Sankhyā: The Indian Journal of Statistics
Special issue on Wavelets
2001, Volume 63, Series A, Pt. 3, pp. 287–308

BAYESIAN NONPARAMETRIC REGRESSION USING WAVELETS

By JEAN-FRANCOIS ANGERS
Université de Montréal, Canada
and
MOHAN DELAMPADY
Indian Statistical Institute, Bangalore, India

SUMMARY. Multi-resolution analysis is used here to derive a wavelet smoother as an estimated regression function for a given set of noisy data. The hierarchical Bayesian approach is employed to model the regression function using a wavelet basis and to perform the subsequent estimations. The Bayesian model selection tool of Bayes factor is used to select the optimal resolution level of the multi-resolution analysis. Error bands are provided as an index of estimation error. The methodology is illustrated with two examples and a simulation study.

1. Introduction

We consider the problem of fitting a general regression function to a set of observations. The observations are assumed to arise from a real valued regression function defined on an interval on the real line. Specifically, we consider the model

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \text{ and } x_i \in \mathcal{T}, \tag{1}$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)' \sim N(0, \sigma^2 I)$, σ^2 unknown and $f(\cdot)$ is a function defined on some index set $\mathcal{T} \subset \mathbb{R}^1$.

Paper received July 2000.
AMS (1991) *subject classification.* Primary 62G08; secondary 62A15, 62F15.
Key words and phrases. Function estimation, Hierarchical Bayes, Model choice.

Proceeding further, we embed this problem in the same set-up as in Angers and Delampady (1992) (AD, hereafter), but instead of the Taylor series expansion employed there, we apply the multi-resolution wavelet analysis. This leads to a representation for f as a linear combination of a set of wavelet functions instead of the sum of a polynomial and a remainder term in AD. This also implies that the regression function f need not be very smooth as was assumed in AD. However, unlike other approaches based on wavelets, we do not transform the data using a discrete wavelet transform. Consequently, we continue to work with the noisy data y_i and not their estimated wavelet coefficients. Also, by employing an appropriate covariance structure for the wavelet coefficients, we reduce the total number of parameters which need to be finally estimated in order to obtain a wavelet smoother.

In the following, we shall first develop a model for the regression function, and then using the hierarchical Bayesian approach we shall derive the Bayesian wavelet smoother. In section 2, we develop a model for f using a wavelet based decomposition. The hierarchical Bayesian approach is then used to describe the prior distribution of the involved parameters. The Bayesian wavelet smoother and its error bands are derived in section 3. Section 4 describes how Bayes factors can be used to determine the most parsimonious model for the given data. Our final results and the methodology are illustrated with two real life data sets in the following section. A simulation study is included to compare the strength of our methodology with that of some of the competitors. A few possible extensions are also indicated here. Finally, the strength of our approach and some comparisons with other approaches are mentioned in the final section.

2. Description of the Model

First, we decompose the regression function into a linear combination of a set of basis functions. We begin with a compactly supported wavelet function $\psi \in \mathcal{C}^s$, the set of real-valued functions with continous derivatives up to order s. Note then that any function f in $\mathcal{L}_2(\mathbb{R})$ has the wavelet decomposition

$$f(x) = \sum_{|k|\leq K_0} \alpha_k \phi_k(x) + \sum_{j\geq 0} \sum_{|k|\leq K_j} \beta_{jk} \psi_{j,k}(x), \tag{2}$$

with

$$\phi_k(x) = \phi(x-k),$$

and

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

where K_j is such that $\phi_k(x)$ and $\psi_{j,k}(x)$ vanish on $\mathcal{T}$ whenever $|k| > K_j$, and ϕ is the scaling function ('father wavelet') corresponding to the 'mother wavelet' ψ. Such K_j's exist (and are finite) since the wavelet function that we have chosen has compact support. For any specified resolution level J, we have

$$\begin{aligned} f(x) &= \sum_{|k|\le K_0} \alpha_k \phi_k(x) + \sum_{j=0}^{J} \sum_{|k|\le K_j} \beta_{jk}\psi_{j,k}(x) + \sum_{j=J+1}^{\infty} \sum_{|k|\le K_j} \beta_{jk}\psi_{j,k}(x) \\ &= g_J(x) + R_J(x), \end{aligned} \tag{3}$$

where

$$\begin{aligned} g_J(x) &= \sum_{|k|\le K_0} \alpha_k \phi_k(x) + \sum_{j\ge 0}^{J} \sum_{|k|\le K_j} \beta_{jk}\psi_{j,k}(x), \text{ and} \\ R_J(x) &= \sum_{j=J+1}^{\infty} \sum_{|k|\le K_j} \beta_{jk}\psi_{j,k}(x). \end{aligned} \tag{4}$$

Therefore, $g_{J+1}(x) - g_J(x) = R_J(x) - R_{J+1}(x)$ for any resolution level J. Delyon and Juditsky (1995) consider a similar model but adopt an approach for analysis which is very different from ours. Some other good references for general wavelet related material and applications are Daubechies (1992), Härdle *et. al.* (1998), Mallat (1998), and Ogden (1997).

2.1. *First stage prior model and bound on J.* To proceed further, first, the above shown wavelet decomposition is exploited to give the prior model for f in (1) using the hierarchical Bayesian approach. The strength of wavelet basis over others is that wavelets can identify local features much more effectively and efficiently. Note in the representation (3) that there are ϕ functions to detect the global features, and once this is done, there are the ψ functions to check for local details. Such a representation is especially useful when we do not assume that the regression function has any global smoothness features, and hence one may not find 'spline type models' very appropriate.

We would like to mention here that, unlike many standard wavelet based procedures, we do not need either the number of observations to be an integer power of 2, or equally spaced observations. The hierarchical Bayesian

approach that we employ can handle general situations quite efficiently. For this reason, unlike most of the other Bayesian approaches to wavelet based smoothing, we do not see the need to apply the 'discrete wavelet transform' first and then work with the wavelet coefficients assuming that they are independent observations. Details on related Bayesian approaches to wavelet based function estimation can be found in Vidakovic (1998) and Müller and Vidakovic (1999).

Note that at the resolution level J, equation (1) can be rewritten as

$$y_i = g_J(x_i) + \eta_i + \varepsilon_i, \tag{5}$$

where $\eta_i = R_J(x_i)$. Since there is usually very little information available in the likelihood function to estimate the (infinitely) many parameters β_{jk}, $j > J$, $|k| \leq K_j$ (arising from the higher levels of resolution), one will need to engage in the difficult task of eliciting a very informative prior on these parameters if the η_i are to be estimated. This will then attract prior robustness issues as well. Therefore, we adopt an approach wherein these remainders η_i are clubbed with the measurement errors ε_i. In this way, our approach will be truly Bayesian, but at the inference stage we can treat these η_i as nuisance quantities and eliminate them by integrating out (rather than estimating) the corresponding β_{jk}.

