On Optimality of Bayesian Wavelet Estimators

FELIX ABRAMOVICH

Tel Aviv University

UMBERTO AMATO and CLAUDIA ANGELINI

Istituto per le Applicazioni del Calcolo

ABSTRACT. We investigate the asymptotic optimality of several Bayesian wavelet estimators, namely, posterior mean, posterior median and Bayes Factor, where the prior imposed on wavelet coefficients is a mixture of a mass function at zero and a Gaussian density. We show that in terms of the mean squared error, for the properly chosen hyperparameters of the prior, all the three resulting Bayesian wavelet estimators achieve optimal minimax rates within any prescribed Besov space $B^s_{p,q}$ for $p \geq 2$. For $1 \leq p < 2$, the Bayes Factor is still optimal for $(2s+2)/(2s+1) \leq p < 2$ and always outperforms the posterior mean and the posterior median that can achieve only the best possible rates for linear estimators in this case.

Key words: Bayes Factor, Bayes model, Besov spaces, minimax estimation, non-linear estimation, non-parametric regression, posterior mean, posterior median, wavelets

1. Introduction

Consider the standard non-parametric regression model:

$$y_i = f\left(\frac{i}{n}\right) + \epsilon z_i, \quad z_i \overset{i.i.d.}{\sim} N(0,1), \qquad i = 1, \ldots, n, \tag{1}$$

where the unknown function f, which we want to recover, belongs to a certain class of functions $\mathcal{F}[0, 1]$. One of the basic techniques in non-parametric regression is the generalized Fourier series approach. Instead of estimating f directly, we expand it into some orthogonal series and then estimate the coefficients of its expansion from the data. The original non-parametric problem is thus transformed to a parametric problem although with an infinite number of parameters. The key point for the efficiency of such an approach is obviously a proper choice of a basis. A 'good' basis should be *sparse* in the sense that a wide variety of possible responses from $\mathcal{F}$ can be approximated well by only a few terms of the expansion. As it is well known, wavelet series allow sparse representation for a large set of function spaces, in particular, for Besov spaces that include among others the Hölder and Sobolev classes of smooth functions, functions of bounded variations, etc. (e.g. Meyer, 1992). For the last decade wavelet-based estimators have been intensively studied in the literature (see Antoniadis, 1997; Vidakovic, 1999; Abramovich *et al*., 2000 for comprehensive reviews). In particular, the well-known results of Donoho & Johnstone (1994, 1998) established the asymptotic optimality (in the minimax sense) of wavelet estimators within the whole range of Besov spaces.

Various Bayesian wavelet estimators have also been proposed in the literature (e.g. Chipman *et al*., 1997; Abramovich *et al*., 1998; Clyde *et al*., 1998; Vidakovic, 1998), see also Müller & Vidakovic (1999) for an overview. Following a Bayesian approach, a prior distribution is imposed on wavelet coefficients of the function and a Bayesian estimator is then obtained by applying a suitable Bayesian rule to the resulting posterior distribution of

the coefficients. Different Bayesian estimators are now widely used and numerous simulation studies showed their good mean squared error (MSE) performance in comparison with other existing wavelet estimators (e.g. Abramovich *et al.*, 1998; Abramovich & Sapatinas, 1999; Antoniadis *et al.*, 2001). However, their asymptotic optimality in the minimax sense have not been studied. This paper is a step to fill this gap. A pure Bayesian would probably generally oppose such a frequentist approach but yet Rubin (1984) argued for the importance of frequentist analysis of Bayesian procedures, in particular, for understanding and validating their results. Despite all criticism, the minimaxity is the mostly used criterion for comparison between various estimators. Diaconis & Freedman (1986), Cox (1993), Freedman (1999), Zhao (2000) among others studied the asymptotical properties of various Bayesian estimators within the minimaxity framework. In this paper, we investigate the asymptotic optimality of several known Bayesian wavelet estimators and derive their convergence rates within a range of Besov spaces. In particular, we find the subsets of Besov spaces where they achieve the optimal minimax rates (up to a log-factor).

The paper is organized as follows. In section 2, some necessary background is given: we start from a short review of wavelets and some relevant aspects of Besov spaces, discuss the prior model on wavelet coefficients and several Bayesian wavelet estimators corresponding to different losses. The main results on their convergence rates and asymptotic optimality are established in section 3. In section 4, we provide a small simulation study to illustrate our results. Some concluding remarks and discussion are made in section 5. All the proofs are given in the appendix.

2. The model

2.1. Short review of wavelet series and Besov spaces

For simplicity of exposition we assume that f is periodic and work with periodic orthonormal wavelet bases on [0,1] generated by a compactly supported scaling function φ and a corresponding mother wavelet ψ (e.g. Daubechies, 1992, section 9.3). Then, f can be expanded as

$$f(t) = w_{-10}\varphi(t) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} w_{jk}\psi_{jk}(t),$$

where $w_{-10} = \int_0^1 f(t)\varphi(t)dt$ and $w_{jk} = \int_0^1 f(t)\psi_{jk}(t)dt$.

Let ψ be a mother wavelet of regularity r. Then, the corresponding wavelet series constitute unconditional bases for Besov spaces $B^s_{p,q}[0,1]$, $\max(0,1/p-1/2) < s < r$, $p, q \geq 1$ (see Meyer, 1992, section 2.9 for rigorous definitions and details), and the Besov norm of f is equivalent to the corresponding sequence space norm:

$$||f||_{B^s_{p,q}} \asymp ||w||_{b^s_{p,q}} = |w_{-10}| + \left\{\sum_{j=0}^{\infty} 2^{j(s+1/2-1/p)q}||w_j||_p^q\right\}^{1/q}, \quad 1 \leq q < \infty,$$

$$||f||_{B^s_{p,\infty}} \asymp ||w||_{b^s_{p,\infty}} = |w_{-10}| + \sup_{j\geq 0}\left\{2^{j(s+1/2-1/p)}||w_j||_p\right\}$$

(e.g. Meyer, 1992, section 6.10; Donoho & Johnstone, 1998). The Besov spaces include, in particular, the well-known Sobolev ($B^m_{2,2}$) and Hölder ($B^s_{\infty,\infty}$) spaces of smooth functions, but in addition less traditional spaces, like the space of functions of bounded variation, sandwiched between $B^1_{1,1}$ and $B^1_{1,\infty}$.

2.2. Prior model

Consider the following white noise model

$$dY(t) = f(t)dt + \sigma_n\, dW(t), \tag{2}$$

where $\sigma_n^2 = \epsilon^2/n$, $f \in B^s_{p,q}[0, 1]$ and W is a standard Wiener process. Brown & Low (1996) showed the asymptotic equivalence between (2) and the original non-parametric regression model (1) under mild conditions.

Let $f \in B^s_{p,q}[0, 1]$, $s > \max(0,1/p-1/2)$, $p, q \geq 1$. Performing the wavelet transform on (2) with a mother wavelet ψ of regularity $r > s$, one has

$$Y_{jk} = w_{jk} + \sigma_n z_{jk}, \quad j \geq -1, \;\; k = 0, \ldots, 2^j - 1, \tag{3}$$

where $Y_{jk} = \int_0^1 \psi_{jk}\, dY(t)$, $\psi_{-10}(t) = \varphi(t)$ and z_{jk} are independent $N(0, 1)$ random variables.

As we have already mentioned, a large variety of different functions allow sparse representation in wavelet series. To capture this characteristic feature of wavelets, Abramovich *et al.* (1998) suggested to place the following prior on the wavelet coefficients w_{jk} of the unknown f:

$$w_{jk} \sim \pi_j N(0, \tau_j^2) + (1 - \pi_j)\delta(0), \quad j \geq 0; \;\; k = 0, \ldots, 2^j - 1, \tag{4}$$

where $0 \leq \pi_j \leq 1$, $\delta(0)$ is a point mass at zero, and w_{jk} are independent. To complete the model a vague prior is placed on the scaling coefficient w_{-10}.

According to the prior (4), every w_{jk} is either zero with probability $1-\pi_j$ or with probability π_j, is normally distributed with zero mean and variance τ_j^2. The probability π_j gives the proportion of non-zero wavelet coefficients at resolution level j while the variance τ_j^2 is a measure of their magnitudes. Note that the prior parameters π_j and τ_j^2 are the same for all coefficients at a given resolution level j.

