Statistica Sinica **17**(2007), 599-633

FREQUENTIST OPTIMALITY OF BAYES FACTOR ESTIMATORS IN WAVELET REGRESSION MODELS

Marianna Pensky and Theofanis Sapatinas

University of Central Florida and University of Cyprus

Abstract: We investigate the theoretical performance of Bayes factor estimators 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 optimality in Besov spaces for a wide variety of error and prior distributions. Furthermore, we provide sufficient conditions that determine whether the underlying regression function belongs to a Besov space *a-priori* with probability one. We also study an adaptive estimator by considering an empirical Bayes estimation procedure of the Bayes factor estimator for a certain combination of error and prior distributions. Simulated examples are used to illustrate the performance of the empirical Bayes estimation procedure based on the proposed Bayes factor estimator, and compared with two recently proposed empirical Bayes estimators. An application to a dataset that was collected in an anaesthesiological study is also presented.

Key words and phrases: Bayesian inference, Besov spaces, empirical Bayes inference, nonparametric regression, optimality, wavelets.

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, which typically amounts to individual assessment of each empirical wavelet coefficient. 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 and Johnstone (1994, 1995, 1998) and Donoho and Johnstone, Kerkyacharian and Picard (1995). See also Vidakovic (1999), Abramovich, Bailey and Sapatinas (2000) and Antoniadis, Bigot and Sapatinas (2001) for comprehensive reviews and appropriate software.

Various Bayes and empirical Bayes approaches for term-by-term wavelet (nonlinear) shrinkage and wavelet thresholding estimators have also been 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 that is 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*) wavelet (nonlinear) shrinkage and wavelet thresholding rules. See, e.g., Chipman, Kolaczyk and McCulloch (1997), Abramovich, Sapatinas and Silverman (1998), Clyde, Parmigiani and Vidakovic (1998), Vidakovic (1998), Angelini and Sapatinas (2004) and Angelini and Vidakovic (2004).

The above papers are devoted to the nonparametric regression model

$$Y_i = f(t_i) + Z_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $t_i = i/n$, f is the unknown response function that is assumed to belong to some functional space $\mathcal{F}$ defined on the unit interval $[0, 1]$, and the errors Z_i are independent and identically distributed (*i.i.d.*) random variables. In most cases, the Z_i's are assumed to be normally distributed with (mean) $E(Z_1) = 0$ and (variance) $\mathrm{Var}(Z_1) = \sigma^2 < \infty$. Recently, the case dealing with non-normal errors in the Bayesian framework was also studied. See, e.g., Clyde and George (2000) and Pensky (2006).

Our focus will be on investigating the optimality of the so-called *Bayes factor* estimator (see Vidakovic (1998)) for a wide variety of error and prior distributions. 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.

The present paper continues the line of investigation of Abramovich, Amato and Angelini (2004), Johnstone and Silverman (2004a, 2005) and Pensky (2006) who examined frequentist optimality (in the minimax sense) of various Bayesian and empirical Bayes wavelet (nonlinear) shrinkage and wavelet thresholding estimators for model (1.1). However, unlike Abramovich et al. (2004) and Johnstone and Silverman (2004a, 2005), this paper is not confined to the case of normal errors and considers a much wider variety of Bayesian models. In this sense, this paper is similar to Pensky (2006), where optimality is studied under very mild restrictions on the errors in (1.1), i.e., it is just assumed that the Z_i's are *i.i.d.* random variables with a unimodal and symmetric (but unknown) distribution having $E(Z_1) = 0$ and $E(Z_1^4) < \infty$. However, although our main results are formulated in a similar way to Pensky (2006), the present paper studies a different set of estimation techniques, namely, Bayes factor estimators which are

based on hard thresholding of wavelet coefficients, while Pensky (2006) investigates posterior mean estimators which lead to a (nonlinear) shrinkage. In spite of the seeming similarity between the two papers, the proofs of the statements of the present paper require a completely different set of techniques. The only exception is Theorem 3 which is concerned with the well-known Bayesian paradox and is independent of a particular estimation technique. However, Corollary 4 is directly tailored to the Bayes factor estimators and is different from Corollaries 4 and 5 in Pensky (2006). Moreover, as we demonstrate below, optimality results for the Bayes factor estimators are derived under milder conditions than the corresponding ones for the posterior mean estimators established in Pensky (2006). The paper of Bochkina and Sapatinas (2005) also adopts the Bayesian paradigm for a wide variety of error and prior distributions. However, it does not deal with model (1.1) directly but instead discusses certain properties of posterior median estimators obtained from recovering a high-dimensional vector observed in white noise; moreover, it does not study optimality of the resulting posterior median estimators. It is also evident from the expressions of the posterior median estimators obtained in Bochkina and Sapatinas (2005), as well as the forms of the posterior mean estimators obtained in Pensky (2006), that the Bayes factor estimator is much easier to evaluate in the majority of cases.

Our work is partly motivated by the results of Abramovich et al. (2004) who investigated posterior mean, posterior median and Bayes factor estimators, but only in the case of normal errors and normal priors. Hence, the results of Abramovich et al. (2004) are much more limited than results of the present paper. Also, due to their sub-optimal choice of tuning parameters, conclusions made in Abramovich et al. (2004) are too pessimistic. However, in the case of normal errors and normal priors, Abramovich et al. (2004) implied that Bayes factor estimators lead to smaller risks than posterior mean or posterior median estimators, and in what follows we confirm that indeed Bayes factor estimators are not only easier to calculate but they also achieve optimality under milder conditions on the actual errors.

Finally, a few words should be said about the relationship between the present paper and those by Johnstone and Silverman (2004a, 2005). Unlike the latter papers, which are concerned with investigation of (adaptive) *empirical Bayes* estimators under normal errors, we consider *Bayes* estimators under a wide variety of error models. Although Section 3.5 studies a modification of the technique of Johnstone and Silverman (2005) when Bayes factor rules are used, investigation of empirical Bayes techniques is not the main goal of this paper and is done mainly for demonstrating a practical benefit of an empirical Bayes version of the proposed Bayes factor estimator in finite sample situations.

It should be mentioned that although all three estimation techniques – posterior mean, posterior median, and Bayes factor estimators – can achieve optimality

as sample size increases, in finite sample situations the posterior mean and posterior median estimators often deliver smaller average mean squared errors than Bayes factor estimators, while the latter ones preserve peak heights better. This is mainly due to the fact that posterior mean and posterior median estimators appear to have slightly better reconstructions over the regions where the underlying function is smooth; this is not surprising since posterior mean estimators are (nonlinear) shrinkage rules while posterior median estimators are soft thresholding rules. Hence, as our simulation study and dataset show, Bayes factor estimators, as hard thresholding rules, are preferable for irregular functions with high peaks.

The rest of the paper is organized as follows. In Section 2 we introduce Bayesian models for the wavelet coefficients. We use some "arbitrary" distribution for the error and a mixture of a point mass at zero and a symmetric, unimodal distribution for the prior, keeping in mind that the actual distribution of the wavelet coefficients is unknown at fine resolution levels, and is asymptotically normal at coarse resolution levels according to the Central Limit Theorem. In Section 3 we discuss assumptions on the error and prior distributions, and provide assertions about optimality of Bayes factor estimators in Besov spaces for various choices of error and prior distributions. Furthermore, we provide sufficient conditions that determine whether the underlying regression function belongs to a Besov space *a-priori* with probability one. An adaptive estimator, based on an empirical Bayes estimation procedure of the Bayes factor estimator for a certain combination of error and prior distributions, is also considered. In Section 4, simulated results are used to illustrate the performance of the empirical Bayes estimation procedure based on the proposed Bayes factor estimator, and compared with two recently proposed empirical Bayes estimators. We also present an application to a dataset that was collected in an anaesthesiological study. Section 5 is reserved for discussion and comparison of the various models on the basis of optimality, as well as performances on the simulated examples and the dataset. Finally, in Section 6 (Appendix), we provide some auxiliary statements as well as the proofs of the theoretical results obtained in Section 3.

2. The Bayesian Model

Consider the nonparametric regression model (1.1) and assume that the underlying response function f belongs to the space of squared integrable functions on $[0,1]$, i.e., $f \in L^2[0,1]$, and that the Z_i's are *i.i.d.* random variables with $E(Z_1) = 0$ and $E(Z_1^4) < \infty$. 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 and Vial (1993), Walter (1994, Chap.7) and Johnstone and Silverman (2004b).

Application of the (boundary corrected) discrete wavelet transform (DWT) to (1.1) yields

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

where $J = \log_2 n$, and ϵ_k and ε_{jk} are uncorrelated random variables due to the unitary property of the DWT. Let $\theta_k = u_k/\sqrt{n}$ 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 optimality results given in subsequent sections (see Lemma 6). 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 and Silverman (2004b, 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}

$$w_{jk} \sim \pi_{j,n}\tau_{j,n}\xi(\tau_{j,n}\cdot) + (1-\pi_{j,n})\delta(0), \quad j = L, L+1, \ldots, \quad k = 0, 1, \ldots, 2^j - 1, \tag{2.1}$$

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. We also assume that ξ is a symmetric probability density function (*pdf*) on $\mathbb{R} = (-\infty, \infty)$ that is unimodal, positive and finite at zero. 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 (2.1), w_{jk} is either zero with probability $(1 - \pi_{j,n})$ or with probability $\pi_{j,n}$ is distributed with the *pdf* ξ 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, however, we impose all conditions on the prior odds ratio

$$\beta_{j,n} = \frac{1-\pi_{j,n}}{\pi_{j,n}}. \tag{2.2}$$

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 values of $\pi_{j,n}$.

Let us now discuss the distribution of the errors ε_{jk}. It follows from (1.1) that

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

Since the Z_i's are *i.i.d.* random variables with $E(Z_1^4) < \infty$, 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 $2^j/n \to 0$ as $n \to \infty$. Hence, if the resolution level is reasonably small ($j \le J_0$ where $J - J_0 \to \infty$ as $n \to \infty$), the errors ε_{jk} are asymptotically $N(0, \sigma^2)$ distributed and, thus, asymptotically independent. On the other hand, at high resolution levels, the errors ε_{jk} are uncorrelated and have some unknown *pdf*'s μ_j which are symmetric and have uniformly bounded (for all j and k) fourth moments, $E(\varepsilon_{jk}^4) \le C_{\sigma,\varepsilon}$. For a more detailed treatment of asymptotic normality, the interested reader is referred to, e.g., Neumann and von Sachs (1995).

The difficulty of using the μ_j in Bayesian inference is that they are unknown. For this reason, we choose general distribution for the errors ε_{jk}, namely,

$$\varepsilon_{jk} \sim \eta_j(\cdot), \tag{2.3}$$

where the η_j are level dependent symmetric *pdf*'s on $\mathbb{R}$ that are unimodal, positive and finite at zero. (For the distribution of errors of the scaling coefficients, ϵ_k, we only assume that it has a finite variance.) As we show later, one does not need the knowledge of the true distribution of the errors ε_{jk} and can obtain optimal estimators of f with a variety error distributions η_j. Moreover, one can consider $\eta_j = \eta$, i.e., the error distribution does not even need to be level dependent.

Note that we have considered the *pdf*'s ξ and η_j to be positive. Narrowing the support of ξ would imply that we ignore large wavelet coefficients which is

inappropriate, since they represent important characteristics of the (possibly) inhomogeneous signal of interest. Similarly, narrowing the supports of the η_j mean that we exclude noise distributions with heavy tails, like double-exponential or Student-t *pdf*'s. Finally, we assume that both ξ and η_j are finite at zero for slightly different reasons. For ξ, we assume that all zero mass is accounted for in the other part of the mixture otherwise the mixture would not be identifiable, whereas in η_j we assume that there are no zero masses to exclude the 'pathological' case of observing data without errors. Note also that due to the unimodality assumption, both ξ and η_j cannot have atom masses at any other points.