Even after treating the η_i's as nuisance parameters, we are still left with the problem of estimating the parameters, α_k and β_{jk} for $|k| \leq K_j$ and $j = 0, 1, \ldots, J$. Let l_ϕ, l_ψ and l_X be the length of the support of $\psi(\cdot)$, $\phi(\cdot)$ and the length of $\mathcal{T}$ respectively. Then the total number of parameters, α's and β's which need to be estimated (from the data and our prior knowledge) is bounded by

$$l_X 2^{J+1} + J(l_\psi + 1) + (l_\phi + l_\psi + 2). \tag{6}$$

Since the total number of observations is n, J should be chosen such that the quantity in (6) is less than n. Failure to do so can make the design matrix (see next subsection) singular and hence the estimated values of α_k and β_{jk} may then be strongly dependent on the prior.

As an example, consider the Daubechies wavelet of order p (*cf.* Daubechies, 1992). The support of ψ and ϕ then are, respectively, $(0, 2p-1)$ and $(1-p, p)$, so that $l_\psi = l_\phi = 2p - 1$. Suppose also that $\mathcal{T} = (0, 1)$ and so $l_X = 1$. Then we have

$$\begin{aligned} \phi_k(x) \neq 0 &\iff 0 \leq x - k \leq 2p - 1 \\ &\iff x - 2p + 1 \leq k \leq x. \end{aligned}$$

However, since $0 \leq x \leq 1$, the maximum possible value for k is $1-(-2p+1)+1 = 2p+1$. Similarly, for any given j,

$$\begin{aligned} \psi_{j,k}(x) \neq 0 &\iff 1-p \leq 2^j x - k \leq p \\ &\iff 2^j x - p \leq k \leq 2^j x + p - 1. \end{aligned}$$

This interval is contained in $(-p, 2^j + p - 1)$. Hence, there are at most $2^j + p - 1 + p - 1 = 2^j + 2p$ possible values for k at any given level j. Consequently, the total number of parameters to be estimated in equation (5) is bounded by

$$\begin{aligned} 2p+1+\sum_{j=0}^{J}(2^j+2p) &= 2p+1+\frac{2^J-1}{2-1}+2p(J+1) \\ &= 2^J+2Jp+4p \\ &= 1\times 2^J + J(2p-1+1)+(2p-1+2p-1+2) \\ &= l_X 2^J + J(l_\psi+1)+(l_\phi+l_\psi+2). \end{aligned}$$

To provide a joint prior distribution on α_k and β_{jk} we assume that they are all independent normal random variables with mean 0. The prior variance of α_k is assumed to be τ^2, whereas to accommodate the decreasing effect of the 'detail' coefficients β_{jk}, we assume that their variance is $\tau^2/2^{2js}$, (*cf.* Abramovich and Sapatinas (1999) for further justification for the choice of the first stage prior variances). Once a joint prior distribution is specified for σ^2 and (the hyperparameter) τ^2, the prior model is complete. Note, further, that since we assign a second stage prior on the variance factor τ^2 of the α and β coefficients, their marginal prior distribution will no longer be normal, but a heavier tailed distribution ensuring a certain degree of prior robustness to our wavelet smoother (*cf.* Berger, 1985).

2.2. *First stage posterior densities.* Let $\gamma = (\alpha', \beta')'$ where $\alpha = (\alpha_k)_{|k|\leq K_0}$, and $\beta = (\beta_{jk})_{0\leq j\leq J, |k|\leq K_j}$. The prior specified above indicates that $\gamma|\tau^2 \sim N(0, \tau^2\Gamma)$ where

$$\Gamma = \begin{pmatrix} I_{2K_0+1} & 0 \\ 0 & \Delta_{M_\beta} \end{pmatrix},$$

where $M_\beta = \sum_{j=0}^{J}(2K_j+1)$ and with Δ being the variance-covariance matrix of β (which is also diagonal, with the diagonal entries as specified in the previous paragraph). Also $(\eta_1, \ldots \eta_n)'|\tau^2 \sim N(0, \tau^2 Q_n)$, where Q_n is formed from the variance structure of $\{\beta_{jk}\}_{J+1\leq j<\infty, |k|\leq K_j}$. Since the β_{jk}

are assumed to be independent, the (i,l) entry of Q_n is given by

$$\begin{aligned}(Q_n)_{i,l} &= \tau^{-2}\mathbb{C}ov(\eta_i,\eta_l)\\ &= \tau^{-2}\mathbb{C}ov\left(\sum_{j\geq J+1}\sum_{|k|\leq K_j}\beta_{jk}\psi_{jk}(x_i),\sum_{p\geq J+1}\sum_{|q|\leq K_j}\beta_{pq}\psi_{pq}(x_l)\right)\\ &= \tau^{-2}\sum_{j\geq J+1}\sum_{p\geq J+1}\sum_{|k|\leq K_j}\sum_{|q|\leq K_j}\mathbb{C}ov(\beta_{jk},\beta_{pq})\psi_{jk}(x_i)\psi_{pq}(x_l)\\ &= \sum_{j\geq J+1}\sum_{|k|\leq K_j}2^{-2js}\psi_{jk}(x_i)\psi_{jk}(x_l).\end{aligned}$$

Since the ψ function is bounded, it can be shown easily that each entry of Q_n is bounded by

$$|(Q_n)_{i,l}| \leq \frac{(l_\psi+1)(\max_z \psi^2(z))}{2^{2Js}(2^{2s}-1)}.$$

Hence, the covariance matrix of $(\eta_1,\ldots,\eta_n)'$ is well defined. In the last section a sensitivity study on the choice of Q_n is also discussed to dispel any fear that our choice of Q_n may be unduly influencing our wavelet smoother.

Let $X=(\Phi',S')$ with the ith row of Φ' being $\{\phi_k(x_i)\}'_{|k|\leq K_0}$ and the ith row of S' being $\{\psi_{jk}(x_i)\}'_{0\leq j\leq J,|k|\leq K_j}$. Then we obtain the following structure. Given γ, σ^2 and τ^2, we have the following linear model for the observation vector $Y=(y_1,\ldots,y_n)'$:

$$Y = X\gamma + u,$$

where $u=\eta+\varepsilon\sim N(0,\Sigma)$ with $\Sigma=\sigma^2 I_n+\tau^2 Q_n$. This follows from the fact that

$$\begin{aligned} Y|\gamma,\eta,\sigma^2,\tau^2 &\sim N(X\gamma+\eta,\sigma^2 I_n),\\ \eta|\tau^2 &\sim N(0,\tau^2 Q_n).\end{aligned} \tag{7}$$

Note that this model is similar to the set-up in AD and to an extent also similar to that in Angers and Delampady (1997), but unlike in those articles, here we do not assume that the regression function is necessarily very smooth. A major departure in this paper, however, is that the remainder terms (here denoted by $(\eta_1,\ldots,\eta_n)'$) are treated as nuisance parameters. In fact, since wavelet basis can approximate any function of $\mathcal{L}_2(\mathbb{R})$, we do not need to complement the basis functions as in AD using the remainder terms. They only appear to complete the model instead.

