1-1-2006

Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise

Marianna Pensky
University of Central Florida

Find similar works at: https://stars.library.ucf.edu/facultybib2000

University of Central Florida Libraries http://library.ucf.edu

FREQUENTIST OPTIMALITY OF BAYESIAN WAVELET SHRINKAGE RULES FOR GAUSSIAN AND NON-GAUSSIAN NOISE[1]

BY MARIANNA PENSKY

University of Central Florida

The present paper investigates theoretical performance of various Bayesian wavelet shrinkage rules in a nonparametric regression model with i.i.d. errors which are not necessarily normally distributed. The main purpose is comparison of various Bayesian models in terms of their frequentist asymptotic optimality in Sobolev and Besov spaces.

We establish a relationship between hyperparameters, verify that the majority of Bayesian models studied so far achieve theoretical optimality, state which Bayesian models cannot achieve optimal convergence rate and explain why it happens.

1. Introduction. Bayesian techniques for shrinking wavelet coefficients have become very popular in the last few years. Starting with the paper by Clyde, Parmigiani and Vidakovic [9], researchers turned to Bayesian methods in wavelet analysis (see, e.g., [1–4, 6–8, 10, 17, 19, 24, 26]).

The majority of the above papers were devoted to the nonparametric wavelet regression model

$$(1.1) \qquad Y_i = f(i/n) + Z_i, \qquad i = 1, \ldots, n,$$

where f is unknown and belongs to some functional space $\mathcal{F}$ and Z_i are i.i.d. random variables. In most cases, the errors Z_i were assumed to be normally distributed with zero mean and unknown variance σ^2. Some papers in engineering fields modeled errors as double-exponential (see, e.g., [20] or [23]); however, this approach was not very popular in statistical applications. The only paper known to the author which deals with nonnormal errors in a Bayesian framework is the one by Clyde and George [8] where the authors conduct comprehensive simulations with Z_i having various distributions. In what follows we just assume that Z_i's are i.i.d., symmetric and have finite fourth moment $EZ_1^4 < \infty$.

Now let us discuss the error model for wavelet coefficients which is used in Bayesian inference. If the Z_i's in (1.1) are i.i.d. normal, the errors in wavelet coefficients are also i.i.d. normal, and most authors adopted this "real" model for

Bayesian inference. However, some of the authors (Clyde and George [8] and Vidakovic [24]) were more adventurous and considered other distributions for error models, namely, distributions that result from a mixture of normal distributions over the scale parameter. Various choices of priors were suggested, and the posterior mean or median was used as an estimator of wavelet coefficients. The procedures were then justified by extensive simulations. Thus, a variety of algorithms have been suggested for applications; all of them have demonstrated good computational performance, but until the present time no attempts have been made to examine frequentist properties of these techniques or to compare various algorithms with each other.

To the best of the author's knowledge, three manuscripts have been produced recently with the objective to assess frequentist properties of Bayesian wavelet shrinkage and thresholding. The first one, by Abramovich, Amato and Angelini [1], explores optimality of Bayesian wavelet estimators in the case of normal errors in (1.1) and a combination of a point mass at zero and a normal density for the prior. The authors consider three choices of Bayesian wavelet estimators: the mean, the median and the mean thresholded on the basis of a Bayes factor rule [24], and then study optimality of the resulting wavelet regression estimators in Besov spaces. Johnstone and Silverman [19] go much further than Abramovich, Amato and Angelini [1], investigating theoretical and computational properties of the *empirical* Bayes thresholding rules. They also assume the errors to be normally distributed and choose priors to be a mixture of an atom probability at zero and a heavy-tailed density. The mixing weight, or sparsity parameter, for each level of the transform is chosen by maximizing marginal likelihood. The estimators can be based on a posterior mean or a posterior median, and they are optimal in Besov spaces. Finally, Autin, Picard and Rivoirard [5] compare a wide variety of Bayes shrinkage and thresholding rules based on both normal and heavy-tailed priors using a novel maxiset approach. However, although Autin, Picard and Rivoirard [5] and Johnstone and Silverman [19] develop breakthrough theory concerning regression with normally distributed errors, they do not discuss the case of a non-Gaussian distribution either for the actual error in (1.1) or as the error model in Bayesian inference.

The present paper is in a sense complementary to the three papers mentioned above as well as to Clyde and George [8]. Our interest was sparked by the excellent paper by Clyde and George [8], who present an amazing volume of simulations for various error distributions and various Bayesian models. All of their Bayesian models produce estimators as efficient (or better) than the ones based on traditional thresholding rules no matter whether the errors in (1.1) are normal or not. Yet, no theoretical study of these estimators is conducted; hence, some of the simulation results are stated but not explained.

The present paper examines various Bayesian models in terms of their frequentist asymptotic optimality without a predominant assumption that errors in (1.1)

are normally distributed. Bayesian estimators for the wavelet coefficients are constructed using posterior means for a variety of error models and priors, and these models are compared in terms of their frequentist optimality. Although theoretical results of the paper are asymptotic, they allow one to explain some features of the simulations conducted by Clyde and George [8], namely, differences in the performances of "normal error–normal prior" model and the models with heavy-tailed error and prior distributions.

The material presented in the paper is interesting from several points of view. First, the relationship between hyperparameters is established, and it is verified that the majority of Bayesian models studied so far in the literature have indeed not only excellent computational properties but also theoretical optimality. Second, it is shown that some Bayesian models cannot achieve the optimal convergence rate and explained why this happens. Third, the prior distributions on wavelet coefficients are compared to see which of them agree with the assumption that the regression function belongs to a particular Besov space. The latter leads to a more comprehensive examination of the mixing weights. For example, we discover that the weights chosen in [1, 3] lead to convergence rates which differ from optimal (e.g., by the logarithmic factor in the case of normal errors and normal priors).

The rest of the paper is organized as follows. In Section 2 we introduce Bayesian models for wavelet coefficients. We use some "arbitrary" error model $\eta_j(\cdot)$ and a mixture of a point mass at zero and a density $\xi(\cdot)$ for a prior, keeping in mind that the actual distribution of 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 $\eta_j(\cdot)$ and $\xi(\cdot)$ and provide assertions about asymptotic optimality of regression estimators in Besov spaces for various choices of $\eta_j(\cdot)$ and $\xi(\cdot)$. In Section 4 we present comparison between various Bayesian models in tabular form and discuss theoretical results of Section 3. We also explain results of simulations conducted in [8]. Section 5 contains some auxiliary statements as well as proofs of the statements in Section 3.

2. The Bayesian model. Consider a standard nonparametric regression model (1.1) and assume that f is square integrable and Z_i are i.i.d. random variables with zero mean and finite variance. Application of the discrete wavelet transform (DWT) based on the wavelet ψ to (1.1) yields

(2.1) $Y_{jk} = w_{jk} + \varepsilon_{jk}, \qquad j = 0, \ldots, J-1, k = 0, \ldots, 2^j - 1, J = \log_2 n,$

where ε_{jk} are uncorrelated random variables due to the unitary property of the DWT. We assume that the wavelet function ψ has finite support and $s > r$ vanishing moments. If one is interested in the error of the resulting estimator of f only at points i/n, one can just use periodized orthogonal wavelets. However, since in this paper we are interested in the L^2-norm based risk as the measure of the quality of the estimation procedure, some kind of boundary correction is necessary. For this

reason, we shall use boundary coiflets introduced in [18, 19], which we describe in detail in Section 3.1.

Denote the scaling and the wavelet coefficients of the original function f by $\tilde{\theta}_k$ and $\tilde{\theta}_{jk}$, respectively, so that f can be reconstructed as

$$(2.2) \qquad f(x) = \sum_{k=0}^{2^L-1} \tilde{\theta}_k 2^{L/2} \varphi(2^L x - k) + \sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} \tilde{\theta}_{jk} \psi_{jk}(x)$$

with $\tilde{\theta}_k = \int_{-\infty}^{\infty} 2^{L/2} \varphi(2^L x - k) f(x)\,dx$ and $\tilde{\theta}_{jk} = \int_{-\infty}^{\infty} \psi_{jk}(x) f(x)\,dx$. Here $\varphi(x)$ is a scaling function corresponding to the mother wavelet $\psi(x)$, $\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$ and the value of L will be defined later. Denote $\theta_{jk} = w_{jk}/\sqrt{n}$ and recall that $\tilde{\theta}_{jk} \approx \theta_{jk}$ (see, e.g., [25]). Later we shall provide a more detailed treatment of the relation between $\tilde{\theta}_{jk}$ and θ_{jk}.