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

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

Taking into account the relation between w_{jk} and θ_{jk} and (2.1)–(2.4), we find that the posterior *pdf* of θ_{jk} given d_{jk} is of the form

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

The Bayes factor estimator of θ_{jk} is derived as follows (see Vidakovic (1998)): after observing d_{jk}, we test $H_0 : \theta_{jk} = 0$ versus $H_1 : \theta_{jk} \neq 0$. If 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

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

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

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

where

$$\zeta_{j,n}(d_{jk}) = \frac{I_j(d_{jk})}{[\sqrt{n}\,\eta_j(\sqrt{n}d_{jk})]}, \tag{2.5}$$

$$I_j(d_{jk}) = \int_{-\infty}^{+\infty}\sqrt{n}\,\eta_j[\sqrt{n}(x - d_{jk})]\ \nu_j\xi(\nu_j x)dx. \tag{2.6}$$

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

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

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 $\zeta_{j,n}(d_{jk}) > \beta_{j,n}$ if and only if $|d_{jk}| > t_{j,n} = \zeta_{j,n}^{-1}(\beta_{j,n})$. Hence, (2.7) is a hard thresholding rule with the threshold $t_{j,n}$, i.e.,

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

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

Lemma 1. *If η_j is the normal or the double-exponential pdf, then $\zeta_{j,n}(d_{jk})$ is strictly increasing in d_{jk} for $d_{jk} > 0$. If both η_j and ξ are Student-t pdf's with a and b degrees of freedom, respectively, with $a > b$, then $\zeta_{j,n}(d_{jk})$ is strictly increasing in d_{jk} for $d_{jk} > 0$ if $\nu_j/\sqrt{n} \to 0$ and $b > a/(a+1)$, while $\zeta_{j,n}(d_{jk}) \le 1 + C_a n/\nu_j^2$ if $\sqrt{n}/\nu_j \to 0$, where the constant $C_a > 0$ depends only on a.*

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 $j \ge J$, $k = 0, 1, \ldots, 2^j - 1$, one has $\hat{\theta}_{jk} = 0$ as $j \ge J$, $k = 0, 1, \ldots, 2^j - 1$, and therefore the estimator $\hat{f}$ of f is of the form

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

Coefficients $\hat{\theta}_{jk}$ are found using (2.7), where $\zeta_{j,n}(d_{jk})$ is defined by (2.5) and (2.6). To complete the construction of the estimator, we need to choose the error model η_j and the prior model ξ, as well as the values of the parameters ν_j and β_{jn}, so that the estimator (2.9) achieves the optimal convergence rate over a wide range of Besov spaces. This is the purpose of the next section.

3. Optimality in Besov Spaces

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

3.1. Optimal convergence rate over Besov spaces

For any estimator $\tilde{f}$ of f based on n observations from model (1.1), define the maximal risk, with respect to the $L^2[0,1]$-loss function, over a function space $\mathcal{F}$ defined on the unit interval $[0,1]$, as

$$R_n(\mathcal{F}, \tilde{f}) = \sup_{f \in \mathcal{F}} E(\|\tilde{f} - f\|_2^2), \tag{3.1}$$

where $\|\cdot\|_2^2$ denotes the $L^2[0,1]$-norm. Donoho and Johnstone (1998) showed that when the Z_i's in (1.1) are independent and normally distributed with $E(Z_1) = 0$ and $\mathrm{Var}(Z_1) = \sigma^2 < \infty$, and when f belongs to a ball $B^r_{p,q}(A)$ of radius $A > 0$ in the Besov space $B^r_{p,q}[0,1]$, then provided $r > 1/p$ and $1 \le p, q \le \infty$,

$$\inf_{\tilde{f}} R_n(B^r_{p,q}(A), \tilde{f}) \asymp n^{-\frac{2r}{2r+1}} \text{ as } n \to \infty, \tag{3.2}$$

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

Note that the normal distribution is a particular case of the distribution of the errors ε_{jk}. Since for the majority of resolution levels ($j \le J_0$ where $J - J_0 \to \infty$ as $n \to \infty$) the errors ε_{jk} asymptotically follow the normal distribution, we can expect to achieve the optimal convergence rate (3.2), as $n \to \infty$, for some choices of prior (2.1) and error (2.3) distributions.

3.2. Assumptions

In what follows, we formulate conditions on the wavelet system $\{\phi, \psi\}$ and the *pdf*'s ξ and η_j, as well as on the parameters ν_j and $\beta_{j,n}$. We point out that these conditions are not always necessary in what follows.

(A0) Let ϕ and ψ be the boundary coiflets introduced in Johnstone and Silverman (2004b), possessing s continuous derivatives, $s-1$ vanishing moments, $s \ge 2$, and based on orthonormal coiflets supported in $[-S+1, S]$, $s < S$. Let also $L \ge \log_2(6S - 6)$.

Let ξ and η_j be symmetric *pdf*'s on $\mathbb{R}$ that are unimodal, positive and finite at zero, that they be three times continuously differentiable everywhere, except possibly at zero, have uniformly bounded fourth moments, and satisfy the conditions

$$\text{(A1)}\ |\xi^{(k)}(x)/\xi(x)| \le C_\xi(1 + |x|^{\lambda_\xi})^k,\ k = 1,2,3,\ \lambda_\xi \ge 0, \tag{3.3}$$

$$\text{(A2)}\ |\eta_j^{(k)}(x)/\eta_j(x)| \le C_\eta(1 + |x|^{\lambda_\eta})^k,\ k = 1,2,3,\ \lambda_\eta \ge 0, \tag{3.4}$$

$$\text{(A3)}\ |\eta_j(x)/\xi(x)| \le C_{\xi,\eta}.$$

The constants λ_η, C_η and $C_{\xi,\eta}$ are assumed to be independent of j which requires some kind of uniformity for the *pdf*'s η_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 the η_j.

In the subsequent development, we consider two general parametric models for ξ and η_j: power exponential models with *pdf*'s proportional to $\exp\{-c|x|^\beta\}$,

$x \in \mathbb{R}$, $\beta > 0$, $c > 0$, and polynomial models with *pdf*'s proportional to $(1 + cx^2)^{-\rho}$, $x \in \mathbb{R}$, $\rho > 0$, $c > 0$. The polynomial models satisfy (A1) and (A2) for all $\rho > 0$ with $\lambda_\xi = 0$ and $\lambda_\eta = 0$, while the power exponential models satisfy (A1) and (A2) with $\lambda_\xi = 0$ and $\lambda_\eta = 0$ if $0 < \beta \leq 1$, and with $\lambda_\xi = \beta - 1$ and $\lambda_\eta = \beta - 1$ if $\beta > 1$. Note that the common normal η_j - normal ξ model satisfies (A1) and (A2) with $\lambda_\eta = 1$ and $\lambda_\xi = 1$.

When (A1) and (A2) hold with $\lambda_\xi = 0$ and $\lambda_\eta = 0$, following Johnstone and Silverman (2004a, 2005), we say that η_j and ξ are *heavy-tailed pdf*'s. The most common examples are the double-exponential and the Student-t *pdf*'s. In this situation, the integral $I_j(d_{jk})$ has the asymptotic expansion

$$I_j(d_{jk}) \sim \nu_j \xi(\nu_j d_{jk}) \quad \text{if} \quad \nu_j/\sqrt{n} \to 0, \tag{3.5}$$

that is valid for any d_{jk} as long as the relation between ν_j and n holds. If λ_ξ is positive, then (3.5) can be used under some restrictions on d_{jk} only (see Lemma 2).

Let also

$$r_p = 0.5 \left[(\frac{1}{p} - \frac{1}{2}) + \sqrt{(\frac{1}{p} - \frac{1}{2})^2 + 2(\frac{1}{p} - \frac{1}{2})} \right] I(1 \leq p < 2). \tag{3.6}$$

Note that $r_p = 0$ when $p \geq 2$, and that $r_p \leq (1 + \sqrt{5})/4$ for any $1 \leq p < 2$.

Remark 1. The assumption (A0) and condition (3.6) are introduced for the sake of obtaining convergence rates for the $L^2[0,1]$-norm based risk function. See Johnstone and Silverman (2004b). All statements of the paper hold for $L = 0$ and any periodic s-regular scaling function ϕ and wavelet ψ with $s > \max(r, r + 1/2 - 1/p)$ if one replaces (3.1) by $R_n(\mathcal{F}, \tilde{f}) = \sup_{f \in \mathcal{F}} \left(1/n \sum_{i=1}^n \mathbb{E}[\tilde{f}(i/n) - f(i/n)]^2 \right)$.

Remark 2. The assumptions of the existence of fourth moments are used for derivation of asymptotic expansions of the integral $I_j(d_{jk})$. These conditions can be dropped and replaced by the conclusions of Lemma 2. These conclusions, however, have to be verified individually for each combination of the error η_j and the prior ξ.

Let $r > 0$, $1 \leq p, q \leq \infty$ and $A > 0$. It is well-known that whenever $f \in B^r_{pq}(A)$, its wavelet coefficients $\tilde{\theta}_{jk}$ satisfy

$$\sum_{k=0}^{2^j - 1} \tilde{\theta}^2_{jk} \leq B \, 2^{-2j(r + \frac{1}{2} - \frac{1}{\min(p,2)})} \tag{3.7}$$

for some $B > 0$. See, e.g., Johnstone (2002). This implies that the properties of Besov spaces vary dramatically for $1 \leq p < 2$ and $p \geq 2$. Indeed, $p \geq 2$ indicates

that the Besov space is spatially homogeneous while $1 \le p < 2$ means that the Besov space is spatially nonhomogeneous. See, e.g., Mallat (1999, Sec.9.2.3). In order to be able to treat both cases together, we introduce the notation $p^* = \min(p, 2)$.

Let

$$\begin{aligned} j_0 &= (2r+1)^{-1} \log_2 n, \qquad J_0 = 0.5\left[\log_2 n + j_1\right], \\ j_1 &= r\left[(r + \tfrac{1}{2} - \tfrac{1}{p^*})(2r+1)\right]^{-1} \log_2 n. \end{aligned} \tag{3.8}$$

We assume that the parameter ν_j is of the form

$$\nu_j = C_1\, 2^{mj}, \tag{3.9}$$

where, for some $\varepsilon > 0$,

$$m = \begin{cases} m_1 = r + \frac{1}{2}, & L \le j \le j_0, \\ m_2 = (r + \frac{1}{2}) - (\frac{1}{p^*} - \frac{1}{2})\left(1 + (2r)^{-1} + \varepsilon\right), & j_0 < j \le j_1, \\ m_3 = r + \frac{1}{2}, & j_1 < j \le J - 1. \end{cases} \tag{3.10}$$

(Note that although we always require $\varepsilon > 0$, in some cases, as we see below, it is necessary to have more restrictive assumptions on ε.) We refer to $L \le j \le j_0$, $j_0 < j \le j_1$, and $j_1 < j \le J - 1$ as *low*, *medium*, and *high* resolution levels, respectively.

Remark 3. The assumptions about ν_j can be translated into the ones on $\tau_{j,n}$ using (2.4), namely, $\tau_{j,n} = C_1 2^{mj}/\sqrt{n}$. This expression coincides with the choice of $\tau_{j,n}$ for the normal η_j – normal ξ model in Abramovich et al. (1998) and Abramovich et al. (2004).

Choose also $\beta_{j,n}$ such that

$$\beta_{j,n} = \left(\frac{\nu_j}{\sqrt{n}}\right)^{a_j}, \quad \text{where} \quad a_j = \begin{cases} a_1, L \le j \le j_0, \\ a_2, j_0 < j \le j_1, \\ a_3, j_1 < j \le J - 1. \end{cases} \tag{3.11}$$

Observe that if $p \ge 2$, then $p^* = 2$ and $j_0 = j_1$, so that the medium resolution levels disappear, and $m = r + 1/2$ for all resolution levels. This situation leads to "almost linear" estimators which deliver optimality in spatially homogeneous Besov spaces. In spatially nonhomogeneous Besov spaces (i.e., $1 \le p < 2$), the medium resolution levels require a larger spread of the prior distribution, which leads to lower values of ν_j and consequently of m. Note also that the resolution level j_1 is chosen so that

$$\sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} \theta_{jk}^2 = O\left(n^{-\frac{2r}{2r+1}}\right).$$

We also assume that $r > r_p$ (see (3.6)), ensuring that $j_1 = o(J)$, so that one can choose resolution level J_0, $j_1 < J_0 < J-1$, such that $J - J_0 \to \infty$. The latter means that the errors ε_{jk} are asymptotically normally distributed for $j \le j_1$, so that $\mu_j(x) = (\sqrt{2\pi}\sigma)^{-1} \exp(-x^2/(2\sigma^2))$ for $j \le j_1$.

Finally, we assume that the η_j do not have significantly lighter tails than the normal distribution with variance equal to the *true* variance of the error, i.e.,

$$\text{(A4)} \quad \varphi(\frac{x}{\sigma}) \le C_\alpha \, [\eta_j(x)]^\alpha, \;\; \alpha > 0, \tag{3.12}$$

where φ is the *pdf* of a $N(0,1)$ random variable, σ is the *true* standard deviation of the error, and $C_\alpha > 0$ is a constant independent of j.

3.3. Optimality of Bayes factor estimators in Besov spaces

We now discuss the combination of error and prior distributions considered below to study optimality of Bayes factor estimators.