From (7) and using standard hierarchical Bayes techniques (*cf.* Lindley and Smith, 1972) and matrix identities (*cf.* Searle, 1982), it follows that

$$\begin{aligned} Y|\sigma^2,\tau^2 &\sim N(0,\sigma^2 I_n+\tau^2\left(X\Gamma X'+Q_n\right)), &(8)\\ \gamma|Y,\sigma^2,\tau^2 &\sim N(AY,B), &(9)\end{aligned}$$

where

$$\begin{aligned} A &= \tau^2\Gamma X'\left(\sigma^2 I_n+\tau^2\left(X\Gamma X'+Q_n\right)\right)^{-1},\\ B &= \tau^2\Gamma-\tau^4\Gamma X'\left(\sigma^2 I_n+\tau^2\left(X\Gamma X'+Q_n\right)\right)^{-1}X\Gamma.\end{aligned}$$

In order to proceed to the second stage calculations, some algebraic simplifications are needed (*cf.* AD). Spectral decomposition yields $X\Gamma X'+Q_n = HDH'$, where $D=\mathrm{diag}(d_1,d_2,\ldots,d_n)$ is the matrix of eigenvalues and H is the orthogonal matrix of eigenvectors. Thus,

$$\begin{aligned}\sigma^2 I_n+\tau^2\left(X\Gamma X'+Q_n\right) &= H\left(\sigma^2 I_n+\tau^2 D\right)H'\\ &= \tau^2 H\left(vI_n+D\right)H',\end{aligned}$$

where $v=\sigma^2/\tau^2$. Using this spectral decomposition, the marginal density of Y given τ^2 and v can be written as

$$\begin{aligned} m(Y\mid\tau^2,v) &= \frac{1}{(2\pi\tau^2)^{n/2}}\frac{1}{\det(vI_n+D)^{1/2}}\\ &\quad\times\exp\left\{-\frac{1}{2\tau^2}Y'H(vI_n+D)^{-1}H'Y\right\}\\ &= \frac{1}{(2\pi\tau^2)^{n/2}}\frac{1}{\prod_{i=1}^n(v+d_i)^{1/2}}\exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^n\frac{s_i^2}{v+d_i}\right\} &(10)\end{aligned}$$

where $s=(s_1,\ldots,s_n)'=H'Y$.

3. Second Stage Prior and Estimation of $g_J(x)$

To derive the wavelet smoother, all that we need to do now is to eliminate the hyper- and nuisance parameters from the first stage posterior distribution, by integrating out these variables with respect to the second stage prior on them. Since it is well known (*cf.* Berger, 1985) that the final Bayes estimator does not depend crucially on the second and higher stage hyper-priors,

these priors can be chosen to simplify computations. Accordingly, the priors on τ^2 and v are chosen as in the next subsection.

An alternative method would be to use an empirical Bayes approach and to estimate σ^2 and τ^2 from equation (8) and to replace σ^2 and τ^2 by their estimates in equation (9) to approximate $\widehat{\gamma}$. In order to maximize equation (8) with respect to σ^2 and τ^2, we proceed as follows:

1. Let $v = \sigma^2/\tau^2$.

2. Solve, in terms of v,

$$\sum_{i=1}^{n} \frac{s_i^2}{(v+d_i)^2} \sum_{j=1}^{n} \left(\frac{d_j - d_i}{v + d_j} \right) = 0.$$

   Denote this solution by $\widehat{v}$.

3. Compute

$$\widehat{\tau}^2 = n^{-1} \sum_{i=1}^{n} (s_i^2)/(\widehat{v} + d_i),$$
$$\widehat{\sigma^2} = \widehat{v}\widehat{\tau}^2.$$

Although this approach is easier to implement than the one used in the rest of this paper, it underestimates the variance of the wavelet estimator $\widehat{Y} = X\widehat{\gamma}$.

3.1 *Prior on τ^2 and v.* If one chooses $\pi_{21}(\tau^2) \propto (\tau^2)^{-c}$, the subsequent integrations with respect to $\pi_{21}(\tau^2 \mid Y, v)$ needed to compute the Bayes estimator of $g_J(x)$ and its posterior variance, can be handled analytically. Hence the numerical integral, unavoidable in a hierarchical Bayesian model, will be reduced to a one dimensional integral.

Furthermore, with this choice of prior on τ^2, the marginal prior on γ will have the form

$$\pi(\gamma) \propto \left(\sum_{i=1}^{m} \frac{\gamma_i^2}{\Gamma_i} \right)^{(m/2)+c-1},$$

where Γ_i is the ith diagonal entry of Γ and m denotes the maximum number of parameters to be estimated (*cf.* equation (6)). This prior density corresponds to the limiting case of a multivariate Student's-t density. Thus we obtain a prior having heavier tails (as recommended in Vidakovic, 1998) than the prior model proposed in Abramovich and Sapatinas (1999).

To arrive at the (second-stage) prior on v we argue as follows. Recall that $v = \sigma^2/\tau^2$. It is then the ratio of two variances – the ratio of the variance of the error terms (ε_i) to the prior variance of α_k. It seems then 'natural' to model this ratio using an F distribution with a and b degrees of freedom. In order to be as noninformative as possible, a and b are chosen such that:

1. the prior variance of v $\left(= \frac{2b^2(a+b-2)}{a(b-4)(b-2)^2}\right)$ is infinite;
2. the Fisher information number $\left(= \frac{a^2(b+2)(b+6)}{2(a-4)(a+b+2)}\right)$ is minimum;
3. the prior mode $\left(= \frac{b(a-2)}{a(b+2)}\right)$ is greater than 0.

This can be done by choosing $2 < b \leq 4$ and $a = 8(b+2)/(b-2)$. Let $\pi_{22}(v)$ denote the resulting prior density.

Once the second stage priors are specified, using equation (9) and taking the expectation with respect to τ^2, we have that

$$\widehat{\gamma} = \Gamma X' H \mathbb{E}\left[(vI_n + D)^{-1} \mid Y\right] s, \tag{11}$$

where the expectation is taken with respect to

$$\pi_{22}(v \mid Y) \propto \frac{v^{a/2}}{(b+av)^{(a+b)/2}} \left(\prod_{i=1}^{n}(v+d_i)\right)^{-1/2} \left(\sum_{i=1}^{n} \frac{s_i^2}{v+d_i}\right)^{-(n+2c)/2}. \tag{12}$$

(Note that in order for $\pi_{22}(v \mid Y)$ to be a proper density, c should be chosen such that $c < b/2$.) Again using equation (9), the posterior expected loss (under squared error loss) of γ can be written as

$$\begin{aligned} \mathbb{V}ar(\gamma \mid Y) \;=\; & \frac{1}{n+2c}\mathbb{E}\left[\sum_{i=1}^{n} \frac{s_i^2}{v+d_i} \mid Y\right] \Gamma \\ & - \frac{1}{n+2c}\Gamma X' H \mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{s_i^2}{v+d_i}\right)(vI_n + D)^{-1} \mid Y\right] H' X \Gamma \\ & + \mathbb{E}\left[\widehat{\gamma}(v)\widehat{\gamma}(v)' \mid Y\right], \end{aligned} \tag{13}$$

where $\widehat{\gamma}(v) = \Gamma X' H (vI_n + D)^{-1} s$.