We shall use a Bayesian technique to construct estimators $\hat{\theta}_{jk}$ of θ_{jk} based on Y_{jk}. Since wavelet representations of a vast majority of functions contain only a few nonnegligible coefficients in their expansions, we place the following prior on the population discrete wavelet coefficient w_{jk}:

$$(2.3) \qquad w_{jk} \sim \pi_{jn} \tau_{jn} \xi(\tau_{jn} x) + (1 - \pi_{jn}) \delta(0),$$

where $0 \le \pi_{jn} \le 1$ if $0 \le j \le J - 1$ and $\pi_{jn} = 0$ if $j \ge J$, $\delta(0)$ is a point mass at zero and w_{jk} are independent. The p.d.f. $\xi(x)$ is even and unimodal. Note that the majority of priors used previously for Bayesian wavelet inference follow the model (2.3). The factors π_{jn} are the prior probabilities that a wavelet coefficient at level j does not vanish. In what follows, however, we shall impose all conditions on the odds

$$(2.4) \qquad \beta_{jn} = (1 - \pi_{jn})/\pi_{jn}.$$

Note that we allow dependence of π_{jn} not only on the resolution level j but also on n. It is most natural since the proportion of coefficients we are intending to keep depends not only on the function f but also on the amount of data available: the larger n is, the more reliable the estimators of coefficients are, the smaller the coefficients we can keep, the larger the value of π_{jn} for a particular resolution level j.

Now let us discuss the distribution of errors. It follows from (1.1) that $\varepsilon_{jk} \approx n^{-1/2} 2^{j/2} \sum_{i=1}^{n} \psi(2^j i/n - k) Z_i$. Since Z_i are i.i.d. with $E Z_1^4 < \infty$, the sequence $2^{j/2} n^{-1/2} \psi(2^j i/n - k) Z_i$ satisfies the Lyapunov condition with $\delta = 2$ provided $2^j/n \to 0$. Hence, if the resolution level is reasonably small ($j < J_0$ where $J - J_0 \to \infty$ as $n \to \infty$), the errors ε_{jk} are asymptotically normally distributed $N(0, \sigma^2)$ and, thus, asymptotically independent, while at large resolution levels they are uncorrelated and have some unknown p.d.f. $\mu(x)$ which is symmetric and has a finite fourth moment. To incorporate both of these features into the distribution of ε_{jk} we write it as

$$(2.5) \qquad \varepsilon_{jk} \sim (1 - \lambda_j)(\sigma \sqrt{2\pi})^{-1} \exp(-x^2/(2\sigma^2)) + \lambda_j \mu(x).$$

Here $\sigma^2 = \int_{-\infty}^{\infty} x^2 \mu(x)\,dx = EZ_1^2$, $0 \le \lambda_j \le 1$, and λ_j are close to 1 at large resolution levels and equal to zero at small resolution levels, namely, $\lambda_j = 1$ if $j = J - 1$ and $\lambda_j = 0$ if $j \le J_0$. For a more detailed treatment of asymptotic normality an interested reader can consult Neumann and von Sachs [21].

The difficulty of using (2.5) in Bayesian inference is that both $\mu(x)$ and λ_j are unknown. For this reason, we shall choose the most general distribution for errors, namely,

$$(2.6) \qquad \varepsilon_{jk} \sim \eta_j(x),$$

where η_j are level-dependent symmetric densities. As we shall show later, one does not need the knowledge of the true distribution of ε_{jk} and can obtain the optimal estimator of f with a variety of error distributions η_j. Moreover, η_j does not even need to be level dependent.

The model (2.6) generalizes the setup of Clyde and George [8], who considered error distributions other than normal in (1.1), namely, the distributions that can be obtained from the normal distribution by mixing over the scale parameter. It also follows the remarkable idea of Vidakovic [24], who assumed that the errors are normally distributed but used a double-exponential distribution for errors which resulted from the hierarchical Bayes approach, hence, conducting Bayesian inference with the error distribution different from the true one.

We shall conduct Bayesian inference for each wavelet coefficient separately. Denote

$$(2.7) \qquad d_{jk} = Y_{jk}/\sqrt{n}, \qquad \nu_j = \sqrt{n}\tau_{jn}.$$

Taking into account the relation between w_{jk} and θ_{jk} and (2.3)–(2.7), we derive that the posterior p.d.f. of θ_{jk} given d_{jk} is of the form

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

Choosing the posterior mean to be an estimator, we arrive at the following estimator of θ_{jk}:

$$(2.8) \qquad \hat{\theta}_{jk} = \frac{I_{1j}(d_{jk})}{I_{0j}(d_{jk}) + \beta_{jn}\sqrt{n}\eta_j(\sqrt{n}d_{jk})}, \qquad j \le J - 1,$$

where

$$(2.9) \qquad I_{ij}(d) = \int_{-\infty}^{\infty} x^i \sqrt{n}\eta_j[\sqrt{n}(x - d)]\nu_j\xi(\nu_j x)\,dx, \qquad i = 0, 1,$$

and $\hat{\theta}_{jk} = 0$ as $j \ge J$, so that the estimator of f is of the form

$$(2.10) \qquad \hat{f}(x) = \sum_{k=0}^{2^L-1} \hat{\theta}_k 2^{L/2}\varphi(2^L x - k) + \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} \hat{\theta}_{jk}\psi_{jk}(x).$$

The objective of the present paper is to formulate conditions under which the estimator of f based on coefficients $\hat{\theta}_{jk}$ is optimal and explore the cases where this optimality cannot be achieved. For any possible estimator $\tilde{f}$ of f based on n observations $X_1, \ldots, X_n$, define the mean integrated square error (MISE) over a function space $\mathcal{F}$ as

$$(2.11) \qquad R_n(\mathcal{F}, \tilde{f}) = \sup_{f \in \mathcal{F}} E \| \tilde{f} - f \|^2_{L^2[0,1]}.$$

Donoho and Johnstone [14] showed that when the errors Z_i in (1.1) are normally distributed and f belongs to a ball $B^r_{p,q}(A)$ in the Besov space $B^r_{p,q}[0, 1]$, then there exist constants C_1 and C_2 independent of n such that

$$(2.12) \qquad C_1 n^{-2r/(2r+1)} \leq \inf_{\tilde{f}} R_n\big(B^r_{p,q}(A), \tilde{f}\big) \leq C_2 n^{-2r/(2r+1)}$$

provided $r > \max(0, 1/p - 1/2)$ and $p, q \geq 1$. Since the Sobolev space $H^r = B^r_{2,2}$, the same rates of convergence hold for $H^r(A)$. Note that since the normal distribution for errors is a particular case of (2.5), the lower bounds in our situation cannot be smaller than (2.12). On the other hand, since for a majority of resolution levels ($j \leq J_0$) the errors ε_{jk} follow the normal model, we can expect to achieve convergence rates (2.12) as $n \to \infty$ for some choices of error models (2.6) and priors (2.3).

In what follows we shall compare various Bayesian models in terms of their ability to achieve optimal convergence rates (2.12) as $n \to \infty$.

3. Asymptotic optimality.

3.1. *Assumptions.* In what follows we shall formulate conditions on the p.d.f.'s $\xi(\cdot)$ and $\eta_j(\cdot)$ and parameters ν_j, β_{jn} and J_0. We warn the readers that not all of the conditions listed below will necessarily appear in every statement later.

