

ON POINTWISE OPTIMALITY OF BAYES FACTOR WAVELET REGRESSION ESTIMATORS

Natalia Bochkina,

Department of Epidemiology and Public Health, Imperial College,
London, W2 1PG, United Kingdom

Email: N.Bochkina@imperial.ac.uk

and

Theofanis Sapatinas,

Department of Mathematics and Statistics, University of Cyprus,
P.O. Box 20537, CY 1678 Nicosia, Cyprus.

Email: T.Sapatinas@ucy.ac.cy

Abstract

We investigate theoretical performance of Bayes factor estimators at a single point in wavelet regression models with independent and identically distributed errors that are not necessarily normally distributed. We compare these estimators in terms of their frequentist pointwise optimality in Besov spaces for some combinations of error and prior distributions. Simulated examples are used to illustrate the performance of the Bayes factor estimation procedure in a fully Bayesian framework, and compared with a recently proposed minimax (projection) wavelet estimator. An application to a dataset that was collected in an anaesthesiological study is also presented.

Keywords: BAYESIAN INFERENCE; BESOV SPACES; NONPARAMETRIC REGRESSION, OPTIMALITY; POINTWISE RISK; WAVELETS

AMS (2000) Subject Classifications: PRIMARY 62G08; SECONDARY 62C10, 62C12, 62C20

1 Introduction

Over the last decade, the nonparametric regression literature has been dominated by *nonlinear wavelet* methods. These methods are based on the idea of thresholding, meaning that if an empirical wavelet coefficient is sufficiently large in magnitude, that is if its magnitude exceeds a predetermined threshold, then the corresponding term in the empirical wavelet expansion is retained (or shrunk towards zero); otherwise it is omitted. The resulting term-by-term wavelet thresholding estimators possess optimal or near-optimal convergence rates, and are typically implemented through fast

algorithms which makes them very appealing in practice (see, e.g., Donoho & Johnstone, 1994, 1995, 1998; Donoho, Johnstone, Kerkyacharian & Picard, 1995; Vidakovic, 1999; Abramovich, Bailey & Sapatinas, 2000; Antoniadis, Bigot & Sapatinas, 2001).

Various Bayesian and empirical Bayes approaches for term-by-term wavelet nonlinear shrinkage and wavelet thresholding estimators have been also proposed. (To introduce terminology, a *shrinkage* rule shrinks empirical wavelet coefficients to zero, whilst a *thresholding* rule shrinks and, in addition, sets to zero all empirical wavelet coefficients below a certain level.) These approaches impose a prior distribution on the wavelet coefficients of the unknown response function, designed to capture the sparseness of wavelet expansions common to most applications. The response function is then estimated by applying a suitable Bayes rule to the resulting posterior distribution of the wavelet coefficients. Different choices of loss function lead to different Bayes rules and hence to different, usually *level-dependent*, nonlinear wavelet shrinkage and wavelet thresholding rules (see, e.g., Chipman, Kolaczyk & McCulloch, 1997; Abramovich, Sapatinas & Silverman, 1998; Clyde, Parmigiani & Vidakovic, 1998; Vidakovic, 1998; Clyde & George, 2000; Angelini & Sapatinas, 2004; Angelini & Vidakovic, 2004).

However, until recently, their frequentist optimality (in the minimax sense) properties have not been studied. Abramovich, Amato & Angelini (2004) investigated optimality of posterior mean, posterior median and Bayes factor estimators in terms of the *global* L^2 -loss function for the combination of normal error and normal prior distributions. Pensky & Sapatinas (2005) and Pensky (2006) studied optimality of Bayes factor and posterior mean estimators respectively with respect to the L^2 -loss function for a wide variety of combination of error and prior distributions. Johnstone & Silverman (2005) explored adaptive optimality of *empirical* Bayes posterior mean and posterior median estimators with respect to a wide range of L^r -loss functions ($0 < r \leq 2$) for normal error and some heavy-tailed prior distributions. The adaptive optimality of an *empirical* Bayes procedure for the Bayes factor estimator with respect to the L^2 -loss function for normal error and some heavy-tailed prior distributions was considered in Pensky & Sapatinas (2005). Recently, Abramovich, Angelini & De Canditiis (2007) explored the optimality of posterior mean, posterior median and Bayes factor estimators in terms of the *pointwise* l^2 -loss function for the combination of normal error and normal prior distributions. They showed that under the considered Bayesian hierarchical model, pointwise optimality is achieved up to a logarithmic factor.

This paper continues the line of investigation of Abramovich, Angelini & De Canditiis (2007); however, our focus will be on investigating optimality of the Bayes factor estimator with respect to the l^2 -loss function. The characteristic of this estimator is that it leads to a *hard* thresholding rule, unlike the posterior mean which leads to a nonlinear *shrinkage* rule and the posterior median which leads to a *soft* thresholding rule. Moreover, the Bayes factor estimator is much easier to evaluate in the majority of cases unlike posterior mean or posterior median estimators (see Bochkina & Sapatinas, 2005; Pensky, 2006). As in Pensky & Sapatinas (2005), who studied optimality of Bayes

factor estimators with respect to the L^2 -loss function, we put very mild restrictions on the errors in the standard nonparametric regression model. Furthermore, we do not assume the distribution of the errors to be known and hence consider a range of error and prior distributions for the wavelet coefficients. Moreover, as we demonstrate below, the use of a more flexible Bayesian hierarchical model improves the pointwise convergence rates and, under certain conditions, achieves pointwise optimality without the extra logarithmic factor appeared in the results of Abramovich, Angelini & De Canditiis (2007).

The paper is organized as follows. In Section 2, we introduce Bayesian models for the wavelet coefficients, extending the previously considered (in the context of pointwise optimality) normal error and normal prior model in Abramovich, Angelini & De Canditiis (2007) to combinations of error and prior distributions with exponential descents. In Section 3, we discuss assumptions on the error and prior distributions, and provide assertions about pointwise optimality of Bayes factor estimators in Besov spaces for certain combinations of error and prior distributions. In Section 4, simulated results are used to illustrate performance of the Bayes factor estimation procedure in a fully Bayesian framework, and compared with a recently proposed minimax (projection) wavelet estimator. We also present an application to a dataset that was collected in an anaesthesiological study. Some concluding remarks are made in Section 5. Finally, in Section 6 (Appendix), we provide some auxiliary statements and the proofs of the theoretical results stated in Section 3.

2 The Bayesian model

Consider the standard nonparametric regression model:

$$Y_i = f(t_i) + Z_i, \quad i = 1, \dots, n, \quad (2.1)$$

where $t_i = i/n$, f is the unknown response function that is assumed to belong to the space of square integrable functions on $[0, 1]$, i.e., $f \in L^2[0, 1]$, and that the Z_i 's are independent and identically distributed (*iid*) random variables with $\mathbb{E}(Z_1) = 0$ and $\mathbb{V}(Z_1) = \sigma^2 < \infty$. We also assume that $\mathbb{E}(Z_1^4) < \infty$.

Then, any $f \in L^2[0, 1]$ can be represented (in the L^2 -sense) by a wavelet series, i.e.,

$$f(t) = \sum_{k \in K_{L-1}} \tilde{\theta}_k \phi_{Lk}(t) + \sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} \tilde{\theta}_{jk} \psi_{jk}(t),$$

where, for some (fixed) *primary resolution* level $L \geq 0$, $\phi_{Lk}(t) = 2^{L/2} \phi(2^L t - k)$, $\psi_{jk}(t) = 2^{j/2} \psi(2^j t - k)$, $\tilde{\theta}_k = \int_{-\infty}^{+\infty} \phi_{Lk}(t) f(t) dt$ and $\tilde{\theta}_{jk} = \int_{-\infty}^{+\infty} \psi_{jk}(t) f(t) dt$; here, ϕ is the *scaling function*, ψ is a corresponding *wavelet function*, and K_{L-1} is the set of indices for which the scaling function ϕ_{Lk} is defined. For suitable choices of ϕ and ψ and appropriate boundary treatments, the corresponding set of ϕ_{Lk} and ψ_{jk} forms an orthonormal set in $L^2[0, 1]$ (see, e.g., Cohen, Daubechies & Vial, 1993; Johnstone & Silverman, 2004).

Application of the (boundary corrected) discrete wavelet transform (DWT) to (2.1) yields

$$\begin{aligned}\mathcal{U}_k &= u_k + \epsilon_k, \quad k \in K_{L-1}, \\ \mathcal{W}_{jk} &= w_{jk} + \varepsilon_{jk}, \quad j = L, L+1, \dots, J-1, \quad k = 0, 1, \dots, 2^j - 1,\end{aligned}$$

where $J = \log_2(n)$ and $\epsilon_k, \varepsilon_{jk}$ are uncorrelated random variables due to the unitary property of the DWT. Denote $\theta_k = u_k/\sqrt{n}$ and $\theta_{jk} = w_{jk}/\sqrt{n}$ and recall that $\tilde{\theta}_k \approx \theta_k$ and $\tilde{\theta}_{jk} \approx \theta_{jk}$ (see, e.g., Vidakovic, 1999). In the appendix, we provide a more detailed treatment of this relationship for the boundary coiflets $\{\phi, \psi\}$, a particular case of a wavelet system used to establish the pointwise optimality results given in subsequent sections (see Lemma 4). In this case, there will be $2^L - 2(S - s - 1)$ scaling coefficients at the primary resolution level L , and, thus, K_{L-1} is the set of indices for which the corresponding scaling function ϕ_{Lk} is defined (see Johnstone & Silverman, 2004, p. 83).

We use the Bayesian framework to construct estimators $\hat{\theta}_k$ of θ_k (based on \mathcal{U}_k) and $\hat{\theta}_{jk}$ of θ_{jk} (based on \mathcal{W}_{jk}) in order to estimate the unknown response function f . Since the wavelet representations of a vast majority of functions contain only a few non-negligible wavelet coefficients in their expansions, similar to the priors used previously in the Bayesian wavelet regression literature, we place the following prior on the wavelet coefficient w_{jk} :

$$w_{jk} \sim \pi_{j,n} \tau_{j,n} h(\tau_{j,n} \cdot) + (1 - \pi_{j,n}) \delta(0), \quad j = L, L+1, \dots, \quad k = 0, 1, \dots, 2^j - 1, \quad (2.2)$$

where $0 \leq \pi_{j,n} \leq 1$ for $L \leq j \leq J-1$ and $\pi_{j,n} = 0$ for $j \geq J$, $\tau_{j,n} > 0$, $\delta(0)$ is a point mass at zero, and w_{jk} are independent random variables. For the prior model h , we consider not only the standard normal probability density function (*pdf*) but also the double-exponential *pdf* with scale parameter 1. To complete the prior specification of f , we place noninformative priors (e.g., the uniform density on \mathbb{R}) on the scaling coefficients $u_k, k \in K_{L-1}$.

According to the prior model (2.2), w_{jk} is either zero with probability $(1 - \pi_{j,n})$ or with probability $\pi_{j,n}$ is distributed with the *pdf* h with scale parameter $\tau_{j,n}$; the proportion $\pi_{j,n}$ indicates whether a value is small or large and can be used to ‘control’ the trade-off between sparse and dense sequences. In what follows, we impose all conditions on the prior odds ratio

$$\beta_{j,n} = (1 - \pi_{j,n})/\pi_{j,n}.$$

Note that we allow dependence of $\pi_{j,n}$ (and hence of $\beta_{j,n}$) not only on the resolution level j but also on n . It is most natural since the proportion of wavelet coefficients we are intending to keep depends not only on the function f itself but also on the amount of data available: when n is larger, the estimators of wavelet coefficients become more reliable and, hence, smaller wavelet coefficients can be distinguished from pure noise. Consequently, for larger n one can keep larger number of wavelet coefficients at a particular resolution level j which leads to the larger the value of $\pi_{j,n}$.

