

Posterior probability intervals for wavelet thresholding

Stuart Barber, Guy P. Nason, and Bernard W. Silverman
University of Bristol, UK.

Abstract

We use cumulants to derive Bayesian credible intervals for wavelet regression estimates. The first four cumulants of the posterior distribution of the estimates are expressed in terms of the observed data and integer powers of the mother wavelet functions. These powers are closely approximated by linear combinations of wavelet scaling functions at an appropriate finer scale. Hence, a suitable modification of the discrete wavelet transform allows the posterior cumulants to be found efficiently for any given data set. Johnson transformations then yield the credible intervals themselves. Simulations show that these intervals have good coverage rates, even when the underlying function is inhomogeneous, where standard methods fail. In the case where the curve is smooth, the performance of our intervals remains competitive with established nonparametric regression methods.

Keywords: Bayes estimation; Cumulants; Curve estimation; Interval estimates; Johnson curves; Nonparametric regression; Powers of wavelets.

1 Introduction

Consider the estimation of a function g from an observed data vector $\mathbf{y} = (y_1, \dots, y_n)'$ satisfying

$$y_i = g(t_i) + \varepsilon_i \quad i = 1, \dots, n,$$

where $t_i = i/n$ and the ε_i are independently $N(0, \sigma^2)$ distributed. There are many methods of estimating smooth functions g , such as spline smoothing (Green and Silverman, 1994), kernel estimation (Wand and Jones, 1995), and local polynomial regression (Fan and Gijbels, 1996). In most cases, such point estimates can be supplemented by interval estimates with some specified nominal coverage probability or significance level.

A recent proposal for estimation of inhomogeneous g is wavelet thresholding (Donoho and Johnstone 1994, 1995). The representation of g in terms of a wavelet basis is typically sparse, concentrating most of the signal in the data into a few large coefficients, whilst the noise is spread “evenly” across the coefficients due to the orthonormality of the wavelet basis. The data is denoised by a thresholding rule of some sort to discard “small” coefficients and retain, possibly with some modification, those coefficients which are thought to contain the signal.

Several authors have described Bayesian wavelet thresholding rules, placing prior distributions on the wavelet coefficients. We consider a means of approximating the posterior distribution of each $g(t_i)$, using the same prior as the `BayesThresh` method of Abramovich, Sapatinas and Silverman (1998). Posterior probability intervals of any nominal coverage probability can then be calculated. Our approach can also be applied to other Bayesian wavelet thresholding rules, such as those surveyed by Chipman and Wolfson (1999), Abramovich and Sapatinas (1999), and Vidakovic (1998).

Address for correspondence: Stuart Barber, Department of Mathematics, University Walk, University of Bristol, Bristol, BS8 1TW, UK. Email: Stuart.Barber@bristol.ac.uk

We briefly review the wavelet thresholding approach to curve estimation in section 2, including two Bayesian methods of wavelet thresholding. In section 3, we derive our WaveBand method of estimating the posterior distribution of $g(t_i)$ given the observed data. We present an example and simulation results in section 4 and make some concluding remarks in section 5.

2 Wavelets and wavelet thresholding

2.1 Wavelets

Wavelets provide a variety of orthonormal bases of $\mathbb{L}_2(\mathbb{R})$, the space of square integrable functions on \mathbb{R} ; each basis is generated from a scaling function, denoted $\phi(t)$, and an associated wavelet, $\psi(t)$. The wavelet basis consists of dilations and translations of these functions. For $j, k \in \mathbb{Z}$, the wavelet at level j and location k is given by

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \quad (1)$$

with an analogous definition for $\phi_{j,k}(t)$. The scaling function is sometimes referred to as the father wavelet, but we avoid this terminology. A function $g(t) \in \mathbb{L}_2(\mathbb{R})$ can be represented as

$$g(t) = \sum_{k \in \mathbb{Z}} \mathcal{C}_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}} \mathcal{D}_{j,k} \psi_{j,k}(t), \quad (2)$$

with $\mathcal{C}_{j_0,k} = \int g(t)\phi_{j_0,k}(t)dt$ and $\mathcal{D}_{j,k} = \int g(t)\psi_{j,k}(t)dt$.

Daubechies (1992) derived two families of wavelet bases which give sparse representations of wide sets of functions. (Technically, by choosing a wavelet ψ with suitable properties, we can generate an unconditional wavelet basis in a wide set of function spaces; for further details, see Abramovich *et al.* (1998).) Generally, smooth portions of g are represented by a small number of coarse scale (low level) coefficients, while local inhomogeneous features such as high frequency events, cusps and discontinuities are represented by coefficients at finer scales (higher levels).

For our nonparametric regression problem, we have discrete data and hence consider the discrete wavelet transform (DWT). Given a vector $\mathbf{g} = (g_1, \dots, g_n)'$, where $g_i = g(t_i)$, the DWT of \mathbf{g} is $\mathbf{d} = W\mathbf{g}$, where W is an orthonormal $n \times n$ matrix and \mathbf{d} is a vector of the discrete scaling coefficient $c_{0,0}$ and $n - 1$ discrete wavelet coefficients $\{d_{j,k} : j = 0, \dots, J - 1, k = 0, \dots, 2^j - 1\}$. These are analogous to the coefficients in (2), with $c_{0,0} \approx \mathcal{C}_{0,0}/\sqrt{n}$ and $d_{j,k} \approx \mathcal{D}_{j,k}/\sqrt{n}$. Our data are restricted to lie in $[0, 1]$, requiring boundary conditions; we assume that g is periodic at the boundaries. Other boundary conditions include the assumption that g is reflected at the boundaries, or the ‘‘wavelets on the interval’’ transform of Cohen, Daubechies and Vial (1993) can be used.

The pyramid algorithm of Mallat (1989) computes \mathbf{d} in $O(n)$ operations, provided $n = 2^J$, $J \in \mathbb{N}$. The algorithm iteratively computes the $\{c_{j,k}\}$ and $\{d_{j,k}\}$ from the $\{c_{j+1,k}\}$. From the vector \mathbf{d} , the inverse DWT (IDWT) can be used to reconstruct the original data. The IDWT starts with the overall scaling and wavelet coefficients $c_{0,0}$ and $d_{0,0}$ and reconstructs the $\{c_{1,k}\}$; it then proceeds iteratively to finer levels, reconstructing the $\{c_{j+1,k}\}$ from the $\{c_{j,k}\}$ and $\{d_{j,k}\}$.

2.2 Curve estimation by wavelet thresholding

Since the DWT is an orthonormal transformation, white noise in the data domain is transformed to white noise in the wavelet domain. If \mathbf{d} is the DWT of \mathbf{g} , and $\mathbf{d}^* = (c_{0,0}^*, d_{0,0}^*, \dots, d_{J-1,2^J-1}^*)'$ the vector of empirical coefficients obtained by applying the DWT to the data \mathbf{y} , then $\mathbf{d}^* = \mathbf{d} + \varepsilon'$, where ε' is a vector of n independent $N(0, \sigma^2)$ variates. Wavelet thresholding assumes implicitly that “large” and “small” $d_{j,k}^*$ represent signal and noise respectively. Various thresholding rules have been proposed to determine which coefficients are “large” and “small”; see Vidakovic (1999) or Abramovich, Bailey and Sapatinas (2000) for reviews. The true coefficients \mathbf{d} are estimated by applying the thresholding rule to the empirical coefficients \mathbf{d}^* to obtain estimates $\hat{\mathbf{d}}$, and the sampled function values \mathbf{g} are estimated by applying the IDWT to obtain $\hat{\mathbf{g}} = W^T \hat{\mathbf{d}}$ where W^T denotes the transpose of W ; symbolically, we can represent this IDWT as the sum

$$\hat{g}(t) = \hat{c}_{0,0} \phi(t) + \sum_{j=0}^J \sum_{k=0}^{2^j-1} \hat{d}_{j,k} \psi_{j,k}(t). \quad (3)$$

Confidence intervals for the resulting curve estimates have received some attention in the wavelet literature. Brillinger (1994) derived an estimate of $\text{var}\{\hat{g}_i\}$ when the wavelet decomposition involved Haar wavelets, and Brillinger (1996) showed that \hat{g}_i is asymptotically normal under certain conditions. Bruce and Gao (1996) extended the estimation of $\text{var}\{\hat{g}_i\}$ to the case of non-Haar wavelets, and gave approximate $100(1 - \alpha)\%$ confidence intervals using the asymptotic normality of \hat{g}_i . Chipman, Kolaczyk and McCulloch (1997) presented approximate credible intervals for their adaptive Bayesian wavelet thresholding method, which we discuss in section 2.3.2, while Picard and Tribouley (2000) derive bias corrected confidence intervals for wavelet thresholding estimators.

