balancing computational costs.

The task of searching over a continuous dictionary with infinitely many models can be extremely challenging. Since the dimensionality of each model, and the model parameters, varies, we develop a reversible jump Markov chain Monte Carlo 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 (increasing $K$ by 1), a death step where we delete a wavelet element (decreasing $K$ by 1), and an update step where we move a wavelet element by changing the index $\lambda_k$ and coefficient $\beta_k$ (leaving $K$ unchanged). For RJ-MCMC algorithms, good proposal distributions are important to speed up convergence. For example, proposing a "birth" of a new wavelet dictionary element from the prior distribution on $(\beta, a, b)$ may simplify the calculation of the Metropolis Hastings ratio, but often results in

slow convergence when it does not 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 develop novel proposal distributions for the birth and death steps that involve residuals and coefficients so that we are more likely to target regions of the function where birth/deaths would successfully improve the function estimate. These proposal distributions can improve convergence in practice since a successful birth is more likely where the residual is large, and removing a wavelet for which the coefficient is small will not change the likelihood dramatically. Because we propose a birth/update for one wavelet at a time, these updates do not require any matrix inversion. Note that we only need to evaluate a new wavelet function via the Daubechies-Lagarias algorithm in the case of a birth step or update step that leads to a new $\lambda$. As we discuss later, this can provide significant computational savings compared to methods for fixed dimensional overcomplete dictionaries. We provide details of the algorithm in the on-line supplement available at `http://www.stat.sinica.edu.tw/statistica`.

After $T$ MCMC iterations post burn-in, each collection of the parameters $\{\boldsymbol{\beta}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, K^{(t)}\}_{t=1}^{T}$ represents a sample from the joint posterior distribution, where $\boldsymbol{\beta}^{(t)} = (\beta_1^{(t)}, \beta_2, \ldots, \beta_{K^{(t)}}^{(t)})$, and $\mathbf{a}^{(t)}$ and $\mathbf{b}^{(t)}$ are defined similarly. At each iteration $t$ we obtain a posterior sample of $f(x)$ from $p(f \mid \mathbf{Y})$ by evaluating

$$\hat{f}^{(t)}(x) = \sum_{k=0}^{K^{(t)}} \beta_{\lambda_k^{(t)}}^{(t)} \psi_{\lambda_k^{(t)}}(x) \tag{4.1}$$

at the posterior draws $\{\boldsymbol{\beta}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, K^{(t)}\}_{t=1}^{T}$. These posterior samples of $f$ provide a full description of the posterior distribution of $f$ given the data $\mathbf{Y}$.

### 4.1. 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,

$$\hat{f}_{\mathrm{AVE}}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{f}^{(t)}(x), \tag{4.2}$$

where $T$ is the number of MCMC iterations after burn-in and $\hat{f}^{(t)}$ represents the estimate from the $t$th MCMC iteration given at (4.1).

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:

$$\hat{f}^{*} = \underset{\hat{f}^{(t)} t \in \{1,\ldots,T\}}{\operatorname{argmin}} \sum_{i=1}^{n} \left[ \hat{f}_{\mathrm{AVE}}(x_i) - \hat{f}^{(t)}(x_i) \right]^2. \tag{4.3}$$

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

$$\hat{\boldsymbol{\beta}}^{(t)} = \mathrm{E}\,(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, K^{(t)}) = (\mathbf{A}^{(t)'}\mathbf{A}^{(t)} + c^{-1}D(\mathbf{a}^{\zeta}))^{-1}\mathbf{A}^{(t)'}\mathbf{Y},$$

where $\mathbf{A}^{(t)}$ is an $n \times K^{(t)}$ matrix with $A_{ij}^{(t)} \equiv \psi_{\lambda_j^{(t)}}(x_i)$ and $D(\mathbf{a}^{\zeta})$ is a $K^{(t)} \times K^{(t)}$ matrix with $(a_k^{(t)})^{\zeta}$ on the diagonal, and zero otherwise.