Let us now discuss the distribution of the errors ε_{jk} . It follows from (2.1) that

$$\varepsilon_{jk} \approx n^{-1/2} 2^{j/2} \sum_{i=1}^n \psi(2^j i/n - k) Z_i.$$

If the Z_i 's are *iid* random variables with $\mathbb{E}(Z_1^4) < \infty$, it is not difficult to see that the sequence $\{n^{-1/2} 2^{j/2} \psi(2^j i/n - k) Z_i\}$ satisfies the Lyapunov condition (see, e.g., Billingsley, 1995, p. 362) provided that $2^j/n \rightarrow 0$ as $n \rightarrow \infty$. Hence, if the resolution level is reasonably small ($j \leq J_0$ where $J - J_0 \rightarrow \infty$ as $n \rightarrow \infty$), the errors ε_{jk} are asymptotically $N(0, \sigma^2)$ distributed and, thus, asymptotically independent. For a more detailed treatment of asymptotic normality, the interested reader is referred to, e.g., Neumann & von Sachs (1995).

We assume that the distribution of the errors ε_{jk} is level-dependent,

$$\varepsilon_{jk} \sim \varphi_j(\cdot), \quad L \leq j \leq J - 1,$$

with the *pdf* φ_j having exponential descents, i.e.,

$$\varphi_j(x) = c_j \exp\{-(|x|/\sigma_j)^\beta\}, \quad 0 < \underline{\sigma} \leq \sigma_j \leq \bar{\sigma} < \infty, \quad c_j > 0, \quad \beta > 0. \quad (2.3)$$

(For the distribution of errors of the scaling coefficients, ϵ_k , we only assume that it has a finite variance σ_{L-1}^2 .) As we shall show later, one does not need the knowledge of the true distribution of the errors ε_{jk} , and can achieve pointwise optimality with the choice (2.3) with either $\beta = 2$ (normal) or $\beta = 1$ (double-exponential). To keep the exposition simple, we do not consider any heavier-tailed *pdf*'s (e.g., Student-*t* distributions) for both φ_j and h . In Section 3, we provide some further explanation about the considered choice of error and prior distributions to be just combinations of the commonly used distributions with exponential descents, namely normal and double-exponential distributions.

In what follows, we conduct Bayesian inference for each wavelet coefficient separately. Denote

$$d_{jk} = \mathcal{W}_{jk}/\sqrt{n} \quad \text{and} \quad \nu_j = \sqrt{n}\tau_{j,n}. \quad (2.4)$$

Taking into account the relation between w_{jk} and θ_{jk} and (2.2)–(2.4), we derive that the posterior *pdf* of θ_{jk} given d_{jk} is of the form

$$p(\theta_{jk} | d_{jk}) = \frac{\sqrt{n} \varphi_j(\sqrt{n}(\theta_{jk} - d_{jk})) \nu_j h(\nu_j \theta_{jk}) + \beta_{j,n} \sqrt{n} \varphi_j(\sqrt{n} d_{jk}) \delta(0)}{\int_{-\infty}^{+\infty} \sqrt{n} \varphi_j(\sqrt{n}(x - d_{jk})) \nu_j h(\nu_j x) dx + \beta_{j,n} \sqrt{n} \varphi_j(\sqrt{n} d_{jk})}.$$

The Bayes factor estimator of θ_{jk} is derived as follows (see Vidakovic, 1998): after observing d_{jk} , we test the hypothesis

$$H_0 : \theta_{jk} = 0 \quad \text{versus} \quad H_1 : \theta_{jk} \neq 0.$$

If the hypothesis H_0 is rejected, θ_{jk} is estimated by d_{jk} , otherwise $\theta_{jk} = 0$, so that the estimator $\hat{\theta}_{jk}$ is given by

$$\hat{\theta}_{jk} = d_{jk} \mathbb{I} \left(\frac{\mathbb{P}(H_1 | d_{jk})}{\mathbb{P}(H_0 | d_{jk})} > 1 \right),$$

where $\mathbb{I}(A)$ denotes the indicator function of the set A . Observe that the posterior odds ratio can be rewritten as

$$\frac{\mathbb{P}(H_1 | d_{jk})}{\mathbb{P}(H_0 | d_{jk})} = \frac{\zeta_{j,n}(d_{jk})}{\beta_{j,n}},$$

where

$$\zeta_{j,n}(d_{jk}) = I_j(d_{jk}) / [\sqrt{n} \varphi_j(\sqrt{n}d_{jk})], \quad (2.5)$$

and

$$I_j(d_{jk}) = \int_{-\infty}^{+\infty} \sqrt{n} \varphi_j[\sqrt{n}(x - d_{jk})] \nu_j h(\nu_j x) dx. \quad (2.6)$$

Rewriting $\hat{\theta}_{jk}$ in view of (2.5), we obtain

$$\hat{\theta}_{jk} = d_{jk} \mathbb{I}(\zeta_{j,n}(d_{jk}) > \beta_{j,n}). \quad (2.7)$$

It is easy to check that $\zeta_{j,n}(d_{jk})$ are even functions of d_{jk} . If, moreover, the functions $\zeta_{j,n}(d_{jk})$ are strictly increasing in d_{jk} for $d_{jk} > 0$, then

$$\zeta_{j,n}(d_{jk}) > \beta_{j,n} \quad \text{if and only if} \quad |d_{jk}| > t_{j,n} = \zeta_{j,n}^{-1}(\beta_{j,n}).$$

Hence, (2.7) is a hard thresholding rule with the threshold $t_{j,n}$, i.e.,

$$\hat{\theta}_{jk} = d_{jk} \mathbb{I}(|d_{jk}| > t_{j,n}). \quad (2.8)$$

Indeed, in the majority of practical cases, it is true that (2.7) gives rise to a hard thresholding rule, as it is confirmed by the following statement.

Proposition 1 (Part of Lemma 1 in Pensky & Sapatinas, 2005). *If φ_j is the normal or the double-exponential pdf, then $\zeta_{j,n}(d_{jk})$ is strictly increasing in d_{jk} for $d_{jk} > 0$.*

Note that under the considered error model, the noninformative priors for the scaling coefficients u_k result in their posterior distributions being proper and their estimates being the corresponding empirical scaling coefficients \mathcal{U}_k , $k \in K_{L-1}$, and thus $\hat{\theta}_k = \mathcal{U}_k / \sqrt{n}$, $k \in K_{L-1}$. Since we assumed that $\pi_{j,n} = 0$ if $j \geq J$, $k = 0, 1, \dots, 2^j - 1$, it implies that $\hat{\theta}_{jk} = 0$ as $j \geq J$, $k = 0, 1, \dots, 2^j - 1$, and therefore the estimator \hat{f} of f is of the form

$$\hat{f}(t) = \sum_{k \in K_{L-1}} \hat{\theta}_k \phi_{Lk}(t) + \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} \hat{\theta}_{jk} \psi_{jk}(t). \quad (2.9)$$

Coefficients $\hat{\theta}_{jk}$ are found using formula (2.7) where the function $\zeta_{j,n}(d_{jk})$ is defined by (2.5) and (2.6). To complete the construction of the estimator, we need to choose the error model φ_j , the prior model h , and the values of the parameters ν_j and $\beta_{j,n}$ so that the estimator (2.9) achieves optimal pointwise convergence rate over a variety of Besov spaces. This is the purpose of the next section.

3 Pointwise optimality

The objective of this paper is to formulate conditions under which the estimator \hat{f} of f , given in (2.9), is pointwise optimal, in the following sense.

3.1 Optimal pointwise convergence rate over Besov spaces

For any possible estimator \tilde{f} of f based on n observations from model (2.1), define the maximal pointwise risk, with respect to the l^2 -loss function, over a function space \mathcal{F} defined on the unit interval $[0, 1]$, as

$$R_n(t_0, \mathcal{F}, \tilde{f}) = \sup_{f \in \mathcal{F}} \mathbb{E} \left(\tilde{f}(t_0) - f(t_0) \right)^2,$$

for any fixed $t_0 \in (0, 1)$. Using the convexity of a Besov ball for $1 \leq p, q \leq \infty$, Lemma 3 in Donoho & Low (1992), and the optimal pointwise convergence rates obtained by Cai (1993) in the Gaussian white noise model, it easily follows that when the Z_i 's in model (2.1) are *iid* normal random variables with $\mathbb{E}(Z_1) = 0$ and $\mathbb{V}(Z_1) = \sigma^2 < \infty$, and when f belongs to a ball $B_{p,q}^r(A)$ of radius $A > 0$ in the Besov space $B_{p,q}^r[0, 1]$, then, provided that $r > 1/p$ and $1 \leq p, q \leq \infty$,

$$\inf_{\tilde{f}} R_n(t_0, B_{p,q}^r(A), \tilde{f}) = O \left(n^{-\frac{2(r-1/p)}{2(r-1/p)+1}} \right) \quad \text{as } n \rightarrow \infty, \quad (3.1)$$

where the infimum is taken over all estimators \tilde{f} of f .

Unlike the global maximal risk with respect to the L^2 -loss function (see Donoho & Johnstone, 1998), the pointwise maximal risk with respect to the l^2 -loss function depends not only on the smoothness index r , but also on the parameter p . Moreover, it converges at a rate slower than the corresponding global rate.

Note that since the normal distribution is a particular case of the distribution of the errors ε_{jk} , the lower bounds in our situation cannot be smaller than (3.1). On the other hand, since for the majority of resolution levels ($j \leq J_0$ where $J - J_0 \rightarrow \infty$ as $n \rightarrow \infty$) the errors ε_{jk} asymptotically follow the normal distribution, we can expect to achieve the optimal pointwise convergence rate (3.1) as $n \rightarrow \infty$ for the considered choices of error φ_j and prior h distributions.

3.2 Assumptions

Now we formulate conditions on the wavelet system $\{\phi, \psi\}$ and the *pdf*'s h and φ_j , as well as on the parameters ν_j and $\beta_{j,n}$.

- (S1) Let ϕ and ψ be the boundary coiflets introduced in Johnstone & Silverman (2004), possessing s continuous derivatives, $s - 1$ vanishing moments, $s \geq 2$, and based on orthonormal coiflets supported in $[-S + 1, S]$, $s < S$. Let also $L \geq \log_2(6S - 6)$.

The distributions h and φ_j considered here are symmetric, positive and unimodal on \mathbb{R} , have uniformly bounded moments of every (polynomial) order. In addition, we consider only those combinations of the distributions which satisfy the following condition

$$(A1) \quad |\varphi_j(x)/h(x)| \leq C_{h,\varphi}.$$

Note that constant $C_{h,\varphi}$ is assumed to be independent of j which requires some kind of uniformity for the *pdf*'s φ_j . The consequence of this restriction is that the asymptotic expressions for the thresholds $t_{j,n}$ will depend on the resolution level j rather than on the particular form of φ_j .

In the subsequent development, we consider the following combinations of error φ_j and prior h distributions:

$$\text{normal } \varphi_j \quad - \quad \text{normal } h, \tag{3.2}$$

$$\text{normal } \varphi_j \quad - \quad \text{double-exponential } h, \tag{3.3}$$

$$\text{double-exponential } \varphi_j \quad - \quad \text{double-exponential } h. \tag{3.4}$$

We do not consider the case double-exponential φ_j - normal h , since assumption (A1) does not hold in this case.