In effect, we consider two kinds of models, with threshold $t_{j,n}$ as defined between (2.7) and (2.8), at high resolution levels:

(a) the models for which the threshold $t_{j,n}$ satisfies the condition

$$\sqrt{n}\, t_{j,n} \ge C_t \, \frac{\nu_j}{\sqrt{n}}, \quad \text{if} \quad \frac{\sqrt{n}}{\nu_j} \to 0, \tag{3.13}$$

where the constant $C_t > 0$ is independent of n and ν_j;

(b) the models for which the threshold $t_{j,n}$ satisfies the condition

$$\sqrt{n}\, t_{j,n} \ge C_t \;\max\left\{\sqrt{\ln(\beta_{j,n}\frac{\sqrt{n}}{\nu_j})}, \sqrt{\ln(\beta_{j,n})}\right\}, \quad \text{if} \quad \frac{\sqrt{n}}{\nu_j} \to 0,$$

where the constant $C_t > 0$ independent of n and ν_j.

The first class of models includes all those with $\lambda_\xi = \lambda_\eta = 0$ in (3.3) and (3.4), the normal η_j – normal ξ model, and the normal η_j - double-exponential ξ model. The second class of models includes, e.g., the normal η_j – polynomial-tailed ξ model. (The case when η_j are heavy-tailed and ξ is normal delivers sub-optimal convergence rates due to the slow convergence of the bias when, e.g., the posterior mean is used as an estimator. See Pensky (2006). Hence, intentionally, we do not consider these models in the subsequent development.) In these cases, we have the following result.

Theorem 1. *Let $\{\phi, \psi, s, L\}$ be as in* (A0)*, and let $f \in B^r_{p,q}(A)$ with $1 \le p, q \le \infty$, $p^* = \min(p, 2)$ and $\max(r_p, 1/p^*) < r < s$. Assume that the error distribution μ_j has l uniformly (in j) bounded moments, where*

$$l > 2 + \left(r + \frac{1}{2}\right)^{-1}. \tag{3.14}$$

Choose ξ and η_j as in Section 3.1 *satisfying* (A1)–(A4), *let $\beta_{j,n}$ be given by* (3.11), *and assume that the threshold t_{jn} satisfies* (3.13).

(i) *If the η_j have exponential descents, i.e.,*

$$\eta_j(x) = C_j \exp\Big(- \Big|\frac{x}{\sigma_j}\Big|^{\beta}\Big), \quad 0 < \underline{\sigma} \le \sigma_j \le \bar{\sigma} < \infty,\ C_j > 0,\ \beta > 0, \tag{3.15}$$

and $\lambda_\xi = 0$ in (3.3) *or ξ is the normal pdf, choose $a_1 < 1$, $a_2 < 1 - (1/p - 1/2)/[\alpha(r + 1/2)(r + 1/2 - 1/p)]$ and $a_3 > 1 + 1/[\alpha(r + 1/2)]$ in* (3.11), *and choose $\varepsilon > [1/(\alpha r(1 - a_2))]I(\beta = 2)$ in* (3.10), *where α is defined by* (3.12). *Then*

$$R_n(B^r_{p,q}(A), \hat{f}) = O\left(n^{-\frac{2r}{2r+1}} (\ln n)^{\frac{2-p^*}{\beta}}\right) \quad as\ \ n \to \infty. \tag{3.16}$$

(ii) *If the η_j have polynomial descents, i.e.,*

$$\eta_j(x) = C_j(1 + |\frac{x}{\sigma_j}|^2)^{-\varrho}, \quad 0 < \underline{\sigma} \le \sigma_j \le \bar{\sigma} < \infty, \quad C_j > 0, \quad \varrho > 0, \tag{3.17}$$

choose $1 - \varrho/(r + 1/2) < a_1 < 1$, $a_2 < 1$ and $a_3 > 0$ in (3.11), *and choose $\varepsilon > 0$ in* (3.10). *Then*

$$R_n(B^r_{p,q}(A), \hat{f}) = O\left(n^{-\frac{2r}{2r+1}+\kappa}\right) \quad as\ \ n \to \infty,$$

where

$$\kappa = \frac{(1 - a_2)(2 - p^*)^2}{8p^*\varrho(r + \frac{1}{2})}\left(1 + \frac{1}{2r} + \varepsilon\right). \tag{3.18}$$

Theorem 1 shows that in spatially homogeneous Besov spaces (i.e., $p^* = 2$), all Bayes factor estimators achieve optimality without a logarithmic factor (note that both the power of $\ln n$ and κ are equal to zero in this case). If $p^* < 2$, the Bayes factor estimator is optimal up to a logarithmic factor if the η_j have exponential descents, and has sub-optimal convergence rates if the η_j have polynomial descents. Observe also that, due to (A4), we have the restriction $0 < \beta \le 2$ in (3.15), so that the power of the logarithmic function is at least $(1 - p/2)$, which is achieved when the η_j are normal *pdf*'s. It is well known that the $(\ln n)^{1-p/2}$ factor in the optimal convergence rate is unavoidable whenever coefficients are treated one-by-one. See, e.g., Donoho and Johnstone (1994) and Cai (1999). This benchmark is met as $\beta = 2$ in (3.15). In general, we have the following results.

Corollary 1. *Let the assumptions of Theorem* 1 *on ϕ, ψ, s, L, f, r, p, q and μ_j hold. Let η_j be double-exponential pdf's satisfying* (3.15), *let ξ be such that $\lambda_\xi = 0$*

in (3.3)*, and that* (A3) *is satisfied. Set* $a_1 < 1$*,* $a_2 < 1$ *and* $a_3 > 1$ *in* (3.11)*, and choose* $\varepsilon > 0$ *in* (3.10)*. Then* $R_n(B^r_{p,q}(A), \hat{f}) = O\left(n^{-2r/(2r+1)} \ (\ln n)^{2-p^*}\right)$ *as* $n \to \infty$*.*

Corollary 2. *Let the assumptions of Theorem* 1 *on* ϕ*,* ψ*,* s*,* L*,* f*,* r*,* p*,* q *and* μ_j *hold. Let* η_j *be normal pdf's satisfying* (3.15)*, and let* ξ *be a normal pdf satisfying* (*A*3) *or a double-exponential pdf. Set* $a_1 < 1$*,* $a_2 < 1-(1/p-1/2)/[\alpha(r+1/2)(r+1/2-1/p)]$*, and* $a_3 > 1+1/[\alpha(r+1/2)]$ *in* (3.11)*, and choose* $\varepsilon > 1/(\alpha r(1-a_2))$ *in* (3.10)*, where* α *is defined by* (3.12)*. Then*

$$R_n(B^r_{p,q}(A), \hat{f}) = O\left(n^{-\frac{2r}{2r+1}} \ (\ln n)^{1-\frac{p^*}{2}}\right) \quad as \ \ n \to \infty. \tag{3.19}$$

Note that the situation when both η_j and ξ have polynomial descents is covered by Theorem 1. Hence, the only case that we have not discussed so far is when $\lambda_\eta > 0$ but (3.13) is invalid, which occurs when the η_j are normal *pdf*'s and ξ has a polynomial descent. In this situation, we cannot guarantee that the threshold t_{jn} satisfies (3.13) and have to use higher values for $\beta_{j,n}$ at high resolution levels in order to achieve optimal convergence rates. In this case, we have the following result.

Theorem 2. *Let the assumptions of Theorem* 1 *on* ϕ*,* ψ*,* s*,* L*,* f*,* r*,* p*,* q *and* μ_j *hold, and choose* ξ *and* η_j *as discussed in Section* 3.1 *to satisfy* (A1)−(A4) *with* $\lambda_\eta > 0$ *and* $\lambda_\xi = 0$*. Let the* η_j *satisfy* (3.15)*, choose* $\beta_{j,n}$ *given by* (3.11) *with* $a_1 < 1$ *and* $a_2 < 1$ *for* $L \le j \le j_1$*, and take*

$$\beta_{j,n} = \exp(\frac{\nu_j}{\sqrt{n}}), \quad \textit{if} \quad j_1 < j \le J-1. \tag{3.20}$$

Choose $\varepsilon > (\alpha r(1-a_2))^{-1} I(\beta = 2)$ *in* (3.10)*, where* α *is defined by* (3.12)*. Then* $R_n(B^r_{p,q}(\mathrm{A}), \hat{f})$ *is given by* (3.16)*.*

Corollary 3. *Let the assumptions of Theorem* 1 *on* ϕ*,* ψ*,* s*,* L*,* f*,* r*,* p*,* q *and* μ_j *hold. Let* η_j *be normal pdf's satisfying* (3.15)*, and let* ξ *be the Student-t pdf. Choose* $\beta_{j,n}$ *given by* (3.11) *with* $a_1 < 1$ *and* $a_2 < 1$ *for* $L \le j \le j_1$*, and by* (3.20) *for* $j_1 < j \le J-1$*. Let* $\varepsilon > (\alpha r(1-a_2))^{-1}$ *in* (3.10)*, where* α *is defined by* (3.12)*. Then* $R_n(B^r_{p,q}(A), \hat{f})$ *is given by* (3.19)*.*

3.4. Does $f \in B^r_{pq}$ *a-priori* with probability one?

In Section 3.3, the conditions to achieve optimal convergence rates were mainly concerned with the choice of the error model η_j. The main assertion about the prior model ξ was that it should not have faster descent at $\pm\infty$ than the error model η_j. However, it is the behaviour of the prior model ξ that

determines whether the regression function f belongs to a Besov space *a-priori* with probability one. Namely, the following sufficient statement proved in Pensky (2006) is valid.

Theorem 3. *Let* $p^* = \min(p, 2)$, $\max(r_p, 1/p^*) < r < s$, $1 \le p, q < \infty$, *and let* ξ, ν_j *and* $\beta_{j,n}$ *be such that*

$$\int_{-\infty}^{+\infty} |x|^{\max(p,q)}\, \xi(x)dx < \infty, \tag{3.21}$$

$$\lim_{n\to\infty} \sum_{j=L}^{J-1} \left[2^{j(r+1/2)}\, \beta_{j,n}^{-1/p} \nu_j^{-1}\right]^{\min(p,q)} < \infty. \tag{3.22}$$

Then $f \in B^r_{p,q}$ *with probability one.*

Since (3.21) requires the prior ξ to have at least $\max(p, q) \ge 1$ finite moments, it immediately eliminates the Cauchy prior from consideration. On the other hand, any prior ξ with exponential descent (e.g., double-exponential) ensures the validity of (3.21). The following result is an application of Theorem 3.

Corollary 4. *Let* ν_j *and* $\beta_{j,n}$ *be determined by* (3.9) *and* (3.11), *respectively, and assume that* ξ *satisfies* (3.21) *for* $1 \le p, q < \infty$. *Let* $p^* = \min(p, 2)$, *assume that* $\max(r_p, 1/p^*) < r < s$, *and take* $a_1 < 0$, $a_3 > 0$ *and*

$$a_2 < -p\varepsilon^{-1}[1 + (2r)^{-1} + \varepsilon], \tag{3.23}$$

where $\varepsilon > 0$ *is defined in* (3.10). *Then* $f \in B^r_{p,q}$ *with probability one.*

Observe that, if $p \ge 2$ (i.e., $p^* = 2$), the medium resolution levels collapse (i.e., $j_0 = j_1$), and the only assumptions we have are $a_1 < 0$ and $a_3 > 0$ (see (3.10) and (3.11)). Note also that Corollary 4 remains valid when $\beta_{j,n}$ has the form (3.20), as $j_1 < j \le J-1$. Hence, it is applicable for the conditions of Theorems 1 and 2 whenever $a_1 < 0$, $a_3 > 0$ and a_2 satisfies (3.23). For example, if one chooses $\varepsilon = (2r)^{-1}$ as in Pensky (2006), then (3.23) becomes $a_2 < -2p(r+1)$. The latter restriction can always be accomplished so that, whenever ξ has exponential descent, $f \in B^r_{p,q}$ with probability one for any values of $\max(r_p, 1/p^*) < r < s$ and $1 \le p, q < \infty$.

Corollary 4 provides a sufficient condition for overcoming the well-known Bayesian paradox, up to a $\log(n)$ factor, when a prior yielding an optimal Bayesian estimator over a certain class of functions (e.g., Sobolev or Besov spaces) lies outside this class. See, e.g., Zhao (2000). This is a stronger result than the statement made in Abramovich et al. (2004, Sec. 3) who considered only the case of the normal ξ model.

3.5. Adaptation: an empirical Bayes estimation procedure

If the errors Z_i in (1.1) are indeed $N(0,\sigma^2)$, we can choose $\eta_j(x) \equiv \eta(x) = (2\pi\sigma^2)^{-1/2}\exp\left(-x^2/(2\sigma^2)\right)$ and study an adaptive estimator, based on an empirical Bayes estimation procedure, similar to the one considered in Johnstone and Silverman (2005).