The hyperparameters of the prior model (4) are assumed to be of the form:

$$\tau_j^2 = c_1 2^{-\alpha j} \;\text{ and }\; \pi_j = \min(1, c_2 2^{-\beta j}), \quad j \geq 0, \tag{5}$$

where α and β are non-negative constants, c_1, $c_2 > 0$. Some intuitive understanding of the model implied by (4) and (5) can be found in Abramovich *et al.* (1998, section 4.2).

Similar priors but with different forms for the hyperparameters π_j and τ_j^2 are considered in Clyde *et al.* (1998). The prior model (4) is also an extreme case of that of Chipman *et al.* (1997) which is the mixture of two normal distributions with zero means but different variances for 'negligible' and 'non-negligible' wavelet coefficients.

Abramovich *et al.* (1998) established a relationship between the hyperparameters of the prior (4) and (5), and the parameters of Besov spaces within which realizations from the prior will fall. Note first that the prior yields the expected number of non-zero wavelet coefficients on the jth level to be $c_2 2^{j(1-\beta)}$. Then, applying the first Borel–Cantelli lemma, in the case $\beta > 1$, the number of non-zero coefficients in the wavelet expansion is finite almost surely and, hence, with probability 1, f will belong to the same Besov spaces as the mother wavelet ψ, i.e. those for which $\max(0,1/p-1/2) < s < r$, $1 \leq p, q \leq \infty$. More fruitful and interesting is, therefore, the case $0 \leq \beta \leq 1$. The case $\beta = 0$ corresponds to the prior belief that all coefficients on all levels have the same probability of being non-zero. This characterizes self-similar processes such as white noise or Brownian motion, the overall regularity depending on the value of α. The case β=1 assumes that the expected number of non-zero wavelet coefficients is the same on each level which is typical, for example, for piecewise polynomial functions (see Abramovich *et al.*, 1998).

Suppose that f is generated from the prior model (4) and (5). Because of the improper nature of the prior distribution of w_{-10}, we consider the prior distribution of f conditioned on any given value for w_{-10}. The following theorem, proved in Abramovich *et al.* (1998), establishes necessary and sufficient conditions for f to fall (with probability 1) in any particular Besov space.

Theorem 1 (Abramovich et al., 1998)
Let ψ be a mother wavelet of regularity r. Consider constants s, p and q such that $\max(0,1/p-1/2) < s < r$, $1 \le p, q \le \infty$. Let the wavelet coefficients w_{jk} of a function f obey the prior model (4) *and* (5), *where $c_1, c_2 > 0$, $\alpha \ge 0$ and $0 \le \beta \le 1$. Then $f \in B^s_{p,q}$ almost surely if and only if either:*

$$s+\frac{1}{2}-\frac{\beta}{p}-\frac{\alpha}{2}<0, \tag{6}$$

or

$$s+\frac{1}{2}-\frac{\beta}{p}-\frac{\alpha}{2}=0 \ \textit{and} \ 0 \le \beta < 1, \quad 1 \le p < \infty, \ q=\infty. \tag{7}$$

Theorem 1 holds for all values of the Besov space parameter q. This should not be surprising due to the embedding properties of Besov spaces. To give some insight on the role of q, Abramovich *et al.* (1998) considered a more delicate dependence of the variance τ_j^2 on the level j by adding a third hyperparameter γ : $\tau_j^2 = c_1 2^{-\alpha j} j^{\gamma}$, and extended the results of theorem 1 for this case (see their theorem 2). More general analogues of theorem 1 for *overcomplete* wavelet dictionaries were obtained in Abramovich *et al.* (2000).

2.3. Bayesian wavelet estimators

Subject to the prior (4) and (5) and the model (3), the posterior distribution of $w_{jk}|Y_{jk}$ is also a mixture of a corresponding posterior normal distribution and $\delta(0)$. Letting Φ be the standard normal cumulative distribution function, the posterior cumulative distribution function $F(w_{jk}|Y_{jk})$ is

$$F(w_{jk} \mid Y_{jk}) = \frac{1}{1+\eta_{jk}}\Phi\left(\frac{w_{jk}-Y_{jk}\tau_j^2/(\sigma_n^2+\tau_j^2)}{\sigma_n\tau_j/\sqrt{\sigma_n^2+\tau_j^2}}\right)+\frac{\eta_{jk}}{1+\eta_{jk}}I(Y_{jk}\ge 0), \tag{8}$$

where the posterior odds ratio for the component at zero is

$$\eta_{jk}=\frac{1-\pi_j}{\pi_j}\frac{\sqrt{\tau_j^2+\sigma_n^2}}{\sigma_n}\exp\left(-\frac{\tau_j^2 {Y_{jk}}^2}{2\sigma_n^2(\tau_j^2+\sigma_n^2)}\right). \tag{9}$$

To derive a Bayesian rule one should specify the loss function. Different losses lead to different Bayesian estimators. The most popular Bayes rule usually considered in the literature corresponds to the L^2-loss and yields the posterior mean (e.g. Chipman *et al.*, 1997; Clyde *et al.*, 1998; Vidakovic, 1998). Using (8) and (9), we then have

$$\hat{w}_{jk} = E(w_{jk} \mid Y_{jk}) = \frac{1}{1+\eta_{jk}} \frac{\tau_j^2}{\tau_j^2 + \sigma_n^2} Y_{jk}. \tag{10}$$

Such a rule is a non-linear smoothing shrinkage. Abramovich *et al*. (1998) suggested the use of the posterior median that corresponds to the L^1-loss and can be obtained in the following closed form

$$\tilde{w}_{jk} = \mathrm{Med}(w_{jk} \mid Y_{jk}) = \mathrm{sign}(Y_{jk}) \max(0, \zeta_{jk}), \tag{11}$$

where

$$\zeta_{jk} = \frac{\tau_j^2}{\sigma_n^2 + \tau_j^2} |Y_{jk}| - \frac{\tau_j \sigma_n}{\sqrt{\sigma_n^2 + \tau_j^2}} \Phi^{-1}\left(\frac{1 + \min(\eta_{jk}, 1)}{2}\right). \tag{12}$$

The quantity ζ_{jk} is negative for all Y_{jk} in some implicitly defined interval $[-\lambda_j^{\mathrm{PM}}, \lambda_j^{\mathrm{PM}}]$, and hence $\mathrm{Med}(w_{jk}|Y_{jk})$ is zero whenever $|Y_{jk}|$ falls below the threshold λ_j^{PM}. The posterior median is therefore a level-dependent 'kill' or 'shrink' thresholding rule with thresholds λ_j^{PM} known also in the literature as the BayesThresh (Abramovich *et al*., 1998). Donoho & Johnstone (1994, 1998) showed that thresholding wavelet coefficients with proper chosen thresholds yields asymptotically optimal (minimax) estimators within Besov spaces and most of the existing wavelet estimators are of this type. In this paper, we investigate the optimality of the thresholds λ_j^{PM}.

Vidakovic (1998) considered another way to obtain a *bona fide* thresholding rule within a Bayesian framework via a hypothesis testing approach. This rule essentially corresponds to the 0/1-loss: after observing Y_{jk}, test the hypothesis $H_0 : w_{jk} = 0$ against a two-sided alternative $H_1 : w_{jk} \neq 0$. If the hypothesis H_0 is rejected, w_{jk} is estimated by Y_{jk}, otherwise $w_{jk} = 0$:

$$\check{w}_{jk} = Y_{jk} I(\eta_{jk} < 1), \tag{13}$$

where the posterior odds ratio $\eta_{jk} = P(H_0 \mid Y_{jk})/P(H_1 \mid Y_{jk})$ is given by (9). Vidakovic (1998) called this thresholding rule Bayes Factor thresholding as the posterior odds ratio is obtained by multiplying the Bayes Factor with the prior odds ratio. From (9), the Bayes Factor rule (13) mimics the level-dependent hard thresholding rule:

$$\check{w}_{jk} = Y_{jk} I(|Y_{jk}| \geq \lambda_j^{\mathrm{BF}}),$$

where

$$(\lambda_j^{\mathrm{BF}})^2 = \frac{2\sigma_n^2(\sigma_n^2 + \tau_j^2)}{\tau_j^2} \log\left(\frac{1-\pi_j}{\pi_j} \frac{\sqrt{\sigma_n^2 + \tau_j^2}}{\sigma_n}\right). \tag{14}$$

To compare the BayesThresh and the Bayes Factor thresholding rules, note that the Bayes Factor is always a 'keep' or 'kill' hard thresholding, while the posterior median is a 'shrink' or 'kill' thresholding, where extent of shrinkage depends on the absolute values of the wavelet coefficients. In addition, the Bayes Factor thresholds Y_{jk} if the corresponding $\eta_{jk} > 1$. From (9), (11) and (12) it follows that the posterior median 'kills' those Y_{jk}, whose

$$\eta_{jk} > 1 - 2\Phi\left(-\frac{\tau_j |Y_{jk}|}{\sigma_n \sqrt{\sigma_n^2 + \tau_j^2}}\right)$$

and, hence, will threshold more coefficients.

