Full Bayesian Wavelet Inference with a Nonparametric Prior

Xue Wang[*,a], Stephen G. Walker[a]

[a] *University of Kent at Canterbury, Canterbury, UK*

Abstract

In this paper we introduce a new Bayesian model for estimating an unknown function in the presence of Gaussian noise. The proposed Bayesian model involves a mixture of a point mass and an arbitrary (nonparametric) symmetric unimodal distribution. Posterior simulation uses slice sampling ideas and the consistency under the proposed model is discussed. In particular, the method is shown to be computationally competitive with some of best Empirical wavelet estimation methods.

Key words:
Stick-breaking priors, Slice sampling, Wavelet shrinkage, Consistency

1. Introduction

Consider the regression model given by:

$$y_i = f(i/n) + \sigma z_i, \qquad i = 1, \ldots, n, \tag{1}$$

where σ is the noise level and $\{z_i\}$ are independent standard normal random variables. The problem of interest is to estimate the unknown regression function $f(\cdot)$, which belongs to a certain class of function $F[0,1]$.

Wavelet based procedures have been shown to be well suited for such settings, and non-parametric estimators of $f(\cdot)$ can be readily obtained by applying various shrinkage rules on the wavelet transformed data.

[*] Corresponding author

Email addresses: X.Wang@kent.ac.uk (Xue Wang), S.G.Walker@kent.ac.uk (Stephen G. Walker)

A variety of shrinkage methods based on classical and Empirical Bayesian statistical models in the wavelet domain have been proposed and studied. See for example Donoho and coauthors (1994, 1995a, 1995b). In this broad context of function estimation, Bayesian wavelet procedures have proved efficient for their capability to incorporate estimation about the unknown signal (e.g. Chipman et al. 1997; Abramovich et al.,1998; Vidakovic and Ruggeri, 2001; Johnstone and Silverman, 2005).

In a Bayesian approach, a prior distribution is constructed on the wavelet coefficients of the function and the function is estimated by applying a suitable Bayesian rule to the resulting posterior distribution of the wavelet coefficients. Different prior distributions are designed to capture the sparseness of wavelet expansions common to most applications. For example, Chipman et al. (1997) proposed a mixture of two weighted normal distributions for the individual wavelet coefficients.

On the other hand, Johnstone and Silverman (2005) proposed a mixture of a mass point and a heavy tailed distribution for a single wavelet coefficient. Wang and Wood (2006) used a mixture of a mass point and a non-central chi-squared distribution for wavelet coefficients in a block. However, all the work in Bayesian wavelet context to date are empirical based approaches, where the priors are designed through some strong prior beliefs and the datasets have been used repeatedly to estimate the hyperparameteres and the function.

In this paper, we intend to apply a full Bayesian model upon the wavelet coefficients to estimate the function (1). The important benefit of using a full Bayesian model compared to say an empirical based approach is that the posterior has the ability to reflect coherent beliefs of the experimenter. This acknowledges that there is more to inference than simply obtaining point estimates. Nevertheless, we do obtain competitive estimates, particularly for smaller samples where we claim improvements. The understanding here is that for empirical Bayes, when samples sizes are smaller the procedure is using "bad" data twice whereas for large samples it is using "good" data twice. Hence, empirical Bayes point estimates can work well in large samples.

We also derive a consistency result which appears new and uses different techniques to usual consistency calculations. This result also suggests that the point mass is essential to consistent posterior inference. Hence, our aim involving this paper is two fold:

- a full Bayes with competitive estimation properties;

- a flexible nonparametric prior for the wavelet coefficients which can cope with all types of underlying functions.

The paper is organized as follows. In Section 2, some background information of the nonparametric prior is given. A full Bayesian model and its algorithm are established in Sections 3 and 4. Numerical illustrations are presented in Section 5 and a discussion appears in Section 6. The appendix contains a discussion of consistency.