More specifically, at each resolution level $j = L,\dots,J-1$, we estimate the parameters ν_j and π_{jn} by maximizing the empirical likelihoods

$$l(\pi_{j,n},\nu_j) = \sum_{k=0}^{2^j-1} \ln\left\{(1-\pi_{j,n})\sqrt{n}\eta(\sqrt{n}d_{jk}) + \pi_{j,n}I_j(d_{jk})\right\}, \tag{3.24}$$

where $I_j(d_{jk})$ is defined in (2.6). Note that, by Lemma 1, the rule (2.7) can be expressed as the hard thresholding rule (2.8). Similarly to Johnstone and Silverman (2005), we want the threshold $t_{j,n}$ to satisfy $t_{j,n} \le (\sigma\sqrt{2\ln n})/\sqrt{n}$, which is equivalent to $\pi_{j,n} \ge \Omega_{j,n}$ with $\Omega_{j,n} = 1/[1+\zeta_{j,n}((\sigma\sqrt{2\ln n})/\sqrt{n})]$. Maximizing (3.24), we obtain the empirical Bayes estimators $\hat\pi_{j,n}$ and $\hat\nu_j$ of $\pi_{j,n}$ and ν_j, respectively, and set $\hat\beta_{j,n} = (1-\hat\pi_{j,n})/\hat\pi_{j,n}$. Choose now $D \ge 0$ and estimate θ_{jk} by using the hard thresholding rule with the modified threshold $\hat t_{j,n}$, i.e.,

$$\hat\theta_{jk} = d_{jk}\, I(|d_{jk}| > \hat t_{j,n}) \equiv d_{jk}\, I(\zeta_{j,n}(d_{jk}) > \hat\tau_{j,n}), \tag{3.25}$$

where

$$\begin{aligned} &\hat t_{j,n} = \zeta_{j,n}^{-1}(\hat\beta_{j,n}), && \hat\tau_{j,n} = \hat\beta_{j,n}, && \text{if } \hat\pi_{j,n} \ge \Omega_{j,n},\\ &\hat t_{j,n} = (\sigma\sqrt{2(1+D)\ln n})n^{-\frac12}, && \hat\tau_{j,n} = \zeta_{j,n}(\sigma\sqrt{2(1+D)\ln n}(n^{-\frac12})), && \text{if } \hat\pi_{j,n} < \Omega_{j,n}. \end{aligned} \tag{3.26}$$

Since the relationship of the posterior median and Bayes factor estimators to a pseudo-threshold defined in Johnstone and Silverman (2004a, 2005) (which serves as the main instrument of their proof) is similar, we modify Theorem 2 of Johnstone and Silverman (2005) to accommodate the Bayes factor estimator. Hence, we have the following result.

Theorem 4. *Let $\{\phi,\psi,s,L\}$ be as in (A0), and let $f \in B^r_{p,q}(A)$ with $1 \le p,q \le \infty$, $p^* = \min(p,2)$ and $\max(r_p, 1/p^*) < r < s$. Let $\hat\theta_{jk}$ be defined by (3.25) with the modified threshold $\hat t_{j,n}$ given by (3.26). Let $D \ge 0$ in (3.26), and let ξ be such that $\lambda_\xi = 0$ in (3.3), $\sup_x(x^2\xi(x)) < \infty$, and*

$$0 < C_1 \le x^{1-\kappa}[\xi(y)]^{-1}\int_y^{+\infty}\xi(x)dx \le C_2 \tag{3.27}$$

for some $C_1, C_2 > 0$, $\kappa \in [1,2]$, and y large enough. Then

$$R_n(B^r_{p,q}(A),\hat f) = O\left(n^{-\frac{2r}{2r+1}}\right) \quad as\ \ n\to\infty. \tag{3.28}$$

As in Johnstone and Silverman (2004a, 2005), the assumption that ξ is a heavy-tailed prior implies that the tails of ξ are exponential or heavier while assumption $\sup_x(x^2\xi(x)) < \infty$ rules out tail behavior heavier than that of a Cauchy. The condition (3.27) is just a mild regularity condition which is valid, e.g., for the double-exponential and Student-t priors used in our development; it is also valid if the quasi-Cauchy prior considered in Johnstone and Silverman (2004a, 2005) is used.

4. Numerical Results and Comparisons

In this section, we illustrate the performance of the empirical Bayes estimation procedure based on the proposed Bayes factor (BF) estimator, and compare it with the empirical Bayes estimation procedures based on the posterior mean (PostMean) and posterior median (PostMed) estimators proposed in Johnstone and Silverman (2005). Simulated samples and a dataset collected in an anaesthesiological study are used for this purpose. The computational algorithms related to wavelet analysis were performed using the WaveLab software (http://www-stat.stanford.edu/software/software.html) and the MatLab version of the EbayesThresh software (http://www-lmc.imag.fr/lmc-sms/Anestis.Antoniadis/EBayesThresh). The entire study was carried out using the Matlab programming environment.

4.1. Simulation study

For the PostMean and PostMed estimators, both double-exponential and quasi-Cauchy priors with normal error were used, while for BF estimators, double-exponential priors with normal errors were considered. All the prior parameters were estimated level-by-level by marginal maximum likelihood from the data. The standard deviation σ of the normal error was estimated by the *median absolute deviation* (as suggested by Donoho and Johnstone (1994) and usually applied in practice), i.e.,

$$\hat{\sigma} = \frac{\text{median}(|\{d_{J-1,k} : k = 0, 1, \ldots, 2^{J-1} - 1\}|)}{0.6745}.$$

In this simulation study, we evaluated the various empirical Bayes wavelet estimators using Daubechies's compactly supported wavelets *Symmlet 8* (see Daubechies (1992, p.198)) and *Coiflet 3* (see Daubechies (1992, p.258)), and primary resolution levels $L = 3$ and 5. We considered various test functions that are standard tests for wavelet estimators. See, e.g., the list of test functions of Section 5 and Appendix I in Antoniadis, Bigot and Sapatinas (2001). For each test function, $M = 500$ samples were generated by adding independent random noise $\varepsilon \sim N(0, \sigma^2)$ to $n = 256$, 512 and 1,024 equally spaced points on [0,1],

representing a range of sample sizes, from low to high. The value of σ was taken to correspond to the values 3 (high noise level), 5 (moderate noise level) and 7 (low noise level) for the (root) signal-to-noise ratio (SNR)

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

The goodness-of-fit for an estimator $\hat{g}$ of g was measured by (a) its average mean squared error (AMSE) from the M simulations, and (b) its average maximal absolute deviation (AMXD) from the M simulations, defined respectively as

$$\mathrm{AMSE}(g) = \frac{1}{nM}\sum_{m=1}^{M}\sum_{i=1}^{n}(\hat{g}_m(t_i)-g(t_i))^2$$

$$\text{and} \quad \mathrm{AMXD}(g) = \frac{1}{M}\sum_{m=1}^{M}\max_{1\le i\le n}|\hat{g}_m(t_i)-g(t_i)|.$$

For brevity, we only report in detail the results for the *Bumps* function using *Symmlet 8* and $L = 3$, and in order to investigate the performance of peak heights, we calculated AMSE and AMXD over the 11 peaks encountered in *Bumps*. Figures 1 and 2 contains the results of this simulation study. As observed in the figures, PostMean and PostMed estimators with quasi-Cauchy priors perform better than the corresponding estimators with double-exponential priors with respect to AMSE in almost all cases, while PostMean estimators perform better that PostMed estimators with respect to AMXD in almost all cases. The BF estimators with double-exponential priors have smaller AMSE and AMXD than PostMean and PostMed estimators with either double-exponential or quasi-Cauchy priors, indicating that they preserve peak heights better. Although not reproduced here, PostMean and PostMed estimators have smaller overall AMSE and AMXD than BF estimators, since they appear to have slightly better reconstructions over the regions where the underlying function is smooth; this is not surprising since PostMean is a (nonlinear) shrinkage rule and PostMed is a soft thresholding rule. On the other hand, although the BF estimators appear to be slightly noisier over the regions where the underlying function is smooth, they are not noisy enough to be visually unpleasant. Moreover, different combinations of test functions with similar characteristics (e.g., *Spikes*), wavelet functions and primary resolution levels yield basically similar results.

In summary, there is evidence that the BF estimator preserves the peak heights better without any substantial cost of inferior treatment of presumably spurious variation elsewhere, and sheds some light on the supposition of

Johnstone and Silverman (2005, p.1712) that hard thresholding rules with suitably estimated thresholds may have computational advantages and may preserve peak heights better in finite sample situations.

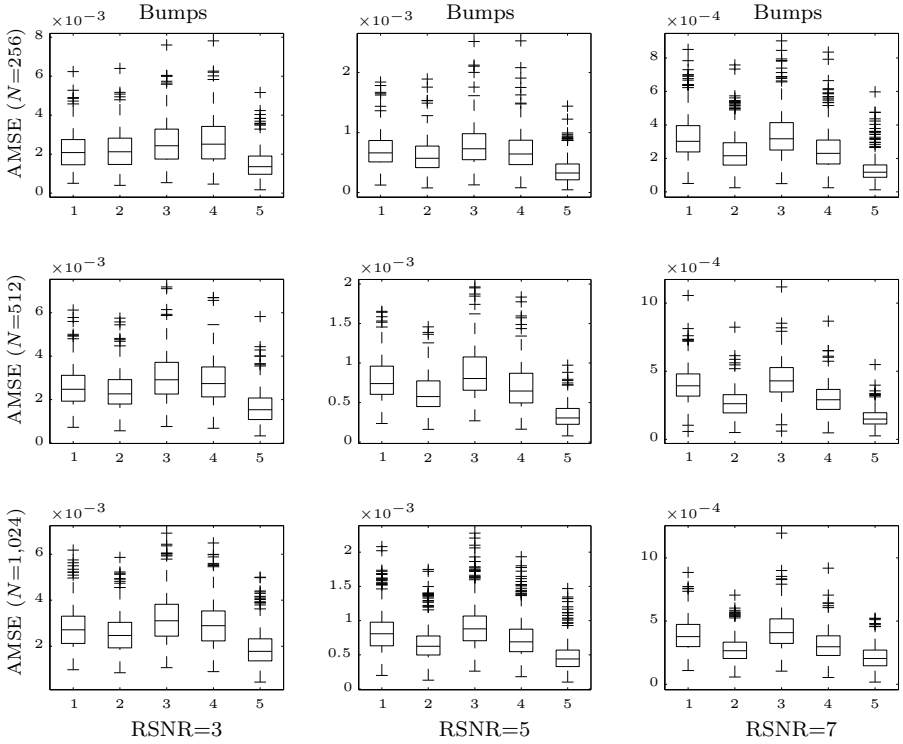

Figure 1. Boxplots of 500 simulation results for the *Bumps* function for all nine combinations of signal points (first row: 256; second row: 512; third row: 1,024) and RSNR's (left column: 3; middle column: 5; right column: 7). In each panel, there are five boxplots indicating the AMSE over the 11 peaks, from left to right, for the estimates produced by (1) PostMean (double-exponential prior), (2) PostMean (quasi-Cauchy prior), (3) PostMed (double-exponential prior), (4) PostMed (quasi-Cauchy prior), (5) BF estimators. See Section 4.1 for more details.

4.2. Inductance plethysmography data

We now consider a dataset from anaesthesiology collected by inductance plethysmography to illustrate the performance of the BF estimator, and to compare it with the PostMean and PostMed estimators. The recordings were made by the Department of Anaesthesia at the Bristol Royal Infirmary and measure the flow of air during breathing. See, e.g., Nason (1996).

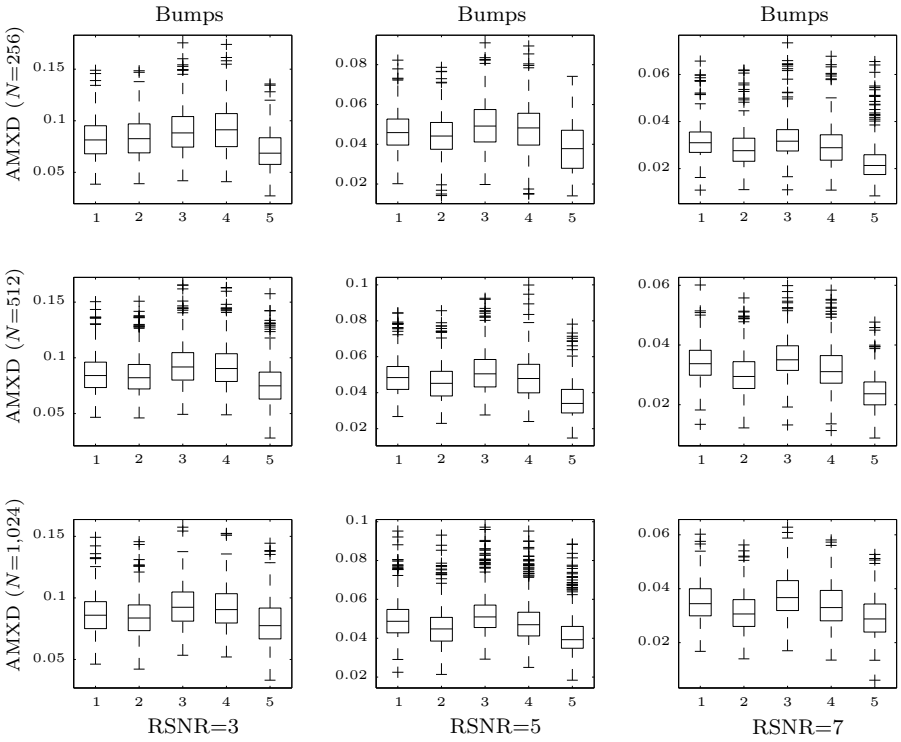

Figure 2. Boxplots of 500 simulation results for the *Bumps* function for all nine combinations of signal points (first row: 256; second row: 512; third row: 1,024) and RSNR's (left column: 3; middle column: 5; right column: 7). In each panel, there are five boxplots indicating the AMXD over the 11 peaks, from left to right, for the estimates produced by (1) PostMean (double-exponential prior), (2) PostMean (quasi-Cauchy prior), (3) PostMed (double-exponential prior), (4) PostMed (quasi-Cauchy prior), (5) BF estimators. See Section 4.1 for more details.