Nevertheless, the lemma below shows that λ_j^{PM} and λ_j^{BF} are 'similar' (we denote it by $\lambda_j^{\mathrm{PM}} \sim \lambda_j^{\mathrm{BF}}$) in the sense that there exist two positive constants $0 < C_1 \le C_2$ such that $C_1\lambda_j^{\mathrm{PM}} \le \lambda_j^{\mathrm{BF}} \le C_2\lambda_j^{\mathrm{PM}}$ for all j. Moreover, it gives the approximations for both threshold sequences.

Lemma 1
$\lambda_j^{\mathrm{PM}} \sim \lambda_j^{\mathrm{BF}} \sim \lambda_j^*$, *where*

$$\lambda_j^* = \begin{cases} \sqrt{\frac{\log n}{n}}, & j \le J_\alpha \\ \sqrt{\frac{j2^{\alpha j}}{n}}, & j > J_\alpha, \end{cases}$$

and $J_\alpha = (1/\alpha)\log_2 n$.

The proof is given in the appendix A1.

Lemma 1 shows that on coarse resolution levels ($j \le J_\alpha$) both λ_j^{PM} and λ_j^{BF} are of the same order as the well-known universal threshold $\lambda_{UN} = \sigma_n\sqrt{2\log n}$ of Donoho & Johnstone (1994), but larger on high levels ($j > J_\alpha$). The universal threshold is known to be optimal within the whole range of Besov spaces, so one would expect that the 'severe' thresholds λ_j^* will 'kill' the significant coefficients present on high levels for spatially inhomogeneous functions. Moreover, the posterior median, in addition, will shrink the 'survivors'. Thus, for such functions the considered Bayesian estimators might not be optimal. The results of the following section will give a rigorous theoretical ground to these preliminary considerations.

3. Main results

In this section we investigate the asymptotic minimax properties of the posterior mean, posterior median and Bayes Factor estimators $\hat{w}$, $\tilde{w}$ and $\check{w}$ defined in section 3. The proofs of all the results are given in the appendix A2.

Consider again the white noise model (2), where $f \in B_{p,q}^s$, $s > \max(0,1/p - 1/2)$, $p, q \ge 1$. Among all possible estimators f^{est} of f define the minimax mean squared error as

$$R(n, B_{p,q}^s) = \inf_{f^{\mathrm{est}}} \sup_{f \in B_{p,q}^s} E||f^{\mathrm{est}} - f||_{L^2[0,1]}^2.$$

In addition, let $R_L(n, B_{p,q}^s)$ be the minimax MSE within the class of linear estimators. Donoho & Johnstone (1998) showed that for large n,

$$R(n, B_{p,q}^s) \asymp n^{-2s/(2s+1)} \tag{15}$$

$$R_L(n, B_{p,q}^s) \asymp n^{-(2s-2/p+2/p_-)/(2s+1-2/p+2/p_-)}, \tag{16}$$

where $p_- = \max(p, 2)$. Hence, for $1 \le p < 2$ that characterizes spatially inhomogeneous functions, linear estimators cannot achieve the optimal rate.

Let ψ be a mother wavelet of regularity $r > s$. Then, the set of the corresponding wavelet coefficients w of f belongs to a Besov ball of some radius R, $b_{p,q}^s(R) = \{w : ||w||_{b_{p,q}^s} \le R\}$ (see section 2.1) and due to the orthonormality of a wavelet basis,

$$R(n, B_{p,q}^s) = \inf_{w^{\mathrm{est}}} \sup_{w \in b_{p,q}^s(R)} E||w^{\mathrm{est}} - w||_{l^2}^2,$$

where w^{est} are the wavelet coefficients of f^{est}.

We derive now the upper bounds for MSE of $\hat{w}$, $\tilde{w}$ and $\check{w}$ and compare them with the optimal (in the minimax sense) ones in (15) and (16).

Theorem 2 (upper bounds)
Let a mother wavelet ψ have regularity r, $max(0, 1/p-1/2) < s < r$, $p, q \geq 1$ and $\alpha > 1$.

1. For $p \geq 2$, let w be any of the $\hat{w}$, $\tilde{w}$ or $\check{w}$. Then,*

$$\sup_{w \in b^s_{p,q}(R)} E||w - w^*||_{l_2}^2 = \mathcal{O}\left(\log n \, n^{-(\alpha-1)/\alpha}\right) + \mathcal{O}\left(n^{-2s/\alpha}\right).$$

2. For $1 \leq p < 2$, let w be either $\hat{w}$ or $\tilde{w}$. Then,*

$$\sup_{w \in b^s_{p,q}(R)} E||w - w^*||_{l_2}^2 = \mathcal{O}\left(\log n \, n^{-(\alpha-1)/\alpha}\right) + \mathcal{O}\left(n^{-(2s+1-2/p)/\alpha}\right),$$

while for the Bayes Factor estimator $\check{w}$ we have

$$\sup_{w \in b^s_{p,q}(R)} E||w - \check{w}||_{l_2}^2 = \mathcal{O}\left(\log n \, n^{-(\alpha-1)/\alpha}\right) + \mathcal{O}\left((\log n)^{(2-p)/p} n^{-(\alpha-\alpha p/2+sp+p/2-1)/\alpha}\right) + \mathcal{O}\left((\log n)^{(2-p)/p}\left(\frac{n}{\sqrt{\log n}}\right)^{-(2s+1-2/p)/(\alpha/2+1/2+s-1/p)}\right).$$

The optimal choice of the hyperparameters α and β of the prior should minimize the upper bounds derived in theorem 2. However, to avoid the paradox, it should also guarantee that the prior (4) and (5) is supported on the assumed Besov space (see theorem 1).

We start from the posterior mean and the posterior median estimators whose asymptotic properties turn out to be similar. The corollary below is an immediate consequence of theorems 1 and 2.

Corollary 1
Let w be one of $\hat{w}$ or $\tilde{w}$. Choose*

1. $\alpha = (2s + 1)$, *any* $0 < \beta \leq 1$ $(q < \infty)$ *or* $0 \leq \beta < 1$ $(q = \infty)$, $p \geq 2$.
2. $\alpha = (2s + 2 - 2/p)$, *any* $1 - p/2 < \beta \leq 1$ $(q < \infty)$ *or* $1 - p/2 \leq \beta < 1$ $(q = \infty)$, $1 \leq p < 2$

Such a choice satisfies the conditions (6) and (7) of theorem 1 and

$$\sup_{w \in b^s_{p,q}(R)} E||w - w^*||_{l_2}^2 = \begin{cases} \mathcal{O}\left(\log n \, n^{-2s/(2s+1)}\right), & p \geq 2 \\ \mathcal{O}\left(\log n \, n^{-(2s+1-2/p)/(2s+2-2/p)}\right), & 1 \leq p < 2. \end{cases}$$

Corollary 1 shows that with the properly chosen hyperparameters of the prior, both the posterior mean and the posterior median estimators are asymptotically optimal up to a log-factor for $p \geq 2$. The following theorem shows that the log-factor for $p \geq 2$ is unavoidable and that the upper bound for $1 \leq p < 2$ also essentially cannot be improved, that is they can achieve only the best possible rate among linear estimators.

Theorem 3
Let again w be either $\hat{w}$ or $\tilde{w}$. Then the choice of α and β from Corollary 1 implies that there exists a positive constant C such that*

$$\sup_{w\in b^s_{p,q}(R)} E||w-w^*||^2_{l_2} \geq \begin{cases} C\left(\frac{\log n}{n}\right)^{2s/(2s+1)}, & p\geq 2 \\ Cn^{-(2s+1-2/p)/(2s+2-2/p)}, & 1 \leq p < 2. \end{cases}.$$

Summarizing, we can conclude that asymptotic minimax properties of the non-linear posterior mean and posterior median estimators are similar to those of optimal linear estimators.