2. Stick-breaking priors

Stick-breaking priors are almost surely discrete random probability measures (RPMs) of the form

$$\mathcal{P} = \sum_{k=1}^{N} q_k \delta_{\theta_k} \tag{2}$$

where the (q_k) are non-negative random weights that sum to unity almost surely, and the (θ_k) are independent and identically distributed random variables from some fixed density function $g(\theta)$. The number of terms N can be either finite or infinite, and for the purposes of this paper we will take it to be $+\infty$.

The random weights (q_k) can be constructed in the following way:

$$q_1 = v_1 \quad \text{and} \quad q_k = v_k \prod_{l<k} (1 - v_l), \quad k \geq 2, \tag{3}$$

where the (v_k) are independent Beta(a_k, b_k) random variables for $a_k, b_k > 0$. See, for example, Ishwaran and James (2001), who show that the sum of the weights is 1 almost surely when

$$\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty. \tag{4}$$

The stick-breaking priors are often used as mixing distribution in mixture models (Lo, 1984) to generate random density functions which can be written as

$$f_{\mathcal{P}}(y) = \int k(y|\theta) d\mathcal{P}(\theta) = \sum_{k=1}^{\infty} q_k k(y|\theta_k) \tag{5}$$

where $k(y|\theta)$ is a continuous probability density function for each θ. There is by now a huge amount of literature on these models and we refer the reader to the recent book on Bayesian nonparametrics (Hjort et al., 2010) for a comprehensive account.

When considering the derivation of the posterior for these models, Markov chain Monte Carlo (MCMC) methods for posterior inference are complicated by the presence of an infinite number of unknown parameters. Recently, there has been interest in developing MCMC methods that only use a finite number of elements in (5) at any iteration of a chain but, by making this number suitably chosen, inference follows the true (infinite-dimensional) model. The retrospective sampler of Papaspiliopoulos and Roberts (2008) uses a carefully constructed Metropolis-Hastings update to ensure this, whereas Walker (2007) uses a slice sampling idea, further developed in Kalli et al. (2011).

The key of the slice sampling idea is the introduction of latent variables which make the infinite model finite. A latent variable u is introduced to (5) such that the joint density of (y, u), given (q_k, θ_k), is given by

$$f_{\mathcal{P}}(y,u) = \sum_{k=1}^{\infty} 1(u < \xi_k)\,(q_k/\xi_k)\,k(y|\theta_k)$$

for some deterministic decreasing sequence (to 0) (ξ_k). This sequence is not a modeling issue since the marginal distribution of interest remains unaltered.

The model in the presence of u becomes finite since the number of k such that $\xi_k > u$ is finite and so, conditional on u, the number of parameters is finite. The parameter u can be easily updated since it is uniformly distributed. Then set of k which are needed in the conditional model is of the form $\{1, \ldots, N\}$ where N is the largest k such that $\xi_k > u$.

Therefore, the Gibbs sampler only operates on finite dimensional spaces but the marginal density of y is the infinite mixture model. Hence, the number of variables required to implement a Gibbs sampler for sampling the correct posterior distribution becomes finite.

Our model is specific to the problem and our interest is in constructing a unimodal density with no other constraints. It is well known by now that this can be achieved by taking $k(\cdot|\cdot)$ to be a uniform kernel. Hence, for us,

$$f_{\mathcal{P}}(y) = \int k(y|\theta)d\mathcal{P}(\theta) = \sum_{k=1}^{\infty} q_k \mathrm{g}(y|\theta_k),$$

where

$$g(y|\theta) = \mathrm{Un}(-\theta, \theta)$$

and Un denotes the uniform distribution.

3. The Bayesian model

Preforming the wavelet transformation on (1), we have

$$y_{jk} = w_{jk} + n^{-1/2}\sigma\epsilon_{jk}, \quad j \geq j_0, k = 1, \ldots, M = 2^j,$$

where j is the resolution level we are interested in and j_0 is some fixed resolution level. The (ϵ_{jk}) are independent standard normal random variables and the noise level σ is assumed known.