To compute these expectations, one can use several techniques. ¿From the many data sets that we have analyzed to obtain our wavelet smoother we have strong empirical evidence showing that $\pi_{22}(v \mid Y)$ is unimodal with its mode away from 0. Therefore, for data sets with moderately large number

of observations one can use the Laplace approximation to evaluate equations (11) and (13) accurately. In order to do so, the first three derivatives of $h(v) = -n^{-1}\log(\pi_{22}(v \mid Y))$ are needed. They are given by

$$\begin{aligned}
h^{(1)}(v) &= -\frac{1}{2n}\left(\frac{a-2}{v} - \frac{a(a+b)}{b+av} - \sum_{i=1}^{n}\frac{1}{v+d_i} + (n+2c)\frac{p_2(v)}{p_1(v)}\right), \\
h^{(2)}(v) &= \frac{1}{2n}\left(\frac{a-2}{v^2} - \frac{a^2(a+b)}{(b+av)^2} - \sum_{i=1}^{n}\frac{1}{(v+d_i)^2}\right. \\
&\qquad \left. +2(n+2c)\frac{p_3(v)}{p_1(v)} - (n+2c)\left(\frac{p_2(v)}{p_1(v)}\right)^2\right), \\
h^{(3)}(v) &= -\frac{1}{n}\left(\frac{a-2}{v^3} - \frac{a^3(a+b)}{(b+av)^3} - \sum_{i=1}^{n}\frac{1}{(v+d_i)^3}\right. \\
&\qquad \left. +3(n+2c)\frac{p_4(v)}{p_1(v)} - 2(n+2c)\frac{p_2(v)p_3(v)}{p_1^2(v)}(n+2c)\left(\frac{p_2(v)}{p_1(v)}\right)^3\right),
\end{aligned}$$

where $p_l(v) = \sum_{i=1}^{n}\frac{s_i^2}{(v+d_i)^l}$.

Since our computations involve only single dimensional integrals there is no compelling need to resort to the above mentioned Laplace approximation. Several versions of the standard Monte-Carlo approach can be employed quite satisfactorily and efficiently also. We would like to note that in our examples we have used both the Laplace approximation and Monte-Carlo, and we do not see any major differences.

3.2 *Prediction error.* To evaluate our methodology we need to derive estimation and prediction errors. Towards this, note that our wavelet smoother yields $\widehat{Y} = X\widehat{\gamma}$. From equation (13), the posterior variance of $\widehat{Y}$ can be expressed easily as $\mathbb{V}ar(\widehat{Y}) = X\mathbb{V}ar(\gamma \mid Y)X'$. However, if we want to obtain the posterior variance of the predicted value of y for any given x, we proceed as follows:

$$\begin{aligned}
\mathbb{V}ar(y \mid Y) &= \mathbb{V}ar(g_J(x) + \eta + \varepsilon \mid Y) \\
&= \mathbb{V}ar(g_J(x) \mid Y) + \mathbb{V}ar(\eta + \varepsilon \mid Y) \\
&= \mathbb{V}ar(x'\gamma \mid Y) + \mathbb{V}ar(\mathbb{E}[\eta + \varepsilon \mid \sigma^2] \mid Y) + \mathbb{E}[\mathbb{V}ar(\eta + \varepsilon \mid \sigma^2) \mid Y] \\
&= x'\mathbb{V}ar(\gamma \mid Y)x + \mathbb{E}[\sigma^2 + \tau^2 q(x) \mid Y] \\
&= x'\mathbb{V}ar(\gamma \mid Y)x + \frac{1}{n+2c}\mathbb{E}\left[\sum_{i=1}^{n}\frac{[v+q(x)]s_i^2}{v+d_i} \mid Y\right],
\end{aligned}$$

where $q(x) = \sum_{j \geq J+1} \sum_{|k| \leq K_j} (2^{-js} \psi_{jk}(x))^2$. Even though the predictive density of $y \mid Y$ is not normal (in fact, it is a mixture of normal densities), we would still like to suggest using

$$\widehat{y} \pm 2\sqrt{\mathbb{V}ar(y \mid Y)} \tag{14}$$

as error bands for $\widehat{y} = \widehat{g}_J(x)$. These are not to be treated strictly as Bayesian credible regions; instead, only as bands for reflecting uncertainty in $\widehat{g}_J(x)$. These error bands are also given in our illustrations in the last section.

4. Bayes Factor and Choice of J

In this section we describe how the optimal level of resolution J is to be determined. As indicated above in (3), the choice of J provides a model for the observations through the choice of the corresponding regression function. Using equation (6) an upper bound on J can be derived as a function of the length of the support of the 'father' and 'mother' wavelet functions and the number of observations. Let J_{max} be this upper bound on J.

Let M_l denote the model, arising from (3) and (4), corresponding to the resolution level l. Our task is to pick the best model for the given data from the set of models:

$$M_l,\ l = 1, 2, \ldots, J_{max}.$$

The well accepted method for deciding between two possible models M_l and $M_{l'}$ is to compute the Bayes factor of M_l relative to $M_{l'}$:

$$B(M_l : M_{l'}) \quad = \quad \frac{m(Y|M_l)}{m(Y|M_{l'})}, \tag{15}$$

where $m(Y|M_i)$ denotes the marginal density of Y under the model M_i, $i = l, l'$. ¿From (8) above it follows that, under M_l, $Y|\sigma^2, \tau^2 \sim N(0, \sigma^2 I_n + \tau^2 (X_l \Gamma_l X_l' + Q_{n,l}))$, where we have shown the dependence of X, Γ and Q_n on l explicitly with subscripts. It follows then that

$$m(Y|M_l) \quad = \quad \int m(Y|M_l, \sigma^2, \tau^2)\, d\pi(\sigma^2, \tau^2),$$

where $\pi(\sigma^2, \tau^2)$ is the joint prior distribution on σ^2 and τ^2.

As in the previous section consider the spectral decomposition of $X_l \Gamma_l X_l' + Q_{n,l}$. Let D_l and H_l be such that $X_l \Gamma_l X_l' + Q_{n,l} = H_l D_l H_l'$. Also, let $d_{l,i}$

be the ith diagonal element of D_l and let $s_l = H_l Y = (s_{l,1}, \ldots, s_{l,n})'$. Then, using equation (12), the marginal density of Y under M_l can be expressed as

$$m_l(Y) = \int_0^\infty \frac{v^{a/2}}{(b+av)^{(a+b)/2}} \left(\prod_{i=1}^n (v + d_{l,i}) \right)^{-1/2} \left(\sum_{i=1}^n \frac{s_{l,i}^2}{v + d_{l,i}} \right)^{-(n+2c)/2} dv.$$

Consequently, to choose the best model M_l, one needs to compute $m_l(Y)$ for $l = 1, \ldots, J_{max}$. Let $M_{J_{max}}$ be the reference model and define the Bayes factors $B_l = m_l(Y)/m_{J_{max}}(Y)$ for $l = 1, \ldots, J_{max}$. (Note that $B_{J_{max}} = 1$.) Then the best value of l (equivalently, the best model M_l) is the one for which B_l is maximum. This method will be illustrated in the next section when we analyze some data sets.