Denote

$$j_1 = \frac{1}{2(r - 1/p + 1/2)} \log_2(n). \tag{3.5}$$

We assume that the parameter ν_j is of the form

$$\nu_j = C_\nu 2^{jm_{(j)}}, \quad \text{where} \quad m_{(j)} = \begin{cases} m_1, & L \leq j \leq j_1, \\ m_2, & j_1 < j \leq J - 1, \end{cases} \tag{3.6}$$

and choose $\beta_{j,n}$ such that

$$\beta_{j,n} = (\nu_j/\sqrt{n})^{a_{(j)}}, \quad \text{where} \quad a_{(j)} = \begin{cases} a_1, & L \leq j \leq j_1, \\ a_2, & j_1 < j \leq J - 1. \end{cases} \tag{3.7}$$

(Note that we allow both hyperparameters $m_{(j)}$ and $a_{(j)}$ to vary with resolution level j .) We refer to $L \leq j \leq j_1$ and $j_1 < j \leq J - 1$ as *low* and *high* resolution levels, respectively.

The considered Bayesian model does not include the Bayesian model of Abramovich, Angelini & De Canditiis (2007) as a particular case. This is due to the dependence of the parameter $\pi_{j,n}$ on the sample size n simultaneously with the dependence on the resolution level j in the model we considered in our development: if $\pi_{j,n}$ is independent of n (i.e., $a_{(j)} = 0$ for all j), this means that it is also independent of j (see definition (3.7)). However, unlike Abramovich, Angelini & De Canditiis (2007), we allow for different behaviour of the hyperparameters at low and high resolution levels. As we shall see below, under some restrictions on our model, the considered Bayes factor

estimators achieve pointwise optimality without the extra logarithmic factor appeared in the results of Abramovich, Angelini & De Canditiis (2007).

The choices of error and prior models given in (3.2), (3.3) and (3.4) are motivated by the repeated use of these distributions in some practical applications, as well as the asymptotic behaviour of the risk when the pointwise convergence rate is not optimal. For example, as shown in Pensky & Sapatinas (2005) who studied convergence rates of Bayes factor estimators with respect to the L^2 -loss function, for distributions with exponential descents the deviation from the optimal behaviour is a factor which grows as a power of the logarithm of the sample size, whereas in the case of the distributions with polynomial descents the deviation is much larger, with a factor being a power of the sample size. Note also that, when the prior model h has faster descent at $\pm\infty$ than φ_j (i.e., when the assumption (A1) does not hold), sub-optimal convergence rates arise with respect to the L^2 -loss function due to the slow convergence of the bias when, e.g., the posterior mean is used as an estimator (see Pensky, 2006). Since we expect to see these types of behaviour for the convergence rates of Bayes factor estimators with respect to the l^2 -loss function, in what follows, we restrict ourselves to study the pointwise optimality of Bayes factor estimators *only* for the combination of error and prior models given in (3.2), (3.3) and (3.4).

3.3 Pointwise optimality of Bayes factor estimators in Besov spaces

The following theorem states under which conditions the considered Bayes factor estimators achieve the optimal pointwise convergence rate under the l^2 -loss function.

Theorem 1. *Let $\{\phi, \psi, s, L\}$ be as in assumption (S1), and let $f \in B_{p,q}^r(A)$ with $1 \leq p, q \leq \infty$, and $1/p < r < s$. Assume that the following restrictions hold for m_1 and m_2 :*

$$m_1 < r - 1/p + 1/2, \quad m_2 > r - 1/p + 1/2, \quad (3.8)$$

and that the following restrictions hold for a_1 , a_2 and φ_j :

- (1) $a_1 \geq 1$;
- (2) if φ_j is double-exponential then $a_2 > 0$;
- (3) if φ_j is normal then $a_2 > \frac{2(r-1/p)}{(r-1/p+1/2)(2m_2-1)}$.

Then, for any $t_0 \in (0, 1)$,

$$R_n(t_0, B_{p,q}^r(A), \hat{f}) = O\left(n^{-\frac{2(r-1/p)}{2(r-1/p)+1}}\right), \quad \text{as } n \rightarrow \infty.$$

Remark 1. The assumption on the hyperparameter a_2 for the double-exponential φ_j is weaker than the corresponding one for the normal φ_j .

Remark 2. For values of a_1 or a_2 violating the assumptions of Theorem 1, the pointwise rate of convergence is no longer the *exact* optimal rate. Following the arguments in the proof of Theorem 1, one can show that, for any $t_0 \in (0, 1)$, $R_n(t_0, B_{p,q}^r(A), \tilde{f}) = O\left(n^{-\frac{2(r-1/p)}{2(r-1/p)+1}} (\log n)^\Delta\right)$, with (i) $\Delta = 1/2$ if $a_1 \geq 1$ and $a_2 = \frac{2(r-1/p)}{(r-1/p+1/2)(2m_2-1)}$, (ii) $\Delta = 1$ if $a_1 < 1$ and φ_j is normal, and (iii) $\Delta = 2$ if $a_1 < 1$ and φ_j is double-exponential.

As it is evident from Theorem 1, under some restrictions on our model, the considered Bayes factor estimators achieve pointwise optimality without the extra logarithmic factor appeared in the results of Abramovich, Angelini & De Canditiis (2007). This result is due to the flexibility of our model with respect to allowing the hyperparameters $a_{(j)}$ and $m_{(j)}$ to be different for low and high resolution levels, and the dependence of the prior odds $\beta_{j,n}$ on the sample size n . As one can see from the proof of Theorem 1 (see Appendix), the most crucial assumption which allows to achieve pointwise optimality without a logarithmic factor is the separation between the low and high resolution levels at the “boundary” level j_1 defined by (3.5).

Remark 3. To make the prior model more flexible, we can divide the low resolution levels into low $\{L, \dots, j_0\}$ and medium $\{j_0 + 1, \dots, j_1\}$ levels, with $j_0 = \kappa \log_2(n)$ for arbitrary $\kappa \in (0, 1/[2(r + 1/2 - 1/p)])$, and consider different values of hyperparameters (a_0, m_0) and (a_1, m_1) . Then, to satisfy Theorem 1, hyperparameters (a_1, m_1) should yield the assumptions of the theorem, and the only additional restriction on the hyperparameters for the low resolution levels is $m_0 < (2\kappa)^{-1}$. Thus, the restrictions on a_1 and m_1 are crucial only for the levels adjacent to the “boundary” level j_1 rather than for all resolution levels coarser than j_1 .

4 Numerical Results

In this section, we illustrate performance of the proposed Bayes factor wavelet estimation procedure (BF) and compare it with the wavelet estimation procedure (PR) proposed in Cai (2003). The PR estimator is a minimax (projection) estimator, constructed by setting all wavelet coefficients above level j_1 to zero. The hyperparameters involved in the BF estimation procedure were estimated in a fully Bayesian framework using the WinBUGS software (see Spiegelhalter, Thomas & Best, 1999) that is freely available from <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>. The wavelet transform was performed using R and the package WAVETHRESH that are freely available from www.r-project.org.

4.1 Simulation Study

Now we present results of the simulation study, with the remainder of this section devoted to the discussion of these results. For the BF estimator, double-exponential prior and normal error were used. The hyperparameters were estimated in a fully Bayesian framework as follows. We used the

non-informative uniform prior for π_j and a weakly informative exponential prior for ν_j , i.e.,

$$\begin{aligned}\pi_j &\sim \text{Uniform}[0, 1], \\ \nu_j &\sim \text{Exponential}(1),\end{aligned}$$

and we used a non-informative scale-invariant prior on the precision parameter σ^{-2} with density $f(x) = 1/x$. To estimate the parameters in the hierarchical model above using the `WinBUGS` software, the improper prior for σ^{-2} was approximated by a proper prior Gamma distribution with *pdf* proportional to $x^{0.001-1}e^{-0.001x} \approx x^{-1}$. Two chains were run for 80000 iterations in each, and the last 50000 thinned by 5 were used for estimating the posterior distributions. For the PR estimator, a range of projection levels j_1 between 4 and 7 was used.

In this simulation study, we evaluated the performance of the BF and PR estimators using Daubechies's compactly supported *ExtremePhase 2* (see Daubechies, 1992, p. 196) and *Coiflet 2* (see Daubechies, 1992, p. 258) wavelet filters, and primary resolution levels $L = 2, 3$ and 4. We have considered three different kind of test functions, defined on the unit interval (representing different types of situations)

- (1) A function that is discontinuous with two jumps

$$f(t) = 4 \sin(4\pi t) - \text{sgn}(t - 0.3) - \text{sign}(0.72 - t), \quad t \in [0, 1],$$

where $\text{sgn}(\cdot)$ is the signum function, named `HeaviSine` (see, e.g., Donoho & Johnstone, 1994);

- (2) A function that is continuous but has a discontinuity in the first derivative

$$f(t) = \exp(-|t - 1/2|), \quad t \in [0, 1],$$

named `Laplace` (see, e.g., Angelini, De Canditiis & Leblanc, 2003);

- (3) A function that has continuous first derivative but there are big jumps in the second derivative

$$\begin{aligned}f(t) = & 0.8 - 30r(t, 0.1) + 60r(t, 0.2) - 30r(t, 0.3) + \\ & 500r(t, 0.35) - 1000r(t, 0.37) + 1000r(t, 0.41) - 500r(t, 0.43) + \\ & 7.5r(t, 0.5) - 15r(t, 0.7) + 7.5r(t, 0.9), \quad t \in [0, 1],\end{aligned}$$

where $r(t, c) = (t - c)^2 \mathbb{I}_{(c, 1]}(x)$, named `Parabolas` (see, e.g., Antoniadis, Bigot & Sapatinas, 2001).

For each test function, $M = 100$ samples were generated by adding independent random noise $\varepsilon \sim N(0, \sigma^2)$ to $n = 256, 512$ and 1024 equally spaced points on $[0, 1]$ (representing a range of sample sizes). The value of σ was taken to correspond to the values $\sqrt{1.5}, \sqrt{3}$ and $\sqrt{5}$ (representing various noise levels) for the (root) signal-to-noise ratio (SNR)

$$\text{SNR}(f, \sigma) = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (f(t_i) - \bar{f})^2}}{\sigma}, \quad \text{where} \quad \bar{f} = \frac{1}{n} \sum_{i=1}^n f(t_i).$$

The goodness-of-fit for an estimator \hat{f} of f was measured by

- (a) its mean squared error (MSE) at point $t_0 \in (0, 1)$ from the M simulations, defined as

$$\text{MSE}(f, t_0) = \frac{1}{M} \sum_{m=1}^M (\hat{f}_m(t_0) - f(t_0))^2;$$

- (b) its mean absolute error (MAE) at point $t_0 \in (0, 1)$ from the M simulations, defined as

$$\text{MAE}(f, t_0) = \frac{1}{M} \sum_{m=1}^M |\hat{f}_m(t_0) - f(t_0)|.$$

For brevity, we only report in detail the results for the `Laplace` function, based on the $\text{MSE}(f, t_0)$ criterion (at point $t_0 = 0.5$) using $n = 1024$, $L = 2$ and *ExtremePhase 2* wavelet filter, for data with $\text{SNR} = \sqrt{1.5}$. For the PR estimator, the projection level was set equal to $j_1 = 5$. Different combinations of test functions, goodness-of-fit measures, sample sizes, primary resolution levels, wavelet filters, projection levels, and SNR values yield similar results in magnitude. Figure 1 contains results of the simulation study. As observed in this figure, the BF estimator, overall, outperforms the PR estimator. In Figure 1, we can see that although at some points the BF estimator has higher MSE compared to the PR estimator (outliers on the boxplot), for the vast majority of the points the error of the PR estimator is consistently higher than that for the BF estimator, including the point of discontinuity of the derivative $t_0 = 0.5$. Quantitatively, the BF estimator produced estimates with smaller MSE at point $t_0 = 0.5$ than the PR estimator: $\text{MSE}(f, 0.5) = 0.00015$ for the BF estimator and $\text{MSE}(f, 0.5) = 0.00367$ for the PR estimator. Similar behaviour is observed for the average value of MSE over all points (0.00016 for the BF estimator and 0.00202 for the PR estimator). Moreover, the BF estimator is less wiggly and it preserves the peak height better than the PR estimator.