The corresponding results for the Bayes Factor estimator are somewhat different for $1 \leq p < 2$ as it is shown in the following corollary:

Corollary 2
Choose

1. $\alpha = (2s + 1)$, *any* $0 < \beta \leq 1$ $(q < \infty)$ *or* $0 \leq \beta < 1$ $(q = \infty)$, $p \geq (2s+2)/(2s+1)$.
2. $\alpha = \alpha^* = s + 1 - 1/p + \sqrt{(s+1-1/p)^2 + (2s+1-2/p)}$,
   any $p(s+1/2-\alpha^*/2) < \beta \leq 1$ $(q < \infty)$ *or* $p(s+1/2-\alpha^*/2) < \beta < 1$ $(q = \infty)$, $1 \leq p < (2s+2)/(2s+1)$.

Such a choice satisfies the conditions (6) and (7) of theorem 1 and

$$\sup_{w\in b^s_{p,q}(R)} E||w-\check{w}||^2_{l_2} = \begin{cases} \mathcal{O}\left(\log n\, n^{-2s/(2s+1)}\right), & p \geq (2s+1)/(2s+2) \\ \mathcal{O}\left(\log n\, n^{-(\alpha^*-1)/\alpha^*}\right), & 1 \leq p < (2s+1)/(2s+2). \end{cases}$$

Hence, with such a choice of α and β, the Bayes Factor estimator achieves the optimal minimax rate (15) (up to a log-factor) for $p \geq (2s+1)/(2s+2)$. In addition, one can easily verify that $\alpha^* \geq 2s+2-2/p$ and, thus, even for $p < (2s+1)/(2s+2)$, $\check{w}$ outperforms the posterior mean, the posterior median and linear estimators.

Again, it is possible to prove that the upper bound in corollary 2 cannot be improved.

Theorem 4
With the choice of α *and* β *given in corollary* 2 *the following lower bound holds*

$$\sup_{w\in b^s_{p,q}(R)} E||w-\check{w}||^2_{l_2} \geq \begin{cases} C\left(\frac{\log n}{n}\right)^{2s/(2s+1)}, & p \geq (2s+2)/(2s+1) \\ Cn^{-(\alpha^*-1)/\alpha^*}, & 1 \leq p < (2s+2)/(2s+1), \end{cases}$$

where α^* *is defined in corollary 2.*

4. Simulation study

In this section, we provide a small simulation study to illustrate both the finite sample and asymptotic properties of the considered Bayesian procedures and compare them with the universal thresholding wavelet estimator of Donoho & Johnstone (1994). We refer to Antoniadis *et al.* (2001) for a comprehensive simulation analysis of various wavelet estimators.

We chose two functions from a battery of test functions used in Antoniadis *et al.* (2001), namely, 'Time Shifted Sine' and 'Angles' (see Fig. 1). 'Time Shifted Sine' is a typical example of a smooth function considered in traditional smoothing estimation. 'Angles' is a piecewise linear continuous function with large jumps in its first derivative and can be viewed as an example of a spatially inhomogeneous function.

For each test function, noisy data were generated for 500 replications by adding independent random noise $\xi_i \sim N(0, \epsilon^2)$ at 1024 data points uniformly spaced on the unit interval.

The values of ϵ were taken to correspond to values 16, 144(4) for the signal-to-noise ratio (SNR), $\int_0^1 (f - \bar{f})^2/\epsilon^2$, where $\bar{f} = \int_0^1 f$. Coiflet 5 and Coiflet 3 mother wavelets were used for 'Time Shifted Sine' and 'Angles', respectively.

The goodness-of-fit of each estimator was measured by its average mean squared error (AMSE) defined as the average of simulated replications of $n^{-1}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2$. We were also interested in the rate of decay of the AMSE as SNR increases. The rate was estimated by a slope of the least squares linear regression on the $\log_{10}(\text{AMSE}) - \log_{10}(\text{SNR})$ scale.

In the initial pilot study we investigated the effect of varying α and β for the posterior mean, posterior median and Bayes Factor estimators. The additional constants c_1 and c_2 in the prior model (5) were numerically estimated by the methods of Abramovich *et al.* (1998). The results were robust towards choices of β, and even different values of $\alpha = 2, 3, 4, 5$ did not have a drastic impact. For 'Time Shifted Sine' the rates of decay were somewhat faster for larger α, while smaller α were preferable for 'Angles'. We chose $\alpha = 4$, $\beta = 0.5$ for 'Time Shifted Sine' and $\alpha = 2$, $\beta = 0.5$ for 'Angles' for further study where we compared the three Bayesian estimators with the universal threshold wavelet estimator. All the four estimators yield

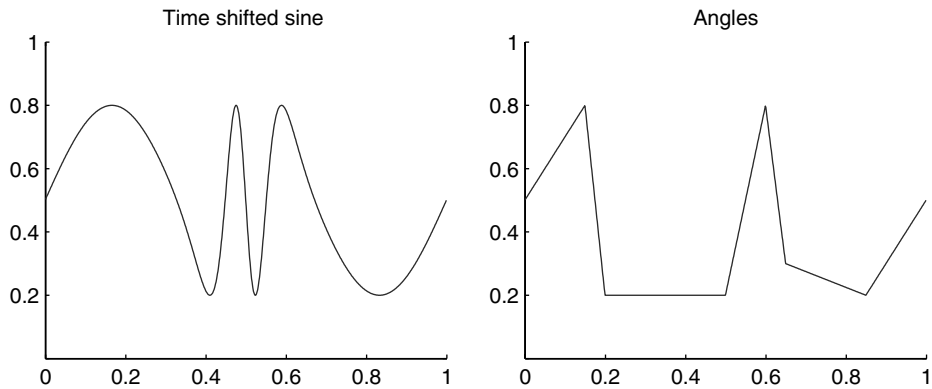

Fig. 1. Test functions used in the simulation study.

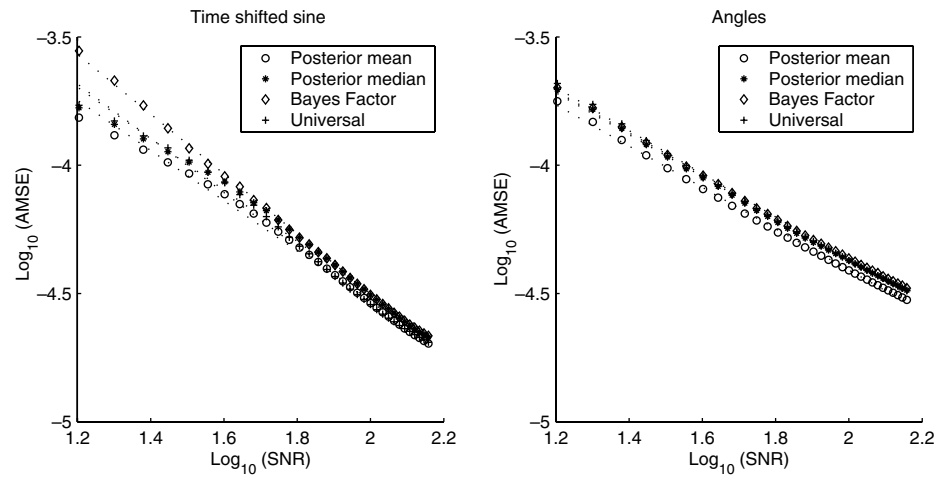

Fig. 2. $\log_{10}(\text{AMSE})$ vs. $\log_{10}(\text{SNR})$ for different wavelet estimators. Dotted lines represent the least squares linear regression fits. Their slopes estimate the convergence rate.

comparable results (see Fig. 2) and the differences in AMSE disappear as SNR increases. Posterior means have somewhat smaller AMSEs, while posterior medians, Bayesian factors and universal threshold estimators are close to it. The rates of decays of AMSEs are also similar – they are faster for smooth 'Time Shifted Sine' (the estimated slopes are within the range −0.88 to −0.95) and slower for 'Angles' (the estimated slopes vary between −0.75 to −0.80). Overall, the simulation results are in accord with the theoretical findings of the previous section.