Let ϕ and ψ be boundary coiflets introduced in [18, 19], possessing $s > r$ vanishing moments and based on orthonormal coiflets supported in $[-S+1, S]$, $s < S$. Assume that $p \geq 1$, $r > \max(1/2, 1/p)$ and

$$(3.1) \quad \begin{aligned} & r \geq r_p = 0.5\big[(1/p - 1/2) + \sqrt{(1/p - 1/2)^2 + 2(1/p - 1/2)}\,\big] I(p < 2), \\ & L \geq \log_2(6S - 6). \end{aligned}$$

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

Let $\xi(x)$ and $\eta_j(x)$ be three times continuously differentiable everywhere except possibly $x = 0$, have finite fourth moments and satisfy the conditions

$$(3.2) \ (A1) \qquad \big|\xi^{(k)}(x)/\xi(x)\big| \leq C_\xi(1 + |x|^{\lambda_\xi})^k, \qquad k = 1, 2, 3, \lambda_\xi \geq 0,$$

$$(3.3) \ (A2) \qquad \big|\eta_j^{(k)}(x)/\eta_j(x)\big| \leq C_\eta(1 + |x|^{\lambda_\eta})^k, \qquad k = 1, 2, 3, \lambda_\eta \geq 0,$$

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

(3.5) (A4) $\qquad x_1^{2+\delta}\eta_j(x_1) \le x_2^{2+\delta}\eta_j(x_2), \qquad x_1 > x_2 > C_\delta > 0, \delta \ge 0.$

Condition (A4) just means that the functions $|x|^{2+\delta}\eta_j(x)$ are nonincreasing for sufficiently large x. Note that the constants λ_η, C_η, $C_{\xi,\eta}$, δ and C_δ are assumed to be independent of j which requires some kind of uniformity of the p.d.f.'s η_j. The consequence of these restrictions is that asymptotic expansions of integrals I_{ij} defined in (2.9) are valid with absolute constants independent of j, so in what follows we shall suppress the index j in $I_{ij}, i = 0, 1$, unless this leads to confusion.

When conditions (A1) and (A2) hold with $\lambda_\xi = 0$ and $\lambda_\eta = 0$, following Johnstone and Silverman [17, 19], we shall say that η_j and ξ are *heavy-tailed* probability densities. The most common examples of the latter are p.d.f.'s of the double-exponential or Student t distributions. In this situation, asymptotic expansions

(3.6) $\quad I_0(d) \sim v_j\xi(v_jd), \qquad |I_1(d)/I_0(d) - d| = O(v_j/n) \qquad$ if $v_j/\sqrt{n} \to 0,$

(3.7) $\quad I_0(d) \sim \sqrt{n}\eta(\sqrt{n}d), \qquad |I_1(d)/I_0(d)| = O(\sqrt{n}/v_j^2) \qquad$ if $v_j/\sqrt{n} \to \infty,$

which are proved later in Lemma 2, are valid for any d as long as the relation between v_j and n holds. If λ_ξ and λ_η are positive, then the expansions (3.6) and (3.7) can be used under some restrictions on d only. There is one important case though, when we can obtain asymptotic expansions for the integrals for any value of d: if ξ and η_j are normal p.d.f.'s and the variances of η_j are bounded from above and from below by a positive value, then

(3.8) $\qquad |I_1(d)/I_0(d) - d| = O(|d|v_j^2/n) \qquad$ if $v_j/\sqrt{n} \to 0,$

(3.9) $\qquad I_0(d) \sim v_j\xi(v_jd) \qquad$ if $v_j/\sqrt{n} \to 0,$

(3.10) $\qquad |I_1(d)/I_0(d)| = O(n|d|/v_j^2) \qquad$ if $v_j/\sqrt{n} \to \infty.$

We denote

(3.11) $$j_0 = (2r + 1)^{-1}\log_2 n,$$

and assume that the parameter v_j is of the form

(3.12) $\qquad v_j = C_1 2^{mj} \qquad$ where $m = r + 1/2$ if $p \ge 2.$

The expression for v_j when $1 \le p < 2$ will be presented later [see (3.28)].

REMARK 1. Assumptions about v_j can be translated into the ones on τ_{jn} using relation (2.7), namely, $\tau_{jn} = C_1 2^{mj}/\sqrt{n}$. This coincides with the choice of τ_{jn} for the normal–normal model in [1, 3].

REMARK 2.   The assumption that $\varphi(x)$ and $\psi(x)$ are boundary coiflets as well as conditions (3.1) are introduced for the sake of obtaining convergence rates for a L^2-norm-based risk function. All statements of the paper will be true for $L = 0$ and an arbitrary s-regular scaling function $\varphi(x)$ and wavelet $\psi(x)$ with $s > \max(r, r + 1/2 - 1/p)$ if one replaces (2.11) by

$$R_n(\mathcal{F}, \tilde{f}) = \sup_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^{n} E[\tilde{f}(i/n) - f(i/n)]^2.$$

REMARK 3.   Conditions on the existence of the fourth moment are purely technical conditions that are used for derivation of asymptotic expansions of the integrals $I_{ij}(\cdot)$. These conditions can be dropped and replaced by requesting that the conclusions of Lemma 2 remain valid. These conclusions, however, have to be verified individually for each combination of $\xi(\cdot)$ and $\eta_j(\cdot)$. In what follows we shall consider the case when η_j and $\xi(\cdot)$ are p.d.f.'s of Student t and Cauchy distributions, respectively. Later, however, we shall see that it is somewhat beneficial to apply a distribution with faster descent at $\pm\infty$ than the Cauchy.

3.2. *Optimality in Sobolev spaces and Besov spaces with $p \geq 2$.*   It is well known that the Sobolev space H^r can be characterized in terms of wavelet coefficients as $\sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} \tilde{\theta}_{jk}^2 (1 + 2^{2jr}) < \infty$ (see, e.g., [11], Section 9.2). Therefore, f belongs to a ball $H^r(A)$ in H^r space with $r > 1/2$:

$$(3.13) \quad f \in H^r(A) \quad \Longleftrightarrow \quad \sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} \tilde{\theta}_{jk}^2 (1 + 2^{2jr}) \leq A, \qquad r > \tfrac{1}{2}.$$

For the Besov ball $B_{p,q}^r(A)$ with $r > 1/p$ there is a similar relationship (see, e.g., [16]):

$$(3.14) \quad f \in B_{p,q}^r(A) \quad \Longrightarrow \quad \sum_{k=0}^{2^j-1} \tilde{\theta}_{jk}^2 \leq \begin{cases} B_1 2^{-2jr}, & \text{if } p \geq 2, \\ B_1 2^{-2j(r+1/2-1/p)}, & \text{if } 1 \leq p < 2, \end{cases}$$

for some $B_1 > 0$. The cases $p \geq 2$ and $1 \leq p < 2$ apply to spatially homogeneous and nonhomogeneous functions, respectively, and, as the reader will see later, the performance of Bayesian models varies greatly in those two cases. We shall start with the spatially homogeneous case $p \geq 2$ and study optimality of Bayesian estimators in this case. Since $H^r = B_{2,2}^r$, this is sufficient to study the general case of $f \in B_{p,q}^r(A)$ with $p \geq 2$; the results for a Sobolev ball will immediately follow.

In this subsection, we assume that $r > 1/2$ and $p \geq 2$. We also assume that ν_j is of the form (3.12) and J_0 are such that

$$(3.15) \qquad\qquad\qquad 2^{J_0} \geq n^{1/(2r)}.$$

Condition (3.15) is quite realistic and agrees with the assumption $J - J_0 \to \infty$ as $n \to \infty$ above, provided $r > 1/2$. In practice, the normality assumption can be checked via level-by-level testing.

In order to make understanding of a variety of assumptions and assertions of this section easier, we refer the reader to Table 1 in Section 4.2. In effect, we consider three kinds of models: (1) models with strictly optimal convergence rates (which include normal ξ–normal η_j or heavy-tailed ξ–heavy-tailed η_j); (2) models optimal up to a log-factor (which include heavy-tailed ξ–normal η_j); (3) suboptimal models (which include normal ξ–heavy-tailed η_j). Within the first set of models we also study what happens when ξ has a faster descent at $\pm\infty$ than η_j, that is, condition (A3) is violated.