At the level j, we place the following mixture of a mass point at zero and symmetric unimodal form on the population discrete wavelet coefficient w_k:

$$\pi(w_k|\gamma, q_l, s_l) = \gamma\mathbf{1}(w_k = 0) + (1-\gamma)\sum_{l=1}^{\infty} q_l g(w_k|s_l). \tag{6}$$

The prior for γ is beta$(1, c^*)$, for some $c^* > 0$, and $\mathbf{1}(w_k = 0)$ is the mass point function at $w_k = 0$. Here, clearly, $\Pr(w_k = 0) = \gamma$, and so determines the prior probability of whether the relevant wavelet coefficient is nonzero and comes from a symmetric unimodal, or zero and arises from a point mass at zero. Hyperparameters (q_l) are given as (3) where the (v_l) are assumed independent and follow a Beta$(1, c)$ distribution, and $g_l(w_k) = (2s_l)^{-1}\mathbf{1}(-s_l < w_k < s_l)$, and the prior for the (s_l), $\pi(s_l)$, assumes they are independent and follow a Ga(a, b) distribution for some fixed (a, b).

Employing the slice sampling idea, we introduce a latent variable u_k that operates on $\xi_l = e^{-\beta l}$, so that the joint density of (w_k, u_k), given (q_l, s_l), is given by

$$\pi(w_k, u_k|\gamma, q_l, s_l) = \frac{\mathbf{1}(u_k < \xi_l)}{\xi_l}\left[\gamma\delta(w_k = 0) + (1-\gamma)\sum_{l=1}^{\infty} q_l g_l(w_k)\right]. \tag{7}$$

Here, if $l < 1$ then the model becomes the point mass at $w_k = 0$.

There is another latent variable needed to make posterior simulation tractable, so more latent variables (d_k) are introduced, which allocate each observation to one component of the mixture model. Therefore the joint density of (w_k, u_k, d_k) given (q_l, s_l) is

$$\pi(w_k, u_k, d_k | q_{d_k}, s_{d_k}) = \frac{\mathbf{1}(u_k < \xi_{d_k})}{\xi_{d_k}} \big[\gamma\delta(w_k = 0) + (1-\gamma) q_{d_k} g_{d_k}(w_k)\big],$$

where $d_k \in \{0, 1, 2, 3, \ldots\}$. This form omits any sums inside the product and the choice of d_k, needed to be sampled within the Gibbs sampler, is from a finite set which is easily found.

Observing all wavelet coefficients at the level j, $(y_1, \ldots, y_M)$ yields a full likelihood

$$\begin{aligned} & f(\mathbf{y}, \mathbf{u}, \mathbf{d} | \mathbf{w}, \mathbf{s}, \gamma, \mathbf{q}) \\ = & \prod_k \exp\left\{\frac{-(y_k - w_k)^2}{2\sigma_n^2}\right\} \frac{\mathbf{1}(u_k < \xi_{d_k})}{\xi_{d_k}} \{\gamma\delta(w_k = 0) + (1-\gamma) q_{d_k} g_{d_k}(w_k)\}. \end{aligned}$$

Hence, the full posterior distribution can be expressed as

$$\begin{aligned} & \pi(\mathbf{w} | \mathbf{y}, \mathbf{u}, \mathbf{d}, \mathbf{s}, \gamma, \mathbf{q}) \\ \propto & \prod_k \frac{\mathbf{1}(u_k < \xi_{d_k})}{\xi_{d_k}} \exp\left\{\frac{-(y_k - w_k)^2}{2\sigma_n^2}\right\} \{\gamma\delta(w_k = 0) + (1-\gamma) q_{d_k} g_{d_k}(w_k)\} \\ \times & \prod_l \pi(s_l) \prod_l \pi(v_1). \end{aligned}$$

In the next section we will describe the Gibbs sampler for estimating this model.