Figure 3 shows a section of plethysmograph recording lasting approximately 80 seconds ($n = 4,096$ signal points). The two main sets of regular oscillations correspond to normal breathing. The disturbed behaviour in the centre of the plot, where the normal breathing pattern disappears, corresponds to the patient vomiting. The same figure contains the curve estimates obtained using the BF, PostMean and PostMed estimators, all with double-exponential prior and normal error models, as suggested by Johnstone and Silverman (2005, p.1718). All the prior parameters were estimated level-by-level by marginal maximum likelihood from the data. The standard deviation σ of the normal error was estimated by the *median absolute deviation* at the finest resolution level. The various estimators were evaluated using Daubechies's compactly supported wavelets *Symmlet* 8

(see Daubechies (1992, p.198)) and *Coiflet* 3 (see Daubechies (1992, p.258)). For all methods, the primary resolution level was $L = 3$.

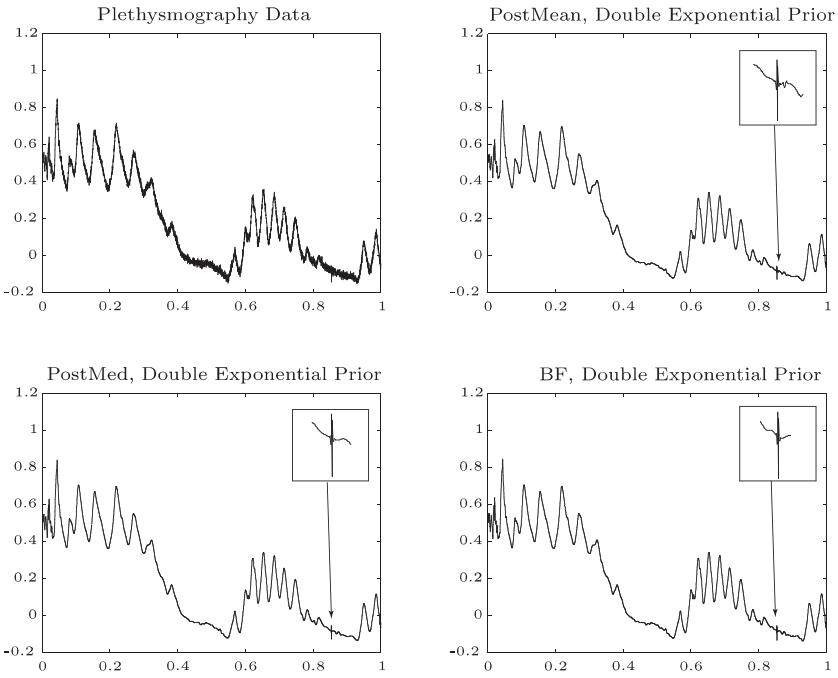

Figure 3. Section of an inductance plethysmography recording lasting approximately 80 seconds (Top Left), and smooth estimates obtained using the PostMean (Top Right), PostMed (Bottom Left), and BF (Bottom Right), all with double-exponential priors. See Section 4.2 for more details.

Quoting Johnstone and Silverman (2005), "... the adaptive smoothing with data of this kind is to preserve features such as peak heights as far as possible, while eliminating spurious rapid variations elsewhere ...". As in Johnstone and Silverman (2005), we have judged the efficacy of the various estimation methods in preserving peak heights simply by looking at the maximum of the various estimates of the height of the first peak in the inductance plethysmography curve. For *Symmlet 8*, the PostMean and PostMed estimators yield the same maximum value of 0.840, while the BF estimator gave 0.845. Similarly, as in Johnstone and Silverman (2005), we have quantified the efficacy of the various estimation methods in dealing with the rapid variation near the point 0.85 (on the x-axis) by the range of the estimated curves over a small interval at this point. The PostMean and PostMed estimators both have a 'glitch' of range 0.073, while the corresponding one for the BF estimator is 0.080. Similar results in magnitude hold for *Coiflet 3*. The PostMean and PostMed estimators yield the same maximum value of 0.830, while the BF estimator gave 0.835. For the

rapid variation near the point 0.85 (on the x-axis), the PostMean and PostMed estimators have a 'glitch' with ranges 0.065 and 0.064, respectively, while the corresponding one for the BF estimator is 0.071. Although we do not reproduce them here, similar results hold when 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 competes well with the PostMean and PostMed estimators on preserving the peak height without any substantial cost of inferior treatment of presumably spurious variation elsewhere.

5. Discussion and Concluding Remarks

Table 1 summarizes the optimality comparison of the various Bayes factor estimators carried out in Section 3. We assume that $\eta_j \equiv \eta$ and consider three choices for the error η and the prior ξ: the normal, the double-exponential and the Student-t distributions. The choices of ξ and η are listed in the first row and in the first column, respectively.

Table 1 presents (asymptotic) expressions for $\Delta = n^{2r/(2r+1)}\, R(n, B^r_{pq}(A))$; hence Δ shows deviation from the "ideal" rate $O\left(n^{-2r/(2r+1)}\right)$. We intentionally do not consider the cases when η has heavier tails than ξ. For these cases, Pensky (2006) showed that for the posterior mean estimators one needs additional assumptions on $\beta_{j,n}$, leading to the situation where f does not belong to the appropriate Besov space with probability one. Moreover, when ξ is the normal *pdf* and η has a heavy-tailed *pdf*, the posterior mean estimators exhibit non-optimal behavior due to a large bias. These combinations of η and ξ are located in the left lower corner of Table 1, left blank.

Table 1 shows that all estimators are optimal if $p \geq 2$ (i.e., $p^* = 2$) while they are optimal up to a logarithmic factor $(\ln n)^{1-p/2}$ when $1 \leq p < 2$ and η is the normal pdf. The $(\ln n)^{1-p/2}$ factor is unavoidable when any fixed threshold is applied to individual wavelet coefficients. See Donoho and Johnstone (1994) and Cai (1999). When η is the double-exponential *pdf*, the corresponding estimators are optimal up to a slightly larger $(\ln n)^{2-p}$ factor. On the other hand if η is the Student-t *pdf*, the corresponding estimators show sub-optimal behavior in spatially nonhomogeneous Besov spaces.

We point out that the optimality results for the Bayes factor estimators discussed above are derived under milder conditions than the corresponding ones for the posterior mean estimators. See Pensky (2006, Table 1): if the posterior mean is used as an estimator, then for the normal η – normal ξ model one needs the restriction that η should have variance which is more than double the variance of the actual error. For the Bayes factor estimators, we only relate the variances of η and of the actual error by (3.12), so that one is able to set variances as $\alpha = 1$. The latter puts restrictions on the values of a_2 in Theorem

1 and Corollary 1. The most immediate consequence of the latter is that the normal η – normal ξ model delivers optimal estimators for any $p \geq 2$, and optimal up to the factor $(\ln n)^{1-p/2}$ for any $1 \leq p < 2$, contrary to the conclusion of Abramovich et al. (2004). The reason for the lack of optimality is that they chose $\pi_{j,n} = \min(1, c2^{-\beta j})$, for some constant $c > 0$, where $\beta \geq 0$ is now a constant independent of n, following the work of Abramovich et al. (1998). Note that this is a highly restrictive assumption that is not true, e.g., for empirical Bayes estimators where $\hat{\pi}_{j,n}$ depends on n. See Section 3.5.

Table 1. Frequentist optimality comparison of Bayes factor estimators.

$\xi(x)$ $\eta(x)$	Normal $\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right)$	Double-Exponential $\frac{1}{2\sigma_1} \exp\left(-\frac{\lvert x\rvert}{\sigma_1}\right)$	Student-t $\frac{\Gamma\left(\frac{\nu_1+1}{2}\right)(\nu_1\pi)^{-\frac{1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\left(1+\frac{x^2}{\nu_1}\right)^{\frac{\nu_1+1}{2}}}$
Normal $\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{x^2}{2\sigma_0^2}\right)$	$\Delta = O\left((\ln n)^{1-\frac{p^*}{2}}\right)$ if $\sigma_0 \leq \sigma_1$	$\Delta = O\left((\ln n)^{1-\frac{p^*}{2}}\right)$	$\Delta = O\left((\ln n)^{1-\frac{p^*}{2}}\right)$ need (3.20)
Double-Exponential $\frac{1}{2\sigma_0} \exp\left(-\frac{\lvert x\rvert}{\sigma_0}\right)$		$\Delta = O\left((\ln n)^{2-p^*}\right)$ if $\sigma_0 \leq \sigma_1$	$\Delta = O\left((\ln n)^{2-p^*}\right)$
Student-t $\frac{\Gamma\left(\frac{\nu_0+1}{2}\right)(\nu_0\pi)^{-\frac{1}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)\left(1+\frac{x^2}{\nu_0}\right)^{\frac{\nu_0+1}{2}}}$			$\Delta = O\left(n^{\kappa}\right)$ $\kappa = \frac{(1-a_2)(2-p^*)^2}{2p^*(\nu_0+1)}\left(\frac{1}{2r} + \frac{\varepsilon}{2r+1}\right)$ if $\nu_0 > \nu_1$
$\Delta = n^{2r/(2r+1)} R_n(B_{pq}^r(A), \hat{f})$; ε and $a_2 < 1$ are defined by (3.10) and (3.11), respectively; $p^* = \min(p, 2)$, $1 \leq p \leq \infty$, $r > 0$.			

The only model in Table 1 which produces sub-optimal estimators when $1 \leq p < 2$ is the Student-t η – Student-t ξ model. Note that the deviation from optimality is $O(n^{\kappa})$, where κ is smaller when ϱ and r are larger and a_2 is closer to one. We, however, cannot take $a_2 = 1$ and, moreover, if $a_2 > 0$, then we cannot guarantee that $f \in B_{p,q}^r$ with probability one. On top of that, ξ should have $\max(p, q)$ finite moments for $f \in B_{p,q}^r$ with probability one. Hence, this combination of models should be avoided.

Let us now compare the sensitivity of the various models to deviations from normality in the actual error. It is easy to see that only the normal η – Student-t ξ model requires the stronger assumption (3.20) on $\beta_{j,n}$. This is quite different from the posterior mean case where models with normal η – heavy-tailed ξ were all sensitive to deviations from normality in the actual error and required very strong conditions on $\beta_{j,n}$, see Pensky (2006). The consequence of (3.20) is that

"almost all" coefficients at high resolution levels will be "killed" because $\beta_{j,n}$ is very large. Therefore, Bayes factor estimators are less sensitive to the deviation from normality in the actual error.

One more observation concerning the Bayes factor estimators is the following: Bayes factor estimators produce better results than the corresponding posterior mean ones. Although the results are similar in both cases, the logarithmic factors are smaller for both normal η and double-exponential η models when Bayes factor estimators are used, and so is the polynomial term $O(n^{\kappa})$ in the case of the Student-t η – Student-t ξ model. Last but not least, an empirical Bayes estimation procedure based on the proposed Bayes factor estimator apparently beats all the corresponding Bayesian estimators in terms of optimal convergence rates (since the parameters ν_j and $\beta_{j,n}$ in the prior are derived from the random sample).

We conclude the discussion by pointing out the suggested choice: it is the normal η – double-exponential ξ model. This model provides the best rates of convergence for Bayes factor estimators (Theorem 1), it is not sensitive to deviation from normality in the actual error (Corollary 1), it ensures that $f \in B^r_{p,q}$ with probability one (Corollary 4), and it is suitable for an (adaptive) empirical Bayes estimation procedure (Theorem 4). On top of that, it allows us to perform all calculations explicitly in closed form. See Johnstone and Silverman (2004a) and Bochkina and Sapatinas (2005). Moreover, as was illustrated in simulated examples as well as in a dataset, it preserves peak heights better, without any substantial cost of inferior treatment of presumably spurious variation elsewhere, see Section 4.