3.2.1. *Models with strictly optimal convergence rates*: normal ξ–normal η_j or heavy-tailed ξ–heavy-tailed η_j.   For those models convergence rates are given by

THEOREM 1.   *Let conditions* (A1) *and* (A2) *be valid with* $\lambda_\xi = \lambda_\eta = 0$, *or, alternatively, conditions* (3.8)–(3.10) *hold for any value of* d. *Let also conditions* (A3) *and* (A4) *hold. If for some positive* ε

$$(3.16) \quad \beta_{jn} = O\big((\sqrt{n}/v_j)^\alpha\big) \quad \text{with } \alpha \leq (2r+1)^{-1}(2+\delta) - 1 - \varepsilon \text{ as } j \leq j_0,$$

then

$$(3.17) \quad R_n\big(B_{p,q}^r(A), \hat{f}\big) = O\big(n^{-2r/(2r+1)}\big), \quad n \to \infty, p \geq 2.$$

COROLLARY 1.   *If* $\eta_j(\cdot)$ *and* $\xi(\cdot)$ *are the p.d.f.'s of the double-exponential or the Student t distributions, or both* $\xi(\cdot)$ *and* $\eta_j(\cdot)$ *are normal p.d.f.'s, satisfying conditions* (A2) *and* (A3) *[and* (A4) *for the t distribution], then for* β_{jn} *given by* (3.16) *the risk* $R_n(B_{p,q}^r(A), \hat{f})$ *is of the form* (3.17).

Note that in both Theorem 1 and Corollary 1 we assume that $\xi(x)$ and $\eta_j(x)$ have finite fourth moments. This is, however, a sufficient, not a necessary condition. Namely, the following statement is valid.

COROLLARY 1*.   *If* $\eta_j(x)$ *are the p.d.f.'s of Student* t_{2m_j+1} *distributions with* m_j *integers*, $M_1 \leq m_j \leq M_2$, *and* $\xi(x)$ *is the p.d.f. of the Cauchy distribution, then* $R_n(B_{p,q}^r(A), \hat{f})$ *is of the form* (3.17) *provided* β_{jn} *is given by* (3.16).

Now suppose condition (A3) is violated and $\xi(x)$ has faster descent to zero than $\eta_j(x)$ as $|x| \to \infty$. Then, in order for convergence rates to still hold, we need to impose extra conditions on β_{jn}, namely, β_{jn} should be small for $j \leq j_0$ (i.e., a priori more coefficients should be kept at low resolution levels). Consider an alternative assumption to (A3):

$$(3.18) \ (A3^*) \qquad |\eta_j(x)/\xi(x)| \leq U(|x|)$$
$$\text{where } U(x_1) \geq U(x_2) \text{ for any } x_1 > x_2 \geq 1.$$

THEOREM 2. *Let the conditions of Theorem* 1 *hold with* (A3) *replaced by* (A3*). *Let also for some* $\varepsilon > 0$

(3.19)
$$
\beta_{jn} = O\left(\min\left(\left[\frac{\sqrt{n}}{v_j}\right]^{(2+\delta)/(2r+1)-1-\varepsilon},\right.\right.
$$
$$
\left.\left.\left[\frac{\sqrt{n}}{v_j}\right]^{1/(2r+1)+\delta-\varepsilon}\frac{2^{-j/2}}{U(2B_1 2^{j/2})}\right)\right), \qquad j \le j_0,
$$

where B_1 *is the constant appearing in* (3.14). *Then* $R_n(B_{p,q}^r(A), \hat{f})$ *is of the form* (3.17).

Note that the constant B_1 is usually unknown. If $U(x)$ is a homogeneous function of some order [which happens, e.g., when $\eta_j(x)$ and $\xi(x)$ are the p.d.f.'s of t distributions], then the value of B_1 has no bearing on β_{jn}. Otherwise, one can replace B_1 by any function of n which grows infinitely as $n \to \infty$, for example, $\ln n$.

It is easy to see that condition (3.19) is more restrictive than (3.16). Condition (3.19) means that we a priori intend to keep many more coefficients than we "kill."

3.2.2. *Models optimal up to a log-factor*: *heavy-tailed* ξ–*normal* η_j. Unfortunately, for some combinations of distributions assumption (3.10) holds only when $\sqrt{n}|\sqrt{n}d|^{\lambda_\eta}/v_j \to 0$. This happens if $\eta_j(x)$ are normal distribution p.d.f.'s and $\xi(x)$ is a heavy-tailed p.d.f., for example, the double-exponential or Student t. In this situation we can ensure somewhat weaker conditions on $|I_1(d)/I_0(d)|$.

LEMMA 1. *If* $\xi(x)$ *and* $\eta_j(x)$ *are even unimodal p.d.f.'s such that* $|I_1(d)| < \infty$ *for any* j *and* d, *then* $|I_1(d)/I_0(d)| \le |d|$.

Note that Theorem 1 and Corollaries 1 and 1* are valid for any distributions of Z_i's in (1.1). However, if (3.10) does not hold for all values of d and the errors in (1.1) are not normally distributed, one needs additional assumptions to achieve near the optimal convergence rate. These can be two different kinds of assumptions: either $\mu(x)$ in (2.5) should decrease reasonably fast at $\pm\infty$ or we should "kill" nearly all coefficients at the highest resolution levels. In particular, we define $\varrho_0(n)$ to be the unique positive solution of the equation

(3.20)
$$
\int_{\varrho_0(n)}^{\infty} x^2 \mu(x)\, dx = n^{-2r/(2r+1)},
$$

where $\mu(x)$ and σ are as in (2.5), and denote

(3.21) $\varrho(n) = \max[\varrho_0(n), \sigma\sqrt{2}\sqrt{2r(2r+1)^{-1}\ln n + 0.5\ln\ln n}]$.

THEOREM 3. *Let conditions* (A1)–(A4) *hold with* $\lambda_\xi = 0$ *and* $\lambda_\eta > 0$ *and let* β_{jn} *satisfy* (3.16). *If* $\mu(\cdot)$ *is such that*

$$(3.22) \qquad \lim_{n \to \infty} n^{-1/4r} (\varrho(n))^{\lambda_\eta} = 0,$$

or β_{jn} *increases quickly as* $j \geq J_0$:

$$(3.23) \qquad \beta_{jn}^{-1} = O\big(\eta_j (2\sqrt{C_\mu} n^{r/(2r+1)}) 2^{rj/(2r+1)}\big), \qquad j \geq J_0,$$

where $C_\mu = \int_0^\infty x^4 \mu(x)\,dx$, *then*

$$(3.24) \quad R_n\big(B_{p,q}^r(A), \hat{f}\big) = O\big(n^{-2r/(2r+1)} (\ln n)^{\lambda_\eta/(2r+1)}\big), \qquad n \to \infty.$$

COROLLARY 2. *If* $\eta_j(\cdot)$ *are normal p.d.f.'s with variances bounded from above and from below by common positive numbers and* $\xi(x)$ *is the p.d.f. of the double-exponential or t distribution, then* (3.24) *is valid provided* (3.16) *and either* (3.22) *or* (3.23) *holds.*

REMARK 4. Note that if the errors are normally distributed, then $\lambda_\eta = 1$ and $\varrho(n) = O(\sqrt{\ln n})$, so that condition (3.22) is unnecessary. If the errors follow the mixture model (2.5), condition (3.22) is valid whenever $\mu(x)$ decreases at an exponential rate or at a rate $|x|^{-a}$ with $a > 3 + 8r^2\lambda_\eta/(2r+1)$ as $|x| \to \infty$. The latter requires relatively fast descent (e.g., $a > 17/3$ for $r = 1$ and $\lambda_\eta = 1$) which may not be true. To overcome this difficulty, one can choose large values for β_{jn} as $j \geq J_0$, suggested by (3.23). This measure will suppress coefficients at higher resolution levels and relax their dependence on a particular form of μ.

3.2.3. *Suboptimal models*: *normal* ξ*–heavy-tailed* η_j. By now, we have addressed all possible models except for the ones where the prior $\xi(x)$ is a normal p.d.f. and the error distributions $\eta_j(x)$ are heavy-tailed. In this situation, assumption (3.8) may be invalid: instead of (3.8), by Lemma 1, we have

$$(3.25) \qquad |I_1(d)/I_0(d) - d| = O(|d|) \qquad \text{if } v_j|v_j d|^{\lambda_\xi}/\sqrt{n} \geq M.$$

This happens when, for example, the $\eta_j(x)$ are p.d.f.'s of the double-exponential or the Student t distribution. In this situation, convergence rates (3.17) or (3.24) do not hold.

THEOREM 4. *Let conditions* (A1) *and* (A2) *be valid with* $\lambda_\xi > 0$ *and* $\lambda_\eta = 0$. *Let conditions* (A3*) *and* (A4) *hold and* β_{jn} *satisfy* (3.19). *Then*

$$(3.26) \qquad R_n\big(B_{p,q}^r(A), \hat{f}\big) = O\big(n^{-2r/(2r+1+\lambda_\xi)}\big), \qquad n \to \infty.$$

Theorem 4 gives an upper bound for the risk. However, in order to conclude that certain combinations of distributions are asymptotically inferior to the others, we are interested in the lower bound for the risk.