4. The Gibbs sampling algorithm

In this section we implement a Gibbs sampler according the model we discussed in the previous section. We require the set of full conditional density functions. The chain can be initialised in the following way. We initialise $\{d_k = k, k = 1 : n\}$ and then simulate $\{u_k, k = 1 : n\}$ from a uniform distribution between 0 and $\xi_{d_k} = e^{-\beta d_k}$. Then let $N_k = \llcorner -\log(u_k)/\beta \lrcorner$, where $\llcorner X \lrcorner$ defines the largest integer less than or equal to X. Define, also, $(N_k)_{\max} = \max_{1 \le i \le n}\{N_i\}$.

Step 1: Updating s. The full conditional distribution of the parameter, s_l when $\max_{d_k=l} |w_{d_k}| \neq 0$, is proportional to

$$\begin{aligned} \pi(s_l|\cdots) &\propto \Big(\prod_{d_k=l} \frac{1}{s_l} 1(s_l > \max_{d_k=l} |w_{d_k}|)\Big) \times \pi(s_l) \\ &\propto s_l^{-n_l} \pi(s_l) 1(s_l > \max_{d_k=l} |w_{d_k}|), \end{aligned}$$

where $n_l = \#\{k|d_k = l, \quad 1 \leq k \leq n\}$. When $\max_{d_k=l} |w_{d_k}| = 0$, then we draw s_l from the prior.

Step 2: Updating q. The prior for q_l is

$$q_l = v_l \prod_{r<l} (1 - v_r),$$

where (v_r) are independent and identically distributed as $\text{Beta}(1, c)$. Hence, the full conditional distribution of q_l is proportional to

$$\pi(q_l|\cdots) \propto \text{Beta}(q_l|1 + n_l, c + n_l^*),$$

where $n_j^* = \#\{k|d_k > l, \quad 1 \leq k \leq n\}$.

Step 3: Updating γ. The prior for γ is $\text{beta}(1, c^*)$ and so the conditional distribution for γ is

$$\pi(\gamma|\cdots) \propto \gamma^{\#(w_k=0)} (1 - \gamma)^{\#(w_k \neq 0)} \pi(\gamma),$$

which is $\text{beta}(1 + \#\{w_k = 0\}, c^* + \#\{w_k \neq 0\})$.

Step 4: Updating (d_k, w_k). The values of d_k can take values between 0 and N_k, which is derived from the value of u_k. We have the joint density of (d_k, w_k) as proportional to

$$\exp\left\{\frac{-(y_k - w_k)^2}{2\sigma_n^2}\right\} \mathbf{1}(0 \leq d_k \leq N_k) \left(p_{d_k}/\xi_{d_k}\right) h_{d_k}(w_k)$$

where

$$p_{d_k} = \begin{cases} \gamma & d_k = 0 \\ \\ (1 - \gamma)\, q_{d_k} & d_k > 0. \end{cases}$$

and $h_l(\cdot)$ is a point mass at 0 if $l = 0$ and otherwise is $g(\cdot|s_l)$. We can easily sample here by sampling d_k, marginalsing over w_k, and then sampling $w_k|d_k$.

5. Numerical Results

In order to examine the numerical performance of the proposed full Bayesian model, we perform a simulation study to compare the proposed approach with some of the recently proposed methods in the literature: namely, BlockJS (Cai, 1999), BAMS (Vidakovic and Ruggeri, 2001), Ebayes (Johnstone and Silverman, 2005), NCP (Wang and Wood, 2006) and BBN (Wang and Walker, 2010).