4.2 Inductance plethysmography data

We now consider a dataset from anaesthesiology collected by inductance plethysmography to illustrate the performance of the proposed BF estimator and compare it with the PR estimator, discussed in Section 4.1. The recordings were made by the Department of Anaesthesia at the Bristol Royal Infirmary and measure the flow of air during breathing. For more details we refer to Nason (1996).

Figure 2 shows a section of plethysmograph recording lasting approximately 80 seconds ($n = 4096$ signal points). The two main sets of regular oscillations correspond to normal breathing. The disturbed behaviour in the center of the plot, where the normal breathing pattern disappears, corresponds to the patient vomiting. In the figure, there is also a zoom in for the rapid variation near point 0.85 (on the x -axis).

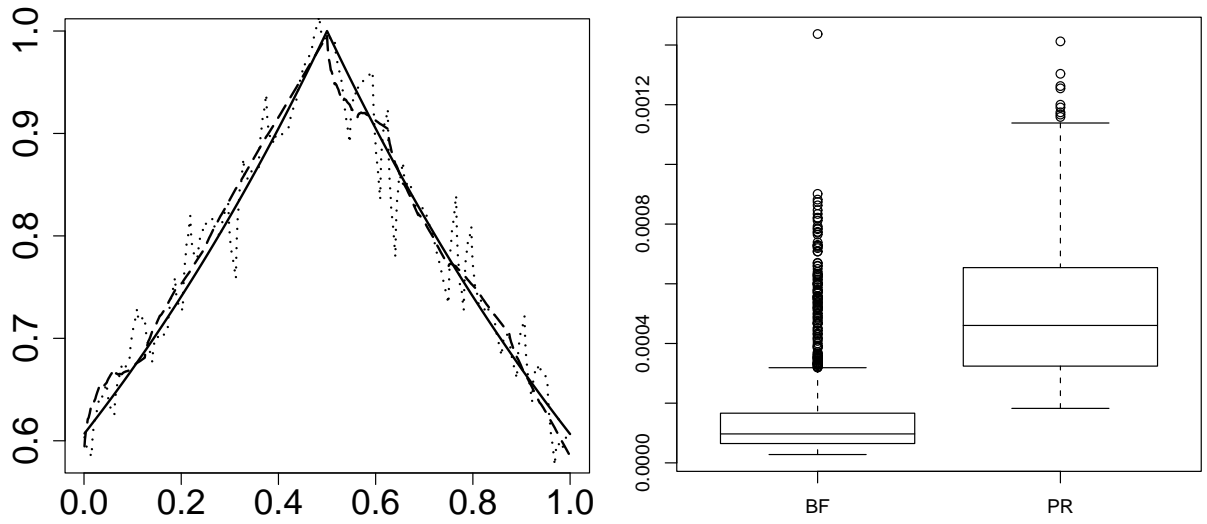


Figure 1: (Left) The Laplace function (solid line) sampled at $n = 1024$ equally spaced points on $[0, 1]$, and the (BF) (dashed line) and PR (dotted line) reconstructions from one simulation with $\text{SNR} = \sqrt{1.5}$ using the *ExtremePhase 2* wavelet filter. (Right) Boxplots of 100 simulation results of the BF and PR estimators for the Laplace function at point $t_0 = 0.5$. See Section 4.1 for more details.

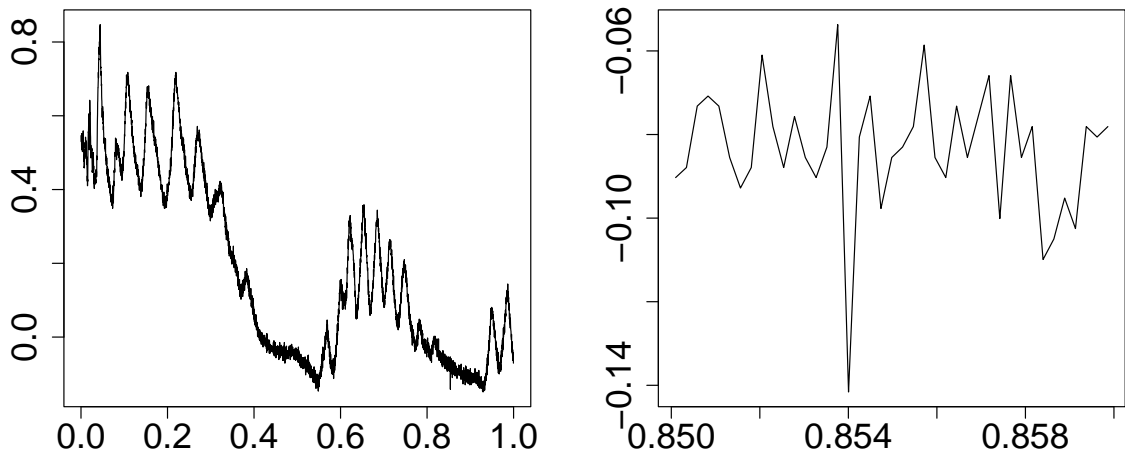


Figure 2: (Left) Section of an inductance plethysmography recording lasting approximately 80 seconds. (Right) Zoom in for the rapid variation near point 0.85. See Section 4.2 for more details.

The posterior distributions of the parameters were obtained using the same hierarchical Bayesian model as described in Section 4.1 from two chains, with the last 50000 iterations thinned by 5 (out of 80000 iterations) in each chain. Plots of the values of the chains for parameters σ^2 and $\pi_{j,n}$ are given in Figure 5, as well as the posterior mean estimates of $\pi_{j,n}$ with 95% credible intervals. Gelman-Rubin statistic modified by Brooks & Gelman (1998) confirms the visual assessment for the convergence of the MCMC. The posterior mean estimate of the variance σ^2 is $\hat{\sigma}^2 = 0.00013$.

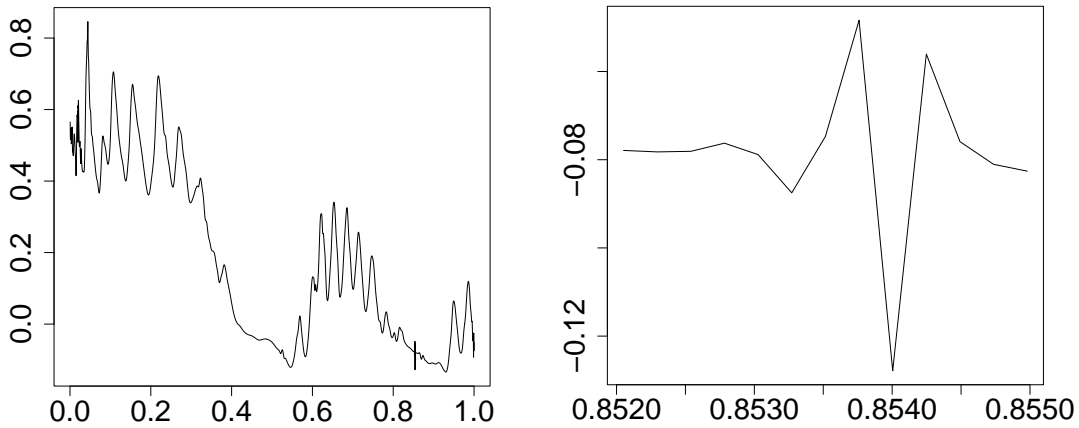


Figure 3: (Left) Smooth estimate obtained using the BF estimator. (Right) Zoom in for the rapid variation near point 0.85 for the panel shown in Figure 2. See Section 4.2 for more details.

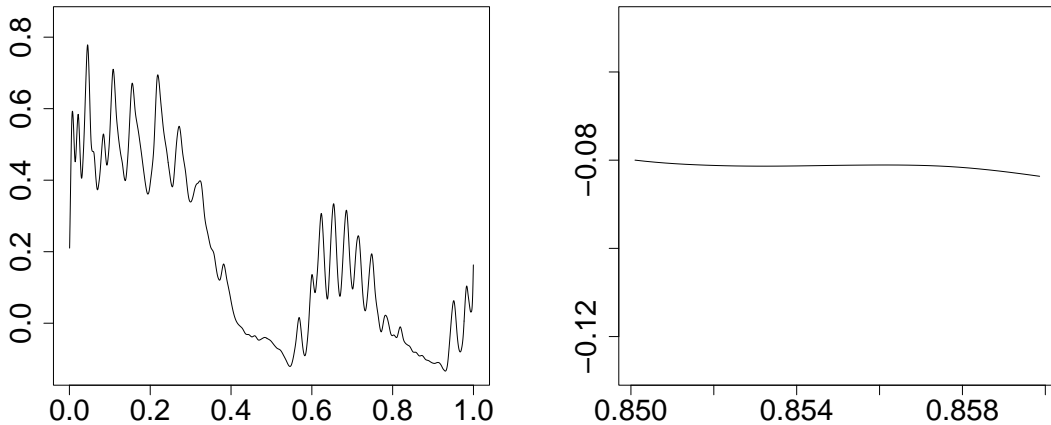


Figure 4: (Left) Smooth estimate obtained using the PR estimator. (Right) Zoom in for the rapid variation near point 0.85 for the panel shown in Figure 2. See Section 4.2 for more details.

Figures 3 and 4 contain the curve estimates obtained using the BF and PR estimators respectively. The various estimators were evaluated using Daubechies's compactly supported *Symmetlet 8* (see Daubechies, 1992, p. 198) and *Coiflet 2* (see Daubechies, 1992, p. 258) wavelet filters. For both methods, the primary resolution level was set equal to $L = 3$, and for the PR estimator the projection level was set equal to $j_1 = 6$. This corresponds to the assumption that the underlying function of interest belongs to Besov space $B_{p,q}^r$ with $r - 1/p = 1/2$.

Smoothing with data of this kind is to preserve features of interest, such as peak heights, as far as possible. At the same time, spurious rapid variations elsewhere should be eliminated. The

efficacy of the various estimation methods in preserving peak heights is most simply judged by the maximum of the various estimates, the height of the first peak in the inductance plethysmography curve. For *Symmlet 8*, the BF estimator yield the maximum value of 0.846, while the PR estimator gave 0.777. The efficacy of the various estimation methods in dealing with the rapid variation near the point 0.85 (on the x -axis) can be quantified by the range of the estimated curves over a small interval at this point. The BF estimator has a ‘glitch’ of range 0.079, while the corresponding one for the PR estimator is almost 0 (see Figures 3 and 4). Note that the results for the PR estimator still remain the same even if the projection level was set equal to $j_1 = 9$.

Similar results in magnitude hold for *Coiflet 2*. The BF estimators yield the maximum value of 0.835, while the PR estimator gave 0.793. For the rapid variation near the point 0.85 (on the x -axis), the BF estimator has a ‘glitch’ of range 0.075, while the corresponding one for the PR estimator is almost 0. Although we do not reproduce them here, similar results in magnitude are also true for both estimators by increasing or decreasing the value of the primary resolution level L . (Note that all the above numbers were rounded to three decimal places.)

In summary, the BF estimator outperforms the PR estimator on preserving the peak height without any substantial cost of inferior treatment of presumably spurious variation elsewhere.