COROLLARY 3.   *If $\eta_j(x) = \eta(x)$ are identical p.d.f.'s of the double-exponential or the Student t distribution and $\xi(x)$ is a normal p.d.f., then for some positive C*

$$(3.27) \qquad R_n\big(H^r(A), \hat{f}\big) \geq Cn^{-2r/(2r+2)}, \qquad n \to \infty.$$

Observe that the lower bound in (3.27) is identical to the upper bound in (3.26) (since $\lambda_\xi = 1$ for the normal p.d.f.) and both are asymptotically larger than the optimal convergence rate (3.17). This is due to the fact that the *bias* of the estimator $\hat{f}$ converges at a slower rate. The latter happens because the model with the above combination of $\xi(\cdot)$ and $\eta(\cdot)$ fails to adapt to sparsity, namely, to the situation when at lower resolution levels we have very few relatively large coefficients. This is not surprising since in Corollary 3 we have a combination of a rather flat error model and a sharp prior p.d.f. which may fail to capture the actual value of the estimated coefficient.

3.3. *Optimality in Besov spaces with $1 \leq p < 2$.*   It follows from the previous section that the models which satisfy Corollary 1 attain the optimal convergence rate in Sobolev spaces and spatially homogeneous Besov spaces $B^r_{p,q}$, $p \geq 2$, with minimal assumptions whether the errors are normally distributed or not. Now, the question is whether all those models also achieve the optimal convergence rate in spatially nonhomogeneous spaces $B^r_{p,q}$, $1 \leq p < 2$. Before giving an answer, we need to introduce new values of parameters for this case. Denote

$$j_1 = [(r + 1/2 - 1/p)(2r + 1)]^{-1} r \log_2 n,$$
$$J_0 = 0.5[\log_2 n + j_1],$$

and let j_0 and v_j be defined by (3.11) and (3.12), respectively, with

$$(3.28) \quad m = \begin{cases} m_1 = r + 1/2 - 0.5(1/p - 1/2), & j \leq j_0, \\ m_2 = (r + 1/2) - (1/p - 1/2)(1 + 1/r), & j_0 < j < j_1, \\ m_3 = r + 1/2, & j \geq j_1. \end{cases}$$

Observe that $m = r + 1/2$ for all resolution levels whenever $p = 2$. Note also that the resolution level j_1 is chosen so that

$$(3.29) \qquad \sum_{j=j_1}^{\infty} \sum_{k=0}^{2^j - 1} \theta_{jk}^2 = O\big(n^{-2r/(2r+1)}\big).$$

The restriction $r > r_p$ ensures that $J - j_1 \to \infty$, so that one can choose resolution level J_0, $j_1 < J_0 < J - 1$, such that $J - J_0 \to \infty$. Assume that

$$(3.30) \ (A5) \qquad \int_{-\infty}^{\infty} z^2 \exp(-z^2/2\sigma^2)[\eta_j(1 + |z|)]^{-2} \, dz \leq C,$$

where σ is the TRUE standard deviation of the error and C is a constant independent of j. Again we have the same three types of models as we had in the case of $p \geq 2$. However, neither of the first two types of models delivers strict optimality when $1 \leq p < 2$. The third type of model is suboptimal even when $p \geq 2$, so we do not consider this type of model here.

3.3.1. *Normal ξ–normal η_j or heavy-tailed ξ–heavy-tailed η_j models.* Here we consider the subset of models studied in Section 3.2.1, namely, the models where the tails of ξ are at least as heavy as the tails of η_j.

THEOREM 5.   *Let $f \in B_{p,q}^r(A)$ with $1 \leq p < 2$ and $r > \max(r_p, 1/p)$. Let ξ and η_j satisfy conditions (A1) and (A2) with $\lambda_\eta = \lambda_\xi = 0$ or let both of them be normal p.d.f.'s. Assume that (A3) and (A4) hold and*

$$(3.31) \qquad \eta_j(x) \leq C(x^2 + 1)^{-\nu/2} \exp(-\lambda|x|^\gamma).$$

If $\beta_{jn} = (\sqrt{n}/\nu_j)^\alpha$ for $j < j_1$ where

$$(3.32) \quad \alpha = \begin{cases} \alpha_1 \in (-\infty, \min(\nu - 3, (2r + 2 - 1/p)^{-1}\nu - 1)), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } j \leq j_0, \\ \alpha_2 \in \left[0, \nu - 2 + \dfrac{r + 1 - 2/p}{2r(r + 1) + 1 - p^{-1}(2r + 2)}\right], \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } j_0 < j \leq j_1, \end{cases}$$

whenever $\gamma = 0$ in (3.31), and $\alpha \geq 0$ whenever $j_0 < j < j_1$ and $\gamma > 0$ in (3.31), then

$$(3.33) \qquad R_n\bigl(B_{p,q}^r(A), \hat{f}\,\bigr) = O\bigl(n^{-2r/(2r+1)+\varepsilon_1}(\ln n)^{\varepsilon_2}\bigr), \qquad n \to \infty.$$

Here

$$(3.34) \quad \begin{aligned} \varepsilon_1 &= 0, \qquad \varepsilon_2 = 8r/[\gamma(4r + p(2r + 1))], &&\text{if } \gamma > 0, \\ \varepsilon_1 &= [\alpha^*(1/p - 1/2)]/[\nu(2r + 1)] > 0, \qquad \varepsilon_2 = -p, &&\text{if } \gamma = 0, \end{aligned}$$

with $\alpha^* = \max(1 + \alpha_1, (1 + \alpha_2)(2r + 2)/r)$. If $\alpha_2 = 0$, then $\alpha^* = (2r + 2)/r$, the minimum possible value, and $\varepsilon_1 = (2r + 2)/[r\nu(2r + 1)]$ for $\gamma = 0$.

Theorem 5 shows that the models where the η_j's have exponential descent achieve optimal convergence rates up to a logarithmic factor while η_j with polynomial descent lead to suboptimal convergence rates. Observe that condition (A5) does not exclude the normal model; it just requires that the variances of η_j be larger than the actual variance of our data. In particular, if $\eta_j(x) = (\sqrt{2\pi\sigma_j^2})^{-1} \exp(-x^2/2\sigma_j^2)$, condition (A5) implies that $\sigma_j^2 > 2\sigma^2$. This does not contradict Abramovich, Amato and Angelini [1], who considered only the case of $\sigma_j \equiv \sigma$.

3.3.2. *Heavy-tailed ξ–normal η_j models.* Now we study the models previously discussed in Section 3.2.2. These models under the assumption that η_j are identical and the actual errors are normally distributed were investigated in depth by Johnstone and Silverman [19], who demonstrated strict optimality of the models in an empirical Bayes setup in a wide variety of Besov spaces. However, since Johnstone and Silverman [19] considered empirical Bayes estimators, we cannot just reproduce their results here.

THEOREM 6. *Let $f \in B_{p,q}^r(A)$ with $1 \le p < 2$ and $r > \max(r_p, 1/p)$ and let conditions* (A1)–(A4) *be valid. Let $\eta_j(\cdot)$ be normal p.d.f.'s and $\lambda_\xi = 0$. Assume also that either condition* (3.22) *or* (3.23) *is valid. If $\beta_{jn} = (\sqrt{n}/v_j)^\alpha$ where $\alpha \ge 0$ whenever $j_0 < j < j_1$, then*

$$(3.35) \qquad R_n\big(B_{p,q}^r(A), \hat{f}\big) = O\big(n^{-2r/(2r+1)}(\ln n)^{\varepsilon_3}\big), \qquad n \to \infty,$$

where $\varepsilon_3 = \max(1/[2r+1], 4r/[4r + p(2r+1)])$.