5. Discussion and concluding remarks

We have investigated the minimax properties of the three Bayesian wavelet estimators: posterior mean, posterior median and Bayes Factor within a range of Besov spaces, where the prior distribution imposed on wavelet coefficients can be adjusted to give realizations within any given Besov space $B^s_{p,q}$. All the three estimators are inherently non-linear: the posterior mean is a non-linear smoothing shrinkage, the posterior median is a 'shrink' or 'kill' thresholding, while the Bayes Factor mimics a 'keep' or 'kill' hard thresholding rule. We showed that for the properly chosen hyperparameters of the prior, up to a log-factor, all of them achieve the optimal convergence rate within any prescribed Besov space $B^s_{p,q}$ for $p \geq 2$. For $1 \leq p < 2$, the situation is, however, different. The posterior mean and posterior median do not outperform linear estimators and do not achieve the optimal rate in this case. The Bayes Factor is optimal for a wider range of Besov spaces ($p \geq (2s + 2)/(2s + 1)$) while for $1 \leq p < (2s + 2)/(2s + 1)$ it is also not optimal although it converges faster than linear estimators.

To understand this phenomenon note that level-dependent thresholds resulted by the posterior median and the Bayes Factor thresholding are too large on high resolution levels. This fact does not affect spatially homogeneous functions ($p \geq 2$) whose wavelet coefficients are concentrated on coarse levels anyway but becomes important for spatially inhomogeneous functions ($1 \leq p < 2$) that are characterized by the presence of significant wavelet coefficients even for large j. The corresponding thresholds are 'too severe' towards them. The posterior median, in addition, even shrinks the 'survivors' that explains why the Bayes Factor performs better in this case. The behaviour of the posterior mean is quite similar to that of the posterior median. The posterior mean is a non-linear shrinkage rule, where the extent of shrinkage increases with j – it yields optimal shrinking on coarse levels but too strong on high. Recent results of Johnstone & Silverman (2002, 2003), Pensky (2003) show that to get the optimal posterior mean and posterior median Bayesian estimators even for $1 \leq p < 2$ one should replace a Gaussian part $N(0, \tau_j^2)$ of the mixture in (4) by heavier-tailed priors, e.g. by a double-scaled exponential density.

It is interesting to compare our results with those of Zhao (2000). Zhao studied the optimality of Bayesian estimation in non-parametric regression within Sobolev spaces ($p = q = 2$ in terms of parameters of Besov spaces) using the standard Fourier basis. She showed that for no Gaussian prior supported on the space of possible distributions, the resulting posterior mean estimator (that obviously coincides with the posterior median in this case) can achieve the optimal minimax rate. Instead, she proposed priors which are compound, or hierarchical *mixtures* of suitable Gaussian distributions and proved the optimality of the posterior mean estimator within Sobolev classes in this case. Our prior, which is a mixture of a single Gaussian distribution and a point mass at zero, is somewhat simpler than that of Zhao and still yields optimal estimators even within a wider range of function spaces.

Acknowledgements

The authors gratefully acknowledge the financial support of the Israeli Ministry of Science and Consiglio Nazionale delle Ricerche. We are delighted to thank Vadim Grinshtein and

Theofanis Sapatinas for valuable remarks. Constructive and helpful comments of the Editor and the two anonymous referees are gratefully acknowledged.

References

Abramovich, F., Bailey, T. C. & Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *The Statistician* **49**, 1–29.

Abramovich, F. & Sapatinas, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian inference in wavelet based models* (eds P. Müller & B. Vidakovic), Lecture Notes in Statistics, vol. 141, 33–50. Springer, New York.

Abramovich, F., Sapatinas, T. & Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. Ser. B* **60**, 725–749.

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

Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion). *J. Ital. Statist. Soc.* **6**, 97–144.

Antoniadis, A., Bigot, J. & Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. *J. Statist. Software* **6**, 1–83.

Brown, L. D. & Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384–2398.

Chipman, H. A., Kolaczyk, E. D. & McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92**, 1413–1421.

Clyde, M., Parmigiani, G. & Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–401.

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

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM, Philadelphia.

Diaconis, P. & Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14**, 1–67.

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

Donoho, D. L. & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

Freedman, D. (1999). On the Bernstein–von Mises theorem with infinite dimensional parameters. *Ann. Statist.* **27**, 1119–1140.

Johnstone, I. M. (1999). *Function estimation in Gaussian noise: sequence models*, unpublished monograph, http://www-stat.stanford.edu/∼imj.

Johnstone, I. M. & Silverman, B. W. (2002). *Needles and hay in haystacks: empirical Bayes estimates of possibly sparse sequences*, unpublished manuscript, http://www.stats.ox.ac.uk/∼silverma/papers.html.

Johnstone, I. M. & Silverman, B. W. (2003). *Empirical Bayes selection of wavelet thresholds*, unpublished manuscript, http://www.stats.ox.ac.uk/∼silverma/papers.html.

Meyer, Y. (1992). *Wavelets and operators*. Cambridge University Press, Cambridge.

Müller, P. & Vidakovic, B. (eds) (1999). *Bayesian inference in wavelet-based models*. Lecture Notes in Statistics, vol. 141, Springer, New York.

Pensky, M. (2003). *Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise*, unpublished manuscript.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151–1172.

Vidakovic, B. (1998). Non-linear wavelet shrinkage with Bayes rules and Bayes Factors. *J. Amer. Statist. Assoc.* **93**, 173–179.

Vidakovic, B. (1999). *Statistical modeling by wavelets*. John Wiley & Sons, New York.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28**, 532–552.

Received July 2002, in final form August 2003

Felix Abramovich, Department of Statistics and Operations Research, Tel-Aviv University, Tel Aviv 69978, Israel.
E-mail: felix@post.tau.ac.il

Appendix

Appendix A1: Proof of lemma 1

For λ_j^{BF} the proof follows directly from (14). For λ_j^{PM} define the function

$$g(\lambda_j) = \frac{1-\pi_j}{\pi_j}\frac{\sqrt{\tau_j^2+\sigma_n^2}}{\sigma_n}\exp\left(-\frac{\tau_j^2\lambda_j^2}{2\sigma_n^2(\tau_j^2+\sigma_n^2)}\right) + 2\Phi\left(-\frac{\tau_j\lambda_j}{\sigma_n\sqrt{\tau_j^2+\sigma_n^2}}\right)$$

The posterior median threshold λ_j^{PM} then satisfies $g(\lambda_j^{\mathrm{PM}}) = 1$. Note that $\Phi(-x) \le c\phi(x)$, $x \ge 0$, where $\phi(\cdot)$ is a standard normal density and c is some suitable positive constant (e.g. $c = 2$).

$$g(\lambda_j) \le g_1(\lambda_j) = \exp\left(-\frac{\tau_j^2\lambda_j^2}{2\sigma_n^2(\tau_j^2+\sigma_n^2)}\right)\left(\frac{1-\pi_j}{\pi_j}\frac{\sqrt{\tau_j^2+\sigma_n^2}}{\sigma_n} + c\right)$$

Let λ_{1j} be the solution of $g_1(\lambda_j) = 1$. Both $g(\lambda_j)$ and $g_1(\lambda_j)$ are decreasing functions of λ_j and, therefore, $\lambda_j^{\mathrm{PM}} \le \lambda_{1j}$. Solving the equation $g_1(\lambda_j) = 1$ one has

$$\lambda_{1j}^2 = \frac{2\sigma_n^2(\sigma_n^2+\tau_j^2)}{\tau_j^2}\log\left(\frac{1-\pi_j}{\pi_j}\frac{\sqrt{\sigma_n^2+\tau_j^2}}{\sigma_n} + c\right),$$

so

$$\lambda_j^{PM} \le \lambda_{1j} \sim \begin{cases} \sqrt{\dfrac{\log n}{n}}, & j \le J_\alpha \\ \dfrac{\sqrt{j2^{\alpha j}}}{n}, & j > J_\alpha \end{cases}$$

To complete the proof recall that on the other hand, $\lambda_j^{\mathrm{PM}} \ge \lambda_j^{\mathrm{BF}}$ and lemma 1 has been already verified for the latter.

Appendix A2: Proofs of the main results

Although theorems 2 and 3 are stated jointly for several estimators, the corresponding proofs are different for each estimator and will be considered separately.

First, without loss of generality assume that $\epsilon = 1$ in the model (1) or, equivalently, $\sigma_n^2 = 1/n$ in the model (2). Define $b_j = \tau_j^2/(\tau_j^2+\sigma_n^2)$, where τ_j^2 is given in (5). Obviously, $c_1/(c_1+1) \le b_j \le 1$ for $j \le J_\alpha$, and $c_1/(c_1+1)\ 2^{-\alpha j}n \le b_j \le c_1\ 2^{-\alpha j}n$ for $j > J_\alpha$, where $J_\alpha = (1/\alpha)\log_2 n$ was defined in lemma 1. Then,

$$b_j \sim \begin{cases} 1, & j \le J_\alpha \\ 2^{-\alpha j}n, & j > J_\alpha \end{cases} \tag{17}$$

We start the proofs from several lemmas. In what follows we use C to denote a generic positive constant, not necessarily the same each time it is used, even within a single equation.