### 4.2. Simultaneous credible bands for $f$

We propose a new method constructing a credible region that 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 defines an envelope along $x$,

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

such that $\Pr(f \in C \mid \mathbf{Y}) \ge 1-\alpha$. In practice, the posterior probability is approximated by using the empirical distribution based on MCMC samples and the condition "for all $x$" is approximated by "for a fine grid $(x^1, \ldots, x^m)$ on the range of $x$", where the $x^j$s could be observed data points, but not necessarily. Our method for constructing credible bands is based on an $L_2$ ball of errors, motivated by work of Cox (1993) and Baraud (2004).

First we start with the ball

$$\{f : \|f - \hat{f}_{\mathrm{AVE}}\|_{\Sigma} \le D_{\alpha}\}, \tag{4.4}$$

where $\|a\|_{\Sigma} = a'\Sigma^{-1}a$ is the $L_2$ norm normalized by the estimated covariance matrix $\Sigma$ of the $f(x^i)$'s, and $D_{\alpha}$ 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 (4.4). 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 (4.2): $D^{(t)} = \|\hat{f}^{(t)} - \hat{f}_{\mathrm{AVE}}\|_{\Sigma}$.
2. Calculate $D_{\alpha}$, the $100(1-\alpha)\%$ quantile of $D^{(t)}$.
3. Let $T_{\alpha}^{D}$ be the collection of indices of MCMC samples of $f$ of which the distance to $\hat{f}_{\mathrm{AVE}}$ is below $D_{\alpha}$: $T_{\alpha}^{D} = \{t : 1 \le t \le T, \quad D^{(t)} \le D_{\alpha}\}$.

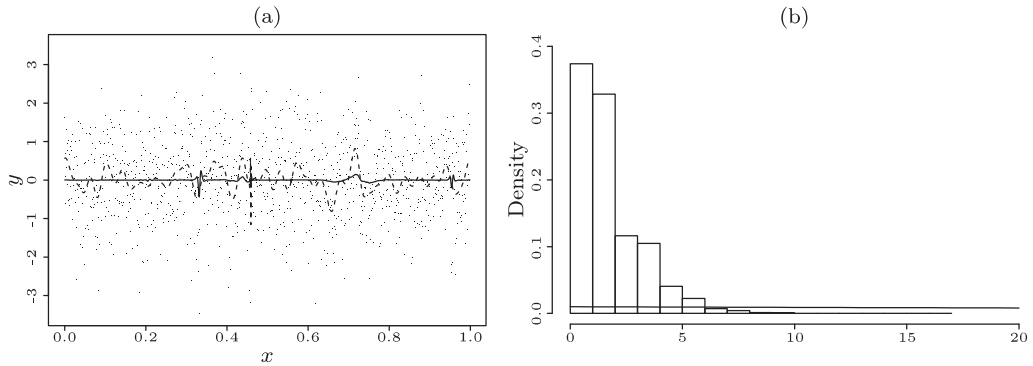


Figure 1. (a) EBayes (Johnstone and Silverman (2004)) with Laplace prior distribution (dash line), and CWD with the Gaussian prior distribution posterior mean (solid line) fits of the null function, and (b) the posterior histogram for $K$ (bars) and the $\mathsf{NB}(1, 0.01)$ prior distribution (solid line).

4. Our simultaneous credible region $C$ is the minimum hyper-rectangle that contains all the posterior samples in $T_\alpha^D$, namely,

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

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

It is straight forward to show that the posterior coverage of $C$ is at least $100(1-\alpha)\%$.

## 5. Examples

We illustrate our Bayesian CWD method and compare it to other approaches in the literature in a series of simulation studies and an example. Throughout, we used the `la8` wavelet (Daubechies least asymmetric family with four vanishing moments), the default in `R`, with $a_0 = 1$ except where noted. The hyperparameters were set to $\delta = 2$ and $\zeta = 1.5$ as discussed previously. The prior for the number of coefficients $K$ was negative binomial with $s = 1$ and $q = 0.01$, which corresponds to 0.01 probability of the null model and 95% percentile at $K = 298$. This prior distribution on $K$ is relatively flat and covers a wide range of possible models (see Figure 1(b)).