3.4. *Optimality of the models with mixture distribution for errors.* In Sections 3.2 and 3.3 we considered only error distributions η_j which satisfy certain assumptions uniformly. There is, however, one more interesting case, namely, that of $\eta_j(\cdot)$ being a mixture distribution mimicking (2.5):

$$(3.36) \qquad \eta_j(x) = (1 - \lambda_j^*)(\sqrt{2\pi}\sigma_0)^{-1}\exp(-x^2/(2\sigma_0^2)) + \lambda_j^*\zeta(x),$$

where $\lambda_j^* = 0$ for $j \le J_0$. Let $\zeta(x)$ be a heavy-tailed distribution (i.e., $\lambda_\zeta = 0$).

THEOREM 7. *Let $\lambda_\zeta = 0$ and $\xi(x) \ge C_{\xi,0}\exp(-x^2/(2\sigma_0^2))$ for any x. Let β_{jn} satisfy the assumptions of Theorem 1 ($p \ge 2$) or Theorems 5 and 6 ($1 \le p < 2$). If $\xi(x)$ is a p.d.f. of the normal distribution, then as $n \to \infty$*

$$
\begin{aligned}
&R_n\big(B_{p,q}^r(A), \hat{f}\big) \\
(3.37) \quad &= \begin{cases} O\big(n^{-2r/(2r+1)}\big), & \text{if } p \ge 2, \\ O\big(n^{-2r/(2r+1)}(\ln n)^{4r/(4r+p(2r+1))}\big), & \text{if } 1 \le p < 2, \end{cases}
\end{aligned}
$$

the second assertion being valid provided $\sigma_0^2 > 2\sigma^2$, where σ^2 is the true error variance given by (2.5). *If $\xi(x)$ is a heavy-tailed p.d.f. ($\lambda_\xi = 0$) and β_{jn} and λ_j^* are such that*

$$(3.38) \qquad \beta_{jn} \ge \beta_0 \quad and \quad \lambda_j^* \le C_\lambda^*, \qquad j \ge J_0,$$

then as $n \to \infty$

$$(3.39) \qquad R_n\big(B_{p,q}^r(A), \hat{f}\big) = O\big(n^{-2r/(2r+1)}(\ln n)^\varepsilon\big)$$

where $\varepsilon = 1/(2r+1)$ if $p \ge 2$, and $\varepsilon = \varepsilon_3$ given by Theorem 6 if $1 \le p < 2$.

The model (3.36) behaves exactly as the model with normal errors for $j \le J_0$. The advantage of adding a heavy-tailed term for $j \ge J_0$ is that even in the case when $\xi(x)$ is a heavy-tailed prior, no additional assumptions on the distribution $\mu(x)$ or β_{jn} are necessary.

3.5. *Does $f \in B_{pq}^r$ a priori? More about prior odds β_{jn}.* So far, the choice of the error model $\eta_j(\cdot)$ was in the limelight. The main assertion about $\xi(\cdot)$ was that it should not have faster descent at $\pm\infty$ than $\eta_j(\cdot)$. However, it is $\xi(\cdot)$ that determines whether the regression function $f(x)$ belongs to a Besov space B_{pq}^r a priori. Namely, the following statement is valid.

THEOREM 8. *If $\xi(\theta)$, v_j and β_{jn} are such that*

$$(3.40) \qquad \int_{-\infty}^{\infty} |\theta|^{\max(p,q)} \xi(\theta) \, d\theta < \infty,$$

$$(3.41) \qquad \lim_{n\to\infty} \sum_{j=L}^{\infty} [2^{j(r+1/2)} \beta_{jn}^{-1/p} v_j^{-1}]^{\min(p,q)} < \infty,$$

then $f \in B_{p,q}^r$ a priori with probability 1.

It is easy to see that condition (3.40) requires $\xi(\cdot)$ to have at least $\max(p,q) \geq 1$ finite moments, which immediately eliminates the Cauchy prior from the list. On the other hand, any prior with exponential descent ensures validity of (3.40).

COROLLARY 4. *Let $p \geq 2$, $q \geq 1$ and $\beta_{jn} = (\sqrt{n}/v_j)^\alpha$ with $\alpha > 0$ as $0 \leq j \leq j_0$. Let also*

$$(3.42) \qquad \lim_{n\to\infty} \sum_{j=j_0+1}^{J-1} \beta_{jn}^{-\min(q/p,1)} < \infty.$$

If the values of v_j are given by (3.12) and $\xi(\theta)$ satisfies assumption (3.40), then $f \in B_{p,q}^r$ a priori with probability 1.

Note that in the case when condition (A3) holds (Theorems 1 and 3), Corollary 4 can be applied. In order for one to be able to choose $\alpha > 0$ in Theorem 1, condition (A4) should hold with $\delta > 2r - 1$. Condition (3.42) is needed since Theorem 1 puts absolutely no restrictions on the values of β_{jn} for $j \geq j_0$. The inequality (3.42) holds when, for example, $\beta_{jn} = (v_j/\sqrt{n})^{\alpha_1}$ with $\alpha_1 > 0$. The restriction (3.23) does not affect (3.42) since it imposes not small but large values of β_{jn} for $j \geq J_0$. However, the case when condition (A3) is violated is troublesome since it calls for smaller values of β_{jn} [see (3.19) in Theorem 2].

COROLLARY 5. *Let $1 \leq p < 2$ and v_j and β_{jn} be determined by Theorem 5 or 6. If (3.40) holds, $\beta_{jn} = (\sqrt{n}/v_j)^\alpha$ for $j < j_1$ where*

$$(3.43) \qquad \alpha = \begin{cases} \alpha_1 \in (2p, \infty), & \text{if } j \leq j_0, \\ \alpha_2 \in (2p(r+1), \infty), & \text{if } j_0 < j \leq j_1, \end{cases}$$

and

$$(3.44) \qquad \lim_{n \to \infty} \sum_{j=j_1+1}^{J-1} [\beta_{jn}^{-1/p}]^{\min(p,q)} < \infty,$$

then $f \in B_{p,q}^r$ *a priori with probability* 1.

Note that Corollary 5 is applicable under the conditions of Theorems 5–7 whenever assumption (3.31) on $\eta_j(\cdot)$ holds with $\gamma > 0$ or ν large enough, so that conditions (3.32) and (3.43) on α can be satisfied simultaneously. Again, assumption (3.44) is necessary since Theorems 5–7 put almost no restrictions on β_{jn} when $j > j_1$.

Corollaries 4 and 5 imply that in order to ensure $f \in B_{p,q}^r$ a priori almost surely, one needs fairly large values β_{jn}, for example, such that the sums $\sum \beta_{jn}^{-\min(q/p,1)}$ are uniformly bounded [see (3.42)]. This fact motivates the choice $\pi_{jn} = 2^{-bj}$ in [1, 3], which is equivalent to $\beta_{jn} \sim 2^{jb}$. In the present paper, however, we choose $\beta_{jn} = (\sqrt{n}/\nu_j)^\alpha$, so that $\beta_{jn} \sim 1$ at the "optimal" resolution level j_0. Is this a reasonable choice? The answer is "yes" if one wants to obtain the optimal convergence rate for a wide range of models.

THEOREM 9. *Let conditions* (A1)–(A4) *be valid. Assume that* $p \geq 2$, ν_j *are defined by* (3.12) *and the densities* $\eta_j(x) = \eta(x)$ *are identical. If* $\pi_{jn} = O(2^{-bj})$ *with* $b > 0$, *then for some positive* C

$$(3.45) \qquad R_n\big(B_{p,q}^r(A), \hat{f}\big) \geq C u(n) n^{-2r/(2r+1)}$$
$$\text{with } u(n) = \big[\eta^{-1}\big(n^{-b/(2r+1)}\big)\big]^{2r/(r+1/2)} \to \infty.$$

In particular, $u(n) \sim (\ln n)^{2r/(2r+1)}$ *if* $\eta(x)$ *is a normal p.d.f.,* $u(n) \sim (\ln n)^{4r/(2r+1)}$ *if* $\eta(x)$ *is a double-exponential p.d.f. and* $u(n) \sim n^{br[(r+1/2)^2(2+\delta)]^{-1}}$ *if* $\eta(x) \sim |x|^{-(2+\delta)}$ *as* $|x| \to \infty$.