Acknowledgements

Marianna Pensky was supported in part by the National Science Foundation (NSF) Grant DMS-0505133. Theofanis Sapatinas would like to thank Marianna Pensky for warm hospitality while visiting Orlando to carry out part of this work. We would also like to thank Anestis Antoniadis for helping us with the use of the EbayesThresh MatLab software and for producing Figure 3. The authors are grateful to Professors Xia-Li Meng and Michelle Liou (Co-Editors), an associate editor and two referees for many useful comments and suggestions.

Appendix. Proofs

A.1. The Bayes factor estimator as a thresholding rule

Proof of Lemma 1. For the sake of convenience, we drop the indices in $\zeta_{j,n}(d_{jk})$, $I_j(d_{jk})$ and η_j. Let $F(x) = \ln(\zeta(x))$ and observe that

$$F'(x) = \frac{n}{I(x)} \int_{-\infty}^{+\infty} \left[\frac{\eta'(\sqrt{n}(x-\theta))}{\eta(\sqrt{n}(x-\theta))} - \frac{\eta'(\sqrt{n}x)}{\eta(\sqrt{n}x)} \right] \eta(\sqrt{n}(x-\theta)) \nu_j \xi(\nu_j \theta) d\theta. \quad \text{(A.1)}$$

If η is the $N(0,\sigma^2)$ *pdf*, then the expression in the square brackets in (A.1) is $\sqrt{n}\theta/\sigma^2$, so that the integral is positive for $x > 0$. Hence, both $F(x)$ and $\zeta(x)$ are strictly increasing for $x > 0$. Similarly, if $\eta(x) = (2\sigma)^{-1}\exp(-|x|/\sigma)$, then the expression in square brackets in (A.1) is $2I(\theta \geq x)/\sigma$, and $F'(x) > 0$.

If both η and ξ are Student-t *pdf*'s with a and b degrees of freedom, respectively, with $a > b$, then $\zeta(x) \leq 1 + C_a n/\nu_j^2$ as $\sqrt{n}/\nu_j \to 0$ by Lemma 2. If $\nu_j/\sqrt{n} \to 0$, we use the asymptotic expansion of Lemma 2 to find $\zeta(x) = \nu_j\xi(\nu_j x)[\sqrt{n}\eta(\sqrt{n}x)]^{-1}\,[1+O(\nu_j^2/n)]$; hence the behaviour of $\zeta(x)$ coincides with that of $q(x) = \xi(\nu_j x)/\eta(\sqrt{n}x)$. By direct calculations, it can be verified that $q'(x) > 0$ for $x > 0$ whenever $(a+1)n(b+\nu_j^2x^2) - (b+1)\nu_j^2(a+nx^2) = n\nu_j^2x^2(a-b) + (ab(n-\nu_j^2 + (bn - a\nu_j^2) > 0$, which is guaranteed by the conditions of Lemma 1.

A.2. Asymptotics of the thresholds

Asymptotics of the thresholds rely on Lemmas 2 and 3. The first one is Lemma 2 in Pensky (2006), while the proof of the second follows easily, so we omit the details.

Lemma 2. *If $\nu_j|d_{jk}|$ is bounded or $\nu_j|\nu_j d_{jk}|^{\lambda_\xi}/\sqrt{n} \to 0$, then for any d_{jk},*

$$I_j(d_{jk}) = \nu_j\xi(\nu_j d_{jk})\left[1 + O\left(n^{-1}\nu_j^2|\nu_j d_{jk}|^{2\lambda_\xi}\right)\right] \qquad as \quad \frac{\nu_j}{\sqrt{n}} \to 0.$$

If $\sqrt{n}|d_{jk}|$ is bounded or $\sqrt{n}|\sqrt{n}d_{jk}|^{\lambda_\eta}/\nu_j \to 0$, then for any d_{jk},

$$I_j(d_{jk}) = \sqrt{n}\,\eta_j(\sqrt{n}d_{jk})\left[1 + O\left(n\nu_j^{-2}|\sqrt{n}d_{jk}|^{2\lambda_\eta}\right)\right] \quad as \quad \frac{\sqrt{n}}{\nu_j} \to 0.$$

Lemma 3. *Let $F_1(x)$ and $F_2(x)$ be strictly increasing functions of x for $x > 0$ such that $F_1(x) \leq F_2(x)$. Let t_i be the solutions of the equations $F_i(x) = b$, $i = 1, 2$. Then, $t_1 \geq t_2$.*

Lemma 4. *Let* (A1)–(A3) *hold and let $\nu_j/\sqrt{n} \to 0$. Then the following hold.*

(i) *If the η_j have pdf's of the form* (3.15)*, and $\lambda_\xi = 0$ in* (3.3) *or ξ is a normal pdf, then the threshold $t_{j,n}$ satisfies*

$$\sqrt{n}\,t_{j,n} = O\left(\left[\ln\left(\nu_j^{-1}\beta_{j,n}\sqrt{n}\right)\right]^{\frac{1}{\beta}}\right) \quad as \;\; n \to \infty. \tag{A.2}$$

(ii) *If the η_j have pdf's of the form* (3.17)*, $\xi(x) = C(1+x^2)^{-\gamma}$ for some constant $C > 0$ with $\gamma < \varrho$, and $\beta_{j,n} = (\nu_j/\sqrt{n})^a$ with $a < 1$, then the threshold $t_{j,n}$*

satisfies

$$\sqrt{n}\,t_{j,n} \sim \begin{cases} (\beta_{j,n}\frac{\sqrt{n}}{\nu_j})^{\frac{1}{2\varrho}}, & \text{if } \ 1-2\varrho \le a < 1, \\ \beta_{j,n}^{\frac{1}{2(\varrho-\gamma)}}\left(\frac{\nu_j}{\sqrt{n}}\right)^{\frac{2\gamma-1}{2(\varrho-\gamma)}}, & \text{if } \ a < 1-2\varrho. \end{cases} \tag{A.3}$$

Proof of Lemma 4. It follows from Lemma 3 that (A.2) and (A.3) will be valid if we replace $\zeta_{j,n}$ by asymptotic lower bounds. For the sake of convenience, we drop the indices in $\zeta_{j,n}$, η_j and σ_j (so that σ here is not the actual standard deviation of the error but just some number between $\underline{\sigma}$ and $\bar{\sigma}$).

(i) When η have exponential descents and ξ is the standard normal *pdf*, then (see Abramovich et al. (2004))

$$\sqrt{n}\,t_{j,n} = \sigma\sqrt{2}\sqrt{1+n^{-1}\sigma^2\nu_j^2}\left[\ln\left((\nu_j\sigma)^{-1}\beta_{j,n}\sqrt{n+\nu_j^2\sigma^2}\right)\right]^{\frac{1}{2}}, \tag{A.4}$$

so that (A.2) is valid. On the other hand if $\lambda_\xi = 0$, then by Lemma 2, $\zeta(x) \ge C_1[\sqrt{n}\eta(\sqrt{n}x)]^{-1}\nu_j\xi(\nu_j x)$ for some constant $C_1 > 0$. Taking into account (A3), we find $\zeta(x) \ge C_2 n^{-1/2}\nu_j \exp\left\{\left[1-(\nu_j/\sqrt{n})^\beta\right]|x\sqrt{n}/\sigma|^\beta\right\}$ for some constant $C_2 > 0$. Note that since $\nu_j/\sqrt{n} \to 0$, we have $1-\nu_j/\sqrt{n} \ge 1/2$, so that $t_{j,n}$ is of the form (A.2).

(ii) When η and ξ have polynomial descents, then for some constant $C_3 > 0$, $\zeta(x) = C_3 n^{-1/2}\nu_j F(x)$ with $F(x) = (1+\nu_j^2x^2)^{-\gamma}(1+nx^2)^\varrho$. Observe now that

$$F(x) \sim \begin{cases} 1, & |x| < \frac{1}{\sqrt{n}}, \\ (nx^2)^\varrho, & \frac{1}{\sqrt{n}} \le |x| \le \frac{1}{\nu_j}, \\ (nx^2)^\varrho(\nu_j^2x^2)^{-\gamma}, & |x| > \frac{1}{\nu_j}. \end{cases} \tag{A.5}$$

Equating $F(x)$ to $C_3^{-1}\sqrt{n}\nu_j^{-1}\beta_{j,n}$ and rewriting the constraints in (A.5) in terms of $\beta_{j,n}$, we arrive at (A.3).

Lemma 5. *Let* (A1)–(A4) *hold and let* $\sqrt{n}/\nu_j \to 0$. *If* $\beta_{j,n} > 2$, *then the following hold.*

(i) *If* $\lambda_\eta = 0$ *in* (3.4), *then the threshold* $t_{j,n} = \infty$.

(ii) *If the* η_j *are normal pdf's satisfying* (3.15) *and* ξ *is either a normal or a double-exponential pdf, then the threshold* $t_{j,n}$ *satisfies* $\sqrt{n}\,t_{j,n} \ge C_t\ \nu_j/\sqrt{n}$, *where the constant* $C_t > 0$ *is independent of* n *and* ν_j.

(iii) *If the* η_j *are normal pdf's satisfying* (3.15) *and* ξ *is a symmetric pdf on* $\mathbb{R}$ *that is unimodal, positive and finite at zero, then the threshold* $t_{j,n}$ *satisfies*

$$\sqrt{n}\,t_{j,n} \ge C_t\ \max\left\{\sqrt{\ln(\beta_{j,n}\frac{\sqrt{n}}{\nu_j})}, \sqrt{\ln(\beta_{j,n})}\right\}, \tag{A.6}$$

where the constant $C_t > 0$ *independent of* n *and* ν_j.

Proof of Lemma 5. (i) follows directly from Lemma 2. (ii) If both η_j and ξ are normal *pdf*'s, then (3.13) follows from (A.4). If the η_j are the $N(0,\sigma^2)$ *pdf*'s and ξ is a double-exponential *pdf* then, using formulae 3.322.2 and 8.254 of Gradshtein and Ryzhik (1980), whenever $|x| < \nu_j\sigma^2/n$ we have $\zeta_{j,n}(x) \leq \nu_j^2(\nu_j^2 - nx^2/\sigma^4)^{-1}$. Equating the last expression to $\beta_{j,n}$, we find that $|x| = \sigma^2\nu_j n^{-1}\sqrt{1-1/\beta_{j,n}} > \sigma^2\nu_j[n\sqrt{2}]^{-1}$. If $|x| \geq \nu_j\sigma^2/n$, then all existing solutions of the equation $\zeta_{j,n}(x) = \beta_{j,n}$ will satisfy (3.13) with $C_t = \sigma^2$. (iii) Note that $\xi(x) \leq \xi(0)$, so that $\zeta_{j,n}(x) \leq \nu_j\xi(0)/(\sqrt{n}\,\eta_j(\sqrt{n}x))$. On the other hand, it is not difficult to see that $\zeta_{j,n}(x) \leq \eta_j(0)/\eta_j(\sqrt{n}x))$. Application of Lemma 3, in view of (3.12), now yields (A.6).

A.3. Proofs of statements in Section 3

The proof of Theorem 1 requires the following lemma which is a straightforward extension of Lemma 4 in Pensky (2006).