5 Conclusions

We investigated theoretical performance of Bayes factor estimators at a single point in wavelet regression models with independent and identically distributed errors that are not necessarily normally distributed. We compared these estimators in terms of their frequentist pointwise optimality (in the minimax sense) in Besov spaces for some combinations of error and prior distributions. The characteristic of the Bayes factor estimator is that it leads to a *hard* thresholding rule, unlike the recently studied posterior mean and posterior median estimators which lead to nonlinear *shrinkage* and *soft* thresholding rules respectively. Moreover, the Bayes factor estimator is much easier to evaluate in the majority of cases unlike posterior mean or posterior median estimators.

We extended the normality assumption about the distribution of errors in the standard nonparametric regression model, to include the double-exponential distribution of errors in the wavelet domain. Furthermore, we showed that optimality can be achieved for different error distributions implying that it is not necessary to know the exact error distribution to achieve pointwise optimality. Moreover, as we demonstrated, the use of a more flexible Bayesian hierarchical model improved the pointwise convergence rate and, under certain conditions, achieved pointwise optimality without the extra logarithmic factor appeared in the results of Abramovich, Angelini & De Canditiis (2007).

As it was illustrated in simulated examples as well as a real-life dataset, the proposed Bayes

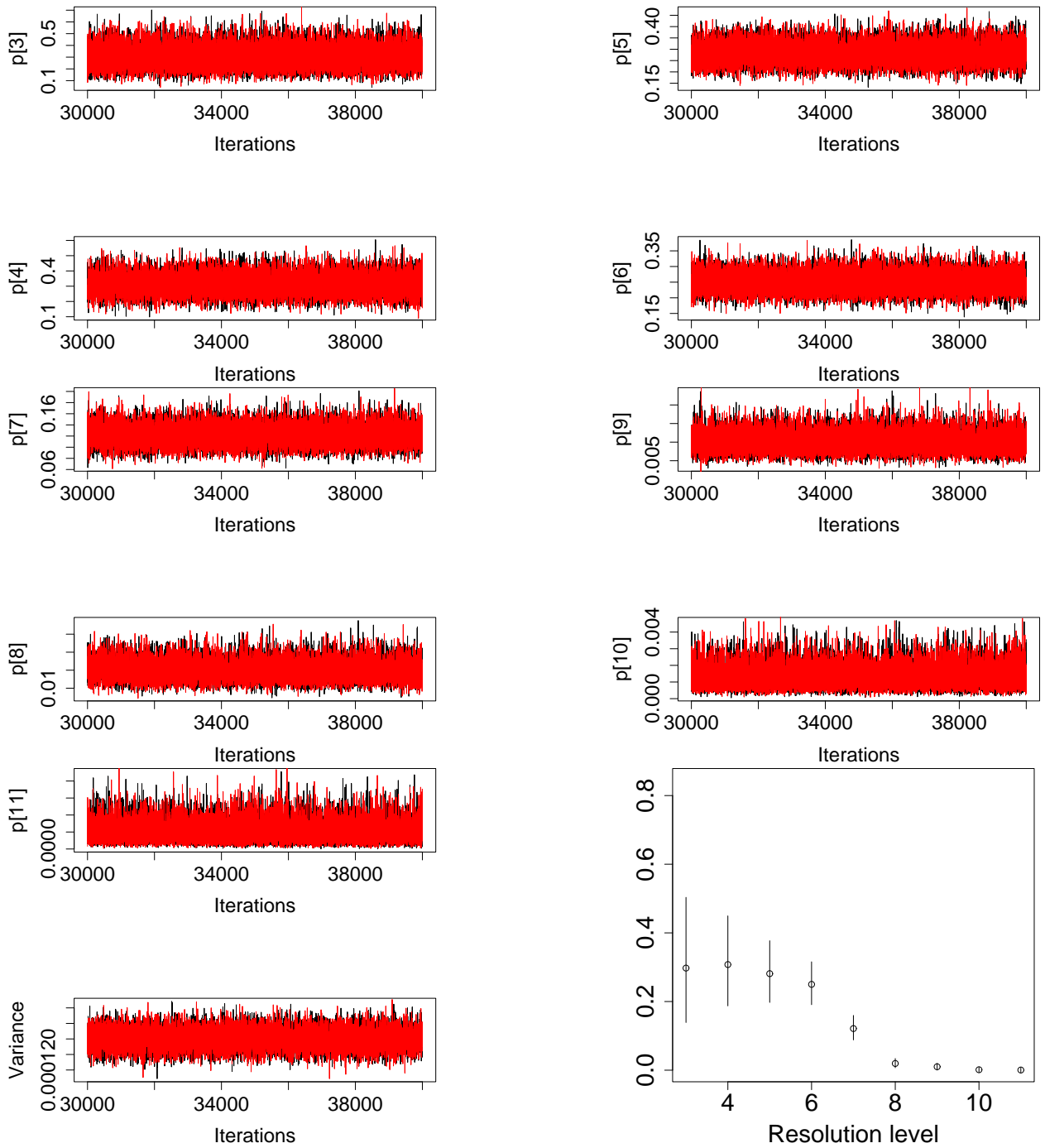


Figure 5: Values of $\pi_{j,n}$ and σ^2 sampled from posterior distribution for the plethysmography data, using the *Symmlet 8* wavelet filter (the two chains are superimposed). See Section 4.2 for more details.

factor estimator with hyperparameters estimated in a fully Bayesian approach preserved peak heights better, without any substantial cost of inferior treatment of presumably spurious variation elsewhere, and outperformed a recently proposed minimax (projection) wavelet estimator.

We conclude this section with some comments on adaptation. Adaptive estimation has become an important part of nonparametric function estimation problems. Adaptation to unknown smoothness is essential because the smoothness parameters of the underlying functions are unknown in virtually all practical situations. In the Gaussian white noise model, Cai (2003) considered adaptation under pointwise risk over Besov spaces; sharp lower bounds on the cost of adaptation were obtained (the minimum cost for adaptation is at least a logarithmic factor) and are shown to be attainable by a (soft-thresholding) wavelet estimator. We have not considered adaptation in our nonparametric regression setup, i.e., to construct a Bayes factor estimator without the knowledge of the parameters of the Besov ball, attaining the adaptive optimal pointwise convergence rate. This is beyond the scope of this article but presents avenues for further research that hopefully will be addressed in the future.

Acknowledgements

Natalia Bochkina would like to thank Theofanis Sapatinas for financial support and warm hospitality while visiting Nicosia to carry out part of this work. The authors would like to thank Marianna Pensky for fruitful discussions. The authors are grateful to Professor Arup Bose (Editor), a Co-Editor, and a referee whose valuable comments and suggestions led to an improvement of this paper.

6 Appendix

Throughout the proof of Theorem 1, we use C to denote a generic positive constant, not necessarily the same each time it is used, even within a single equation. (Auxiliary lemmas with proofs are given in the following sections.)

6.1 Proof of Theorem 1.

Since the wavelet basis is orthonormal, for any fixed $t_0 \in (0, 1)$,

$$\begin{aligned} R_n(t_0, B_{p,q}^r(A), \hat{f}) &= \mathbb{E}[\hat{f}(t_0) - f(t_0)]^2 \\ &= \mathbb{E} \left[\sum_{k \in K_{L-1}} (\hat{\theta}_k - \tilde{\theta}_k) \phi_{Lk}(t_0) + \sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} (\hat{\theta}_{jk} - \tilde{\theta}_{jk}) \psi_{jk}(t_0) \right]^2. \end{aligned}$$

Now, we can decompose the risk above into the following terms:

$$\begin{aligned}
R_n(t_0, B_{p,q}^r(A), \hat{f}) &= \mathbb{E} \left[\sum_{k \in K_{L-1}} (\hat{\theta}_k - \theta_k) \phi_{Lk}(t_0) + \sum_{k \in K_{L-1}} (\theta_k - \tilde{\theta}_k) \phi_{Lk}(t_0) \right. \\
&+ \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} (\hat{\theta}_{jk} - \theta_{jk}) \psi_{jk}(t_0) + \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} (\theta_{jk} - \tilde{\theta}_{jk}) \psi_{jk}(t_0) \\
&\left. + \sum_{j=J}^{\infty} \sum_{k=0}^{2^j-1} \tilde{\theta}_{jk} \psi_{jk}(t_0) \right]^2.
\end{aligned}$$

Now, we can apply the elementary inequality $\mathbb{E}(\sum_{i=1}^n X_i)^2 \leq [\sum_{i=1}^n (\mathbb{E}|X_i|^2)]^2$ to bound the risk:

$$\begin{aligned}
R_n(t_0, B_{p,q}^r(A), \hat{f}) &\leq \left[\sum_{k \in K_{L-1}(t_0)} 2^{L/2} [\mathbb{E}(\hat{\theta}_k - \theta_k)^2]^{1/2} \|\phi\|_{\infty} + \sum_{k \in K_{L-1}(t_0)} 2^{L/2} |\tilde{\theta}_k - \theta_k| \|\phi\|_{\infty} \right. \\
&+ \sum_{j=L}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} [\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2]^{1/2} \|\psi\|_{\infty} + \sum_{j=L}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} |\tilde{\theta}_{jk} - \theta_{jk}| \|\psi\|_{\infty} \\
&\left. + \sum_{j=J}^{\infty} \sum_{k \in K_j(t_0)} 2^{j/2} |\tilde{\theta}_{jk}| \|\psi\|_{\infty} \right]^2 \\
&= [Q_{11} + Q_{12} + Q_{21} + Q_{22} + Q_3]^2, \tag{6.1}
\end{aligned}$$

where, for any function g , $\|g\|_{\infty} = \sup_x |g(x)|$ and $K_j(t_0) = \{k : 0 \leq k \leq 2^j - 1 \text{ and } \psi_{jk}(t_0) \neq 0\}$, for $j \geq J$, with $K_{L-1}(t_0) = \{k : k \in K_{L-1} \text{ and } \phi_{Lk}(t_0) \neq 0\}$. For the boundary coefficients stated in assumption (S1), the cardinality of $K_j(t_0)$ is less or equal to $2S - 1$, $j \geq L - 1$, which is independent of j (see Johnstone & Silverman, 2004). Note also that, by construction, ϕ and ψ are bounded functions, i.e., $\phi, \psi \in L^{\infty}[0, 1]$.

Term $Q_{11} + Q_{12}$ in (6.1) is bounded by

$$\begin{aligned}
C \sum_{k \in K_{L-1}} 2^{L/2} [\mathbb{V}(\hat{\theta}_k)]^{1/2} + C \sum_{k \in K_{L-1}} 2^{L/2} |\tilde{\theta}_k - \theta_k| &\leq C n^{-1/2} \sigma_{L-1} + C n^{-r} \\
&= O(n^{-1/2}) + o(n^{-(r-1/p)}) = o\left(n^{-\frac{(r-1/p)}{2(r-1/p)+1}}\right),
\end{aligned}$$

due to (6.15) and the fact that $\mathbb{V}(\hat{\theta}_k) = O(n^{-1})$.