BlockJS is a classical block thresholding procedure with a fixed block size $L = \log_2(n)$ and a fixed threshold level. BAMS is an empirical Bayes threshold method which imposes a mixture of a double exponential distributin and a point mass as the prior for each individual wavelet coefficient. The posterior mean is used here as the shrinkage rule. Ebayes is an empirical Bayes threshold of individual wavelet coefficients based on a mixture of a heavy tailed distribution and a point mass as the prior. The mixture of a "Cauchy" distribution and a point mass as the prior and the posterior mean as the threshold rule are considered. NCP is a Bayesian block shrinkage approach based on the block sum of squares with a fixed block size. It imposes a mixture of a non-central chi-squared distribution and a point mass. The"power" prior as the distribution of the hyperparameter and posterior mean as the shrinkage rule are used here and the block size is fixed as $L = 2$. BBN is an Bayesian block wavelet shrinkage method based on a multinormal distribution, where the block size and the shrinkage level at each resolution level are chosen adaptively by the data. All these Bayes rules are empirical, in the sense that the prior is estimated from the data.

In practice the noise level σ_n in (1), which we assumed known for simplicity, needs to be estimated from the data and for this we will use the following robust estimator of σ given in Donoho and Johnstone (1994). This estimator σ is based on the noisy wavelet coefficients $(y_{j,k})$ at the highest resolution level J, so

$$\hat{\sigma} = \frac{1}{0.6745}\text{median}\left(|y_{J,k}| : 1 \leq k \leq 2^J, J = \log_2(n) - 1\right).$$

Four functions, 'HeaviSine','Blocks', 'Bumps' and 'Doppler', representing different level of spatial variability, are used as test functions for the purposes of simulation studies. Each test function was rescaled to achieve different signal-to-noise ratios (SNR), and the standard normal noise was added to the functions. The average MSE for the estimator $\hat{f}$ of f defined as

$$MSE_f = \frac{1}{n} \sum_{i=0}^{n-1} \left[\left\{ \hat{f}(i/n) - f(i/n) \right\}^2 \right]. \tag{8}$$

Before the intensive simulation study, we performed a preliminary simulation study to examine the general performance of the proposed full Bayeisan method with the range of sample sizes from 64 to 2048. We found that the proposed method is highly competitive with the best of the existing classic threshold methods, empirical Bayes block and term-by-term methods, when the sample size is small, e.g 64, 128 and 256. The proposed method tends to underperform the best of the empirical Bayes methods when the sample size becomes large. The reason, we believe, is that the most of the empirical Bayes methods use the "good" data twice to estimate the unknown function. Hence, we will concentrate on the small sample sizes (64, 128, 256 and 512) in the study.

The average MSE (AMSE) results with 100 simulation runs for the four test functions with SNR=7 at different sample sizes (64, 128, 256 and 512) are provided in Table 1. The simulation results show that the proposed full Bayesian method performs constantly well over the whole range of signals and sample sizes we considered here, while the performance of the other methods involve fluctuating. The simulation study with difference SNRs shows the similar patterns as the SNR=7.

6. Discussion

In terms of function estimation using wavelets, we have a competitive Bayesian model for small to moderate sample sizes (≤ 256).

The merit of our model is that it is full Bayes and hence posterior distributions are meaningful and can be used in the standard way, e.g. decision making, as they represent posterior beliefs about the function of interest. The key, in our mind, is the nonparametric component, which can adequately capture the distribution of the nonzero coefficients, however they may appear for each of the different type of function we estimate. Hence we have good "across the board" estimation.

We have also included a consistency result which effectively states that $\mathrm{E}\gamma^n \to 1$ as $n \to +\infty$ as a sufficient condition.

Table 1: Simulation results of six methods (BlockJS, BAMS, EBCmean, NCP , BBN and FBayes) with 100 simulation runs, where AMSE was obtained with SNR=7 and sample sizes N=(64, 128, 256 or 512). An asterisk is used to denote the best in a column.