### 5.1. Null simulation

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 applied the CWD method using the Gaussian prior distribution on

$\beta_\lambda$ to a data set of $n = 1,024$ observations generated from the null function $f(x) = 0$ with noise $\sigma = 1$. We set $a_1 = 500$ (approximately $n/2$) so that each wavelet was supported by an adequate number of data points. The results from the 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 EBayes (Johnstone and Silverman (2005a)), an empirical Bayes method, carried out using their `R` package `EBayesThresh` (Johnstone and Silverman (2005b)). Using the best combination of settings from Johnstone and Silverman (2005a), we applied EBayes using the Laplace prior distribution with the translation-invariant wavelet transform, an overcomplete representation. We used `la8` with $J = 4$ (the default), where $J$ is the number of levels in the multi-resolution expansion, leading to $nJ$ wavelet coefficients. EBayes shrinks and thresholds $nJ$ wavelet coefficients, while keeping the $n$ scaling coefficients without any shrinkage or thresholding. In this example, even though EBayes thresholded all but one of the wavelet coefficients to zero, the estimate still appears "bumpy" due to the included 1,024 scaling coefficients.

### 5.2. Wavelet test functions

We carried out a second simulation study using four standard test functions from Donoho and Johnstone (1994): `bumps`, `blocks`, `doppler`, and `heavisine`. For each test function, 100 replications were generated with two levels of SNR (signal-to-noise ratio), 3 (high) and 7 (low). In each replicate, the data were simulated at 1,024 equally spaced points in $[0, 1]$.

For the CWD method, we used the default choice of wavelet in `R` (`la8`) for all functions except for `blocks`, where we used the Haar wavelet. Unlike many other wavelet methods, we did not assume a boundary correction here, since some of the functions (e.g. `doppler`) are clearly not periodic. We set the lower bound for the scale parameter at $a_0 = 1$ and the upper bound at $a_1 = 500$, roughly $n/2$, as in the null model simulation. The following results are based on one million iterations of the RJ-MCMC algorithm, with a burn-in period of 500,000 iterations and the remaining iterations used for posterior inference.

We compared the CWD with EBayes (using the same settings as described in the null simulation in Section 5.1) using the average mean squared error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i) - f(x_i)\right]^2. \tag{5.1}$$

For CWD we calculated point estimates of $f$ based on the posterior mean (PM) in (4.2) for both the normal and Laplace prior distributions for $\boldsymbol{\beta}$, along with the model selection (MS) estimate from (4.3) with the normal prior. Figure 2 shows