Therefore, if l_* is such that $B_{l_*} = \max_{1 \le l \le J_{max}} B_l$, then the model considered would be

$$\begin{aligned} f(x) &= \sum_{|k| \le K_0|} \alpha_k \phi_k(x) + \sum_{j=0}^{l_*} \sum_{|k| \le K_j} \beta_{kj} \psi_{j,k}(x) + \sum_{j=l_*}^{\infty} \sum_{|k| \le K_j} \beta_{kj} \psi_{j,k}(x) \\ &= g_{l_*}(x) + R_{l_*}(x), \end{aligned}$$

and γ is estimated by

$$\widehat{\gamma} = \Gamma_{l_*} X_{l_*}' H_{l_*} \mathbb{E}[(vI_n + D_{l_*})^{-1} \mid Y] s_{l_*}.$$

5. Illustrative Examples

5.1 *Data analysis.* We illustrate our methodology with two examples in this section. Both these use real life data sets. We have chosen the wavelet function 'Daubechies of order 2' for illustration. Other wavelet functions have also been tried. For larger data sets we find that higher order Daubechies wavelet functions provide increased smoothing. In these examples, the hyperparameters have been chosen after conducting sensitivity analysis as described in AD.

EXAMPLE 1. This example is based on a data set from Ma and Zidek (1988). The data represent the monthly precipitation (rain plus one-tenth of snow) in inches from March 1965 to December 1966 in Vancouver (Canada).

In Figure 1, the posterior density of v is given. Note that even with only 22 observations and vague priors on τ^2 and v, this posterior distribution is

very informative on v. Hence we are able to employ the Laplace approximation (*cf.* Robert, 1994) to compute $\widehat{\gamma}$ and the corresponding posterior covariance matrix.

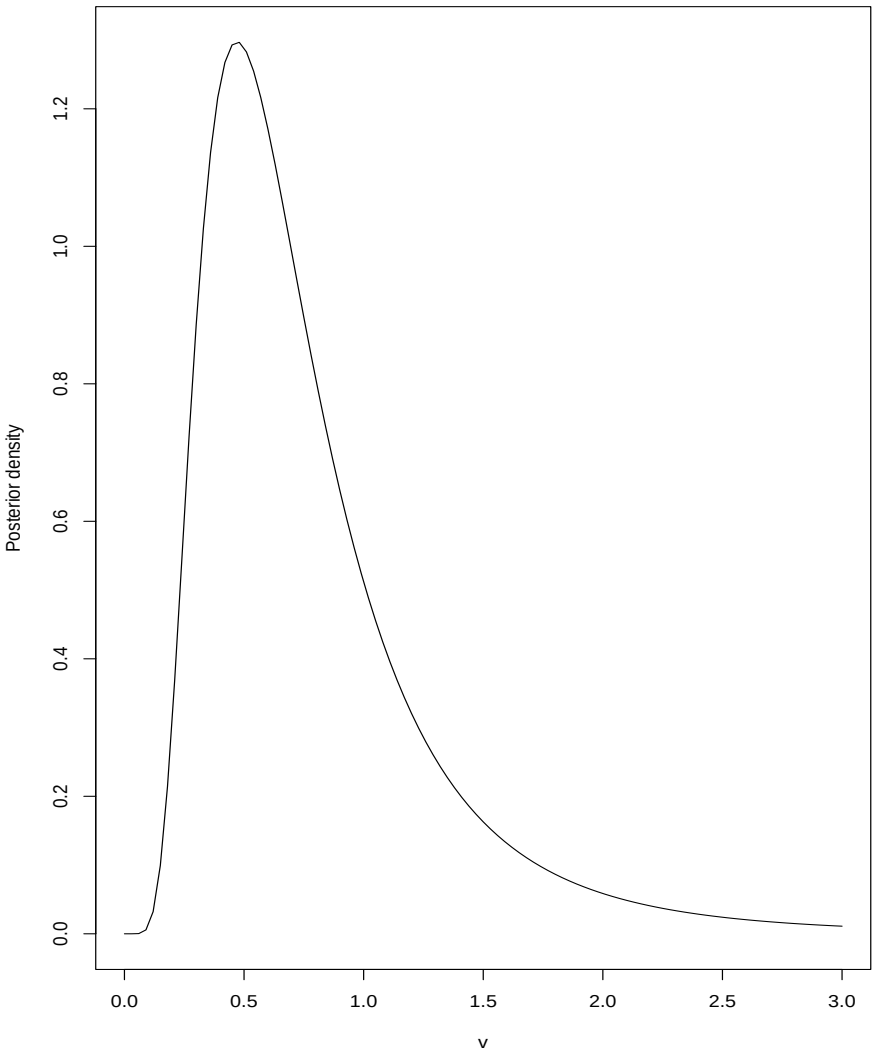

Figure 1. Posterior density of v for the first example

In Figure 2, our Bayesian wavelet smoother $(\hat{g}_J)$ for the rain data is plotted for $b = 2.5, 3, 3.5, 4$ and $c = (b-1)/2$. As suggested in Section 3, we have chosen $a = 8(b+2)/(b-2)$. Since only 22 observations are available

J is 1. Figure 2 indicates that the choice of the particular value of the hyperparameter b has very little influence on the predicted value of y_i. In fact, the maximum coefficient of variation is only 3.1%.

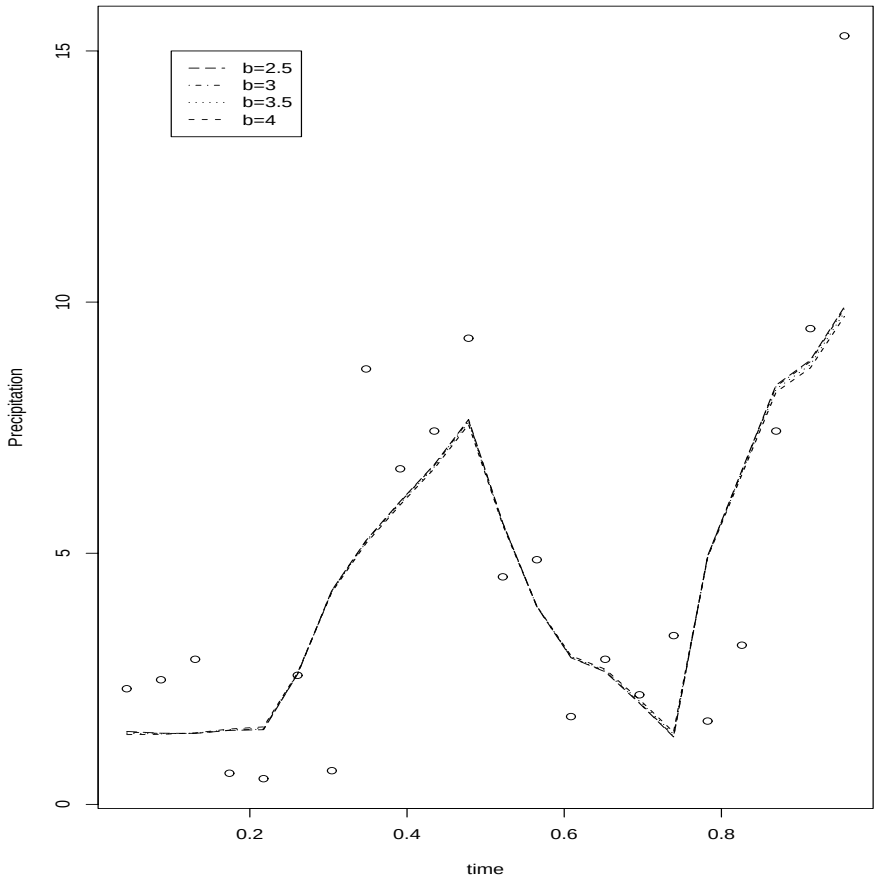

Figure 2. Sensitivity analysis of the hyperparameter b

In Figure 3, our estimator (solid line) with $b = 4$ along with the error bands (dashed lines) ($\pm 2\sqrt{v_i}$) are plotted. It can be seen that most of the observations fall inside the error bands. We have also shown (dotted line) the standard wavelet estimator (Daubechies of order 2 with hard thresholding) using the S library WaveThresh (*cf.* Nason and Silverman, (1994) or the web site: *www.stats.bris.ac.uk/pub/reports/wavelets/wavelets.html*). Since the number of observations is not a power of 2, Wavethresh needs the aid of the regressogram technique before computing the wavelet coefficients