2.3 Bayesian wavelet regression

2.3.1 Introduction

Several authors have proposed Bayesian wavelet regression estimates, involving priors on the wavelet coefficients $d_{j,k}$, which are updated by the observed coefficients $d_{j,k}^*$ to obtain posterior distributions $[d_{j,k} | d_{j,k}^*]$. Point estimates $\hat{d}_{j,k}$ can be computed from these posterior distributions and the IDWT employed to estimate \mathbf{g} in the usual fashion outlined above. Some proposals have included priors on σ^2 ; we restrict ourselves to the situation where σ^2 can be well estimated from the data.

The majority of Bayesian wavelet shrinkage rules have employed mixture distributions as priors on the coefficients, to model the notion that a small proportion of coefficients contain substantial signal. Chipman *et al.* (1997) and Crouse, Nowak and Baraniuk (1998) considered mixtures of two normal distributions, while Abramovich *et al.* (1998), Clyde, Parmigiani and Vidakovic (1998) and Clyde and George (2000) used mixtures of a normal and a point mass. Other proposals include a mixture of a point mass and a t -distribution used by Vidakovic (1998), and an infinite mixture of normals considered by Holmes and Denison (1999). More thorough reviews are given by Chipman and Wolfson (1999) and Abramovich and Sapatinas (1999).

2.3.2 Adaptive Bayesian wavelet shrinkage

The ABWS method of Chipman *et al.* (1997) places an independent prior on each $d_{j,k}$:

$$[d_{j,k} | \gamma_{j,k}] \sim \gamma_{j,k} N(0, c_j^2 \nu_j^2) + (1 - \gamma_{j,k}) N(0, \nu_j^2), \quad (4)$$

where $\gamma_{j,k} \sim \text{Bernoulli}(p_j)$; hyperparameters p_j , c_j , and ν_j are determined from the data by empirical Bayes methods. At level j , the proportion of non-zero coefficients is represented by p_j , while ν_j and $c_j \nu_j$ represent the magnitude of negligible and non-negligible coefficients respectively.

Given $d_{j,k}$, the empirical coefficients are $d_{j,k}^* \sim N(d_{j,k}, \sigma^2)$ independently, so the posterior distribution $[d_{j,k} | d_{j,k}^*]$ is a mixture of two normal components independently for each (j, k) . Chipman *et al.* (1997) use the mean of this mixture as their estimator $\hat{d}_{j,k}$, and similarly use the variance of $[d_{j,k} | d_{j,k}^*]$ to approximate the variance of their ABWS estimate of \mathbf{g} . Estimating each coefficient by the mean of its posterior distribution is the Bayes rule under the \mathbb{L}_2 loss function. They plot uncertainty bands of $(\hat{g}_i \pm 3\text{sd}\{\hat{g}_i\})$. These bands are useful in representing the uncertainty in the estimate \hat{g}_i , but are based on an assumption of normality despite \hat{g}_i being a sum of random variables with mixture distributions.

2.3.3 BayesThresh

The BayesThresh method of Abramovich *et al.* (1998) also places independent priors on the coefficients:

$$d_{j,k} \sim p_j N(0, \tau_j^2) + (1 - p_j) \delta(0), \quad (5)$$

where $0 \leq p_j \leq 1$ and $\delta(0)$ is a probability mass at zero. This is a limiting case of the ABWS prior (4).

The hyperparameters are assumed to be of the form $\tau_j^2 = 2^{-\alpha j} C_1$ and $p_j = \min\{1, 2^{-\beta j} C_2\}$ for non-negative constants C_1 and C_2 chosen empirically from the data and α and β selected by the user. Abramovich *et al.* (1998) show that choices of α and β correspond to choosing priors within certain Besov spaces, incorporating prior knowledge about the smoothness of $g(t)$ in the prior. Chipman and Wolfson (1999) also discuss the interpretation of α and β . In the absence of any such prior knowledge, Abramovich *et al.* (1998) show that the default choice $\alpha = 0.5$, $\beta = 1$ is robust to varying degrees of smoothness of $g(t)$. The prior specification is completed by placing a non-informative prior on the scaling coefficient $c_{0,0}$, which thus has the posterior distribution $N(c_{0,0}^*, \sigma^2)$, and is estimated by $c_{0,0}^*$.

The resulting posterior distribution of $d_{j,k}$ given the observed value of $d_{j,k}^*$ is again independent for each (j, k) and is given by

$$[d_{j,k} | d_{j,k}^*] \sim \omega_{j,k} N(d_{j,k}^* r_j^2, \sigma^2 r_j^2) + (1 - \omega_{j,k}) \delta(0), \quad (6)$$

where $\omega_{j,k} = (1 + \xi_{j,k})^{-1}$, with

$$\xi_{j,k} = \frac{1 - p_j}{p_j} \frac{\sqrt{\tau_j^2 + \sigma^2}}{\sigma} \exp \left\{ \frac{-\tau_j^2 d_{j,k}^*}{2\sigma^2(\tau_j^2 + \sigma^2)} \right\}$$

and $r_j^2 = \tau_j^2 / (\sigma^2 + \tau_j^2)$. The BayesThresh approach minimises the \mathbb{L}_1 loss in the wavelet domain by using the posterior median of $[d_{j,k} | d_{j,k}^*]$ as the point estimate $\hat{d}_{j,k}$; Abramovich *et al.* (1998) show

that this gives a level-dependent true thresholding rule. The `BayesThresh` method is implemented in the `WaveThresh` package for `SPlus` (Nason, 1998).

3 WaveBand

3.1 Posterior cumulants of the regression curve

We now consider the posterior density of $[g_i|\mathbf{y}]$ for each $i = 1, \dots, n$. From (3), we see that $[g_i|\mathbf{y}]$ is the convolution of the posteriors of the wavelet coefficients and the scaling coefficient,

$$[g_i|\mathbf{y}] = [c_{0,0}|c_{0,0}^*]\phi(t_i) + \sum_j \sum_k [d_{j,k}|d_{j,k}^*]\psi_{j,k}(t_i); \quad (7)$$

this is a complicated mixture, and is impractical to evaluate analytically. Direct simulation from the posterior is extremely time consuming. Instead we estimate the first four cumulants of (7) and fit a distribution which matches these values. Evaluating cumulants of $[g_i|\mathbf{y}]$ requires the use of powers of wavelets $\psi_{j,k}^r(t)$ for $r = 2, 3$, and 4 , which we discuss in section 3.2. We then fit a parametric distribution to $[g_i|\mathbf{y}]$ in section 3.3.

If the moment generating function of a random variable X is written $M_X(t)$, then the cumulant generating function is $K_X(t) = \log M_X(t)$. We write $\kappa_r(X)$ for the r^{th} cumulant of X , given by the r^{th} derivative of $K_X(t)$, evaluated at $t = 0$. Note that the first r moments uniquely determine the first r cumulants and vice-versa. Further details of cumulants and their properties and uses can be found in Barndorff-Nielsen and Cox (1989) or Stuart and Ord (1994, chapter 3).