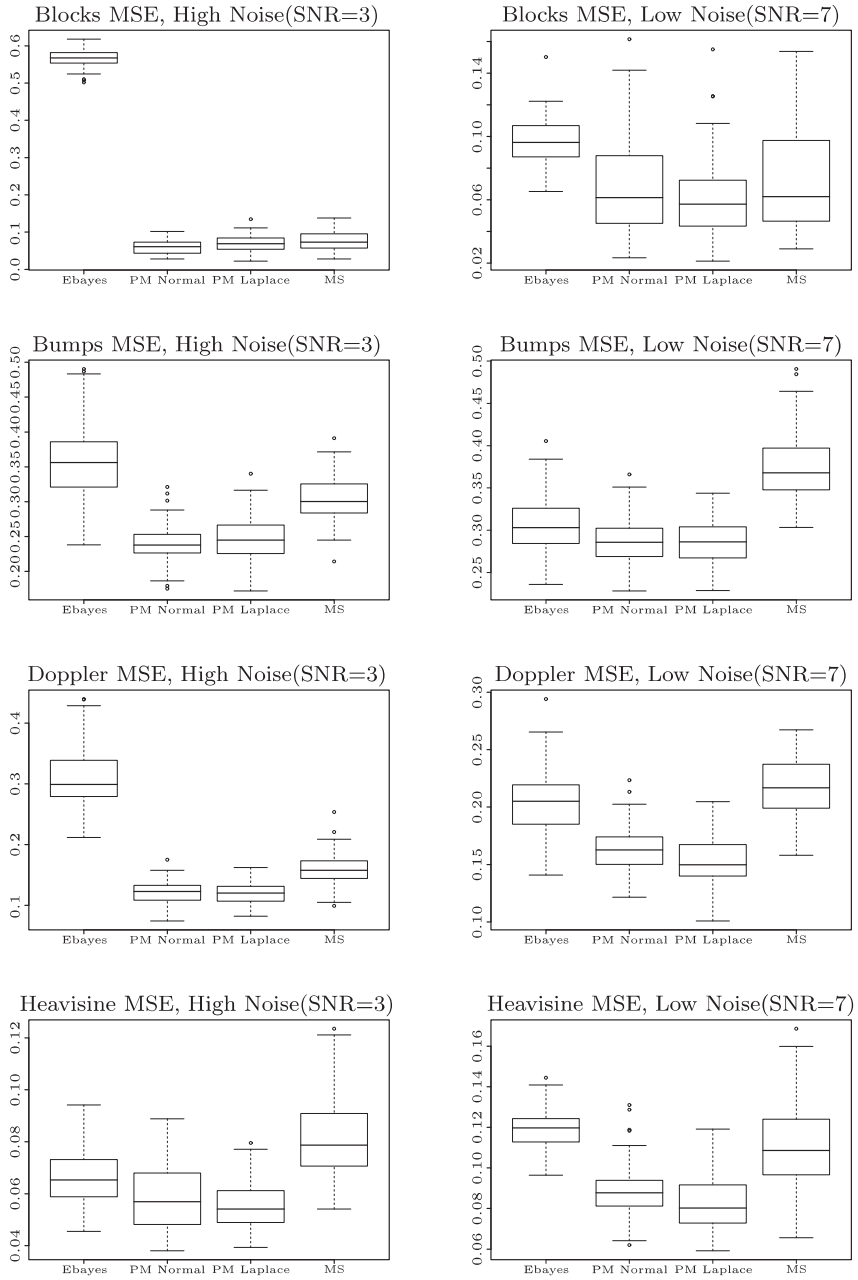


Figure 2. Box plots for mean squared error for the four test functions using the EBayes method of Johnstone and Silverman (2005a), and the continuous wavelet dictionary (CWD) method with the posterior mean (PM) under the Gaussian and Laplace prior distributions, and with model selection (MS) under the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

that the PM estimates have smaller MSE than EBayes for all four functions and for both noise levels. The heavy-tailed Laplace prior with the CWD led to an additional reduction in MSE except for in `bumps`, where it is roughly the same as the CWD with Gaussian prior distributions. Using model selection under squared error loss, we find that the EBayes estimate is better only for `bumps` and `doppler` in the low noise scenario and for `heavisine` in the high noise scenario. When we compared the number of non-zero coefficients in $\hat{f}$, however, the CWD method clearly gave a much sparser representation than EBayes (See Figure 3). We note that the EBayes summaries for $K$ do not include the 1,024 coefficients from the scaling function, which are not shrunk or thresholded. The on-line supplement includes figures illustrating the fitted functions for the two methods.

### 5.3. Ethanol example

When the data are not equally spaced, EBayes and other methods based on the regular discrete wavelet transform cannot be directly applied. Several wavelet-based methods for non-equally spaced data have been proposed, including Kovac and Silverman (2000), Nason (2002) and Amato, Antoniadis and Pensky (2006). The CWD based method can also be applied directly to non-equally spaced data sets without interpolation or imputing missing data. To illustrate this point, we applied our method to a well-studied data set, the `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 applied our CWD method using the Gaussian and Laplace prior distributions with 4, 8, and 10 vanishing moments of the least asymmetric Daubechies' wavelets (`la8, la16` and `la20` in `R`). We set the upper bound $a_1$ to 100, which is roughly one half of the inverse of the median distance between observations. Because the sample size is smaller, we used $m = 10$ in the Daubechies-Lagarias algorithm. We show the estimated posterior mean curves defined in (4.2) in Figure 4, and the 95% simultaneous credible band with `la16` in Figure 5 using the Gaussian prior distribution on the coefficients. In calculating the $L_2$ distance, the posterior samples $f^{(t)}$ were evaluated at 512 equally-spaced grid points covering the range of the data using the Daubechies-Lagarias algorithm.