4. Discussion.

4.1. *Discussion of the simulations.* The case of the present paper is an occasion where extensive finite sample simulations have been carried out a few years before theoretical properties of the estimators have been studied. Clyde and George [8] discuss wavelet regression with errors being a scale mixture of normal errors and the priors on wavelet coefficients of the forms

$$\theta_{jk}|\lambda_{jk}^*, \gamma_{jk} \sim N(0, \sigma^2 c_j \gamma_{jk}/\lambda_{jk}^*), \qquad \lambda_{jk}^* \sim h^*, \qquad \gamma_{jk} \sim \text{Bernoulli}(\pi_j).$$

They consider threshold shrinkage and multiple shrinkage estimators, both based on the posterior mean, with the difference that the first one is calculated conditional on γ_{jk} and data, while the second one is conditional on data only. Similarly to

Johnstone and Silverman [17, 19], the authors estimate hyperparameters by maximizing marginal likelihood. They show that their estimators are computationally competitive with standard classical thresholding methods and also perform well in the case of nonnormally distributed errors. It is easy to check that the multiple shrinkage estimators of Clyde and George [8] coincide with the ones proposed in the present paper. Simulation study shows (see Figure 2 of their paper) that multiple shrinkage estimators are superior to the threshold shrinkage ones; therefore, the authors use multiple shrinkage estimators in their later simulations.

Clyde and George [8] suggest several choices for the prior and error distributions (normal prior–normal errors, normal prior–Student t_ν errors, t_ν prior–t_ν errors, etc.), but the only cases that made it to the actual simulations are normal prior–normal errors (N), Cauchy prior–t_5 errors (C5) and uncorrelated but dependent t_5 errors and prior (T5). Note that all of the three models satisfy Corollary 1 and are asymptotically optimal in B_{pq}^r with $p \geq 2$ without additional assumptions on the error. Hence, though no studies for the less favorable cases are presented in the paper by Clyde and George [8], they probably have been eliminated as the ones producing inferior results.

Clyde and George [8] run simulations for the by now traditional test functions "blocks," "bumps," "doppler" and "heavisine" proposed by Donoho and Johnstone [12] and report results for three types of actual error distributions: normal, t_5 errors in wavelet domain and t_5 errors in data domain (see Figures 3, 5 and 6 of their paper). Before discussing Clyde and George's [8] findings we want to draw the reader's attention to the fact that their enumeration of resolution levels is just the opposite from ours: their highest resolution level is the most coarse while ours is the finest one.

Clyde and George [8] state that empirical Bayes wavelet estimators based on the models N, C5 and T5 are in the majority of cases superior to the traditional thresholding rules. However, they have no tool to explain the discrepancy in the performances of these three models, and discrepancies in the performance of various Bayesian models in general. We now are ready to make this final step.

While performing simulations with normal errors the authors discovered that model N gives somewhat better precision than the T5 and C5 models. The results reverse when the error has the Student t_5 distribution in the wavelet domain: models T5 and C5 lead to smaller MSE than model N. Clyde and George [8] remark that "performance of all the estimators tends to be worse under heavy-tailed error distributions" and that "the performance of N worsens the most." However, this is not surprising in view of the findings of Section 3.2. Note that heavy-tailed error distributions lead to increase in variance, and model N is the most vulnerable to this increase. Recall that at high resolution levels the variances of Bayesian estimators of the wavelet coefficients are bounded by a constant times n/ν_j^4 for the models with $\lambda_\eta = 0$, while for the model N they are proportional to n/ν_j^2 times the variance of the wavelet coefficient. This is the reason why the growth of the variance decreases efficiency of model N the most. The same phenomenon (although

somewhat milder due to the central limit theorem) carries over to the case when the errors have t_5 distributions in the data domain.

The above also explains why Bayesian models show the least discrepancy for the "bumps" test function. This function, as Figure 4 of [8] shows, has the smallest value of prior variance which translates into higher ν_j in our notation. This leads to a higher proportion of the bias component in the overall error, and, by far, the highest value of the overall error among the four test functions. Hence, the variance component has the lowest weight in the overall error for the "bumps" function, so that all three models, N, C5 and T5, show similar performance in this case.

4.2. *Discussion and summary of the models.* Table 1 summarizes comparison of the Bayesian models carried out in the previous section. We assume that $\eta_j(\cdot) \equiv \eta(\cdot)$ are identical and consider three choices for $\xi(\cdot)$ and $\eta(\cdot)$: the normal, the double-exponential and the t distribution.

The choices of ξ and η are listed in the first row and in the first column, respectively. The table presents asymptotic expressions for $n^{2r/(2r+1)} R(n, B^r_{pq}(A))$, which we denote by Δ_1 when $p \geq 2$ and Δ_2 when $1 \leq p < 2$; hence, Δ_1 and Δ_2 show deviation from the "ideal" rate $O(n^{-2r/(2r+1)})$. We also introduce the common parameter $\varsigma = 4r + p(2r+1)$. The case of the mixture distribution for errors is not covered by the table since it mimics the normal η–normal ξ or normal η–heavy-tailed ξ, depending on whether ξ is normal or heavy-tailed, with the only difference that the restriction (3.22) or (3.23) on the unknown distribution μ or β_{jn}, respectively, is unnecessary. To be more specific, we distinguish the situations when we have a mixed model (2.5) or a normal model [$\lambda_j \equiv 0$ in (2.5)] for errors. If both models give the same results with no additional assumptions, we leave the cell of the table unmarked; otherwise, if assumption (3.22) or (3.23) is required, we mark the cell with $\diamondsuit$. The cells where the additional assumption (3.19) on β_{jn}, $j \leq j_0$, is required are marked with $\triangle$.

Table 1 provides a comprehensive comparison of the models. As one can see, for spatially homogeneous Besov spaces ($p \geq 2$), the models of choice are the models which provide the optimal convergence rate, N–N, DE–DE, DE–T and T–T, where N, DE and T stand for normal, double-exponential and t distributions, respectively; the first letter gives the choice of η and the second that of ξ. These models provide optimal convergence rates with very few restrictions [only (A3) and (A4)], and no additional assumptions are required if the errors in (1.1) are not normally distributed. The models N–DE and N–T provide optimal convergence rates up to a logarithmic factor; however, the main flaw of these models is that they are sensitive to slow descent of the error distribution μ at $\pm\infty$: if (3.23) is not guaranteed, additional assumptions (3.22) on β_{jn} for $j > J_0$ are necessary. To avoid this feature of N–DE and N–T models one can use models with the mixture error distribution (3.36), where $\zeta(\cdot)$ is a heavy-tailed p.d.f. We shall denote these models by Nζ–N, Nζ–DE and Nζ–T, depending on the choice of ξ. Note that

TABLE 1
Comparison of various models

$\xi(x)$ / $\eta(x)$	Normal $\dfrac{1}{\sqrt{2\pi}\sigma_1}\exp\!\left(-\dfrac{x^2}{2\sigma_1^2}\right)$	Double-exponential $\dfrac{1}{2\sigma_1}\exp\!\left(-\dfrac{|x|}{\sigma_1}\right)$	t-distribution $\dfrac{\Gamma((v_1+1)/2)(v_1\pi)^{-1/2}}{\Gamma(v_1/2)(1+(x^2/v_1))^{(v_1+1)/2}}$
Normal $\dfrac{1}{\sqrt{2\pi}\sigma_0}\exp\!\left(-\dfrac{x^2}{2\sigma_0^2}\right)$	$\Delta_1 = O(1)$ if $\sigma_0 \le \sigma_1$ or (3.19) $\Delta_2 = O((\ln n)^{4r/\varsigma})$ if $\sqrt{2}\sigma < \sigma_0 \le \sigma_1$	$\Delta_1 = O((\ln n)^{1/(2r+1)})$ $\Delta_2 = O((\ln n)^{\kappa})$ $\kappa = \max\left(\dfrac{4r}{\varsigma}, \dfrac{1}{2r+1}\right)$ $\diamond$	$\Delta_1 = O((\ln n)^{1/(2r+1)})$ $\Delta_2 = O((\ln n)^{\kappa})$ $\kappa = \max\left(\dfrac{4r}{\varsigma}, \dfrac{1}{2r+1}\right)$ $\diamond$
Double-exponential $\dfrac{1}{2\sigma_0}\exp\!\left(-\dfrac{|x|}{\sigma_0}\right)$	$\Delta_1 = O(n^{2r/((2r+1)(2r+2))})$ $\triangle$	$\Delta_1 = O(1)$ if $\sigma_0 \le \sigma_1$ or (3.19) $\Delta_2 = O((\ln n)^{8r/\varsigma})$ if $\sigma_0 \le \sigma_1$	$\Delta_1 = O(1)$ $\Delta_2 = O((\ln n)^{8r/\varsigma})$
t-distribution $\dfrac{\Gamma((v_0+1)/2)(v_0\pi)^{-1/2}}{\Gamma(v_0/2)(1+(x^2/v_0))^{(v_0+1)/2}}$	$\Delta_1 = O(n^{2r/((2r+1)(2r+2))})$ $\triangle$	$\Delta_1 = O(1)$ $\triangle$	$\Delta_1 = O(1)$ if $v_0 \ge v_1$ or (3.19) $\Delta_2 = O(n^{\varpi}(\ln n)^{-p})$ if $v_0 \ge v_1$ $\varpi = \dfrac{2r+2}{r(2r+1)(1+v_0)}$

$\Delta_i = n^{2r/(2r+1)} R_n(B^r_{pq}(A), \hat{f})$ with $i=1$ if $p \ge 2$ and $i=2$ if $1 \le p < 2$.