Methods	HeaviSine				BlockJS			
	64	128	256	512	64	128	256	512
BlockJS	0.932	0.712	0.592	0.521	1.245	0.873	0.807	0.676
BAMS	0.922	0.775	0.698	0.641	0.873	0.722	0.635	0.568
EBCmean	0.446	0.295	0.211	0.146	0.652	0.406	0.274	0.189
NCP	0.549	0.336	0.269	0.170	0.729	0.467	0.312	0.232
BBN	0.494	0.334	0.269	0.171	0.628	0.492	0.313	0.234
FBayes	0.488	0.349	0.303	0.218	0.664	0.499	0.376	0.326
Mathods	Bumps				Doppler			
	64	128	256	512	64	128	256	512
BlockJS	1.876	1.2	0.957	0.794	1.091	0.832	0.682	0.569
BAMS	1.011	0.862	0.744	0.768	1.192	1.042	0.906	0.655
EBCmean	1.077	0.940	0.632	0.469	0.956	0.514	0.385	0.230
NCP	0.804	0.749	0.677	0.426	0.775	0.481	0.351	0.195
BBN	0.831	0.747	0.677	0.504	0.717	0.519	0.351	0.226
FBayes	0.841	0.802	0.631	0.505	0.728	0.509	0.405	0.239

We can exploit the good small sample properties of our method by introducing blocking ideas to larger samples. Performing the wavelet transform on (1), we have

$$y_{jk} = w_{jk} + n^{-1/2}\sigma z_{jk}, \qquad j \geq j_0, k = 0, \ldots, 2^j - 1. \tag{9}$$

For each fixed resolution level $j \geq j_0$, let $L \geq 1$ be the possible length of each block, and $M = 2^j/L$ be the number of blocks. Let $\mathbf{y}_b = (y_{(b-1)L+1}, \ldots, y_{bL})$ represent observations in the b-th block on level j, and similarly define $\mathbf{w}_b = (w_{(b-1)L+1}, \ldots, w_{bL})$ and $\mathbf{z}_b = (z_{(b-1)L+1}, \ldots, z_{bL})$. Hence, we can write

$$\mathbf{y}_b = \mathbf{w}_b + n^{-1/2}\sigma \mathbf{z}_b. \tag{10}$$

We can place a mixture of a mass point at zero and a symmetric unimodal form on the wavelet coefficients in the same block:

$$\pi(\mathbf{w}_b|\gamma, q_l, s_l) = \gamma \mathbf{1}(\mathbf{w}_b = 0) + (1-\gamma)\sum_{l=1}^{\infty} q_l g(\mathbf{w}_b|\mathbf{s}_{bl}), \tag{11}$$

where, we can take,

$$g(\mathbf{w}_b|\mathbf{s}_{bl}) = \prod_{j=1}^{L} g_{bjl}(w_{bj}).$$

So instead of introducing the latent variables (u_k) and (d_k) for every coefficient, we introduce them only for each block: so rather than the choices being over $k = 1, \ldots, n$, it is over $k = 1, \ldots, M$. Hence,

$$\pi(\mathbf{w}_b, u_b, d_b|q_{d_b}, s_{d_b}) = \frac{\mathbf{1}(u_b < \xi_{d_b})}{\xi_{d_b}}\big[\gamma \mathbf{1}(\mathbf{w}_b = 0) + (1-\gamma) q_{d_b} g_{d_b}(\mathbf{w}_b)\big],$$

and now for a block, either all the coefficients are assigned to be zero, or they are all assigned to be non–zero.

7. Appendix: Bayesian consistency

Here we establish sufficient conditions on the prior for the consistency of the model.

The posterior probability of A_ϵ^c where $A_\epsilon = \{w : \sup_{k\in\{1,\ldots,n\}} |w_k - w_{0k}| < \epsilon\}$, given $\mathbf{y}$, is given by

$$\begin{aligned}
&\Pr(w \in A_\epsilon^c|\mathbf{y}) = J_n/I_n \\
= &\frac{\int_{A_\epsilon^c} q_n(dw_1,\ldots,dw_n)\exp\left\{-\sum_{k=1}^n \frac{1}{2\sigma_n^2}(y_k - w_k)^2\right\} / \exp\left\{-\sum_{k=1}^n \frac{1}{2\sigma_n^2}(y_k - w_{0k})^2\right\}}{\int_{R^n} q_n(dw_1,\ldots,dw_n)\exp\left\{-\sum_{k=1}^n \frac{1}{2\sigma_n^2}(y_k - w_k)^2\right\} / \exp\left\{-\sum_{k=1}^n \frac{1}{2\sigma_n^2}(y_k - w_{0k})^2\right\}}.
\end{aligned}$$