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 that result, we performed a leave-one-out cross validation study

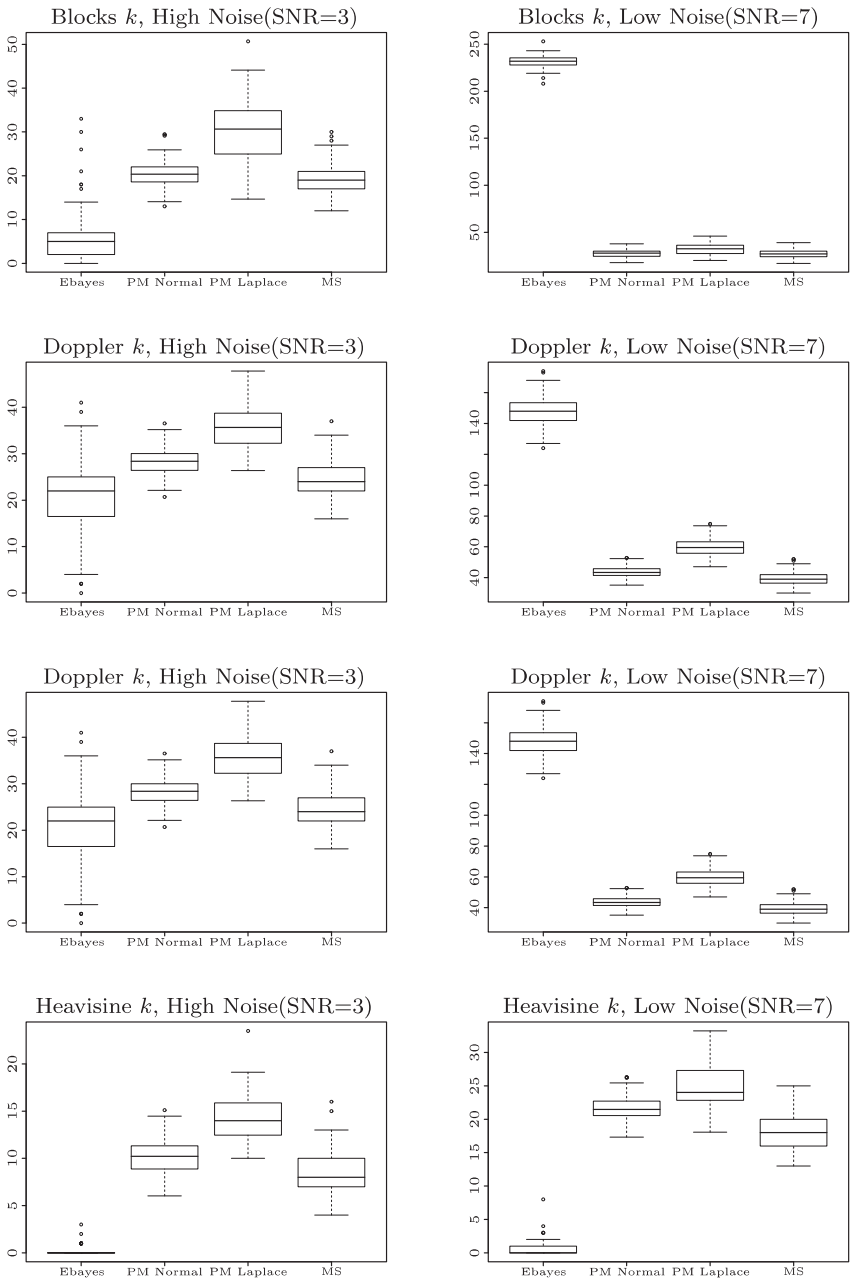


Figure 3. Box plots for the number or average number (for PM) of non-zero wavelet coefficients for the four test functions using the EBayes method Johnstone and Silverman (2005a), and continuous wavelet dictionary (CWD) method with the posterior mean (PM) under the normal and Laplace priors, and with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