On the other hand, term Q_3 in (6.1) is bounded by

$$\begin{aligned}
C \sum_{j=J}^{\infty} \sum_{k \in K_j(t_0)} 2^{j/2} |\tilde{\theta}_{jk}| &\leq C \sum_{j=J}^{\infty} 2^{j/2} 2^{-j(r-1/p+1/2)} = O(2^{-J(r-1/p)}) \\
&= O(n^{-(r-1/p)}) = o\left(n^{-\frac{(r-1/p)}{2(r-1/p)+1}}\right),
\end{aligned}$$

due to (6.17). By Lemma 4, term Q_{22} in (6.1) is dominated by $C n^{-(r-1/p)}$. Therefore, now we need to evaluate the contribution to $R_n(t_0, B_{p,q}^r(A), \hat{f})$ made by term Q_{21} in (6.1):

$$Q_{21} = \|\psi\|_\infty \sum_{j=L}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} [\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2]^{1/2} =: \|\psi\|_\infty (R_1 + R_2), \quad (6.2)$$

with terms

$$\begin{aligned} R_1 &= \sum_{j=L}^{j_1} \sum_{k \in K_j(t_0)} 2^{j/2} [\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2]^{1/2}, \\ R_2 &= \sum_{j=j_1+1}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} [\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2]^{1/2}, \end{aligned} \quad (6.3)$$

corresponding to *low* and *high* resolution levels, respectively. Let us now construct an asymptotic upper bound for each of the terms.

Low resolution levels. We can use Lemma 3 to bound R_1 from above:

$$R_1 \leq \sqrt{2} \sum_{j=L}^{j_1} \sum_{k \in K_j(t_0)} 2^{j/2} \min(t_{j,n}, |\theta_{jk}|) + O(2^{j_1/2} n^{-1/2}), \quad (6.4)$$

since $\kappa_{2,j} = \int_{-\infty}^{+\infty} x^2 \varphi_j(x) dx = c_j \sigma_j^3 \int_{-\infty}^{+\infty} z^2 e^{-|z|^\beta} dz \leq C$, due to σ_j being bounded (see (2.3)). The last term achieves the optimal pointwise convergence rate: $n^{-1/2} 2^{j_1/2} = n^{-(r-1/p)/(2(r-1/p)+1)}$, so we need to study the behaviour of the first term.

To bound $t_{j,n}$ from above, we apply the last statement of Lemma 2. To use Lemma 2, we need to check the assumption that $\nu_j/\sqrt{n} \rightarrow 0$, as $n \rightarrow \infty$. Note that, $\nu_j/\sqrt{n} = C_\nu 2^{jm_1} n^{-1/2} \leq C_\nu 2^{j_1 m_1} n^{-1/2} = C_\nu n^{(m_1/(r-1/p+1/2)-1)/2} \rightarrow 0$ as $n \rightarrow \infty$, since $j \leq j_1$ and, according to assumption (3.8), $m_1/(r+1/2-1/p) - 1 < 0$. Therefore, the assumption of the last statement of Lemma 2 is satisfied for the low resolution levels.

According to Lemma 2, for the low resolution levels with $a_1 \geq 1$, the threshold $t_{j,n}$ is bounded by $C_1 n^{-1/2}$, therefore the first term of R_1 is bounded by

$$\begin{aligned} \sqrt{2} \sum_{j=L}^{j_1} \sum_{k \in K_j(t_0)} 2^{j/2} \min(t_{j,n}, |\theta_{jk}|) &\leq C \sum_{j=L}^{j_1} \min\left(2^{j/2} n^{-1/2}, 2^{j/2} 2^{-j(r+1/2-1/p)}\right) \\ &= O\left(\min\left(n^{-1/2} 2^{j_1/2}, 2^{-L(r-1/p)}\right)\right) = O\left(n^{-(r-1/p)/(2(r-1/p)+1)}\right). \end{aligned}$$

High resolution levels. For high resolution levels, first note that

$$\begin{aligned} \mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2 &= \mathbb{E}(d_{jk} - \theta_{jk})^2 \mathbb{I}(|d_{jk}| > t_{j,n}) + \theta_{jk}^2 \mathbb{P}(|d_{jk}| \leq t_{j,n}) \\ &\leq \mathbb{E}(d_{jk} - \theta_{jk})^2 \mathbb{I}(|d_{jk}| > t_{j,n}) + \theta_{jk}^2. \end{aligned}$$

Therefore,

$$R_2 \leq \sum_{j=j_1+1}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} |\theta_{jk}| + \sum_{j=j_1+1}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} [\mathbb{E}(d_{jk} - \theta_{jk})^2 \mathbb{I}(|d_{jk}| > t_{j,n})]^{1/2}.$$

According to Lemma 4, the first term is bounded by

$$C \sum_{j=j_1+1}^{J-1} 2^{j/2} 2^{-j(r-1/p+1/2)} = C \sum_{j=j_1+1}^{J-1} 2^{-j(r-1/p)} = O\left(n^{-(r-1/p)/(2(r-1/p)+1)}\right).$$

Consider now the second term separately for normal and double-exponential *pdf*'s φ_j . Note that at high resolutions levels $j_1 + 1 \leq j \leq J - 1$, where $\nu_j/\sqrt{n} \rightarrow \infty$ due to assumption (3.8), we get

$$\nu_j/\sqrt{n} = C_\nu 2^{jm_2} n^{-1/2} > C_\nu 2^{j_1 m_2} n^{-1/2} = C_\nu n^{m_2/2(r-1/p+1/2)-1/2} \rightarrow \infty.$$

If φ_j is the double-exponential *pdf* and $a_2 > 0$, then $\beta_{j,n} = \left(\frac{\nu_j}{\sqrt{n}}\right)^{a_2} \rightarrow \infty$, and thus, by Lemma 1, we have

$$\zeta_{j,n}(x) = \frac{\int_{-\infty}^{+\infty} \sqrt{n} \varphi_j(\sqrt{n}(x-y)) \nu_j h(\nu_j y) dy}{\sqrt{n} \varphi_j(\sqrt{n}x)} = \frac{\sqrt{n} \varphi_j(\sqrt{n}x)(1+o(1))}{\sqrt{n} \varphi_j(\sqrt{n}x)} = 1+o(1) < \beta_{j,n},$$

implying that $\mathbb{I}(|d_{jk}| > t_{j,n}) = \mathbb{I}(\zeta_{j,n}(d_{jk}) > \beta_{j,n}) = 0$. Hence, in this case, the second term in the upper bound for R_2 is zero, and the Bayes factor estimator achieves the optimal pointwise rate of convergence.

If now φ_j is the normal *pdf*, then

$$\begin{aligned} \mathbb{E}(d_{jk} - \theta_{jk})^2 \mathbb{I}(|d_{jk}| > t_{j,n}) &= \sqrt{n} \int_{|x| > t_{j,n}} (x - \theta_{jk})^2 \varphi_j(\sqrt{n}(x - \theta_{jk})) dx \\ &= n^{-1} \int_{|w + \sqrt{n}\theta_{jk}| > \sqrt{n}t_{j,n}} w^2 \varphi_j(w) dw. \end{aligned} \quad (6.5)$$

For $j \geq j_1 + 1$, by Lemma 4,

$$\sqrt{n}|\theta_{jk}| \leq C\sqrt{n}2^{-j(r-1/p+1/2)} \leq C\sqrt{n}2^{-j_1(r-1/p+1/2)} = O(1). \quad (6.6)$$

If $\sqrt{nt_{j,n}} \leq C$, then the integral above is a constant, implying that (6.5) is bounded by Cn^{-1} , and the corresponding sum is bounded by

$$Cn^{-1/2} \sum_{j=j_1+1}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} = O\left(n^{-1/2} 2^{J/2}\right) = O(1),$$

i.e., it does not tend to zero and thus it is slower than the optimal pointwise convergence rate. To achieve pointwise optimality, we consider only the cases where $\sqrt{nt_{j,n}} \rightarrow \infty$, i.e., such that $a_2 > 0$ which implies that

$$\sqrt{nt_{j,n}} \geq \sigma_j \sqrt{\log(\beta_{j,n})} = \sigma_j \sqrt{a_2 \log(\nu_j/\sqrt{n})} \rightarrow \infty,$$

since $\nu_j/\sqrt{n} \rightarrow \infty$ for $j_1 + 1 \leq j \leq J - 1$. For $\sqrt{nt_{j,n}} \rightarrow \infty$, $|\theta_{jk}|/t_{j,n} \rightarrow 0$ due to (6.6) for $j \geq j_1 + 1$, therefore we can write

$$n^{-1} \int_{|w + \sqrt{n}\theta_{jk}| > \sqrt{n}t_{j,n}} w^2 \varphi_j(w) dw \leq Cn^{-1} (t_{j,n} \sqrt{n})^3 \varphi_j(t_{j,n} \sqrt{n}),$$

which, according to the first statement of Lemma 2 and due to $x^3\varphi_j(x)$ decreasing for large positive x , is bounded by

$$\begin{aligned} Cn^{-1}(t_{j,n}\sqrt{n})^3\varphi_j(t_{j,n}\sqrt{n}) &\leq Cn^{-1}[\log(\beta_{j,n})]^{3/2}c_j\exp\left\{-\left[(\log(\beta_{j,n}))^{1/2}\right]^2\right\} \\ &= Cn^{-1}\left[a_2\log(C_\nu 2^{jm_2}/\sqrt{n})\right]^{3/2}\beta_{j,n}^{-1} \\ &\leq C(\log n)^{3/2}n^{-1+a_2/2}2^{-a_2m_2j}. \end{aligned}$$

Now, by substituting this bound into the sum R_2 , we get

$$\begin{aligned} R_2 &\leq O\left(n^{-(r-1/p)/(2(r-1/p)+1)}\right) + Cn^{-1/2+a_2/4}(\log n)^{3/4}\sum_{j=j_1+1}^{J-1}2^{j(1-m_2a_2)/2} \\ &= O\left(n^{-(r-1/p)/(2(r-1/p)+1)}\right) + \begin{cases} C(\log n)^{3/4}n^{-1/2+a_2/4+\frac{1-m_2a_2}{4(r-1/p+1/2)}}, & 1-m_2a_2 < 0, \\ C(\log n)^{1+3/4}n^{-1/2+a_2/4}, & 1-m_2a_2 = 0, \\ C(\log n)^{3/4}n^{a_2/4-m_2a_2/2}, & 1-m_2a_2 > 0, \end{cases} \\ &= O\left(n^{-(r-1/p)/(2(r-1/p)+1)}\right) + \begin{cases} C(\log n)^{3/4}n^{-1/2+\frac{1}{4(r-1/p)+2}+\frac{a_2}{4}\left(1-\frac{m_2}{(r-1/p+1/2)}\right)}, & a_2 > 1/m_2, \\ C(\log n)^{1+3/4}n^{-1/2+1/4m_2}, & a_2 = 1/m_2, \\ C(\log n)^{3/4}n^{a_2(1/2-m_2)/2}, & a_2 < 1/m_2. \end{cases} \end{aligned}$$

For $a_2 > 1/m_2$, $1 - m_2/(r - 1/p + 1/2)$ is negative since $m_2 > r - 1/p + 1/2$ by assumption (3.8), and thus the rate $O(n^{-(r-1/p)/(2(r-1/p)+1)})$ is achieved. For $a_2 = 1/m_2$, the convergence rate is also faster than the rate $O(n^{-(r-1/p)/(2(r-1/p)+1)})$, since $-1 + 1/2m_2 < -1 + 1/(2(r - 1/p) + 1)$.

For $a_2 < 1/m_2$, the convergence rate is faster than the rate $O(n^{-(r-1/p)/(2(r-1/p)+1)})$ if $a_2(1/2 - m_2) < -1 + 1/[2(r - 1/p) + 1]$, i.e., if $a_2 > (1 - 1/(2(r - 1/p) + 1))/(m_2 - 1/2)$ since $1/2 - m_2 < 1/2 - (r - 1/p + 1/2) = -(r - 1/p) < 0$. These conditions on a_2 are compatible if and only if $(1 - 1/(2(r - 1/p) + 1))/(m_2 - 1/2) < 1/m_2$, which holds under the assumptions of Theorem 1, since

$$\begin{aligned} \frac{1 - 1/(2(r - 1/p) + 1)}{m_2 - 1/2} - \frac{1}{m_2} &= \frac{m_2 - m_2/(2(r - 1/p) + 1) - m_2 + 1/2}{m_2(m_2 - 1/2)} \\ &= \frac{r - 1/p + 1/2 - m_2}{m_2(2m_2 - 1)(r - 1/p + 1/2)} < 0. \end{aligned}$$

Therefore, combining all the cases, we have that R_2 achieves the rate $O(n^{-(r-1/p)/(2(r-1/p)+1)})$ if $a_2 > \frac{r-1/p}{(m_2-1/2)(r-1/p+1/2)}$.