Lemma 2
Let $\alpha > 1$. *Then*,

$$\sigma_n^2 \sum_{j=0}^{\infty} 2^j b_j^2 = \mathcal{O}\left(n^{-(\alpha-1)/\alpha}\right).$$

Proof of lemma 2.
From (17) we have

$$\sigma_n^2 \sum_{j=0}^{\infty} 2^j b_j^2 \le \sigma_n^2 \left(\sum_{j=0}^{J_\alpha} 2^j + \sum_{j=J_\alpha+1}^{\infty} n^2 2^{-j(2\alpha-1)} \right) = \mathcal{O}\left(n^{-(\alpha-1)/\alpha}\right).$$

Lemma 3
For any function $f \in B^s_{p,\mathrm{q}}[0,1]$

$$\sum_{j=0}^{\infty} (1-b_j)^2 \sum_{k=0}^{2^j-1} w_{jk}^2 = \begin{cases} \mathcal{O}\left(n^{-2s/\alpha}\right), & p \ge 2 \\ \mathcal{O}\left(n^{-(2s+1-1/p)/\alpha}\right), & 1 \le p < 2 \end{cases}$$

Proof of lemma 3

$$\sum_{j=0}^{\infty} (1-b_j)^2 \sum_{k=0}^{2^j-1} w_{jk}^2 = \sum_{j=0}^{\infty} \left(\frac{1}{2^{-\alpha j} n + 1} \right) \sum_{k=0}^{2^{2j}-1} w_{jk}^2$$

Let $s' = s$ for $p \ge 2$ and $s' = s + 1/2 - 1/p$ otherwise. Then, for sequences from Besov balls

$$\sum_{k=0}^{2^j-1} w_{jk}^2 \le C 2^{-2js'} \tag{18}$$

(e.g. Johnstone, 1999, lemma 9.3). Applying (18) for $p \ge 2$ we have

$$\begin{aligned} \sum_{j=0}^{\infty} \left(\frac{1}{2^{-\alpha j} n + 1} \right) \sum_{k=0}^{2^{2j}-1} w_{jk}^2 &\le C \sum_{j=0}^{\infty} \left(\frac{1}{2^{-\alpha j} n + 1} \right)^2 2^{-j2s} \\ &\le C \left(\frac{1}{n^2} \sum_{j=0}^{J_\alpha} 2^{2(\alpha-s)j} + \sum_{j=J_\alpha+1}^{\infty} 2^{-2sj} \right) = \mathcal{O}(n^{-2s/\alpha}), \end{aligned}$$

while for $1 \le p < 2$,

$$\begin{aligned} \sum_{j=0}^{\infty} \left(\frac{1}{2^{-\alpha j} n + 1} \right) \sum_{k=0}^{2^{2j}-1} w_{jk}^2 &\le C \left(\frac{1}{n^2} \sum_{j=0}^{J_\alpha} 2^{-(2s+1-2/p-2\alpha)j} + \sum_{j=J_\alpha+1}^{\infty} 2^{-(2s+1-2/p)j} \right) \\ &= \mathcal{O}(n^{-(2s+1-2/p)/\alpha}). \end{aligned}$$

Lemma 4
Let $\alpha > 1$. *Then, for any function* $f \in B^s_{p,\mathrm{q}}[0,1]$

$$\sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1} w_{jk} E\left(\frac{\eta_{jk} Y_{jk}}{1+\eta_{jk}} \right) = \begin{cases} \mathcal{O}\left(\log n \, n^{-(\alpha-1)/\alpha}\right) + \mathcal{O}\left(n^{-2s/\alpha}\right), & p \ge 2 \\ \mathcal{O}\left(\log n \, n^{-(\alpha-1)/\alpha}\right) + \mathcal{O}\left(n^{-(2s+1-1/p)/\alpha}\right), & 1 \le p < 2 \, . \end{cases}$$

Proof of lemma 4. As $\eta_{jk}/(1+\eta_{jk}) \le \min(\eta_{jk},1)$, we have

$$\sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1} w_{jk} E\left(\frac{\eta_{jk} Y_{jk}}{1+\eta_{jk}}\right) \le C\left(\sum_{j=0}^{J_\alpha}\sum_{k=0}^{2^j-1} \min\left(w_{jk} E(\eta_{jk} Y_{jk}), w_{jk}^2\right) + \sum_{j=J_\alpha+1}^{\infty} (1-b_j)^2 \sum_{k=0}^{2^j-1} w_{jk}^2\right) := A_1 + A_2 \tag{19}$$

Consider the first term A_1 in (19). The straightforward calculus implies

$$E(\eta_{jk} Y_{jk}) = \frac{1-\pi_j}{\pi_j}\frac{\sqrt{\sigma_n^2+\tau_j^2}}{\sigma_n}\left(\frac{\sigma_n^2+\tau_j^2}{\sigma_n^2+2\tau_j^2}\right)^{3/2} \exp\left(-\frac{w_{jk}^2\tau_j^2}{2\sigma_n^2(\sigma_n^2+2\tau_j^2)}\right) w_{jk}$$

and solving the equation $w_{jk} E(\eta_{jk} Y_{jk}) = w_{jk}^2$ for $j \le J_\alpha$ one has $w_{jk}^2 \sim \log n/n$. Thus,

$$A_1 \le C \sum_{j=0}^{J_\alpha} 2^j \frac{\log n}{n} = \mathcal{O}\left(\log n\, n^{-(\alpha-1)/\alpha}\right)$$

The upper bound for A_2 in (19) immediately follows from lemma 3.

Proof of theorem 2

1. *Posterior mean.* From (10) we have

$$\hat{w}_{jk} = b_j \frac{Y_{jk}}{1+\eta_{jk}}, \quad j \ge 0, \ \ k = 0, \ldots, 2^j - 1,$$

where the posterior odds ratio η_{jk} is given by (9). Then, for any sequence of wavelet coefficients $w = (w_{jk}) \in b^s_{p,q}(R)$,

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\left(\hat{w}_{jk} - w_{jk}\right)^2 = \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\left(b_j\left(\frac{Y_{jk}}{1+\eta_{jk}} - w_{jk}\right) - (1-b_j)w_{jk}\right)^2$$
$$\le 2\left\{\sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1} E\left(\frac{Y_{jk}}{1+\eta_{jk}} - w_{jk}\right)^2 + \sum_{j=0}^{\infty}(1-b_j)^2 \sum_{k=0}^{2^j-1} w_{jk}^2\right\} := 2(A_1 + A_2).$$

Consider the first term A_1:

$$A_1 = \sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1}\left\{E\left(\frac{Y_{jk}}{1+\eta_{jk}}\right)^2 + w_{jk}^2 - 2w_{jk}E\left(\frac{Y_{jk}}{1+\eta_{jk}}\right)\right\}$$
$$= \sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1}\left\{E\left(\frac{Y_{jk}}{1+\eta_{jk}}\right)^2 - w_{jk}^2 + 2w_{jk}E\left(\frac{\eta_{jk}Y_{jk}}{1+\eta_{jk}}\right)\right\}$$
$$\le \sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1}\left\{(EY_{jk}^2 - w_{jk}^2) + 2w_{jk}E\left(\frac{\eta_{jk}Y_{jk}}{1+\eta_{jk}}\right)\right\}$$
$$= \sigma_n^2 \sum_{j=0}^{\infty} b_j^2 2^j + 2\sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1} w_{jk} E\left(\frac{\eta_{jk}Y_{jk}}{1+\eta_{jk}}\right) := B_1 + B_2.$$

To complete the proof apply lemma 2, lemma 4 and lemma 3 to get the upper bounds for B_1, B_2 and A_2, respectively.