(*cf.* Härdel *et al.*, 1998, Section 10.8). It can be further noted that our proposed methodology seems to lead to a smoother wavelet estimator than the standard procedure.

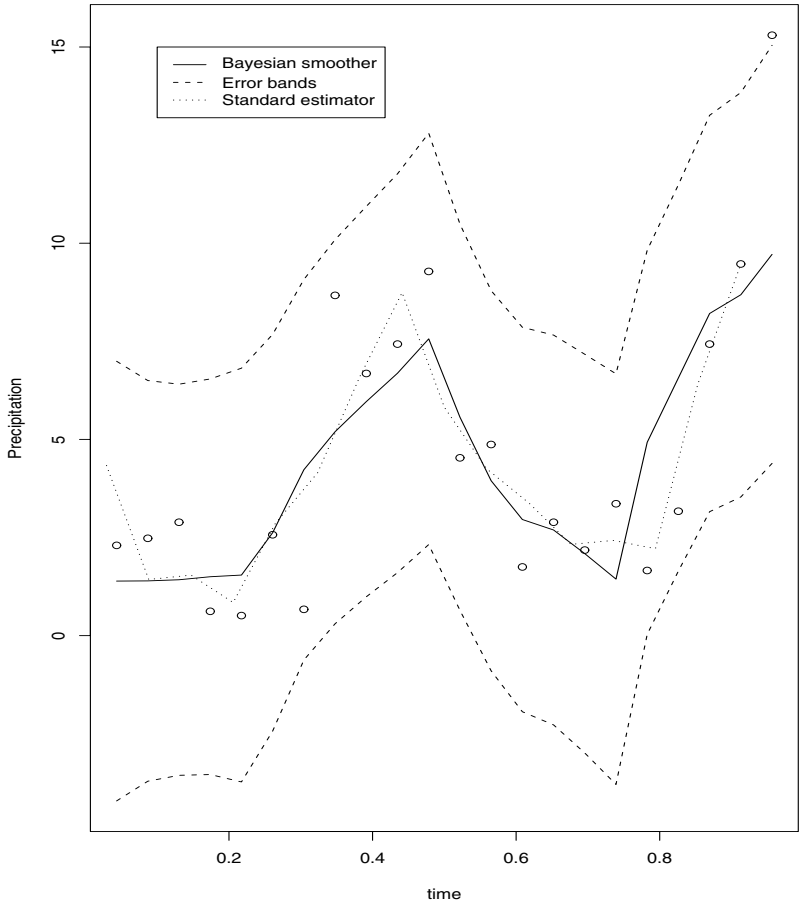

Figure 3. Standard and Bayesian ($b = 4$) estimators

In Figure 4, alternative forms for the covariance matrix of the remainder term are considered. Relevance of such extensions will be discussed in our last section. The estimated values of y_i obtained with the covariance matrix discussed in Section 2 (our proposal) are shown by the curve with solid line. The other covariance matrices considered are:

1. the covariance kernel given in AD with $c = 5$, that is; $\mathbb{C}ov(\eta_i, \eta_j) = \exp\{-c|x_i - x_j|\}/2^{2Js}$,

2. a band matrix with diagonal entries equal to 2^{-2Js}, the first lower and upper diagonal entries equal to $2^{-2(J+1)s}$ 0 everywhere else;

3. a diagonal matrix with 2^{-2Js} as diagonal entries.

The estimates of y_i obtained by these different covariance matrices are plotted in Figure 4 and they correspond, respectively, to the dashed line, the dotted line and the dot-dashed line. It is worth noting that, as with the choice of the hyperparameter b, the different forms for the covariance structure do not lead to very different smoothers, once the optimal resolution level is determined.