For the numerator, let us first consider

$$J_{n1} = \int_{A_1} q_n(dw_1,\ldots,dw_n)\exp\left\{-\frac{1}{2\sigma_n^2}\sum_{k=1}^n \left[(y_k - w_k)^2 - (y_k - w_{0k})^2\right]\right\},$$

where $A_1 = \{w : |w_1 - w_{01}| > \epsilon\}$ and $\sigma_n = n^{-1/2}\sigma$. We have that

$$y_k = w_{0k} + \sigma_n z_k$$

where the (z_k) are independent standard normal, so

$$\sum_{k=1}^n \left[(y_k - w_k)^2 - (y_k - w_{0k})^2\right] \geq (y_1 - w_1)^2 - (y_1 - w_{01})^2 - \sum_{k=2}^n \sigma_n^2 |z_k|^2.$$

Therefore, with $k = 1$, and with $w \in A_1$, we have

$$\begin{aligned}
(y_k - w_k)^2 - (y_k - w_{0k})^2 &= -2(w_{0k} + \sigma_n\epsilon_k)w_k + w_k^2 + 2(w_{0k} + \sigma_n\epsilon_k)w_{0k} - w_{0k}^2 \\
&= (w_k - w_{0k})^2 + 2\sigma_n z_k(w_{0k} - w_k) \\
&\geq |(w_k - w_{0k})^2| - 2\sigma_n|z_k||(w_{0k} - w_k)| \\
&\geq \epsilon^2 - 2\epsilon\sigma_n|z_k|,
\end{aligned}$$

and hence, with $\sigma_n^2 = n^{-1}\sigma^2$ and the fact that

$$\sum_{k=1}^n \sigma_n^2 |z_k|^2 \to 0 \text{ a.s.}$$

It is easy to see that

$$J_{n1} < q_n(A_1)\exp\left[-c(n/2)\epsilon^2\right] \text{ a.s.}$$

for all large n and for some $c > 0$. Hence

$$J_n \leq \sum_{i=1}^{n} J_{ni} < c_1 n \exp(-nc_2)$$

a.s. for all large n and for some constants $c_1, c_2 > 0$.

For the denominator, let us assume without loss of generality that for some fixed but finite M, it is that $w_{0k} \neq 0$ for $k = 1, \ldots, M$ and $w_{0k} = 0$ for $k = M+1, \ldots, n$. For studying I_n, let, for any $\delta > 0$,

$$B_\delta = \left\{ w : \sum_{k=1}^{M} |w_k - w_{0k}| < \delta \quad \& \quad w_k = 0, \quad \forall\, k = M+1, \ldots, n \right\}.$$

So we have

$$I_n > \int_{B_\delta} q_n(dw_1, \ldots, dw_n) \exp\left\{ -\frac{1}{2\sigma_n^2} \sum_{k=1}^{n} \left[(y_k - w_k)^2 - (y_k - w_{0k})^2 \right] \right\}.$$

Since we know that $y_k = w_{0k} + \sigma_n z_k$, where the (z_k) are independent standard normal random variables, it is that

$$\begin{aligned}(y_k - w_k)^2 - (y_k - w_{0k})^2 &= -2(w_{0k} + \sigma_n \epsilon_k) w_k + w_k^2 + 2(w_{0k} + \sigma_n z_k) w_{0k} - w_{0k}^2 \\ &= (w_k - w_{0k})^2 - 2\sigma_n z_k (w_k - w_{0k}) \\ &= (w_k - w_{0k})(w_k - w_{0k} - 2\sigma_n z_k).\end{aligned}$$

Given that $w \in B_\delta$, we therefore have for an arbitrary $\delta > 0$, the denominator follows

$$I_n \geq c_3 \int_{B_\delta} q_n(dw_1, \ldots, dw_n) \exp\left(-nc_4\delta^2\right)$$

a.s. for all large n for some constants $c_3, c_4 > 0$. Here q_n is the probability model for the $(w_1, \ldots, w_n)$ described in the paper and specifically we note that $\gamma \sim \text{beta}(1, c^*)$, where we will need to determine c^*. Recall that

$$\Pr(w_k = 0 | \gamma) = \gamma$$

independently for all $k = 1, \ldots, n$.