2. Posterior median. Define $w'_{jk} = b_j Y_{jk} I(|Y_{jk}| \geq \lambda_j^{\mathrm{PM}})$ and $w''_{jk} = b_j w_{jk}$. Then,

$$\begin{aligned}\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\left(\tilde{w}_{jk} - w_{jk}\right)^2 &= \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\Big((\tilde{w}_{jk} - w'_{jk}) + (w'_{jk} - w''_{jk}) + (w''_{jk} - w_{jk})\Big)^2 \\ &\leq 3\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\Big[(w''_{jk} - w_{jk})^2 + E(\tilde{w}_{jk} - w'_{jk})^2 + E(w'_{jk} - w''_{jk})^2\Big] \\ &:= 3(A_1 + A_2 + A_3),\end{aligned}$$

The upper bound for A_1 follows immediately from lemma 3. Consider now the second term A_2. Note that $\tilde{w}_{jk} = w'_{jk} = 0$ for $|Y_{jk}| < \lambda_j^{\mathrm{PM}}$ and $\tilde{w}_{jk} \to w'_{jk}$ monotonically as $|Y_{jk}| \to \infty$. Hence, the difference between the two thresholding rules is maximal at $|Y_{jk}| = \lambda_j^{\mathrm{PM}}$ and, therefore, $\max_{Y_{jk}}|\tilde{w}_{jk} - w'_{jk}| = b_j\lambda_j^{\mathrm{PM}}$. Then,

$$A_2 \leq \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}(\lambda_j^{\mathrm{PM}})^2 b_j^2 \leq C\left(\sum_{j=0}^{J_\alpha} 2^j\frac{\log n}{n} + \sum_{j>J_\alpha} j2^j 2^{-\alpha j}\right) = \mathcal{O}\Big(n^{-(\alpha-1)/\alpha}\log n\Big),$$

Finally, for the last term A_3, lemma 1 of Donoho & Johnstone (1994) for hard thresholding yields

$$E(w_{jk} - Y_{jk}I(|Y_{jk}| \geq \lambda_j^{\mathrm{PM}}))^2 \leq C\begin{cases}(\lambda_j^{\mathrm{PM}})^2 + \sigma_n^2, & w_{jk}^2 > (\lambda_j^{\mathrm{PM}})^2 \\ w_{jk}^2 + \sigma_n\lambda_j^{\mathrm{PM}}\phi\Big(\lambda_j^{\mathrm{PM}}/\sigma_n\Big), & w_{jk}^2 \leq (\lambda_j^{\mathrm{PM}})^2.\end{cases} \tag{20}$$

Moreover, as $x\phi(x) \leq c/x^2$ for some constant c and $\sigma_n = o(\lambda_j^{\mathrm{PM}})$ for all j,

$$A_3 \leq C\left(\sum_{j=0}^{\infty} b_j^2 \min\left(2^j(\lambda_j^{\mathrm{PM}})^2, \sum_{k=0}^{2^j-1} w_{jk}^2\right) + \sum_{j=0}^{\infty} b_j^2 2^j\frac{\sigma_n^4}{(\lambda_j^{\mathrm{PM}})^2}\right) := B_1 + B_2. \tag{21}$$

Then, from lemma 1, (17) and (18),

$$\begin{aligned}B_1 &\leq C\sum_{j=0}^{J_\alpha} 2^j(\lambda_j^{\mathrm{PM}})^2 + \sum_{j>J_\alpha} b_j^2\sum_{k=0}^{2^j-1} w_{jk}^2 \leq C\sum_{j=0}^{J_\alpha} 2^j\frac{\log n}{n} + \sum_{j>J_\alpha} n^2 2^{-2j(\alpha+s')} \\ &= \mathcal{O}\Big(\log n\, n^{-(\alpha-1)/\alpha}\Big) + \mathcal{O}\Big(n^{-2s'/\alpha}\Big)\end{aligned}$$

while

$$B_2 \leq C\left(\sum_{j=0}^{J_\alpha} 2^j\frac{1}{n\log n} + \sum_{j>J_\alpha} n^2 2^{-j(3\alpha-1)} j^{-1}\right) = \mathcal{O}\Big((\log n)^{-1} n^{-(\alpha-1)/\alpha}\Big).$$

3. Bayes Factor. Due to the embedding properties of Besov spaces ($B_{p,q}^s \subset B_{p,\infty}^s$) it is sufficient to prove the result for $q=\infty$. Similar to (21), from (20) we have:

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E(w_{jk} - \tilde{w}_{jk})^2 \leq C\left(\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\min\Big((\lambda_j^{\mathrm{BF}})^2, w_{jk}^2\Big) + \sum_{j=0}^{\infty} 2^j\frac{\sigma_n^4}{(\lambda_j^{\mathrm{BF}})^2}\right) := A_1 + A_2.$$

Then, from lemma 1

$$A_1 \leq C\left(\sum_{j=0}^{J_\alpha} 2^j\frac{\log n}{n} + \sum_{j>J_\alpha}\sum_{k=0}^{2^j-1}\min\left(j\frac{2^{\alpha j}}{n^2}, w_{jk}^2\right)\right) := B_1 + B_2,$$

where $B_1 = \mathcal{O}(\log n \quad n^{-(\alpha-1)/\alpha})$. For $p \geq 2$, apply (18) to get $B_2 \leq \sum_{j>J_\alpha}\sum_{k=0}^{2^j-1} w_{jk}^2 = \mathcal{O}\Big(n^{-2s/\alpha}\Big)$.

More delicate analysis is required to find an upper bound for B_2 within $b^s_{p,q}(R)$ for $1 \le p < 2$. Consider the upper bound for $\sum_{k=0}^{2^j-1} \min\left((\lambda_j^{\mathrm{BF}})^2, w_{jk}^2\right)$ subject to $\sum_{k=0}^{2^j-1} |w_{jk}|^p \le C_j^p$, where $C_j = 2^{-j(s+1/2-1/p)}$. For $1 \le p < 2$ the extreme vectors are permutations of the 'spike' $(C_j, 0, \ldots, 0)$ if $\lambda_j^{\mathrm{BF}} \ge C_j$ and permutations of the vector $(\lambda_j^{\mathrm{BF}}, \ldots, \lambda_j^{\mathrm{BF}}, 0, \ldots, 0)$, where the number of nonzero entries λ_j^{BF} is $(C_j/\lambda_j^{\mathrm{BF}})^p$, otherwise. The corresponding upper bounds are then C_j^2 and $C_j^p(\lambda_j^{\mathrm{BF}})^{2-p}$ respectively. Let J^* satisfy $\lambda_j^{\mathrm{BF}} = C_j$, where recall that $\lambda_j^{\mathrm{BF}} \sim \sqrt{j2^{\alpha j}}/n$ for $j \ge J_\alpha$. Then $J^* = (\alpha/2 + s + 1/2 - 1/p)^{-1} \log_2(n/\sqrt{\log n})(1 + o(1))$. Note that for α satisfying theorem 1, $J^* \ge J_\alpha$. Apply now the above upper bounds to get the upper bound for B_2:

$$
\begin{aligned}
B_2 &\le \sum_{j=J_\alpha+1}^{J^*} 2^{j\alpha(2-p)/2} n^{-(2-p)} j^{(2-p)/2} 2^{-jp(s+1/2-1/p)} + \sum_{j>J^*} 2^{-j(2s+1-2/p)} \\
&= \mathcal{O}\left((\log n)^{(2-p)/p} n^{-(\alpha-\alpha p/2+sp+p/2-1)/\alpha}\right) \\
&\quad + \mathcal{O}\left((\log n)^{(2-p)/p} \left(\frac{n}{\sqrt{\log n}}\right)^{-(2s+1-2/p)/(\alpha/2+1/2+s-1/p)}\right).
\end{aligned}
$$

Finally, for any p

$$
A_2 \le C\left(\sum_{j=0}^{J_\alpha} 2^j \frac{1}{n \log n} + \sum_{j>J_\alpha} 2^{-j(\alpha-1)} j^{-1}\right) = \mathcal{O}\left((\log n)^{-1} n^{-(\alpha-1)/\alpha}\right).
$$

Proof of theorem 3

1. Posterior mean. Note that $\left(w_{jk} - E\frac{Y_{jk}}{1+\eta_{jk}}\right) w_{jk} \ge 0$ and, therefore, one obtains the following lower bound for the risk:

$$
\begin{aligned}
\sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} E\left(w_{jk} - \hat{w}_{jk}\right)^2 &= \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} E\left(w_{jk} - b_j \frac{Y_{jk}}{1+\eta_{jk}}\right)^2 \ge \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \left(w_{jk} - b_j E\left(\frac{Y_{jk}}{1+\eta_{jk}}\right)\right)^2 \\
&= \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \left(b_j \left(w_{jk} - E\frac{Y_{jk}}{1+\eta_{jk}}\right) + w_{jk}(1-b_j)\right)^2 \\
&\ge \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \left[b_j^2 \left(E\frac{\eta_{jk} Y_{jk}}{1+\eta_{jk}}\right)^2 + (1-b_j)^2 w_{jk}^2\right] \qquad (22)
\end{aligned}
$$

Consider first the case $p \ge 2$. From (22),

$$
\begin{aligned}
\sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} E\left(w_{jk} - \hat{w}_{jk}\right)^2 &\ge \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} b_j^2 \left(E\frac{\eta_{jk} Y_{jk}}{1+\eta_{jk}}\right)^2 \\
&\ge \frac{1}{4} \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} b_j^2 \min\left(E(\eta_{jk} Y_{jk}), w_{jk}\right)^2, \qquad (23)
\end{aligned}
$$

where we have used the fact that $\frac{\eta_{jk}}{\eta_{jk}+1} \ge \min(\eta_{jk}/2; 1/2)$.