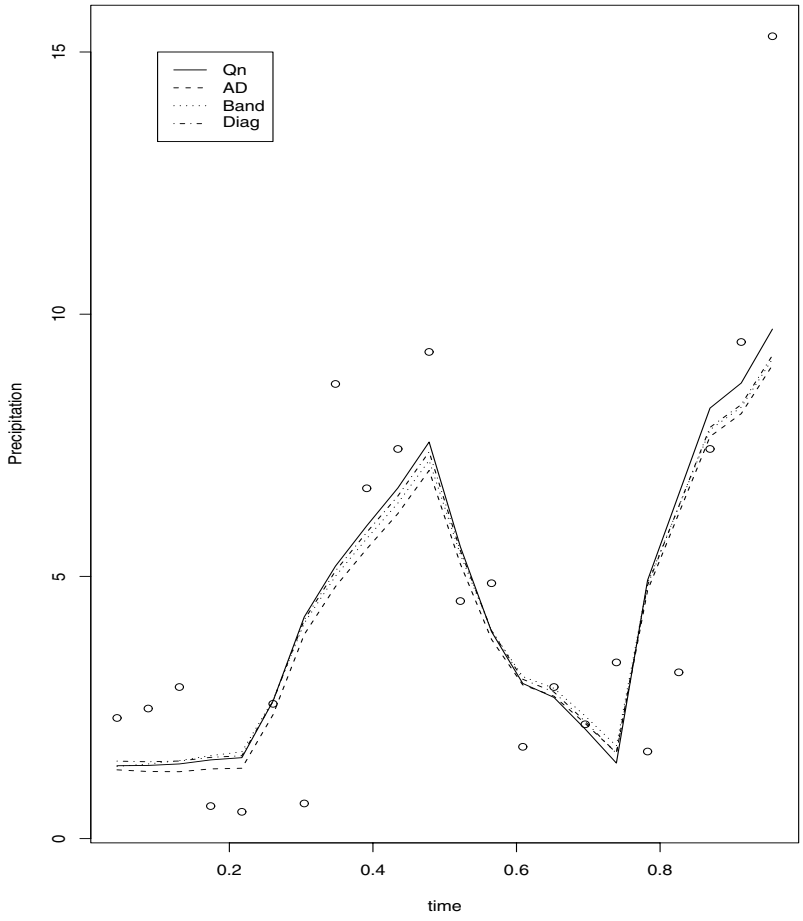

Figure 4. Sensitivity to the remainder's covariance matrix

EXAMPLE 2. The second example is based on data provided by Prof. Abraham Verghese of the Indian Institute of Horticultural Research, Bangalore, India (personal communication). The variable of interest y that we have chosen from the data set is the weekly average humidity level. The observations were made from June 1, 1995 to December 13, 1998. (For some reason, the observations were not recorded on the same day of the week everytime.) We have chosen time (day of recording the observation) as the covariate x. (Any other available covariate can be used also since wavelet based smoothing with respect to any arbitrary covariate (measured in some general way) can be handled with our methodology.) Since we have 185 observations here, the maximum possible value for J is 6. Using Bayes factors, as described in Section 3, we checked if the model corresponding to the choice $J = 5$ is more appropriate than that corresponding to $J = 6$. Doing so, the Bayes factor in favor of $J = 5$ came out to be only 0.5257. Consequently, we chose the model corresponding to $J = 6$. In Figure 5, we have plotted our wavelet smoother (solid line) along with its ($\pm 2v_i$) error bands (dotted lines). The hyperparameters were chosen to be $b = 4$, $a = 8(b+2)/(b-2) = 24$ and $c = (b-1)/2 = 1/2$.

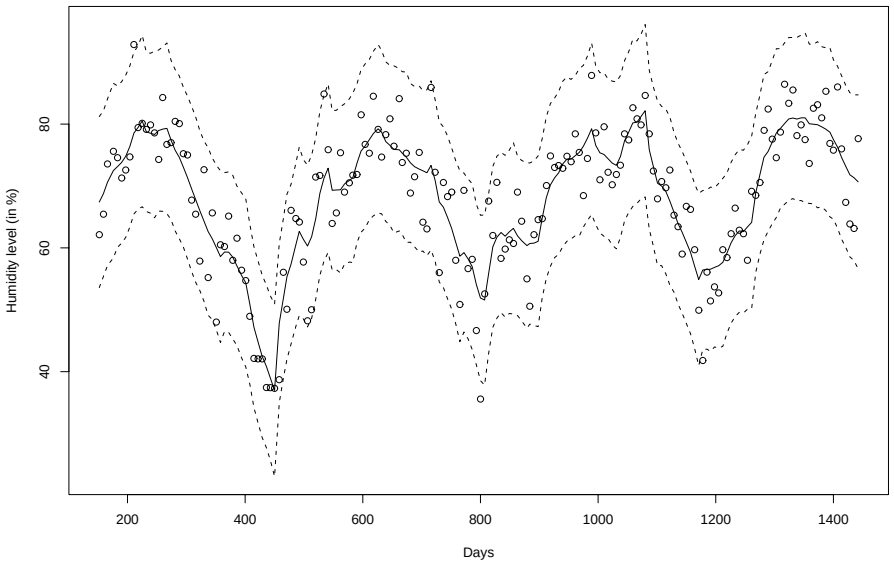

Figure 5. Wavelet smoother and its error bands for the Humidity data

Even though this data set is is larger that the first one with $n = 185$ observations, the numerical computations required to obtain our results can

still be performed on a regular desktop computer. This indicates that the matrix structure of our model is such that the singular value decomposition of the required $n \times n$ matrix does not seem to be an issue (at least for moderate size data sets).

In Figure 6, a sensitivity analysis of the order of the Daubechies wavelets is illustrated. The curves in this figure illustrate the effect of using Daubechies wavelets of order 2 (solid line), 3 (dashed line), 4 (dotted line) and 5 (dot-dashed line). Although the estimator of $g_J(x)$ becomes smoother as the order increases, the different curves are similar in shape and present the same basic features.

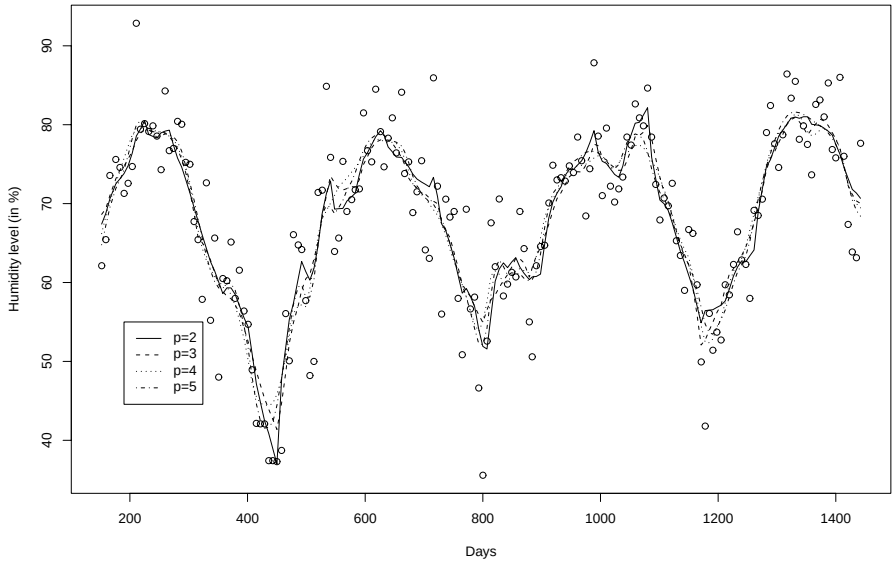

Figure 6. Sensitivity analysis with respect to the order of the Daubechies' wavelet

5.2. *Simulation study.* To investigate the strength of our methodology, we conduct a simulation study here, and compare our results with those of some competitors. In particular, we compare (i) the standard VisuShrink approach (*cf.* Donoho and Johnstone, 1994) which achieves a low variance at the expense of bias, (ii) the SureShrink approach (*cf.* Donoho and Johnstone, 1995) with soft-threshold function which minimizes an estimate of the expected mean squared error (MSE), (iii) the adaptive Bayesian wavelet shrinkage approach (ABWS, *cf.* Chipman *et al.*, 1997) which minimizes the MSE

Table 1: Simulation results for the test-functions

	Blocks			Bumps		
Method	Variance	bias^2	IMSE	Variance	bias^2	IMSE
Visu	0.0719	0.6122	0.6840	0.1165	1.4543	1.5707
Sure	0.1369	0.0856	0.2225	0.2660	0.4167	0.6827
ABWS	0.0874	0.0121	0.0995	0.2228	0.1267	0.3495
BNPR	0.0716	0.2577	0.3293	0.0701	0.2584	0.3285
	Doppler			HeaviSine		
Method	Variance	bias^2	IMSE	Variance	bias^2	IMSE
Visu	0.0523	0.4327	0.4850	0.0339	0.0864	0.1204
Sure	0.0946	0.1340	0.2285	0.0416	0.0534	0.0949
ABWS	0.1006	0.0640	0.1646	0.0442	0.0433	0.0874
BNPR	0.0116	0.0315	0.0431	0.0341	0.0515	0.0856

using the posterior mean) and (iv) our proposed estimator (BNPR). The standard test-functions of Donoho and Johnstone (1994), namely, 'Bumps', 'Blocks', 'Doppler' and 'HeaviSine' are used as the regression test-functions.