The first four cumulants each have a direct interpretation; $\kappa_1(X)$ and $\kappa_2(X)$ are the mean and variance of X respectively, while $\kappa_3(X)/\kappa_2^{3/2}(X)$ is the skewness and $\kappa_4(X)/\kappa_2^2(X) + 3$ is the kurtosis. If X is normally distributed then both $\kappa_3(X)$ and $\kappa_4(X)$ are zero. We make use of two standard properties of cumulants. If X and Y are independent random variables and a and b are real constants, then

$$\kappa_r(aX + b) = \begin{cases} a\kappa_1(X) + b & r = 1 \\ a^r \kappa_r(X) & r = 2, 3, \dots, \end{cases} \quad (8)$$

$$\text{and} \quad \kappa_r(X + Y) = \kappa_r(X) + \kappa_r(Y) \quad r \in \mathbb{Z}. \quad (9)$$

Applying (8) and (9) to (7), we can see that the cumulants of $[g_i|\mathbf{y}]$ are given by

$$\kappa_r(g_i|\mathbf{y}) = \kappa_r(c_{0,0}|c_{0,0}^*)\phi_{j,k}^r(t_i) + \sum_j \sum_k \kappa_r(d_{j,k}|d_{j,k}^*)\psi_{j,k}^r(t_i). \quad (10)$$

Once the cumulants of the wavelet coefficients are known, this sum can be evaluated directly using the IDWT when $r = 1$, but for $r = 2, 3$, and 4 , the computation involves powers of wavelets.

The posterior distributions (6) for the wavelet coefficients using `BayesThresh` are of the form $Z \sim pX + (1 - p)\delta(0)$, where $X \sim N(\mu, \nu^2)$, $0 \leq p \leq 1$, and $\delta(0)$ is a point mass at zero. Then

$\mathbb{E}(Z^r) = p\mathbb{E}(X^r)$ so the moments of (6) are easily obtained. From these, we can derive the cumulants of $[d_{j,k}|d_{j,k}^*]$:

$$\begin{aligned}\kappa_1(d_{j,k}|d_{j,k}^*) &= \omega_{j,k}d_{j,k}^*r_j^2, \\ \kappa_2(d_{j,k}|d_{j,k}^*) &= \omega_{j,k}r_j^2\{(d_{j,k}^*)^2r_j^2(1-\omega_{j,k})+\sigma^2\}, \\ \kappa_3(d_{j,k}|d_{j,k}^*) &= \omega_{j,k}(1-\omega_{j,k})d_{j,k}^*r_j^4\{(d_{j,k}^*)^2r_j^2(1-2\omega_{j,k})+3\sigma^2\}, \\ \kappa_4(d_{j,k}|d_{j,k}^*) &= \omega_{j,k}(1-\omega_{j,k})r_j^4[(d_{j,k}^*)^4r_j^4\{1-6\omega_{j,k}(1-\omega_{j,k})\}+6(d_{j,k}^*)^2r_j^2\sigma^2(1-2\omega_{j,k})+3\sigma^4].\end{aligned}$$

Since the scaling coefficient $c_{0,0}$ has a normal posterior distribution, the first two cumulants are $\kappa_1(c_{0,0}|c_{0,0}^*) = c_{0,0}^*$ and $\kappa_2(c_{0,0}|c_{0,0}^*) = \sigma^2$; all higher order cumulants are zero.

3.2 Powers of wavelets

The parallel between (10) and the sum (3) evaluated by the IDWT is clear, and we shall describe a means of taking advantage of the efficient IDWT algorithm to evaluate (10). For Haar wavelets, this is trivial; the Haar scaling function and mother wavelet are $\phi(t) = I(0 \leq t < 1)$ and $\psi(t) = I(0 \leq t < 1/2) - I(1/2 \leq t < 1)$, where $I(\cdot)$ is the indicator function, hence $\psi^3(t) = \psi(t)$ and $\psi^2(t) = \psi^4(t) = \phi(t)$. Moreover, $\psi_{j,k}^2(t) = 2^{j/2}\phi_{j,k}(t)$, $\psi_{j,k}^3(t) = 2^j\psi_{j,k}(t)$, and $\psi_{j,k}^4(t) = 2^{3j/2}\phi_{j,k}(t)$, all terms which can be included in a modified version of the IDWT algorithm which incorporates scaling function coefficients. Since representing $\psi_{j,k}^r(t)$ by scaling functions and wavelets works in the Haar case, we consider a similar approach for other wavelet bases.

Following Herrick (2000, pp. 69), we approximate a general wavelet $\psi_{j_0,0}^r$, ($0 \leq j_0 \leq J - m$), by

$$\psi_{j_0,0}^r(t) \approx \sum_l e_{j_0+m,l} \phi_{j_0+m,l}(t) \quad (11)$$

for $r = 2, 3, 4$, where m is a positive integer; the choice of m is discussed below. We use scaling functions rather than wavelets as the span of the set of scaling functions at a given level j is the same as that of the union of $\phi(t)$ and wavelets at levels $0, 1, \dots, j-1$. Moreover, if scaling functions $\phi_{j,k}(t)$ are used to approximate some function $h(t)$, and both ϕ and h have at least ν derivatives, then the mean squared error in the approximation is bounded by $C2^{-\nu j}$, where C is some positive constant; see, for example, Vidakovic (1999, p.87).

To approximate $\psi_{j_0,k}^r(t)$ for some fixed j_0 , we first compute $\psi_{j_0,0}^r$ using the cascade algorithm (Daubechies, 1992, pp. 205), then take the DWT of $\psi_{j_0,0}^r$ and set the coefficients $\{e_{m_0,l}\}$ to be equal to the scaling function coefficients $\{c_{m_0,k}\}$ at level m_0 , where $m_0 = j_0 + m$. Recall that the wavelets at level j are simply shifts of each other; from (1) $\psi_{j,k}(t) = \psi_{j,0}(t - 2^{-j}k)$, hence

$$\psi_{j_0,k}^r(t) \approx \sum_l e_{m_0,l-2^m k} \phi_{m_0,l}(t). \quad (12)$$

As we are assuming periodic boundary conditions, the $\{e_{m_0,l}\}$ can be cycled periodically.

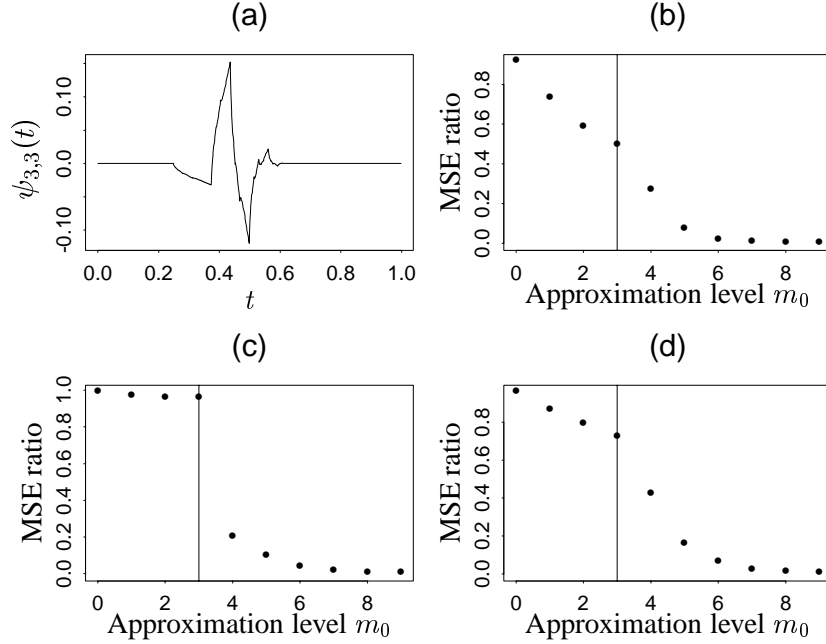


Figure 1: *Daubechies' extremal phase wavelet with two vanishing moments* $\psi_{3,3}(t)$ (panel a) and accuracy of estimating $\psi_{3,3}^r$ for $r = 2, 3, 4$ by scaling functions at level m_0 . The mean square errors of the estimates, divided by the norm of $\psi_{3,3}^r$, are plotted against m_0 : panels (b), (c), and (d) show results for $r = 2, 3$, and 4 respectively. The vertical lines are at level 3, the level at which $\psi_{3,3}$ exists.