Lemma 6. *Let* $\{\phi, \psi, s, L\}$ *be as in* (A0), *and let* $1 \leq p, q \leq \infty$, $p^* = \min(p, 2)$, *and* $\max(r_p, 1/p^*) < r < s$. *If* $f \in B^r_{p,q}(A)$ *then, for some constants* A_0, A_1, A_2, $A_3 > 0$, *we have*

$$\sum_{k\in K_{L-1}} (\tilde\theta_k - \theta_k)^2 \leq A_0\, n^{-2r} \quad \textit{and} \quad \sum_{j=L}^{J-1}\sum_{k=0}^{2^j-1} (\tilde\theta_{jk} - \theta_{jk})^2 \leq A_1\, n^{\frac{-2r}{(2r+1)}} \tag{A.7}$$

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

$$\sum_{k=0}^{2^j-1} \theta_{jk}^2 \leq A_2\, 2^{-2j(r+\frac{1}{2}-\frac{1}{p^*})} \quad \textit{and} \quad \sum_{k=0}^{2^j-1} |\theta_{jk}|^p \leq A_3\, 2^{-pj(r+\frac{1}{2}-\frac{1}{p^*})}. \tag{A.8}$$

Proof of Theorem 1. Since the wavelet basis is orthonormal,

$$R_n(B^r_{p,q}(A), \hat f) = \sum_{k\in K_{L-1}} E(\hat\theta_k - \tilde\theta_k)^2 + \sum_{j=L}^{J-1}\sum_{k=0}^{2^j-1} E(\hat\theta_{jk} - \tilde\theta_{jk})^2 + \sum_{j=J}^{\infty}\sum_{k=0}^{2^j-1} \tilde\theta_{jk}^2. \tag{A.9}$$

Observe that the first term in (A.9) is bounded by

$$2\sum_{k\in K_{L-1}} \left[\mathrm{Var}(\hat\theta_k) + (\theta_k - \tilde\theta_k)^2\right] = O(n^{-1}) + O(n^{-2r}) = o\left(n^{-\frac{2r}{2r+1}}\right)$$

due to (A.7), while the last term in (A.9) is bounded by $A2^{-2J(r+1/2-1/p^*)} = O(n^{-2r/(2r+1)})$ due to (3.7) and $r > r_p$. The second term in (A.9) is dominated by

$$\sum_{j=L}^{J-1}\sum_{k=0}^{2^j-1} E(\hat\theta_{jk} - \tilde\theta_{jk})^2 \leq 2\sum_{j=L}^{J-1}\sum_{k=0}^{2^j-1} E(\hat\theta_{jk} - \theta_{jk})^2 + A_1\, n^{\frac{-2r}{(2r+1)}} \tag{A.10}$$

due to (A.7). Thus, we need to evaluate the contribution to $R_n(B^r_{p,q}(A), \hat{f})$ made by the first term in (A.10). Let $R = \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2 = R_1 + R_2 + R_3$ with terms

$$R_1 = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2, \quad R_2 = \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2,$$
$$R_3 = \sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2,$$

corresponding to *low*, *medium* and *high* resolution levels, respectively. We examine each term in turn.

Low resolution levels. Since for $j \le j_1$ the errors are asymptotically normal (see (3.8)), we can use a result of Donoho and Johnstone (1994) to get

$$R_i = O\left(\sum_j \sum_{k=0}^{2^j-1} \left[(t^2_{j,n} + n^{-1}\sigma^2) I(|\theta_{jk}| \ge t_{j,n}) \right.\right.$$
$$\left.\left. + \ (\theta^2_{jk} + t_{j,n}\sigma n^{-1/2}\, \varphi(t_{j,n}\sqrt{n}\sigma^{-1})) I(|\theta_{jk}| < t_{j,n}) \right] \right), \quad i = 1, 2, \text{(A.11)}$$

where, as before, $\varphi(x)$ is the $N(0,1)$ *pdf* and the sum is taken over $L \le j \le j_0$ for $i = 1$, and $j_0 + 1 \le j \le j_1$ for $i = 2$. Hence, $R_1 = O(R_{11} + R_{12} + R_{13})$, where

$$R_{11} = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} n^{-1}\sigma^2, \quad R_{12} = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} t_{j,n}\sigma n^{\frac{-1}{2}} \varphi(t_{j,n}\frac{\sqrt{n}}{\sigma}),$$
$$R_{13} = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} \min(\theta^2_{jk}, t^2_{j,n}). \qquad \text{(A.12)}$$

Note that $R_{11} = O(2^{j_0} n^{-1}) = O\left(n^{-2r/(2r+1)}\right)$, and that $R_{12} = \sum_{j=L}^{j_0} 2^j \sigma^2 n^{-1} t_{j,n}$ $\sqrt{n}\sigma^{-1}\, \varphi(t_{j,n}\sqrt{n}\sigma^{-1}) = O\left(n^{-2r/(2r+1)}\right)$ since $|x\varphi(x)| < 1$ for any $x \in \mathbb{R}$. In order to derive an upper bound for R_{13}, observe that if $\beta_{j,n}$ is of the form (3.11) with $a_1 < 1$, then $\beta_{j,n}\sqrt{n}/\nu_j \to \infty$. If $\eta_j(x)$ is of the form (3.17) then, by Lemma 4, we have $t_{j,n} = O\left(n^{-1/2}(\beta_{j,n}\sqrt{n}/\nu_j)^{1/2\varrho}\right)$. Since the threshold is smaller when η_j has thinner tails, then for $\eta_j(x) < C(1+x^2)^{-\rho}$ we have $t_{j,n} = o\left(n^{-1/2}(\beta_{j,n}\sqrt{n}/\nu_j)^{1/2\varrho}\right)$. Hence, $R_{13} = O(\sum_{j=L}^{j_0} 2^j n^{-1}[n\nu_j^{-2}]^{(1-a_1)/(2\varrho)}) = O(n^{-1+(1-a_1)/(2\varrho)} 2^{j_0[2\varrho-(2r+1)(1-a_1)]/(2\varrho)}) = O(n^{-2r/(2r+1)})$ provided $a_1 > 1 - \varrho/(r+1/2)$. Note that when η_j have exponential descents, one can choose ϱ to be arbitrarily large, i.e., $\varrho/(r+1/2) > 1 - a_1$ is always true. Combining expressions for R_{11}, R_{12} and R_{13} we find that $R_1 = O\left(n^{-2r/(2r+1)}\right)$.

Medium resolution levels. Since for $p \geq 2$ the medium resolution levels collapse, we consider only the case $1 \leq p < 2$. Observe also that m_2 is chosen so that $\nu_j/\sqrt{n} \to 0$ when $j_0 + 1 \leq j \leq j_1$. Using (A.11), partition R_2 as $R_2 = O(R_{21} + R_{22} + R_{23})$ as at (A.12). Let $\gamma_{j,n} = \sqrt{n}t_{j,n}$ and observe that $\gamma_{j,n} \geq \gamma_0$, where $\gamma_0 > 0$ is a constant. Then using the second inequality in (A.8),

$$\begin{aligned} R_{21} &= \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} n^{-1}\sigma^2 I(|\theta_{jk}| > t_{j,n}) \\ &= O\Bigg(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} n^{-1+\frac{p}{2}} n^{-\frac{p}{2}} I\left(n^{-1} < \theta_{jk}^2 \gamma_{j,n}^{-2}\right) \Bigg) \\ &= O\Bigg(n^{-1+\frac{p}{2}} \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} |\theta_{jk}|^p \gamma_0^{-p} \Bigg) = O\left(n^{-1+\frac{p}{2}} 2^{-pj_0(r+\frac{1}{2}-\frac{1}{p})}\right) \\ &\quad O\left(n^{-2r/(2r+1)}\right). \end{aligned}$$

To examine R_{22}, note that for some constant $C > 0$ we have $[\eta_j(\sqrt{n}t_{j,n})]^{-1} \geq C[\eta_j(\sqrt{n}t_{j,n})]^{-1}\xi(\nu_j t_{j,n}) = C\nu_j^{-1}\sqrt{n}\beta_{j,n} = C(\nu_j^{-1}\sqrt{n})^{1-a}$. Hence, by (A4), we have $\varphi(t_{j,n}\sqrt{n}/\sigma) \leq \sigma C_\alpha [\eta_j(t_{j,n}\sqrt{n})]^\alpha = \sigma C_\alpha \beta_{j,n}^\alpha$, so that

$$R_{22} = O\Bigg(\sum_{j=j_0+1}^{j_1} 2^j n^{-1} \gamma_{j,n} \beta_{j,n}^\alpha \Bigg). \tag{A.13}$$

Consider two separate cases. If the η_j are of the form (3.17), then by (A.2), $\gamma_{j,n} = O([\ln n]^{1/\beta})$. Hence using (3.11), we rewrite (A.13) as

$$\begin{aligned} R_{22} &= O\Bigg(\sum_{j=j_0+1}^{j_1} n^{-1-\frac{\alpha(1-a_2)}{2}} [\ln n]^{\frac{1}{\beta}} 2^{j(1+m_2\alpha(1-a_2))} \Bigg) \\ &= O\left(n^{-1-\frac{\alpha(1-a_2)}{2}} [\ln n]^{\frac{1}{\beta}} 2^{j_1(1+m_2\alpha(1-a_2))}\right) = O\left(n^{-\frac{2r}{2r+1}}\right), \end{aligned}$$

provided

$$1 + \frac{\alpha(1-a_2)}{2} - \frac{r(1+m_2\alpha(1-a_2)}{(r+\frac{1}{2}-\frac{1}{p})(2r+1)} > \frac{2r}{2r+1},$$

which is equivalent to $m_2 < (r+1/2) - (1/p - 1/2)\left(1 + (2r)^{-1} + (\alpha r(1-a_2))^{-1}\right)$. Note that if $\beta < 2$ in (3.15), the value of α can be chosen arbitrary large, so that $\varepsilon > (\alpha r(1-a_2))^{-1}$ can be chosen arbitrary small. If the η_j have the polynomial descents (3.17), then $\gamma_{j,n} \sim (\beta_{j,n}\sqrt{n}/\nu_j)^{1/2\varrho}$ and $\varphi(\gamma_{j,n}/\sigma) = o(n^{-\upsilon})$ for any $\upsilon > 0$, so that (A.13) becomes $R_{22} = O\left(n^{-2r/(2r+1)}\right)$.

Let us now examine $R_{23} = R_{231} + R_{232}$, where

$$R_{231} = \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} \theta_{jk}^2 I(|\theta_{jk}| < t_{j,n}), \quad R_{232} = \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} t_{j,n}^2 I(|\theta_{jk}| \geq t_{j,n}). \tag{A.14}$$

To analyze R_{231}, observe that by Lemma 6,

$$R_{231} = O\Big(\sum_{j=j_0+1}^{j_1} t_{j,n}^{2-p} \sum_{k=0}^{2^j-1} |\theta_{jk}|^p \Big) = O\Big(\sum_{j=j_0+1}^{j_1} t_{j,n}^{2-p} 2^{-j(r+\frac{1}{2}-\frac{1}{p})p} \Big). \tag{A.15}$$

Consider again two cases. If the η_j are of the forms (3.15) then, by (A.2), $t_{j,n} = O(n^{-1/2}(\ln n)^{1/\beta})$, and (A.15) becomes

$$R_{231} = O\Big(\sum_{j=j_0+1}^{j_1} n^{-1+\frac{p}{2}} 2^{-j(r+\frac{1}{2}-\frac{1}{p})p} (\ln n)^{\frac{2-p}{\beta}} \Big) = O\left(n^{-\frac{2r}{2r+1}} (\ln n)^{\frac{2-p}{\beta}} \right). \tag{A.16}$$

If the η_j have the polynomial descents (3.17) and the $\beta_{j,n}$ are of the form (3.11) with $a_2 > 1 - 2\varrho$, then $t_{j,n} \sim n^{-1/2}(\sqrt{n}/\nu_j)^{(1-a_2)/(2\varrho)}$. Then

$$\begin{aligned} R_{231} &= O\Big(\sum_{j=j_0+1}^{j_1} n^{-1+\frac{p}{2}} 2^{-j(r+\frac{1}{2}-\frac{1}{p})p} \Big(\frac{\sqrt{n}}{\nu_j} \Big)^{\frac{(1-a_2)(2-p)}{2\varrho}} \Big) \\ &= O\left(n^{-1+\frac{p}{2}+\frac{(1-a_2)(2-p)}{4\varrho}} \, 2^{-j_0\left[\frac{m_2(1-a_2)(2-p)}{2\varrho} + p\left(r+\frac{1}{2}-\frac{1}{p}\right)\right]} \right) \\ &= O\left(n^{-\frac{2r}{2r+1}+\kappa} \right), \end{aligned} \tag{A.17}$$

where $\kappa = (4\varrho)^{-1}(1-a_2)(2-p)[1 - 2m_2/(2r+1)]$. Plugging m_2 given by (3.10) into κ results in (3.18).

To finish, we need to examine R_{232}. Recall that $\sqrt{n}t_{j,n} = \gamma_{j,n}$. By (A.14) and (A.15), we have

$$\begin{aligned} R_{232} &= \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} n^{-1} \gamma_{j,n}^2 I(\theta_{jk}^2 > \frac{\gamma_{j,n}^2}{n}) = \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} n^{-1+\frac{p}{2}} \gamma_{j,n}^2 (|\theta_{jk}| \gamma_{j,n}^{-1})^p \\ &= \sum_{j=j_0+1}^{j_1} t_{j,n}^{1-\frac{p}{2}} \sum_{k=0}^{2^j-1} |\theta_{jk}|^p = O\left(R_{231}\right). \end{aligned}$$

The latter implies that the overall error R_2 is of the form (A.16) or (A.17) when the η_j are given by (3.15) or (3.17), respectively.