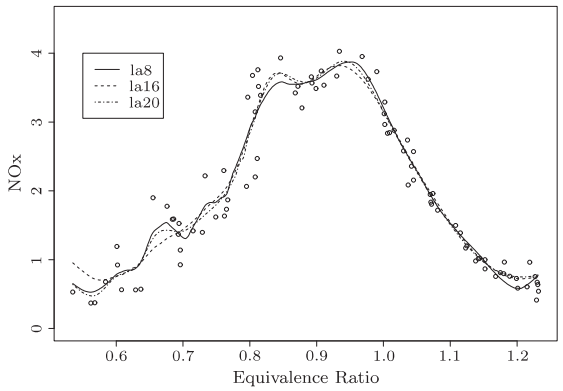


Figure 4. The posterior mean using the CWD with the Gaussian prior distribution for the ethanol data from Brinkman (1981).

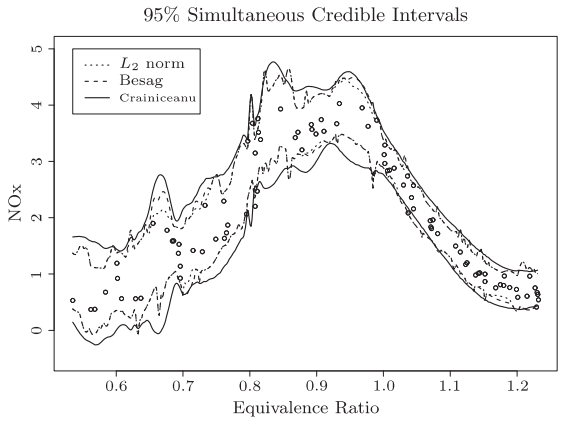


Figure 5. Simultaneous credible bands ($\alpha = 0.05$) for `la16`.

and calculated the cross validation score

$$\text{CV-score} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}_{(-i)}(x_i) - Y_i\right]^2, \tag{5.2}$$

where $\hat{f}_{(-i)}$ is the estimated posterior mean of $f$ using all of the data except the $i$th point. With no attempts to optimize our hyperparameters, the CV score from CWD with `la16` ranks second out of the 60 combinations reported in Nason (2002), and the estimated function looks 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 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.

We also constructed 95% simultaneous credible bands for $f(x)$ with `la16` using methods of Besag, Green, Higdon and Mengersen (1995) and Crainiceanu, Ruppert, Carroll, Joshi and Goodner (2007). Our credible band actually covers about 95.75% of the posterior sample; other methods give comparable results in this respect. From Figure 5 we can see that while all credible bands take a similar shape, our $L_2$ loss based credible region is narrower than Crainiceanu's, although the difference between our credible band and Besag's is negligible.

We calculated the area of the credible region by numerical integration:

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

where $x^1, \ldots, x^n$ are the grid locations where the functions are being evaluated. The $L_2$ based method had the smallest area of 0.69, while Besag's method gave 0.70, and Crainiceanu was the largest with 0.92. In general with the same coverage rate a smaller credible region is more desirable.

## 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 the 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 showed 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 permits wavelet methods to be used in a greater variety of applications. We also introduced a new approach for constructing simultaneous credible bands in the overcomplete setting that appears to give narrower bands than other existing methods.

The price to pay for increased flexibility is computational cost. The major cost is in the wavelet evaluation using the Daubechies-Lagarias (DL) algorithm. For the examples in the wavelet test function simulation study with $n = 1,024$, the running time was about 1-1.5 hours per function when the DL algorithm was used with the `la8` wavelet (all running times are based on a single Intel 2.6 gigahertz processor). In contrast, the running time was only about 5-6 minutes using the Haar wavelet since the DL algorithm is not required for evaluation of the Haar wavelet. Running times for other wavelets that can be computed analytically should be on the order of 5-6 minutes, as with the Haar wavelet.