Owing to the localised nature of wavelets, the coefficients $\{e_{m_0+1,l}\}$ used to approximate $\psi_{j_0+1,0}^r(t)$ can be found by inserting 2^{m_0} zeros into the vector of $\{e_{m_0,l}\}$:

$$e_{m+1,l} = \begin{cases} \sqrt{2}e_{m_0,l} & l = 0, \dots, 2^{m_0-1} - 1 \\ 0 & l = 2^{m_0-1}, \dots, 2^{m_0+1} - 2^{m_0-1} - 1 \\ \sqrt{2}e_{m_0,l-2^{m_0}} & l = 2^{m_0+1} - 2^{m_0-1}, \dots, 2^{m_0+1} - 1. \end{cases}$$

Approximation (11) cannot be used for wavelets at the m finest levels $J - m, \dots, J - 1$. We represent these wavelets by both scaling functions and wavelets at the finest level of detail, level $J - 1$, using the block-shifting technique outlined above to make the computations more efficient.

In all the cases we have examined, $m = 3$ has been sufficient for a highly accurate approximation; examples are shown in figures 1 and 2. Consider approximating the powers of Daubechies' extremal phase wavelet with two vanishing moments; this wavelet is extremely non-smooth and so can be regarded as a near worst-case scenario for the approximation. Panel (a) of figure 1 shows $\psi_{3,3}(t)$, while panels (b)-(d) show the mean square error in our approximation divided by the \mathbb{L}_2 norm of the function being approximated, $\sum_i \left\{ \sum_l e_{m_0,l} \phi_{m_0,l}(t_i) - \psi_{3,0}^r(t_i) \right\}^2 / \sum_i \left\{ \psi_{3,0}^r(t_i) \right\}^2$ for $r = 2, 3, 4$ respectively. In each plot (b)-(d), the vertical line is at resolution level $j_0 = 3$, the level at which the wavelet whose power is being estimated exists. The approximation is excellent for $m_0 = j_0 + 3$ in each case, with little improvement at level $m_0 = j_0 + 4$.

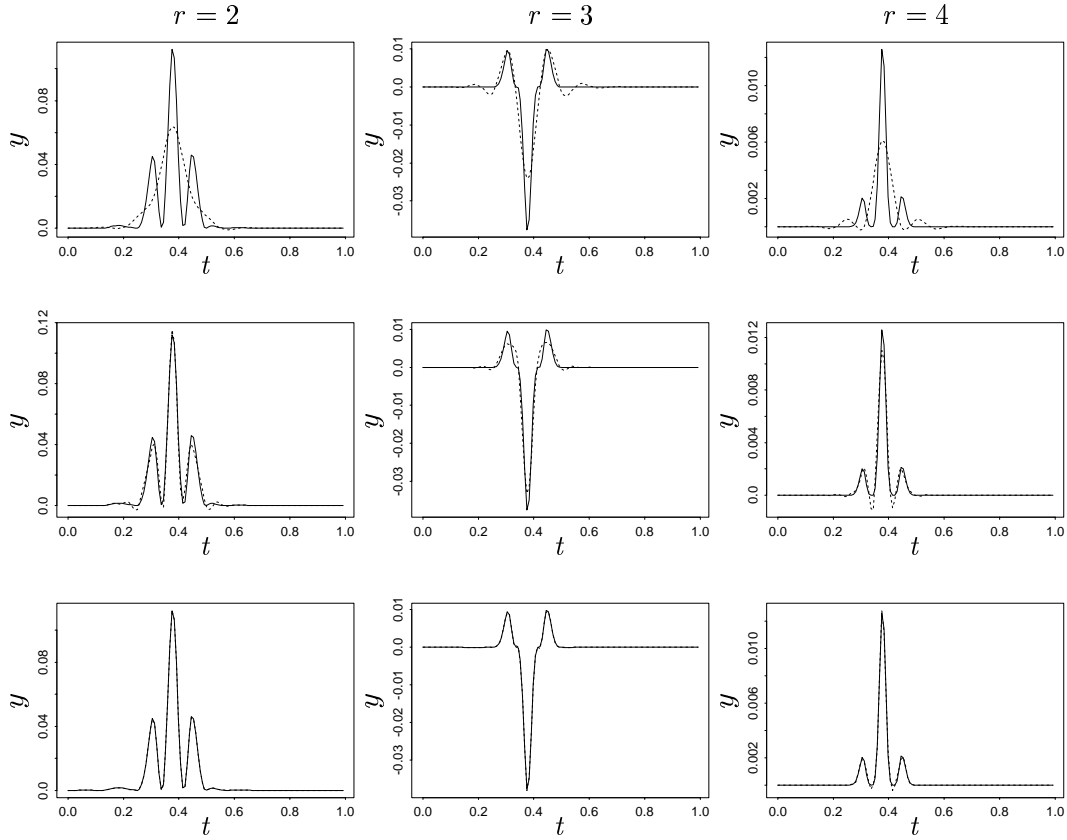


Figure 2: *Approximations to powers of Daubechies' least asymmetric wavelet with eight vanishing moments; the powers are indicated at the top of each column. Solid lines are wavelet powers and dotted lines show approximations using scaling functions at level m_0 . From top to bottom, graphs show approximation at levels $m_0 = 4$, $m_0 = 5$, and $m_0 = 6$; the original wavelet is at level $j_0 = 3$.*

Figure 2 shows approximations to powers of Daubechies' least asymmetric wavelet with eight vanishing moments as the approximation level m_0 increases. Again, the base wavelet is at level $j_0 = 3$, and approximations are shown for $m_0 = j_0 + 1$ (top row), $m_0 = j_0 + 2$, and $m_0 = j_0 + 3$ (bottom row). In each case, the solid line is the wavelet power and the dotted line is the approximation.

Using (12), we can now re-write (10) as

$$\begin{aligned}
 \kappa_r(g_i|\mathbf{y}) &= \kappa_r(c_{0,0}|c_{0,0}^*)\phi_{j,k}^r(t_i) + \sum_{j,k} \kappa_r(d_{j,k}|d_{j,k}^*)\psi_{j,k}^r(t_i) \\
 &= \kappa_r(c_{0,0}|c_{0,0}^*)\phi_{j,k}^r(t_i) + \sum_{j,k} \left\{ \kappa_r(d_{j,k}|d_{j,k}^*) \sum_l e_{j+3,l} \phi_{j+3,l}(t_i) \right\} \\
 &= \sum_{j,k} \rho_{j,k} \phi_{j,k}(t_i),
 \end{aligned}$$

for suitable coefficients $\rho_{j,k}$, and use a modified version of the IDWT algorithm which incorporates scaling function coefficients to evaluate this sum.

3.3 Posterior distribution of the regression curve

We must now estimate $[g_i|y]$ from its cumulants. Edgeworth expansions give poor results in the tails of the distribution, where we require a good approximation. Saddlepoint expansions improve on this, but require the cumulant generating function $K_{[g_i|y]}(r)$, while we only have the first four cumulants. Therefore, we approximate $[g_i|y]$ by a suitable parametric distribution.