High resolution levels. We need only consider the case $1 \le p < 2$. Note that if $j_1 + 1 \le j \le J - 1$, the errors are not normally distributed any more and we cannot use (A.11). Partition R_3 as $R_3 = R_{31} + R_{32}$, where

$$R_{31} = \sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} E\left[(d_{jk} - \theta_{jk})^2 I(\zeta_{j,n}(d_{jk}) \ge \beta_{j,n})\right],$$
$$R_{32} = \sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} \theta_{jk}^2 P(\zeta_{j,n}(d_{jk}) < \beta_{j,n}),$$

and consider two cases here. If $\lambda_\xi = 0$ then, by Lemma 2, $\zeta_{j,n}(d_{jk}) = O(1)$ while, by (3.11), we have $\beta_{j,n} \to \infty$, so that $\zeta_{j,n}(d_{jk}) < \beta_{j,n}$ for any d_{jk} and $R_{31} = 0$. Hence, in this case, by Lemma 6,

$$R_{32} = R_3 = O\left(\sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} \theta_{jk}^2\right) = O\left(\sum_{j=j_1+1}^{J-1} 2^{-2j(r+\frac{1}{2}-\frac{1}{p^*})}\right)$$
$$= O\left(n^{-\frac{2r}{2r+1}}\right). \tag{A.18}$$

If $\lambda_\xi > 0$ and (3.13) holds, then R_{32} has the form (A.18) as before, so that we need only derive an asymptotic expression for R_{31}.

Since $I(\zeta_{j,n}(d_{jk}) \ge \beta_{j,n}) = I(|d_{jk}| > t_{j,n}) \le I(|\theta_{jk}| > t_{j,n}/2) + I(|d_{jk} - \theta_{jk}| > t_{j,n}/2)$, we obtain $R_{31} = O(R_{311} + R_{312})$. Here, since by (A.6) and (3.11) we have $t_{j,n}\sqrt{n} > 1$, similarly to (A.18) we obtain

$$R_{311} = \sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} E(d_{jk} - \theta_{jk})^2 I(|\theta_{jk}| > \frac{t_{j,n}}{2})$$
$$= O\Big(\sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} n^{-1} I(n^{-1} < 2\theta_{jk}^2)\Big)$$
$$= O\Big(\sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} \theta_{jk}^2\Big) = O\Big(n^{-\frac{2r}{2r+1}}\Big). \tag{A.19}$$

For the second term, we have

$$R_{312} = \sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} E\left[(d_{jk} - \theta_{jk})^2 I(|d_{jk} - \theta_{jk}| > \frac{t_{j,n}}{2})\right]$$
$$= \sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} J_{jn}(d_{jk}), \tag{A.20}$$

where, by (3.13) and the assumptions of Theorem 1,

$$J_{jn}(d_{jk}) = \int_{-\infty}^{+\infty} z^2\sqrt{n}\mu_j(\sqrt{n}z)I(\sqrt{n}|z| \geq 0.5\, C_t\nu_j n^{-\frac{1}{2}})dz$$
$$= n^{-1}\Big(\frac{2\sqrt{n}}{C_t\nu_j}\Big)^{l-2} \int_{-\infty}^{+\infty} |y|^s\mu_j(y)dy = O\Big(n^{-1}\Big(\frac{\sqrt{n}}{\nu_j}\Big)^{l-2}\Big). \qquad \text{(A.21)}$$

Plugging (A.21) into (A.20), we obtain

$$R_{312} = O\Big(\sum_{j=j_1+1}^{J-1} 2^j n^{-1}\Big(\frac{\sqrt{n}}{\nu_j}\Big)^{l-2}\Big)$$
$$= O\Big(\sum_{j=j_1+1}^{J-1} n^{-\frac{2r}{2r+1}} n^{\frac{(l-2)(r+\frac{1}{2})-1}{2r+1}}\, 2^{j(1-(r+\frac{1}{2})(l-2))}\Big)$$
$$= O\Big(n^{-\frac{2r}{2r+1}} n^{\frac{(l-2)(r+\frac{1}{2})-1}{2r+1}}\, 2^{-j_1[(l-2)(r+\frac{1}{2})-1]}\Big) = O\left(n^{-\frac{2r}{2r+1}}\right), \qquad \text{(A.22)}$$

since $(l-2)(r+1/2) > 1$ by (3.14) and $2^{-j_1} < n^{-1/(2r+1)}$. Hence, by (A.19) and (A.22), $R_3 = O\left(n^{-2r/(2r+1)}\right)$. This completes the proof of Theorem 1.

Proof of Corollary 1. Note that the α in (A4) can be chosen as large as one wishes and $\lambda_\eta = 0$. Then validity of Corollary 1 follows directly from Theorem 1.

Proof of Corollary 2. It follows easily by using Lemma 5.

Proof of Theorem 2. The difference between the conditions of Theorem 2 and Theorem 1 is that (3.13) is no longer valid. Hence we need to re-examine the error at high resolution levels. Observe that the only part of the error which is affected by this is R_{312}. Note that by (3.20) and (A.6), similarly to (A.21), we find

$$J_{jn}(d_{jk}) = \int_{-\infty}^{+\infty} z^2\sqrt{n}\mu_j(\sqrt{n}z)I\left(\sqrt{n}|z| \geq 0.5\, C_t\sqrt{\ln(\beta_{j,n})}\right)dz$$
$$= O\left(n^{-1}[\ln(\beta_{j,n})]^{2-l}\right) = O\left(n^{-1}(\frac{\sqrt{n}}{\nu_j})^{l-2}\right);$$

the proof is completed by repeating (A.22). This completes the proof of Theorem 2.

Proof of Corollary 3. It follows easily by using Lemma 5.

Proof of Corollary 4. It follows from (3.22), (3.10) and (3.11) that $f \in B^r_{p,q}$ with probability one if

$$\lim_{n\to\infty} \sum_{j=L}^{J-1} \left[\Big(\frac{\sqrt{n}}{\nu_j}\Big)^{a_j} 2^{sj}\right]^{\min(1,\frac{p}{q})} < \infty, \qquad \text{(A.23)}$$

where $s = (1-p^*/2)(1+(2r)^{-1}+\varepsilon)I(j_0 < j \le j_1)$ and $\varepsilon > 0$ is defined by Theorem 1. Since $\lim_{n\to\infty}\sum_{j=L}^{j_0}(\sqrt{n}/\nu_j)^{a_1} < \infty$ and $\lim_{n\to\infty}\sum_{j=j_1+1}^{J-1}(\sqrt{n}/\nu_j)^{a_3} < \infty$ for any $a_1 < 0$ and $a_3 > 0$, the only part of (A.23) which needs to be examined is

$$S = \lim_{n\to\infty}\sum_{j=j_0+1}^{j_1}\left[\left(\frac{\sqrt{n}}{\nu_j}\right)^{a_2}2^{sj}\right]^{\min(1,\frac{p}{q})}. \tag{A.24}$$

Let $b = -a_2$ and observe that we can consider $1 \le p < 2$ only (since otherwise the term disappears), so that $S_j = (\nu_j/\sqrt{n})^b 2^{sj} = n^{-b/2}2^{Bj}$ with $B = (1+(2r)^{-1}+\varepsilon)(1-p/2)+b[(r+1/2)-(1/p-1/2)(1+(2r)^{-1}+\varepsilon)$. It is easily seen now that $B > 0$ and (A.24) holds whenever $n^{-b/2}2^{Bj_1} \equiv n^{B_1}$ is uniformly bounded. Here, $B_1 = -b/2 + Br/[(2r+1)(r+1/2-1/p)]$ by (3.8). The proof is completed by checking that (3.23) ensures $B_1 < 0$.

Proof of Theorem 4. Due to the embedding properties of Besov spaces (i.e., $B^r_{p,q}(A) \subset B^r_{p,\infty}(A)$ for $1 \le q \le \infty$), it is sufficient to prove the results for $q = \infty$. The proof follows directly by modifying the proof of Theorem 2 of Johnstone and Silverman (2005). Note also that the proof of Theorem 2 of Johnstone and Silverman (2005) is just a modification of Theorem 1 of Johnstone and Silverman (2004a), so it is necessary to check that it is valid in the case of the Bayes factor estimator. For simplicity, we choose $\sigma = 1$, suppress the index j, and note that in our case $q = 2$ in the notation of Johnstone and Silverman (2004a, 2005). Johnstone and Silverman (2004a, 2005) introduced a function $\beta(x) = I(x)/\varphi(x) - 1$, where $\varphi(x)$ is the standard normal *pdf*, and defined a pseudo-threshold $\zeta(\pi) = \beta^{-1}(1/\pi)$, i.e., $\beta(\zeta(\pi)) = 1/\pi$. Further, in the proof of Theorem 1, Johnstone and Silverman (2004a) used their Lemma 3 which states that, for $0 < \pi \le 1$,

$$1 + \beta\left(t^{PM}(\pi)\right) < \beta(\zeta(\pi)) < 2 + \beta\left(t^{PM}(\pi)\right), \tag{A.25}$$

where $t^{PM}(\pi)$ is the threshold due to the posterior median estimator. Recall that the Bayes factor threshold $t^{BF}(\pi)$ is derived by solving the equation $I(t^{BF})/\varphi(t^{BF}) = (1-\pi)/\pi$, which is equivalent to $\beta(t^{BF}) = 1/\pi - 2$. Hence, (A.25) is modified to become $1+\beta(t^{BF}(\pi)) < \beta(\zeta(\pi)) \le 2+\beta(t^{BF}(\pi))$. With this modification, all of the steps in the proof of Theorem 1 of Johnstone and Silverman (2004a) go through, the only things which are affected are the constants. Note also that, under the conditions of Theorem 4, $\tilde{\theta}_{jk} \in b^r_{p,\infty}(A)$ (Besov sequence ball of radius A), so that $\|\tilde{\theta}_{j\cdot}\|_p \le A2^{-j(r+1/2-1/p)}$ for all j. Hence we can apply Theorem 2 of Johnstone and Silverman (2005) to get

$$R_n(B^r_{p,\infty}(A), \hat{f}) \le c\left(A^{\frac{2}{2r+1}}n^{-\frac{2r}{2r+1}} + n^{-1}(\ln n)^{\nu} + A^2 n^{-2(r+\frac{1}{2}-\frac{1}{p})}\right),$$

where $\nu = 0$ if $D > 0$, and $\nu = 4 - p^*/2$ if $D = 0$. Observe that the second term is asymptotically smaller than the first one, and the last term is $O(n^{-2r/(2r+1)})$ whenever $r > r_p$, resulting in (3.28).

References

Abramovich, F., Amato, U. and Angelini, C. (2004). On optimality of Bayesian wavelet estimators. *Scand. J. Statist.* **31**, 217-234.

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

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

Angelini, C. and Sapatinas, T. (2004). Empirical Bayes approach to wavelet regression using ε-contaminated priors. *J. Statist. Comput. Simulation* **74**, 741-764.

Angelini, C. and Vidakovic, B. (2004). Γ-minimax wavelet shrinkage: a robust incorporation of information about energy of a signal in denoising applications. *Statist. Sinica* **14**, 103-125.

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. *J. Statist. Soft.* **6**, Issue 6, 1-83.

Billingsley, P. (1995). *Probability and Measure*. 3rd edition. Wiley, New York.

Bochkina, N. and Sapatinas, T. (2005). On the posterior median estimators of possibly sparse sequences. *Ann. Inst. Statist. Math.* **57**, 315-351.

Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898-924.

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

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. Roy. Statist. Soc. Ser. B* **62**, 681-698.

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

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1**, 54-81.

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

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-456.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

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

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 301-337.

Gradshtein, I. S. and Ryzhik, I. M. (1980). *Tables of Integrals, Series and Products.* Academic Press, New York.

Johnstone, I. M. (2002). Function estimation in Gaussian noise: sequence models. Manuscript. (available at http://www-stat.stanford.edu/∼imj)

Johnstone, I. M. and Silverman, B. W. (2004a). Finding needles hay in haystacks: Risk bounds for empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**, 1594-1649.

Johnstone, I. M. and Silverman, B. W. (2004b). Boundary coiflets for wavelet shrinkage in function estimation. *J. Appl. Probab.* **41A**, 81-98.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33**, 1700-1752.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B* **58**, 463-479.

Mallat, S. G. (1999). *A Wavelet Tour of Signal Processing.* 2nd edition. Academic Press, San Diego.

Neumann, M. H. and von Sachs, R. (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. In *Wavelets and Statistics* (Edited by A. Antoniadis and G. Oppenheim), 301-329. Lecture Notes in Statistics **103**. Springer-Verlag, New York.

Pensky, M. (2006). Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise. *Ann. Statist.* **34**, 769-807.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* **93**, 173-179.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Wiley, New York.

Walter, G. G. (1994). *Wavelets and Other Orthogonal Systems with Applications.* CRC Press, Boca Raton.

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

Department of Mathematics, University of Central Florida, Orlando, FL 32816-1364, U.S.A.

E-mail: mpensky@pegasus.cc.ucf.edu

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

E-mail: T.Sapatinas@ucy.ac.cy

(Received November 2005; accepted September 2006)