While the CWD is much more computationally intensive than EBayes (which runs in seconds), the reductions in MSE were significant, with the MSE under EBayes ranging from 15% to 60% higher than the MSE using the CWD (using the same wavelet to generate the dictionary). The running times for the non-equispaced ethanol example with $n = 88$ and 100,000 iterations of the RJ-MCMC algorithm using wavelets `la8, la16` and `la20` were 3, 19 and 40 minutes, respectively. These running times could be approximately cut in half by reducing $m$ to 5 (from $m = 10$) in the DL algorithm. In many practical applications, non-equispaced designs often have small sample sizes (i.e., a few hundred samples). Using `la8`, the current algorithm can run in about 10-20 minutes.

We note that using other methods for overcomplete representations based on a fixed dimensional dictionary, such as the Dantzig selector (Candes and Tao (2007)), LASSO (Tibshirani (1996)) or greedy methods (Barron, Cohen, Dahmen and DeVore (2008)), will be just as expensive (more or less) if the DL algorithm is used to precompute the wavelet dictionary elements on a fine grid. For example, using a $1,000 \times 1,000$ grid for $a$ and $b$ will be the same order of magnitude as running one million iterations of our RJ-MCMC algorithm. However, the correlation between the dictionary elements in the $1,000 \times 1,000$ grid (or even coarser $100 \times 100$ grid) of $a$ and $b$ led us to empirical failure when we tried to use the Dantzig selector. The high degree of correlation among the dictionary elements violates theoretical conditions needed for optimality of many algorithms in the "large $p$ small $n$" paradigm (Candes and Tao (2007), Donoho and Elad (2003) and Donoho et al. (2006)). Stochastic search variable selection algorithms have also been used in the overcomplete setting (Wolfe et al. (2004)); our approach may be viewed as a limiting version of such algorithms for continuous dictionaries. Stochastic search algorithms for finite dictionaries typically involve only birth and death steps for adding or deleting a dictionary element. When there are high correlations among dictionary elements, it is often difficult to adequately explore the multiple modes in the model space (due to the over-completeness of the dictionary) using only birth and death steps. Because of the update step in our RJ-MCMC algorithm, we are able to smoothly transition from one dictionary element to a new one, making it easier to explore multiple modes of the posterior distribution.

As a final note, the calculations for the simulations studies and examples using the DL algorithm were done using an `R` package under development by the first author. We expect that further improvements in running time will be possible by considering more efficient $O(n)$ algorithms (e.g., Muñoz, Ertlé and Unser (2002)) to evaluate wavelet functions at arbitrary points. Incorporating these algorithms could significantly improve the efficiency of our code for Bayesian function estimation using continuous wavelet dictionaries, and is a promising direction for future research.

**Acknowledgement**

The authors would like to thank Brani Vidakovic for suggesting the Daubechies-Lagarias algorithm for evaluating continuous wavelets. The authors would also like to thank an associate editor and anonymous referees for extremely helpful comments and suggestions. The authors acknowledge support of the National Science Foundation (NSF) through grants DMS–0342172 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.

## References

Abramovich, F., Sapatinas, T. and Silverman, B. W. (2000). Stochastic expansions in an overcomplete wavelet dictionary, *Probab. Theory Related Fields* **117**, 133-144.

Amato, U., Antoniadis, A. and Pensky, M. (2006). Wavelet kernel penalized estimation for non-equispaced design, *Statist. Comput.* **16**, 37-55.

Baraud, Y. (2004). Confidence balls in Gaussian regression, *Ann. Statist.* **32**, 528-551.

Barron, A. R., Cohen, A., Dahmen, W. and DeVore, R. A. (2008). Approximation and learning by greedy algorithms, *Ann. Statist.* **36**, 64-94.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems, *Statist. Sci.* **10**(1), 3-41.

Brinkman, N. (1981). Ethanol – a single-cylinder engine study of efficiency and exhaust emissions, *SAE Transcations* **90**, 1410-1424.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$, *Ann. Statist.* **35**(6), 2313-2351.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998). Atomic decomposition by basis pursuit, *SIAM, J. Sci. Comput.* **20**, 33-61.