A family of distributions is the set of transformations of the normal curve described by Johnson (1949). As well as the normal distribution, Johnson curves fall into three categories;

- (a) the lognormal (S_L), $z = \gamma + \delta \log(x - \xi)$, with $\xi < x$,
- (b) the unbounded (S_U), $z = \gamma + \delta \sinh^{-1}\{(x - \xi)/\lambda\}$, and
- (c) the bounded (S_B), $z = \gamma + \delta \log\{(x - \xi)/(\xi + \lambda - x)\}$, with $\xi < x < \xi + \lambda$,

to which Hill, Hill and Holder (1976) added a limiting case of S_B curves where $\kappa_3 = \sqrt{\kappa_4 + 2}$, referred to as S_T curves. In each case z has a standard normal distribution, while x is the Johnson variable. The Johnson curves provide as rich a family of distributions as the better known Pearson curves, and have the practical advantage of easily available software (Hill *et al.*, 1976). As noted by these authors, when the moments of a distribution are computed theoretically, as in our case, the standard objections to fitting a curve by moments on the grounds of inefficiency do not apply. Percentage points of Johnson curves can be computed using algorithms by Hill (1976).

4 An example and simulation results

4.1 Example

We apply our WaveBand method to the piecewise polynomial test function of Nason and Silverman (1994) sampled on $n = 512$ data points $t_i = i/512$ for $i = 1, \dots, 512$. This function is defined as

$$g_p(t) = \begin{cases} 4t^2(3 - 4t) & t \in [0, \frac{1}{2}) \\ \frac{4}{3}(4t^2 - 10t + 7) - \frac{3}{2} & t \in [\frac{1}{2}, \frac{3}{4}) \\ \frac{16}{3}t(t - 1)^2 & t \in [\frac{3}{4}, 1), \end{cases}$$

and shown in figure 3 (solid line). Figure 3 also shows data formed by adding independent normally distributed noise to $g_p(t)$. The noise has mean zero and root signal to noise ratio (rsnr) 3; the rsnr is defined to be the ratio of the standard deviation of the data points $g_p(t_i)$ to the standard deviation of the noise, σ .

Dotted lines in figure 3 mark upper and lower endpoints of 99% credible intervals for $g_p(t_i)$ for each t_i calculated using our WaveBand method, using default parameters of $\alpha = 0.5$, $\beta = 1$, and Daubechies' least asymmetric wavelet with eight vanishing moments. In this example, the credible intervals include the true value of $g_p(t_i)$ in 490 of 512 cases, an empirical coverage rate of 95.7%. The brief spikes which occur in the bands are typical of wavelet regression methods; they can be smoothed out by using different α and β , but this risks oversmoothing the data.

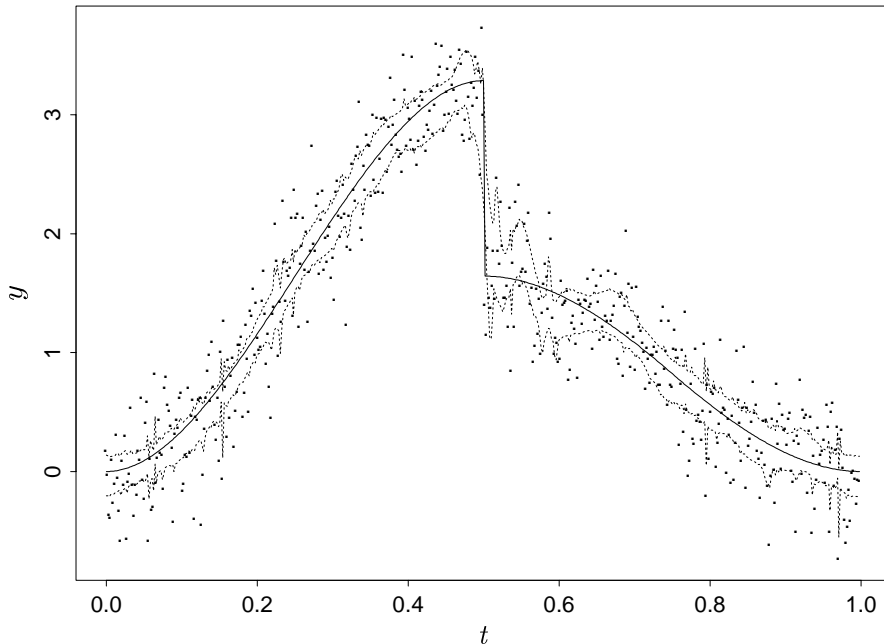


Figure 3: *Pointwise 99% WaveBand interval estimates (dotted line) for the piecewise polynomial signal (solid line). Dots indicate data on $n = 512$ equally spaced points with the addition of independent normally distributed noise with mean zero and rsnr 3.*

Figure 4 shows the mean, variance, skewness, and kurtosis of $[g_p(t_i)|\mathbf{y}]$ for each point t_i . The mean is itself an estimate of $g_p(t_i)$ and the variance is generally low except near the discontinuity at $t = 0.5$. We use the usual definitions of skewness and kurtosis, $\sqrt{\beta_1} = \kappa_3/\kappa_2^{3/2}$ and $\beta_2 = \kappa_4/\kappa_2^2 + 3$, respectively; for reference, a normal random variable has skewness zero and kurtosis three. For many values of t_i , the posterior distribution $[g_p(t_i)|\mathbf{y}]$ has significant skewness and kurtosis; the values of $\sqrt{\beta_1}$ range from approximately -7.9 to 7.5, and β_2 from approximately 1.7 to almost 120, indicating that for some t_i , the posterior distribution has heavy tails. (For comparison the $F_{6,4}$ distribution has skewness 7.1 and kurtosis 79.) The prior (5) has kurtosis $3/p_j$; note that in our example the largest values of kurtosis occur when the function is smooth; in these cases the majority of the wavelet coefficients are zero, and hence p_j is small.

4.2 Simulation results

4.2.1 Inhomogeneous signals

We investigated the performance of our WaveBand credible intervals by simulation on the piecewise polynomial example of Nason and Silverman (1994) (denoted “PPoly”) and the standard “Blocks”, “Bumps”, “Doppler” and “HeaviSine” test functions of Donoho and Johnstone (1994), shown in figure 5. For each test function, 100 simulated data sets of length $n = 1024$ were created with rsnr

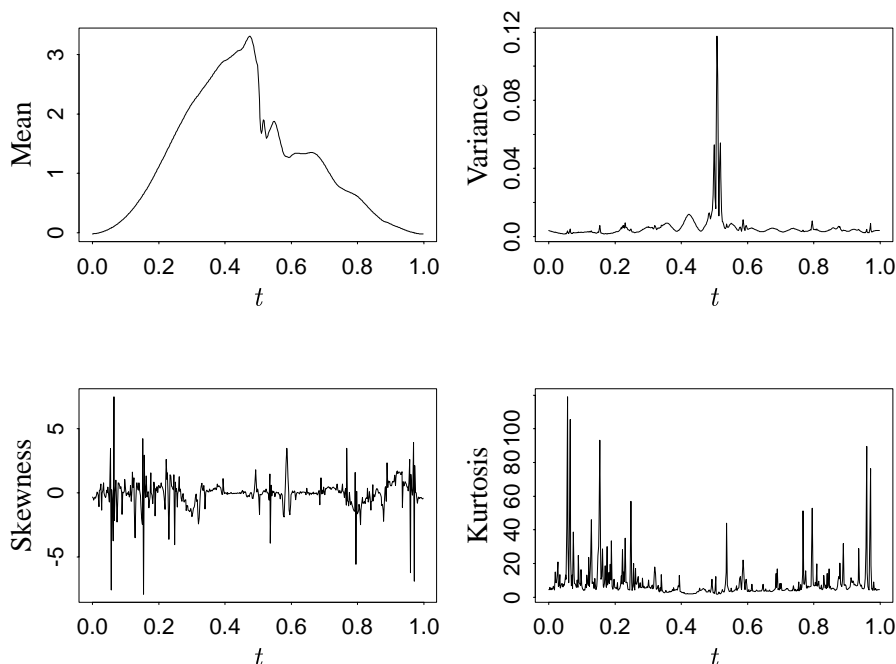


Figure 4: *Mean, variance, skewness and kurtosis for the piecewise polynomial data shown in figure 3.*

4 and the WaveBand and ABWS credible intervals evaluated at each data point for nominal coverage probabilities 0.90, 0.95, and 0.99. The default hyperparameters $\alpha = 0.5$ and $\beta = 1$ were used for WaveBand and both methods used Daubechies' least asymmetric wavelet with eight vanishing moments. Table 1 shows the coverage rates and interval widths averaged over all points t_i and the 100 replications, with standard error shown in brackets. The WaveBand intervals have higher empirical coverage rates in each case, although still below the nominal coverage probabilities. Average widths of the WaveBand credible intervals are always greater than those of the ABWS intervals; one reason is that the ABWS method typically has a lower estimated variance. Moreover, the WaveBand posterior of $[g_i | \mathbf{y}]$ is typically heavier-tailed than the normal distribution, as we saw in the piecewise polynomial example.