Combining all the terms, we have that $Q_{11} + Q_{12} + Q_{21} + Q_{22} + Q_3 = O(n^{-(r-1/p)/(2(r-1/p)+1)})$, and, thus, using (6.1), the optimal pointwise convergence rate $O(n^{-2(r-1/p)/(2(r-1/p)+1)})$ is achieved. This completes the proof of Theorem 1. \square

6.2 The Bayes factor estimator as a thresholding rule

Proposition 1 is part of Lemma 1 in Pensky & Sapatinas (2005). For completeness, we provide below a sketch of proof for this result.

Proof of Proposition 1. For the sake of convenience, we drop the indices in $\zeta_{j,n}(d_{jk})$, $I_j(d_{jk})$, ν_j and φ_j . Denote $F(x) = \log(\zeta(x))$ and observe that

$$F'(x) = \frac{n}{I(x)} \int_{-\infty}^{+\infty} \left[\frac{\varphi'(\sqrt{n}(x-\theta))}{\varphi(\sqrt{n}(x-\theta))} - \frac{\varphi'(\sqrt{n}x)}{\varphi(\sqrt{n}x)} \right] \varphi(\sqrt{n}(x-\theta)) \nu h(\nu\theta) d\theta. \quad (6.7)$$

If φ is the $N(0, \sigma^2)$ pdf, then the expression in the square brackets in (6.7) is equal to $\sqrt{n}\theta/\sigma^2$, so that the integral is positive for $x > 0$. Hence, both $F(x)$ and $\zeta(x)$ are strictly increasing for $x > 0$. Similarly, if $\varphi(x) = (2\sigma)^{-1} \exp(-|x|/\sigma)$, then the expression in square brackets in (6.7) is equal to $2\mathbb{I}(\theta \geq x)/\sigma$, and $F'(x) > 0$. This completes the proof of Proposition 1. \square

6.3 Asymptotics of the thresholds

To prove Theorem 1, we referred to the following lemmas. Lemmas 1 and 2 can be easily established by working along the same lines of the proofs in Lemmas 2, 4 and 5 in Pensky & Sapatinas (2005). However, to keep the exposition self-contained, we provide below a sketch of proofs for these results.

Lemma 1. *The following statements hold*

(i) *If h is a double-exponential pdf, then, for any x ,*

$$I_j(x) = \nu_j h(\nu_j x) \left[1 + O(n^{-1} \nu_j^2) \right], \quad \text{as } \nu_j / \sqrt{n} \rightarrow 0. \quad (6.8)$$

(ii) *If φ_j is a double-exponential pdf, then, for any x ,*

$$I_j(x) = \sqrt{n} \varphi_j(\sqrt{n}x) \left[1 + O(n \nu_j^{-2}) \right], \quad \text{as } \sqrt{n} / \nu_j \rightarrow 0. \quad (6.9)$$

Proof of Lemma 1. We only provide the proof of (6.9). The proof of (6.8) is conducted in a similar manner. Using a Taylor series expansion for arbitrary x and a change of variables, we obtain:

$$\begin{aligned} I_j(x) &= \int_{-\infty}^{+\infty} \sqrt{n} \varphi_j \left(\sqrt{n}x - \frac{\sqrt{n}}{\nu_j} y \right) h(y) dy \\ &= \sqrt{n} \int_{-\infty}^{+\infty} \left[\varphi_j(\sqrt{n}x) - \frac{\sqrt{n}}{\nu_j} y \varphi_j'(\sqrt{n}x) + \frac{n}{2\nu_j^2} y^2 \varphi_j''(\sqrt{n}x) \right. \\ &\quad \left. - \frac{n\sqrt{n}}{6\nu_j^3} y^3 \varphi_j'''(\sqrt{n}x) + \dots \right] h(y) dy \\ &= \sqrt{n} \varphi_j(\sqrt{n}x) \left[1 + C \frac{n}{2\nu_j^2} \int_{-\infty}^{+\infty} y^2 h(y) dy + o\left(\frac{n}{\nu_j^2}\right) \right], \end{aligned} \quad (6.10)$$

since h is a symmetric function. This completes the proof of Lemma 1. \square

Lemma 2. *The following statements hold.*

(i) If φ_j is the pdf of the form (2.3) and $\sqrt{n}/\nu_j \rightarrow 0$, then

$$\sqrt{n}t_{j,n} \geq \sigma_j \max \left\{ \left[\log \left(\frac{\varphi_j(0)\beta_{j,n}\sqrt{n}}{h(0)\nu_j} \right) \right]^{1/\beta}, [\log(\beta_{j,n})]^{1/\beta} \right\}.$$

(ii) Let the assumption (A1) hold. If φ_j is the pdf of the form (2.3) with $\beta = 1$ or 2 (i.e., φ_j is normal or double-exponential) and $\nu_j/\sqrt{n} \rightarrow 0$, then,

$$\sqrt{n}t_{j,n} \leq \sigma_j \left[2 \log \left(\frac{C_1\beta_{j,n}\sqrt{n}}{\nu_j} \right) \right]^{1/\beta} \mathbb{I}(a_{(j)} \leq 1), \quad (6.11)$$

for some constant $C_1 > 0$.

[Note that since the threshold $t_{j,n}$ is nonnegative, the last statement of Lemma 2 implies that $t_{j,n} = 0$ for $a_{(j)} > 1$.]

Proof of Lemma 2. (i) Note that the symmetry and unimodality of φ_j implies that $\varphi_j(x) \leq \varphi_j(0)$ for any x . Therefore, the equation for the threshold $t_{j,n}$ (see expression above (2.8)) can be rewritten as follows:

$$\begin{aligned} \beta_{j,n} = \zeta_{j,n}(t_{j,n}) &= \frac{\int_{-\infty}^{+\infty} \sqrt{n} \varphi_j(\sqrt{n}(t_{j,n} - x)) \nu_j h(\nu_j x) dx}{\sqrt{n} \varphi_j(\sqrt{n}t_{j,n})} \\ &\leq \frac{\int_{-\infty}^{+\infty} \sqrt{n} \varphi_j(0) \nu_j h(\nu_j x) dx}{\sqrt{n} \varphi_j(\sqrt{n}t_{j,n})} = \frac{\varphi_j(0)}{\varphi_j(\sqrt{n}t_{j,n})}. \end{aligned}$$

Similarly, by symmetry and unimodality of h , we have

$$\begin{aligned} \beta_{j,n} = \zeta_{j,n}(t_{j,n}) &= \frac{\int_{-\infty}^{+\infty} \sqrt{n} \varphi_j(\sqrt{n}x) \nu_j h(\nu_j(t_{j,n} - x)) dx}{\sqrt{n} \varphi_j(\sqrt{n}t_{j,n})} \\ &\leq \frac{\int_{-\infty}^{+\infty} \sqrt{n} \varphi_j(\sqrt{n}x) \nu_j h(0) dx}{\sqrt{n} \varphi_j(\sqrt{n}t_{j,n})} = \frac{\nu_j h(0)}{\sqrt{n} \varphi_j(\sqrt{n}t_{j,n})}. \end{aligned}$$

Rearranging the terms, we have

$$\varphi_j(\sqrt{n}t_{j,n}) \leq \min \left\{ \beta_{j,n}^{-1} \varphi_j(0), \beta_{j,n}^{-1} h(0) \nu_j / \sqrt{n} \right\}. \quad (6.12)$$

Substituting the power exponential function φ_j given in (2.3) into (6.12), we obtain the first statement of Lemma 2.

(ii) When h and φ_j are the standard normal pdfs, then (see Abramovich, Amato & Angelini, 2004)

$$\sqrt{n}t_{j,n} = \sigma_j \sqrt{2} \sqrt{1 + \frac{\sigma_j^2 \nu_j^2}{n}} \left[\log \left(\frac{\beta_{j,n} \sqrt{n + \nu_j^2 \sigma_j^2}}{\nu_j \sigma_j} \right) \right]^{1/2} \mathbb{I} \left(\frac{\beta_{j,n} \sqrt{n + \nu_j^2 \sigma_j^2}}{\nu_j \sigma_j} > 1 \right), \quad (6.13)$$

so that (6.11) is valid. On the other hand, if h is double-exponential, then by Lemma 1 as $\nu_j/\sqrt{n} \rightarrow 0$ and for any x ,

$$\zeta_{j,n}(x) \geq C_0 \frac{\nu_j}{\sqrt{n}} \frac{h(\nu_j x)}{\varphi_j(\sqrt{n}x)},$$

for some constant $C_0 > 0$. Taking into account assumption (A1), we derive that

$$\zeta_{j,n}(x) \geq C_0 C_{h,\varphi}^{-1} \frac{\nu_j}{\sqrt{n}} \frac{\varphi_j(\nu_j x)}{\varphi_j(\sqrt{n}x)} \geq C_1 \frac{\nu_j}{\sqrt{n}} \exp \left\{ \left[1 - \left(\frac{\nu_j}{\sqrt{n}} \right)^\beta \right] \left| \frac{x\sqrt{n}}{\sigma_j} \right|^\beta \right\},$$

for some constant $C_1 > 0$. Note that since $\nu_j/\sqrt{n} \rightarrow 0$, we have $1 - \nu_j/\sqrt{n} \geq 1/2$, so that

$$\zeta_{j,n}(x)\sqrt{n}/\nu_j \geq C_1 \exp \left\{ \frac{1}{2} \left| \frac{x\sqrt{n}}{\sigma_j} \right|^\beta \right\}.$$

Now take $x = t_{j,n}$. Since the right hand side is bounded from below by $C_1 > 0$, if $\beta_{j,n}\sqrt{n}/\nu_j = (\nu_j/\sqrt{n})^{a_{(j)}-1} \rightarrow 0$ (i.e., $a_{(j)} > 1$), then $t_{j,n} = 0$. If $a_{(j)} \leq 1$, and thus $\beta_{j,n}\sqrt{n}/\nu_j = 1$ ($a_{(j)} = 1$) or $\beta_{j,n}\sqrt{n}/\nu_j \rightarrow \infty$ ($a_{(j)} < 1$), then $t_{j,n}$ is of the form (6.11). This proves the second statement of the lemma, and hence the proof of Lemma 2 is completed. \square

Lemma 3. *Assume that $d_{jk} \sim \sqrt{n}\varphi_j(\sqrt{n}(x - \theta_{jk}))$ and $\hat{\theta}_{jk} = d_{jk}\mathbb{I}(|d_{jk}| > t_{j,n})$ for some $t_{j,n} \geq 0$. Then, for any $m > 0$ such that $\kappa_{m,j} < \infty$, the following inequality holds:*

$$\mathbb{E}|\hat{\theta}_{jk} - \theta_{jk}|^m \leq \gamma_m \left(\min \{t_{j,n}^m, |\theta_{jk}|^m\} + \kappa_{m,j}n^{-m/2} \right), \quad (6.14)$$

where $\kappa_{m,j} = \int_{-\infty}^{+\infty} |x|^m \varphi_j(x) dx$, and $\gamma_m = 1$ if $0 < m \leq 1$ and $\gamma_m = 2^{m-1}$ if $m > 1$.