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets, *J. Roy. Statist. Soc. Ser. B* **62**, 681-698.

Coifman, R. R. and Meyer, Y. (1990). Orthonormal wave packet bases. Preprint.

Coifman, R. R., Meyer, Y. and Wickerhauser, M. (1992). Adapted waveform analysis, wavelet packets and applications, *ICIAM 1991, Proceedings of the Second International Conference on Industrial and Applied Mathematics*, 41-50.

Coifman, R. R. and Wickerhauser, M. (1992). Entropy-based algorithms for best-basis selection, *IEEE Trans. Inform. Theory* **38**, 713-718.

Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression, *Ann. Statist.* **21**, 903-923.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J. Joshi, A. and Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors, *J. Comput. Graph. Statist.* **16**(2), 265-288.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, number 61 *in* CBMS-NSF Series in Applied Mathematics, *SIAM*, Philadelphia.

Donoho, D., Elad, M. and Temlyakov, M. (2006). Stable recovery of sparse overcomplete representation in the presence of noise, *IEEE Trans. Inform. Theory* **52**, 6-18.

Donoho, D. L. and Elad, M. (2003). Maximal sparsity representation via $l_1$ minimization, *Proc. Nat. Acad. Sci.* **100**, 2197-2202.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**, 425-455.

Donoho, D. and Johnstone, I. (1998). Minimax estimation via wavelet shrinkage, *Ann. Statist.* **26**, 879-921.

Dutilleux, P. (1989). An implementation of the "algorithme á trous" to compute the wavelet transform. In *Wavelets: Time frequency methods and phase space*, Inverse problems and theoretical imaging, (Edited by J.-M. Combes, A. Grossman and P. Tchamitchain), 298-304, Springer-Verlag, Berlin.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**, 711-732.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences, *Ann. Statist.* **32**, 1594-1649.

Johnstone, I. M. and Silverman, B. W. (2005a). EbayesThresh: R programs for empirical Bayes thresholding, *J. Statist. Soft.* **12**, 1-38.

Johnstone, I. M. and Silverman, B. W. (2005b). Empirical Bayes selection of wavelet thresholds, *Ann. Statist.* **33**, 1700-1752.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *J. Amer. Statist. Assoc.* **90**(431), 928-934.

Kovac, A. and Silverman, B. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding, *J. Amer. Statist. Assoc.* **95**, 172-183.

Kovačević, J. and Chebira, A. (2007). Life beyond bases: The advent of frames (part i), *IEEE signal Processing Magazine* **24**(4), 86-104.

Lewicki, M. and Sejnowski, T. (1998). Learning overcomplete representations, *Neuron Computation*.

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

Mallat, S. and Zhang, Z. (1993). Matching pursuit in a time-frequency dictionary, *IEEE Trans. Inform. Theory* **41**, 3397-3415.

Muñoz, A., Ertlé, R. and Unser, M. (2002). Continuous wavelet transform with arbitrary scales and O(N) complexity, *Signal Processing* **82**, 749-757.

Nason, G. P. (2002). Choice of wavelet smoothness, primary solution and thresholding wavelet shrinkage, *Statist. Comput.* **12**, 219-227.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In *Wavelets and Statistics of Lecture Notes in Statist* **103**, (Edited by A. Antoniadis and G. Oppenheim), 281-300. Springer-Verlag, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*, Computational & Graphical Statistics, John Wiley & Sons, New York, NY.

Wolfe, P. J., Godsill, S. J. and Ng, W.-J. (2004). Bayesian variable selection and regularisation for time-frequency surface estimation, *J. Roy. Statist. Soc. Ser. B* **66**, 575-589.

Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, U.S.A.
E-mail: jen-hwa.chu@channing.harvard.edu

Department of Statistical Science, Duke University, Durham, NC 27708, U.S.A.
E-mail: clyde@stat.duke.edu

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, U.S.A.
E-mail: liangf@uiuc.edu

(Received December 2007; accepted December 2008)