The performance of the WaveBand intervals improves as the nominal coverage rate increases. In part, this can be attributed to the kurtosis of the posterior distributions. As the nominal coverage rates increase, the limits of the credible intervals move out into the tails of the posterior distributions. With heavy-tailed distributions, a small increase in the nominal coverage rate can produce a substantially wider interval.

Figure 6 examines the empirical coverage rates of the nominal 99% credible intervals in more detail. Solid lines denote the empirical coverage rates of the WaveBand intervals for each t_i , and dotted lines give the equivalent results for the ABWS intervals. Coverage varies greatly across each signal; unsurprisingly, the coverage is much better where the signal is smoother and less variable. This can be seen most clearly in the results for the piecewise polynomial and HeaviSine test functions,

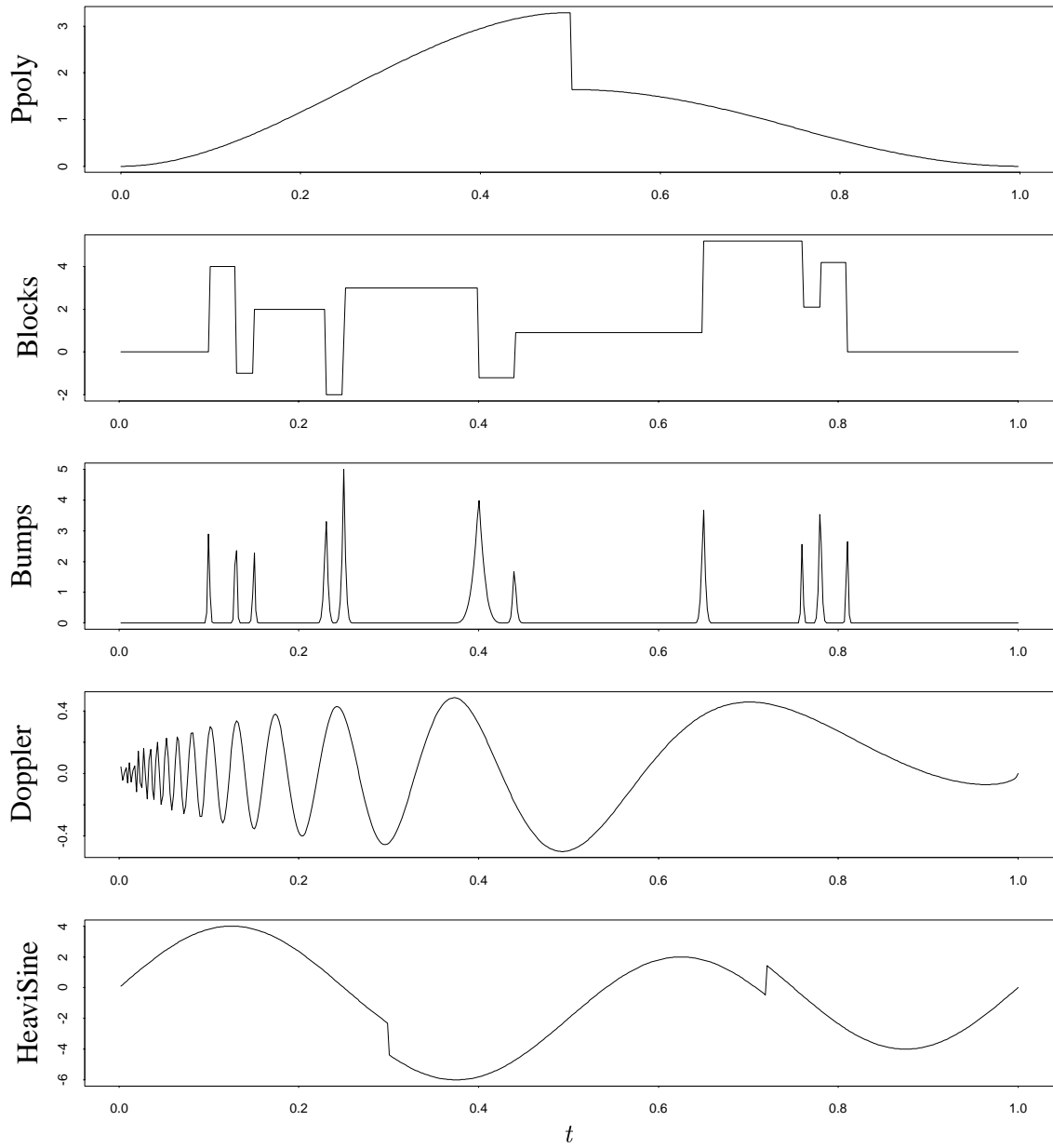


Figure 5: The piecewise polynomial of Nason and Silverman (1994) and the test functions of Donoho and Johnstone (1994).

	Nominal coverage probability					
	0.90		0.95		0.99	
	CR	Width	CR	Width	CR	Width
PPoly						
WB	0.804 (.006)	0.532 (.004)	0.892 (.004)	0.685 (.003)	0.966 (.002)	1.084 (.005)
ABWS	0.657 (.010)	0.408 (.005)	0.728 (.009)	0.486 (.006)	0.828 (.008)	0.638 (.007)
Blocks						
WB	0.825 (.003)	1.536 (.005)	0.899 (.002)	1.246 (.004)	0.971 (.001)	2.202 (.007)
ABWS	0.775 (.004)	1.121 (.004)	0.842 (.003)	1.325 (.005)	0.923 (.002)	1.741 (.006)
Bumps						
WB	0.865 (.002)	1.357 (.004)	0.927 (.001)	1.676 (.004)	0.980 (.001)	2.408 (.006)
ABWS	0.735 (.003)	1.107 (.003)	0.810 (.003)	1.319 (.004)	0.904 (.002)	1.733 (.005)
Doppler						
WB	0.847 (.004)	0.966 (.004)	0.911 (.002)	0.765 (.003)	0.965 (.001)	1.460 (.005)
ABWS	0.808 (.005)	0.693 (.004)	0.871 (.004)	0.827 (.005)	0.936 (.002)	1.090 (.006)
HeaviSine						
WB	0.749 (.009)	0.487 (.005)	0.851 (.007)	0.629 (.004)	0.952 (.003)	0.997 (.006)
ABWS	0.613 (.011)	0.371 (.005)	0.684 (.010)	0.442 (.006)	0.790 (.009)	0.581 (.008)

Table 1: *Simulation results comparing mean coverage rates (CR) and interval widths for ABWS and WaveBand (denoted WB) credible intervals on the piecewise polynomial, Blocks, Bumps, Doppler, and HeaviSine test functions. Each test function was evaluated on $n = 1024$ points, the rsnr was 4, and 100 replications were done in each case. Standard errors are given in brackets. All methods were employed with Daubechies' least asymmetric wavelet with 8 vanishing moments, and the default hyperparameters $\alpha = 0.5$ and $\beta = 1$ were used for WaveBand.*