Therefore, we have

$$I_n > c_3 q_n(B_\delta)\, \exp(-c_4 n \delta^2).$$

We now need to investigate $q_n(B_\delta)$ and actually we show that

$$q_n(B_\delta) > c_5$$

for all large n for some $c_5 > 0$. Hence, we have $I_n > c_6 \exp(-nc_4\delta^2)$ a.s. for all large n for any $\delta > 0$ and so this combines with the upper bound for the numerator and yields the desired consistency result, by choosing δ small enough that $\delta c_4 < \epsilon^2$.

The probability of B_δ is easily seen to be bounded below by

$$q_n(B_\delta) \geq \mathrm{E}\left(\gamma^{n-M}(1-\gamma)^M\right) \times \Pr\left(\sum_{k=1}^{M} |w_k - w_{0k}| < \delta, \quad w_k \neq 0 \quad \forall\, k = 1, \ldots, M\right).$$

Now the probability part of this expression is independent of n and so is bounded away from 0. The expectation part is given by

$$\frac{c^*\Gamma(n-M+1)\Gamma(M+c^*)}{\Gamma(n+1+c^*)}.$$

This can be shown to converge to 1 when we take $c^* = \xi/(n-\xi)$ for some $\xi > 0$. Therefore, for some constant c_5, $q_n(B_\delta) > c_5$ for all large n.

In conclusion, we have that

$$\Pr(w \in A_\epsilon^c|\mathbf{y}) \leq C_1 n \exp(-nC_2)$$

a.s. for all large n, for constants $C_1, C_2 > 0$.

References

[1] Antoniadia, A. and Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. *J. Statist. Software*, **6**, 1-83.

[2] Chipman, H. A., Kolaczyk, E.D. and McCulloch, R.E. (1997). Adaptive Bayesian wavelet shrinkage.*J. Am. Statist. Ass.*, **92**, 1413-1421.

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

[4] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D (1995a). Wavelet shrinkage: asymptopia (with discussion)? *J. R. Statist. Soc. B*, **57**, 301-365.

[5] Donoho, D. L. and Johnstone, I. M. (1995b). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200-1224.

[6] Hjort, N. L, Holmes, C., Müller, P. and Walker, S. (2010). Bayesian Nonparametric. Ed. *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.

[7] Ishwaran, H. and James, L. F. (2001). Gibbs Sampling methods for stick-breaking priors. *J. Am. Statist. Ass.*, **96**, 161-173.

[8] Kalli, M., Griffin, J.E. and Walker, S.G. (2011). Slice sampling mixture models. *Statistics and Computing*, **21**, 93-105.

[9] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: density estimates. *Ann. Statist.*, **12**, 351-357.

[10] Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169-186.

[11] Johnstone I. and Silverman B. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.*, **33**, 1700–1752.

[12] Walker, S. G. 2007. Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, **36**, 45-54.

[13] Wang, X. and Wood, A. T. A. (2006). Empirical Bayes block shrinkage of wavelet coefficients via the noncentral χ^2 distribution. *Biometrika* **93**, 705-722.

[14] Wang, X. and Walker, S. (2010). A penalised data-driven block shrinkage approach to empirical Bayes wavelet estimation *Stat. Probab. Lett.*, **80**, 990-996.

[15] Vidakovic, B. and Ruggeri, F. (2001). BAMS method: theory and simulations. *Sankhyā: The Indian Journal of Statistics*, **63**, 234-249.