Define $J'_s = (2s+1)^{-1} \log_2(n/\log n)$ and consider the following wavelet sequence

$$
w_{jk} = \begin{cases} c\sqrt{\dfrac{\log n}{n}}, & 0 \le j \le J'_s, \quad 0 \le k < 2^j; \\ 0, & j > J'_s, \quad 0 \le k < 2^j. \end{cases} \qquad (24)
$$

Using the definition of $||w||_{b^s_{p,q}}$ given in section 2.1 one can easily verify that it is always possible to choose the normalized constant c such that $||w||_{b^s_{p,q}} \le R$.

Recall that $\alpha = 2s+1$, so $J'_s < J_\alpha$, where J_α was defined in lemma 1 and, therefore, $b_j^2 \geq c_1^2/(c_1+1)^2$ for $j \leq J'_s$. In the proof of lemma 4 we argue that for $w_{jk} = \sqrt{\log n/n}$, $E(\eta_{jk} Y_{jk}) \sim w_{jk}, \quad j \leq J_\alpha$. Then, (23) implies

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\left(w_{jk} - \hat{w}_{jk}\right)^2 \geq C \sum_{j=0}^{J'_s} b_j^2\, 2^j\, \frac{\log n}{n} \geq C\left(\frac{\log n}{n}\right)^{2s/(2s+1)}$$

Analogously, for $1 \leq p < 2$, from (22) one also has

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\left(w_{jk} - \hat{w}_{jk}\right)^2 \geq \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} (1-b_j)^2 w_{jk}^2. \tag{25}$$

Consider the sequence of wavelet coefficients with $w_{jk} = c\, n^{-(s+1/2-1/p)/\alpha}\delta_{k0}$, $j \leq J_\alpha$ and zero otherwise, where δ_{k0} is a Kronecker delta. Then, obviously, $||w||_{b^s_{p,q}} < \infty$. On the other hand, for $\alpha = 2s + 2 - 2/p$ from (25) one has

$$\begin{aligned}\sum_{j=0}^{J_\alpha}\sum_{k=0}^{2^j-1} E\left(w_{jk} - \hat{w}_{jk}\right)^2 &\geq C \sum_{j=1}^{J_\alpha}\left(\frac{1}{2^{-\alpha j} n + 1}\right)^2 n^{-(2s+1-2/p)/(2s+2-2/p)} \\ &\geq C\, n^{-(2s+1-2/p)/(2s+2-2/p)}.\end{aligned}$$

2. Posterior median. Consider the case $p \geq 2$. One can easily verify that

$$\begin{aligned}&(w''_{jk} - w_{jk})E(\tilde{w}_{jk} - w'_{jk}) \geq 0, \quad (w''_{jk} - w_{jk})E(w''_{jk} - w'_{jk}) \geq 0, \\ &E\Big((w''_{jk} - w'_{jk})(\tilde{w}_{jk} - w'_{jk})\Big) \geq 0\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\left(\widetilde{w}_{jk} - w_{jk}\right)^2 &= \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\Big((\tilde{w}_{jk} - w'_{jk}) + (w'_{jk} - w''_{jk}) + (w''_{jk} - w_{jk})\Big)^2 \\ &\geq \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E(w'_{jk} - w''_{jk})^2 = \sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1} E\Big(Y_{jk} I(|Y_{jk}| \geq \lambda_j^{\mathrm{PM}}) - w_{jk}\Big)^2\end{aligned}$$

Consider again the sequence $w \in b^s_{p,q}(R)$ from (24). Note that $w_{jk} \sim \lambda_j^{\mathrm{PM}}$, $j \leq J'_s = (2s+1)^{-1}\log_2(n/\log n)$, so from (6.11) of Johnstone (1999), $E\left(Y_{jk} I(|Y_{jk}| \geq \lambda_j^{\mathrm{PM}}) - w_{jk}\right)^2 \geq C(\lambda_j^{\mathrm{PM}})^2$. Then, the straightforward calculus yields

$$\sum_{j=0}^{\infty} b_j^2 \sum_{k=0}^{2^j-1} E\Big(Y_{jk} I(|Y_{jk}| \geq \lambda_j^{PM}) - w_{jk}\Big)^2 \geq C \sum_{j=0}^{J'_s} 2^j \left(\frac{\log n}{n}\right) = C\left(\frac{\log n}{n}\right)^{2s/(2s+1)}$$

Analogously, for $1 \leq p < 2$,

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} E\left(\tilde{w}_{jk} - w_{jk}\right)^2 \geq \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} (w''_{jk} - w_{jk})^2 = \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} (1-b_j)^2 w_{jk}^2$$

and the rest of the proof is exactly as for the posterior mean case above.

Proof of corollary 2

For $p \geq 2$ the proof follows immediately from theorems 1 and 2. For $1 \leq p < 2$, the optimal α is obtained as the minimum between the solutions of $(2\alpha - \alpha p + 2sp + p - 2)/(2\alpha) = 1 - 1/\alpha$ and $(2s + 1 - 2/p)/(1/2 + \alpha/2 + s - 1/p) = 1 - 1/\alpha$, i.e. between $2s+1$ and α^* defined in the corollary depending on whether $p \geq (2s+2)/(2s+1)$ or not. The restrictions on β in the corollary, as usual, are the result of (6) and (7) in theorem 1.

Proof of theorem 4

For $p \geq (2s+2)/(2s+1)$ we apply the arguments similar to those used in the proof of theorem 3. Consider again the sequence $w \in b^s_{p,q}(R)$ defined in (24). For $j \leq J_s'$, $w_{jk} \sim \lambda_j^{\mathrm{BF}}$, so $E(Y_{jk} I(|Y_{jk}| \geq \lambda_j^{\mathrm{BF}}) - w_{jk})^2 \geq C(\lambda_j^{\mathrm{BF}})^2$ [(6.11) of Johnstone, 1999]. Then,

$$\sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} E\left(w_{jk} - \breve{w}_{jk}\right)^2 \geq C \sum_{j=0}^{J_s'} 2^j \left(\frac{\log n}{n}\right) \geq C \left(\frac{\log n}{n}\right)^{2s/(2s+1)}$$

For $1 \leq p < (2s+2)/(2s+1)$, define $\tilde{J} = (\alpha^*/2 + s + 1/2 - 1/p)^{-1} \log_2 n$ and consider the sequence with the only non-zero entry $w_{\tilde{J}0} = R\, n^{-(s+1/2-1/p)/(\alpha^*/2+s+1/2-1/p)}$. Obviously, $||w||_{b^s_{p,q}} = R$. Recall that $\tilde{J} > J^* \geq J_{\alpha^*}$, where J^* was defined in the proof of theorem 2 for the Bayes Factor. Hence, $w^2_{\tilde{J}0} \leq (\lambda_{\tilde{J}}^{\mathrm{BF}})^2$ and from (6.11) of Johnstone (1999) we have $E(Y_{\tilde{J}k} I(|Y_{\tilde{J}k}| \geq \lambda_j^{\mathrm{BF}}) - w_{\tilde{J}k})^2 \geq C w^2_{\tilde{J}0}$. Then,

$$\sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} E\left(w_{jk} - \breve{w}_{jk}\right)^2 \geq C w^2_{\tilde{J}0} \geq C n^{-(2s+1-2/p)/(\alpha^*/2+1/2+s-1/p)} = C n^{-(\alpha^*-1)/\alpha^*}$$