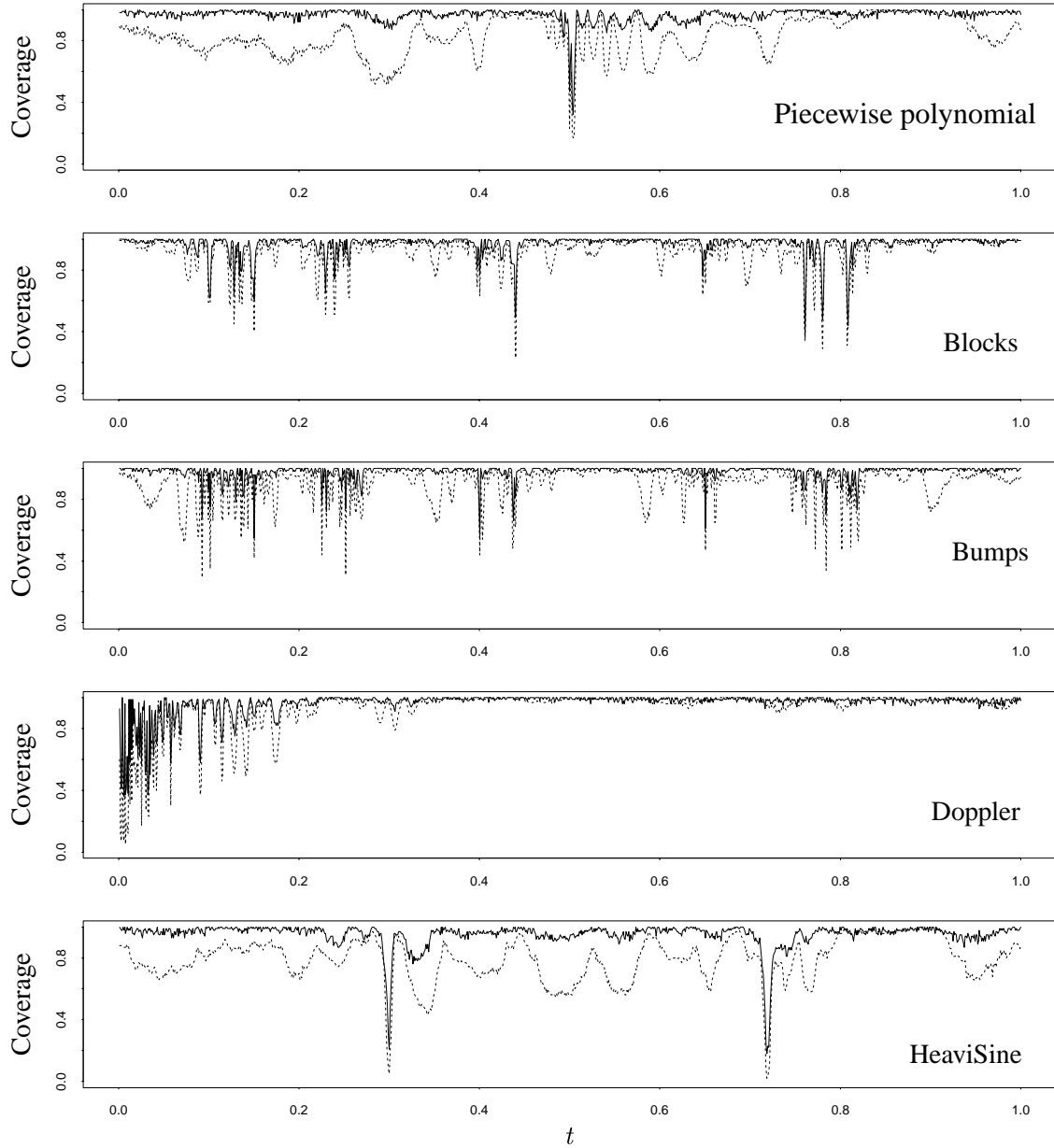


Figure 6: Empirical coverage rates of nominal 99% interval estimates for the indicated test functions evaluated at $n = 1024$ equally spaced data points. In each case, 100 simulated data sets with a rsnr of 4 were used. Solid and dotted lines indicate coverage rates for the WaveBand method and ABWS methods respectively.

with sharp drops in performance near the discontinuities, and in the excellent coverage in the lower-frequency portion of the Doppler signal and the long constant parts of the Bumps signal.

4.2.2 Smooth signals

The major advantage of wavelet thresholding as a nonparametric regression technique is the ability to model inhomogeneous signals such as those considered in section 4.2.1. However, wavelet thresholding can also be used successfully on smooth signals, and we now consider such an example, the function $g_s(t) = \sin(2\pi t) + 2(t - 0.5)^2$. Table 2 shows the performance of ABWS and WaveBand credible intervals and confidence intervals using smoothing splines for $g_s(t)$.

The smoothing spline estimate of \mathbf{g} can be written $\widehat{\mathbf{g}}_{\text{spline}} = S\mathbf{y}$, where S is an $n \times n$ matrix, so $\text{var}\{\widehat{\mathbf{g}}_{\text{spline}}\} = \sigma^2 S S^T$. Hence we can construct an approximate $100(1 - \alpha)\%$ confidence interval for g_i as $\left(\widehat{g}_{\text{spline}}(t_i) \pm z_{\alpha/2} \sigma s_i\right)$ where $s_i = \sqrt{(S S^T)_{i,i}}$.

The signal $g_s(t)$ was evaluated on $n = 128$ equally spaced data points and 100 data sets with root signal to noise ratios of two and six were generated. WaveBand intervals were computed using both the default hyperparameters $\alpha = 0.5$, $\beta = 1$ and also using values $\alpha = 4$, $\beta = 1$, corresponding to a smoother prior. Both WaveBand and ABWS used Daubechies' least asymmetric wavelet with 8 vanishing moments. For each simulated data set, the spline smoothing parameter λ was chosen by cross-validation using the `smooth.spline` function in `Splus`. Table 2 reports average coverage rates and interval widths calculated as in table 1. The WaveBand results with a smooth prior are somewhat better, as would be expected. However, the default prior still gives respectable performance which is generally comparable with the performance of the smoothing splines.

5 Discussion

5.1 An alternative prior

The approach which we have used in constructing our posterior probability intervals can, in principle, be applied to other Bayesian thresholding methods. The only requirement is that the first four cumulants of the posterior distribution of the wavelet coefficients must exist, although independence of the posteriors given the data is also desirable as computations for the dependent case would be more complex.

As an example, we consider a double exponential prior; the heavier tails of this prior are suitable for representing the possibility of a few large wavelet coefficients. Johnstone and Silverman (personal communication) have shown that this prior has superior asymptotic properties to the prior (5). It also has the property that $|d_{j,k}^* - \widehat{d}_{j,k}|$ is bounded no matter how large $\widehat{d}_{j,k}$ is, as does the standard soft thresholding function; this is not the case for the prior (5). Moulin and Liu (1999) have suggested the use of generalised Gaussian priors, of which the double exponential is a special case, on the grounds that heavier tailed distributions are more robust to prior misspecification.