Proof of Lemma 3. By definition of $\hat{\theta}_{jk}$, we have

$$\begin{aligned} \mathbb{E}|\hat{\theta}_{jk} - \theta_{jk}|^m &= \mathbb{E}|d_{jk} - \theta_{jk}|^m \mathbb{I}(|d_{jk}| > t_{j,n}) + |\theta_{jk}|^m \mathbb{P}(|d_{jk}| \leq t_{j,n}) \\ &\leq \int_{\mathbb{R}} \sqrt{n}|x|^m \varphi_j(x\sqrt{n}) dx + |\theta_{jk}|^m = n^{-m/2} \kappa_{m,j} + |\theta_{jk}|^m. \end{aligned}$$

On the other hand, by first representing $\hat{\theta}_{jk} - \theta_{jk}$ as a sum of $\hat{\theta}_{jk} - d_{jk}$ and $d_{jk} - \theta_{jk}$, and then applying the definition of $\hat{\theta}_{jk}$ together with the elementary inequality $(a + b)^m \leq \gamma_m(a^m + b^m)$ for $a, b \geq 0$, $\gamma_m = 1$ if $0 < m \leq 1$ and $\gamma_m = 2^{m-1}$ if $m > 1$, we get

$$\begin{aligned} \mathbb{E}|\hat{\theta}_{jk} - \theta_{jk}|^m &\leq \gamma_m \left\{ \mathbb{E}|\hat{\theta}_{jk} - d_{jk}|^m + \mathbb{E}|d_{jk} - \theta_{jk}|^m \right\} = \gamma_m \left\{ \mathbb{E}|d_{jk}|^m \mathbb{I}(|d_{jk}| \leq t_{j,n}) + \kappa_{m,j}n^{-m/2} \right\} \\ &\leq \gamma_m \left\{ t_{j,n}^m + \kappa_{m,j}n^{-m/2} \right\}. \end{aligned}$$

Combining these two inequalities together, we obtain (6.14). This completes the proof of Lemma 3. \square

Lemma 4. *Let $\{\phi, \psi, s, L\}$ be as in assumption (S1), and let $1 \leq p, q \leq \infty$ and $1/p < r < s$. If $f \in B_{p,q}^r(A)$, then for some constants $A_0, A_1, A_2, A_3 > 0$, we have:*

$$\sum_{k \in K_{L-1}(t_0)} |\tilde{\theta}_k - \theta_k| \leq A_0 n^{-r}, \quad (6.15)$$

$$\sum_{j=L}^{J-1} \sum_{k \in K_j(t_0)} 2^{j/2} |\tilde{\theta}_{jk} - \theta_{jk}| \leq A_1 n^{-(r-1/p)} \quad (6.16)$$

and, for $L \leq j \leq J - 1$,

$$\sum_{k \in K_j(t_0)} |\tilde{\theta}_{jk}| \leq A_2 2^{-j(r-1/p+1/2)}, \quad (6.17)$$

$$\sum_{k \in K_j(t_0)} |\theta_{jk}| \leq A_3 2^{-j(r-1/p+1/2)}, \quad (6.18)$$

where $K_j(t_0) = \{k : 0 \leq k \leq 2^j - 1 \text{ and } \psi_{jk}(t_0) \neq 0\}$, $L \leq j \leq J - 1$, and $K_{L-1}(t_0) = \{k : k \in K_{L-1} \text{ and } \phi_{Lk}(t_0) \neq 0\}$.

Proof of Lemma 4. Under the conditions of the lemma, due to the embedding properties of Besov spaces (i.e., $B_{p,q}^r(A) \subset B_{p,\infty}^r(A)$, for $1 \leq q \leq \infty$), the equivalence between the Besov norm of the function f on $[0, 1]$ and the corresponding sequence norm of its wavelet coefficients, and using Proposition 5 of Johnstone & Silverman (2004), we get:

$$2^{j(r-1/p+1/2)} \left(\sum_{k=0}^{2^j-1} |\tilde{\theta}_{jk} - \theta_{jk}|^p \right)^{1/p} \leq A C(\phi, \psi, p, r) 2^{-r(J-j)}, \quad L-1 \leq j \leq J-1,$$

which implies that

$$\sum_{k=0}^{2^j-1} |\tilde{\theta}_{jk} - \theta_{jk}|^p \leq \left(A C(\phi, \psi, p, r) 2^{-j(1/2-1/p)} n^{-r} \right)^p, \quad L-1 \leq j \leq J-1.$$

(Here, we abused notation and $j = L - 1$ refers to replacing $\psi_{L-1,k}$ by ϕ_{Lk} , so that $\tilde{\theta}_{L-1,k} = \tilde{\theta}_k$ and $\theta_{L-1,k} = \theta_k$.)

First consider the case $L \leq j \leq J - 1$. Applying Hölder's inequality, we obtain

$$\sum_{k \in K_j(t_0)} |\tilde{\theta}_{jk} - \theta_{jk}| \leq \left[\sum_{k=0}^{2^j-1} |\tilde{\theta}_{jk} - \theta_{jk}|^p \right]^{1/p} K^{1-1/p} \leq A C(\phi, \psi, p, r) K^{1-1/p} 2^{-j(1/2-1/p)} n^{-r},$$

where $K = \text{Card}(K_j(t_0))$ is the cardinality of $K_j(t_0)$ (see also the discussion after (6.1)). So, in each of the above cases, summing the corresponding terms over j with factor $2^{j/2}$, we obtain the bound $A_1 n^{-(r-1/p)}$. Therefore, the second statement of Lemma 4 is proved.

Following the above arguments in the case $j = L - 1$, and replacing $K_j(t_0)$, $L \leq j \leq J - 1$, with $K_{L-1}(t_0)$, we easily obtain the bound $A_0 n^{-r}$. Therefore, the first statement of Lemma 4 is proved.

To prove the third statement of the lemma, we use again the above mentioned embedding properties of Besov spaces, the equivalence between the Besov norm of the function f on $[0, 1]$ and the corresponding sequence norm of its wavelet coefficients. Using equation (20) of Johnstone & Silverman (2004), we have:

$$\begin{aligned} \sum_{k \in K_j(t_0)} |\tilde{\theta}_{jk}| &\leq K^{1-1/p} \left(\sum_{k=0}^{2^j-1} |\tilde{\theta}_{jk}|^p \right)^{1/p} &\leq A K^{1-1/p} 2^{-j(r-1/p+1/2)} \\ & &= A_2 2^{-j(r-1/p+1/2)}, \quad L \leq j \leq J-1. \end{aligned}$$

This completes the proof of the third statement of Lemma 4.

To prove the last statement of the lemma, note that from the above we get:

$$\begin{aligned} \sum_{k \in K_j(t_0)} |\theta_{jk}| &\leq \sum_{k \in K_j(t_0)} |\tilde{\theta}_{jk} - \theta_{jk}| + \sum_{k \in K_j(t_0)} |\tilde{\theta}_{jk}| \\ &\leq AC(\phi, \psi, p, r) K^{1-1/p} 2^{-j(1/2-1/p)} n^{-r} + A_2 2^{-j(r-1/p+1/2)} \\ &\leq A_3 2^{-j(r-1/p+1/2)}, \quad L \leq j \leq J-1. \end{aligned}$$

Thus, the last statement of Lemma 4 is proved. This completes the proof of Lemma 4. \square

References

- [1] ABRAMOVICH, F., AMATO, U. & ANGELINI, C. (2004). On optimality of Bayesian wavelet estimators. *Scandinavian Journal of Statistics*, **31**, 217–234.
- [2] ABRAMOVICH, F., ANGELINI, C. & DE CANDITIIS, D. (2007). Pointwise optimality of Bayesian wavelet estimators. *Annals of the Institute of Statistical Mathematics*, **59** (to appear).
- [3] ABRAMOVICH, F., BAILEY, T.C. & SAPATINAS, T. (2000). Wavelet analysis and its statistical applications. *The Statistician*, **49**, 1–29.
- [4] ABRAMOVICH, F., SAPATINAS, T. & SILVERMAN, B.W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, **60**, 725–749.
- [5] ANGELINI, C. & SAPATINAS, T. (2004). Empirical Bayes approach to wavelet regression using ε -contaminated priors. *Journal of Statistical Computation and Simulation*, **74**, 741–764.
- [6] ANGELINI, C. & VIDAKOVIC, B. (2004). Γ -minimax wavelet shrinkage: a robust incorporation of information about energy of a signal in denoising applications. *Statistica Sinica*, **14**, 103–125.
- [7] ANGELINI, C., DE CANDITIIS, D. & LEBLANC, F. (2003). Wavelet regression estimation in nonparametric mixed effects models. *Journal of Multivariate Analysis*, **85**, 267–291.
- [8] ANTONIADIS, A., BIGOT, J. & SAPATINAS, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. *Journal of Statistical Software*, **6**, Issue 6, 1–83.
- [9] BILLINGSLEY, P. (1995). *Probability and Measure*. 3rd Edition, New York: Wiley & Sons.
- [10] BOCHKINA, N. & SAPATINAS, T. (2005). On the posterior median estimators of possibly sparse sequences. *Annals of the Institute of Statistical Mathematics*, **57**, 315–351.
- [11] BROOKS, S.P. & GELMAN A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

- [12] CAI, T.T. (2003). Rates of convergence and adaptation over Besov spaces under pointwise risk. *Statistica Sinica*, **13**, 881–902.
- [13] CHIPMAN, H.A., KOLACZYK, E.D. & MCCULLOCH, R.E. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, **92**, 1413–1421.
- [14] CLYDE, M. & GEORGE, E.I. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, Series B*, **62**, 681–698.
- [15] CLYDE, M., PARMIGIANI, G. & VIDAKOVIC, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–401.
- [16] COHEN, A., DAUBECHIES, I. & VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, **1**, 54–81.
- [17] DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*, Philadelphia: SIAM.
- [18] DONOHO, D.L. & JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–456.
- [19] DONOHO, D.L. & JOHNSTONE, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90**, 1200–1224.
- [20] DONOHO, D.L. & JOHNSTONE, I.M. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **26**, 879–921.
- [21] DONOHO, D.L. & LOW, M.G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Annals of Statistics*, **20**, 944–970.
- [22] DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. & PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 301–337.
- [23] JOHNSTONE, I.M. & SILVERMAN, B.W. (2004). Boundary coefficients for wavelet shrinkage in function estimation. *Journal of Applied Probability*, **41A**, 81–98.
- [24] JOHNSTONE, I.M. & SILVERMAN, B.W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, **33**, 1700–1752.
- [25] NASON, G.P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Series B*, **58**, 463–479.
- [26] NEUMANN, M.H., & VON SACHS, R. (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. In *Wavelets and Statistics*, Antoniadis, A. & Oppenheim, G. (Eds.), Lecture Notes in Statistics, **103**, pp. 301–329, New York: Springer-Verlag.

- [27] PENSKEY, M. & SAPATINAS, T. (2005). Frequentist optimality of Bayes factor estimators in wavelet regression models. *Technical Report, TR-08-05*, Department of Mathematics and Statistics, University of Cyprus, Cyprus. (Under revision for *Statistica Sinica*).
- [28] PENSKEY, M. (2006). Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise. *Annals of Statistics*, **34**, 769–807.
- [29] SPIEGELHALTER, D.J., THOMAS, A., & BEST, N. (1999). WinBUGS Version 1.4, User Manual. MRC Biostatistics Unit, Cambridge.
- [30] VIDAKOVIC, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, **93**, 173-179.
- [31] VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. New York: Wiley & Sons.