For this, $n = 1024$ points were selected from the regression function and standard normal noise was added to the function values. A total of 1000 trials were conducted and, as in Chipman *et al.* (1997), the Haar wavelets were used for 'Blocks', Daubechies wavelets of order 3 for 'Bumps', and Daubechies wavelets of order 8 for Doppler and HeaviSine. (The results for the first three estimators are taken from Chipman *et al.* (1997).)

Estimated Integrated Mean Squared Error (IMSE) of each of the estimator $\widehat{f}$ is used for comparisons as follows:

$$\begin{aligned} IMSE &= \int_{\mathcal{T}} \mathbb{E}[(\widehat{f}(x) - f(x))^2]dx \\ &= \int_{\mathcal{T}} \mathbb{V}ar(\widehat{f}(x))dx + \int_{\mathcal{T}} \text{bias}^2(\widehat{f}(x))dx. \end{aligned}$$

Table 1 gives estimates of the IMSE for the different estimators of the test-functions. It can be seen that the BNPR estimator performs excellently. It generally has a smaller variance than the others, but a somewhat larger bias than ABWS. However, except for 'Blocks', this translates into a smaller IMSE than that of the other estimators. While ABWS seems to be the choice for 'Blocks', it is BNPR which comes out on top for 'Doppler'. In these test cases, the other estimators don't seem to compete well with these two except in the case of 'Blocks'.

6. Discussion

We have developed a Bayesian wavelet smoother in this paper which seems to be different from what is currently available. We have also provided error bands as an index of estimation error. The starting point of the approach taken here is similar to that in AD but there are substantial differences in the actual methodology. The foremost among them being the treatment of the remainder R_J. It was crucial in AD to estimate this remainder since it decided whether the regression function should be a polynomial or not. Here, however, the remainder is only needed to complete the model since g_J with a large enough resolution level J can approximate any (reasonable) regression function satisfactorily. The other major difference, of course, is the use of wavelets instead of any other set of basis functions to represent a general regression function. As mentioned in the Introduction, the strength of wavelet basis over others is that wavelets can identify local features much more effectively and efficiently.

There are many differences between our proposed approach and the standard wavelet based procedures. The main difference is that the analysis is based on the data domain rather than the wavelet domain, so that assumptions which lead to easy application of the discrete wavelets transforms (DWT) are not needed. Our methodology may be perceived to be computationally intensive, requiring the singular value decomposition (SVD) of an $n \times n$ matrix. However, note that this computation is required only once and further, the involved matrix has special structure rendering this computation not very complex. An indication of this was seen in Example 2, where the entire set of computations required to bring out the smoother for a moderate sized data set could easily be done on a desktop computer. One other major difference between our approach and many other wavelet based approaches is that we do not need to impose additional thresholding since it is incorporated into the model choice criterion discussed in Section 4. The reason for this, we feel, is that the α and β parameters that we end up estimating finally are determined by a model selection rule such as Bayes factor. Finally, based on a simulation study using standard test-functions, our wavelet smoother seems to perform very well compared to the other available wavelet based estimators.

Acknowledgement. This research is partially supported by a grant of the first author from NSERC, Canada, and much of the work was done when the first author was visiting the Indian Statistical Institute, Bangalore. The authors wish to thank Prof. Abraham Verghese for allowing the use of his

data set in one of the illustrations. The authors would also like to thank the referees for their useful comments which have brought about substantial improvements on a previous draft.

References

ABRAMOVICH, F. AND SAPATINAS, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian Inference in Wavelet-Based Models*, *Lecture Notes in Statistics*, **141**, P. Müller and B. Vidakovic (Eds.), 33–50, Springer, New York.

ANGERS, J-F. AND DELAMPADY, M. (1992) Hierarchical Bayesian estimation and curve fitting, *The Canadian Journal of Statistics*, **20**, 35–49.

– – –– (1997). Hierarchical Bayesian curve fitting and model choice for spatial data, *Sankhyā*, Series B, **59** 28–43.

BERGER, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

CHIPMAN, H.A., KOLACZYK, E.D. AND MCCULLOCH, R.E. (1997). Adaptative Bayesian wavelet shrinkage, *Journal of the American Statistical Association* **92**, 1413–1421.

CLYDE, M., PARMIGIANI, G. AND VIDAKOVIC, B. (1998). Multiple shrinkage and subset selection in wavelets, *Biometrika*, **85** 391–401.

CLYDE, M. AND GEORGE, E.I. (2000). Robust empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet-Based Models*, *Lecture Notes in Statistics*, **141**, P. Müller and B. Vidakovic (Eds.), 309–322, Springer, New York.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

DELYON, B. AND JUDITSKY, A. (1995). Estimating wavelet coefficients. In *Wavelets and Statistics*, *Lecture Notes in Statistics*, **103**, A. Antoniadis and G. Oppenheim (Eds.), 151–167, Springer, New York.

DONOHO, D.L. AND JOHNSTONE, I.M. (1994), Ideal spatial adaptation via wavelet shrinkage, *Biometrika*, **81**, 424–455.

– – –– (1995), Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statisitcal Association*, **90**, 1200–1224.

HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. AND TSYBAKOV, A. (1998). *Wavelets, Approximation, and Statistical Application. Lecture Notes in Statistics*, **129**, Springer, New York.

JOHNSTONE, I.M. AND SILVERMAN, B.W. (1998). Empirical Bayes approaches to mixture problems and wavelet regression. *Technical Report*, Department of Mathematics, University of Bristol, U.K.

KOVAC, A. AND SILVERMAN, B.W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding, *Journal of the American Statisitcal Association*, **95**, 172–183.

LINDLEY, D.V. AND SMITH, A.F.M. (1972) Bayes estimates for the linear model, *Journal of the Royal Statistical Society*, B, 34, 1–41.

MA, P.H. AND ZIDEK, J.V. (1988) Data for statistics research and instruction. *Technical Report*, Department of Statistics, University of British Columbia, Vancouver.

MALLAT, S. (1998). *A Wavelet Tour of Signal Processing.* Academic Press, San Diego.

Müller, P. and Vidakovic, B. (Eds.) (1999). *Bayesian Inference in Wavelet-Based Models. Lecture Notes in Statistics*, **141**, P. Müller and B. Vidakovic, (Eds.), 33–50, Springer, New York.

Ogden, R.T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis.* Birkhäuser, Boston.

Nason, G.P. and Silverman, B.W. (1994) The discrete wavelet transform in S, *Journal of Computational and Graphical Statistics*, **3**, 163–191.

Robert, C.P. (1994). *The Bayesian Choice.* Springer, New York.

Searle, S.R. (1982) *Matrix Algebra useful for Statistics.* Wiley, New York.

Vidakovic, B. (1998) Nonlinear wavelet shrinkage with Bayes rules and Bayes factors, *Journal of the American Statistical Association*, **93**, 173–179.

Jean-François Angers
Dép. de mathématiques et de statistique
Université de Montréal
C.P. 6128, succursale "Centre-ville"
Montréal, Qc H3C 3J7
E-mail:jean-fraccois.angers@UMontreal.CA

Mohan Delampady
Statistics and Mathematics Unit
Indian Statistical Institute
RVCE Post
Bangalore, India 560059
E-mail:mohan@isibang.ac.in