Define $\gamma(u) = \frac{1}{2}ae^{-a|u|}$, and consider the random variable $X \sim N(\theta, 1)$, $\theta \sim (1 - p)\delta(0) + pG$, where G is a double exponential variate with density function γ . Given an observation x of X , the

$RSNR = 2$

	Nominal coverage probability					
	0.90		0.95		0.99	
	CR	Width	CR	Width	CR	Width
WB ¹	0.939 (.008)	0.300 (.002)	0.970 (.005)	0.358 (.003)	0.990 (.003)	0.471 (.004)
WB ²	0.906 (.009)	0.229 (.003)	0.960 (.006)	0.344 (.003)	0.990 (.002)	0.530 (.004)
Spline	0.886 (.012)	0.250 (.004)	0.942 (.008)	0.298 (.005)	0.980 (.004)	0.391 (.007)
ABWS	0.557 (.017)	0.166 (.006)	0.617 (.017)	0.198 (.007)	0.706 (.017)	0.260 (.009)

$RSNR = 4$

	Nominal coverage probability					
	0.90		0.95		0.99	
	CR	Width	CR	Width	CR	Width
WB ¹	0.913 (.009)	0.153 (.001)	0.945 (.007)	0.183 (.002)	0.989 (.003)	0.241 (.002)
WB ²	0.850 (.010)	0.143 (.002)	0.919 (.007)	0.181 (.002)	0.981 (.002)	0.276 (.002)
Spline	0.878 (.011)	0.133 (.002)	0.934 (.008)	0.159 (.002)	0.979 (.005)	0.209 (.003)
ABWS	0.493 (.016)	0.078 (.003)	0.550 (.016)	0.093 (.003)	0.643 (.017)	0.122 (.004)

Table 2: Simulation results comparing the mean coverage rate (CR) and interval widths for WaveBand, smoothing spline, and ABWS methods on the function $g_s(t) = \sin(2\pi t) + 2(t - \frac{1}{2})^2$; standard errors are given in brackets. White noise with the indicated rsnr was added to the function evaluated on $n = 128$ equally spaced t_i , and 100 replications were carried out. WB¹ denotes results using WaveBand hyperparameters $\alpha = 4$, $\beta = 1$, corresponding to a smooth prior on the wavelet coefficients, while results denoted WB² used the default $\alpha = 0.5$, $\beta = 1$. All wavelet methods used Daubechies' least asymmetric wavelet with eight vanishing moments.

posterior probability that $\theta \neq 0$ is $p^{-1}(\gamma \circ \phi)(x)$, where ϕ is the density function of a standard normal random variable, not a scaling function, and \circ denotes convolution. Write Φ for the distribution function associated with ϕ , and let $y = x - a$ and $z = x + a$; then $(\gamma \circ \phi)(x)$ is

$$(\gamma \circ \phi)(x) = \frac{a}{2} e^{a^2/2} [e^{ax} \Phi(-z) + e^{-ax} \Phi(y)].$$

The posterior distribution of θ given an observed value x of X and that θ is drawn from the non-zero part of the distribution is

$$f_p(\theta | \theta \neq 0, X = x) = \frac{e^{-ax}}{K} \phi(\theta - y) I(\theta > 0) + \frac{e^{ax}}{K} \phi(\theta - z) I(\theta < 0), \quad (13)$$

where $K = e^{ax} \Phi(-z) + e^{-ax} \Phi(y)$, and $I(\cdot)$ is the indicator function. The r^{th} moment about zero of (13) is $K^{-1} e^{-ax} J_y(r) + K^{-1} e^{ax} J_z(r)$, where $J_y(r) = \int_{-y}^{\infty} (\nu + y)^r \phi(\nu) d\nu$ and $J_z(r) = \int_{-\infty}^{-z} (\nu + z)^r \phi(\nu) d\nu$. For $r = 1, 2, 3, 4$, these quantities are

$$\begin{aligned} J_y(1) &= \phi(y) + y\Phi(y), & J_z(1) &= -\phi(y) + z\Phi(-z), \\ J_y(2) &= y\phi(y) + (y^2 + 1)\Phi(y), & J_z(2) &= -z\phi(z) + (z^2 + 1)\Phi(-z), \\ J_y(3) &= (y^2 + 2)\phi(y) + (y^3 + 6y)\Phi(y), \\ J_z(3) &= -(z^2 + 2)\phi(z) + (z^3 + 6z)\Phi(-z), \\ J_y(4) &= (y^3 + 5y)\phi(y) + (y^4 + 6y^2 + 3)\Phi(y), \quad \text{and} \\ J_z(4) &= -(z^3 + 5z)\phi(z) + (z^4 + 6z^2 + 3)\Phi(-z). \end{aligned}$$

5.2 Closing remarks

We have derived an approximation to the posterior distribution of the `BayesThresh` estimate of the unknown regression function value $g(t_i)$ given the data \mathbf{y} for $t_i = i/n, i = 0, \dots, n - 1$. This approximation can be used to assess interval estimates of $g(t_i)$ with any desired nominal coverage probability. Our simulations indicate that the `WaveBand` method produces reasonable interval estimates, especially for higher nominal coverage rates.

Several authors have proposed Bayesian thresholding rules in which the priors on the wavelet coefficients are dependent. Investigation of these thresholding rules may provide a route to the evaluation of simultaneous posterior credible bands. However, the computation involved would be significantly more demanding as the cumulants of a sum of dependent random variables do not share property (9). Also, the independence of the posterior distributions means that the computations involved are of order n ; this would no longer be the case if the priors were dependent.

Software to implement our methods is available from the first author and will be included in the next release of the `WaveThresh` package by Nason (1998), the current release of which is available from <http://www.stats.bris.ac.uk/~wavethresh>.

Acknowledgments

The authors are grateful for the support of the EPSRC (grant GR/M10229) and by Unilever Research; GPN was also supported by EPSRC Advanced Research Fellowship AF/001664. The authors wish to thank Eric Kolaczyk and Thomas Yee for programs that implement the ABWS and spline methods respectively. The authors are also grateful for the constructive comments of the Joint Editor and two referees.

References

- Abramovich, F., Bailey, T. C. and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *The Statistician* **49**, 1–29.
- Abramovich, F. and Sapatinas, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In Müller, P. and Vidakovic, B., editors, *Bayesian Inference in Wavelet Based Models*, volume 141 of *Lecture Notes in Statistics*, pages 33–50. Springer-Verlag, New York.
- Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc. B* **60**, 725–749.
- Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.
- Brillinger, D. R. (1994). Some river wavelets. *Environmetrics* **5**, 211–220.
- Brillinger, D. R. (1996). Uses of cumulants in wavelet analysis. *J. Nonparam. Statist.* **6**, 93–114.
- Bruce, A. G. and Gao, H. Y. (1996). Understanding WaveShrink: variance and bias estimation. *Biometrika* **83**, 727–745.
- Chipman, H., Kolaczyk, E. and McCulloch, R. (1997). Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.* **92**, 1413–1421.
- Chipman, H. A. and Wolfson, L. J. (1999). Prior elicitation in the wavelet domain. In Müller, P. and Vidakovic, B., editors, *Bayesian Inference in Wavelet Based Models*, volume 141 of *Lecture Notes in Statistics*. Springer-Verlag, New York.
- Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc. B* **62**, 681–698.
- Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–402.
- Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *App. Comp. Harm. Analysis* **1**, 54–81.
- Crouse, M., Nowak, R. and Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Processing* **46**, 886–902.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.
- Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.
- Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.* **90**, 1200–1224.

- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman & Hall, London.
- Herrick, D. R. M. (2000). *Wavelet Methods for Curve Estimation*. PhD thesis, University of Bristol, Bristol, U.K.
- Hill, I. D. (1976). Algorithm AS100: Normal-Johnson and Johnson-Normal transformations. *Appl. Statist.* **25**, 190–192. Available from the StatLib archive.
- Hill, I. D., Hill, R. and Holder, R. L. (1976). Algorithm AS99: Fitting Johnson curves by moments. *Appl. Statist.* **25**, 180–189. Available from the StatLib archive.
- Holmes, C. C. and Denison, D. G. T. (1999). Bayesian wavelet analysis with a model complexity prior. In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., editors, *Bayesian Statistics*, volume 6, pages 769–776. Oxford University Press, Oxford.
- Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* **36**, 149–176.
- Mallat, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Am. Math. Soc.* **315**, 69–87.
- Moulin, P. and Liu, J. (1999). Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. *IEEE Transactions on Information Theory* **45**, 909–919.
- Nason, G. P. (1998). Wavethresh3 Software. Department of Mathematics, University of Bristol, UK. Available from <http://www.stats.bris.ac.uk/~wavethresh>.
- Nason, G. P. and Silverman, B. W. (1994). The discrete wavelet transform in S. *J. Comp. Graph. Stat.* **3**, 163–191.
- Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28**, 298–335.
- Stuart, A. and Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics, volume 1: Distribution Theory*. Edward Arnold, London.
- Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Am. Statist. Ass.* **93**, 173–179.
- Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley & Sons, Inc., New York.
- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.