$\varsigma = 4r + p(2r+1)$, σ^2 is the true error variance.

$\diamond$ requires assumption (3.22) or (3.23) for non-Gaussian errors in (1.1).

$\triangle$ requires assumption (3.19) on β_{jn}, $j \le j_0$.

models Nζ–DE and Nζ–T behave similarly to N–DE and N–T but do not require assumption (3.22) or (3.23).

The situation changes if one considers spatially nonhomogeneous Besov spaces ($1 \le p < 2$). In these cases, the model T–T becomes suboptimal while N–N, N–DE, N–T, DE–DE and DE–T still provide optimality up to a logarithmic factor. The model N–N, however, requires the very restrictive assumption $\sigma_0^2 > 2\sigma^2$, which can cripple empirical Bayes inference on parameters of the model. The models N–DE and N–T still require additional assumptions on β_{jn} for $j > J_0$ in this case. Hence, the models of choice in this situation are DE–DE, DE–T (although they have a slightly higher power of the log-factor), Nζ–DE, Nζ–T, or N–DE and N–T if one is sure that the error distribution has fairly fast decay.

The T–DE model provides optimal convergence rates in spatially homogeneous Besov spaces but only under condition (3.19) which, as one can see from Corollary 4, prevents $f \in B_{pq}^r$ a priori. Hence, this model is inferior to DE–DE, DE–T and T–T and, for this reason, has not been studied in the case of $p < 2$. Note that N–N, DE–DE and T–T require condition (3.19) if $\sigma_0 > \sigma_1$ or $\nu_0 < \nu_1$.

The least advantageous choices in terms of asymptotic convergence rates are DE–N and T–N. Not only do these models fail to provide optimal convergence rates but they also require assumption (3.19) which prevents $f \in B_{pq}^r$ a priori.

Section 3.5 allows us to supplement our comparison with some restrictions on $\xi(\cdot)$. In order that $f \in B_{pq}^r$, condition (3.40) requires p.d.f. ξ to have at least $\max(p, q)$ finite moments. Hence, condition (3.40) prevents $f \in B_{p\infty}^r$ whenever ξ is a p.d.f. of the t distribution. In general, Theorem 8 calls for prior distributions with faster descent at $\pm\infty$. However, it should be noted that the considerations of Theorem 8 have no bearing on the asymptotic convergence rates of the estimator.

Finally, some recommendations for a choice of a model can be drawn. If the errors in (1.1) are normally distributed, one can use DE–DE, DE–T, N–DE and N–T models which ensure optimality (up to a logarithmic factor) in both spatially homogeneous and nonhomogeneous Besov spaces. However, if the normality assumption is violated and errors may have a heavy-tailed distribution, DE–DE, DE–T, Nζ–DE and Nζ–T will be safer choices. In addition, if one wants to be true to the Bayesian spirit and make sure that $f \in B_{pq}^r$ a priori for any possible q, then the models where ξ decreases exponentially (i.e., DE–DE and Nζ–DE) may seem more preferable.

5. Proofs.

PROOF OF LEMMA 1. Since $I_0(d)$ is an even and $I_1(d)$ is an odd function of d, consider the case $d > 0$. Note that

$$0 \le I_1(d) = dI_0(d) - \int_{-\infty}^{\infty} (d-x)\sqrt{n}\eta[\sqrt{n}(d-x)]\nu_j\xi(\nu_j x)\,dx$$

(5.1)

$$= dI_0(d) - \int_{-\infty}^{\infty} z\sqrt{n}\eta(\sqrt{n}z)\nu_j\xi(\nu_j(d-z))\,dz \le dI_0(d)$$

since the integral in (5.1) is nonnegative. Thus, $I_1(d) \leq d I_0(d)$. □

Proof of Theorem 1 is based on the following lemmas.

LEMMA 2. *If $v_j|d|$ is bounded or $v_j|v_j d|^{\lambda_\xi}/\sqrt{n} \to 0$, then as $v_j/\sqrt{n} \to 0$*

(5.2) $$I_0(d) = v_j \xi(v_j d)[1 + O(n^{-1} v_j^2 |v_j d|^{2\lambda_\xi})],$$

(5.3)
$$\frac{I_1(d)}{I_0(d)} = d - \eta_2 \frac{v_j}{n} \frac{\xi'(v_j d)}{\xi(v_j d)} \left[1 + O\left(\frac{v_j^2 |v_j d|^{2\lambda_\xi}}{n}\right)\right]$$
$$= d - O\left(\frac{v_j |v_j d|^{\lambda_\xi}}{n}\right).$$

If $\sqrt{n}|d|$ is bounded or $\sqrt{n}|\sqrt{n}d|^{\lambda_\eta}/v_j \to 0$, then as $\sqrt{n}/v_j \to 0$

(5.4) $$I_0(d) \sim \sqrt{n}\eta(\sqrt{n}d)[1 + O(n v_j^{-2} |\sqrt{n}d|^{2\lambda_\eta})],$$

(5.5) $$\frac{I_1(d)}{I_0(d)} \sim \xi_2 \frac{\sqrt{n}}{v_j^2} \frac{\eta'(\sqrt{n}d)}{\eta(\sqrt{n}d)}\left[1 + O\left(\frac{n}{v_j^2}|\sqrt{n}d|^{2\lambda_\eta}\right)\right] = O\left(\frac{\sqrt{n}}{v_j^2}|\sqrt{n}d|^{\lambda_\eta}\right).$$

Here $\eta_2 = \int_{-\infty}^{\infty} x^2 \eta(x)\,dx$, $\xi_2 = \int_{-\infty}^{\infty} x^2 \xi(x)\,dx$.

PROOF. We shall give the proofs for (5.4) and (5.5); the proofs of (5.2) and (5.3) are conducted in a similar manner. Change variables $y = v_j x$ in (2.9) and use Taylor series expansion:

$$\frac{v_j^i}{\sqrt{n}} I_i(d) = \int_{-\infty}^{\infty} y^i \eta\left(\sqrt{n}d - \frac{\sqrt{n}}{v_j}y\right)\xi(y)\,dy$$
$$= \int_{-\infty}^{\infty} y^i \left[\eta(\sqrt{n}d)\right.$$
$$\left. - \frac{\sqrt{n}}{v_j}y\eta'(\sqrt{n}d) + \frac{n}{2v_j^2}y^2\eta''(\sqrt{n}d)\right.$$
$$\left. - \frac{n\sqrt{n}}{6v_j^3}y^3\eta'''(\sqrt{n}d) + \cdots\right]\xi(y)\,dy.$$

Letting $i = 0$ and $i = 1$ in (5.6) and taking into account that $\xi(\cdot)$ is an even function, we obtain (5.4) and (5.5). □

LEMMA 3. *Let $\varrho(n)$ be defined by (3.20) and (3.21) and let σ be as in (2.5). Then*

(5.6) $$E(d_{jk} - \theta_{jk})^{2i} = O(n^{-i}), \qquad i = 1, 2;$$

(5.7) $$P(\sqrt{n}|d_{jk} - \theta_{jk}| > a\sqrt{\ln n}) = o(n^{-a^2/(2\sigma^2)}), \qquad j < J_0;$$

(5.8) $$E[(d_{jk} - \theta_{jk})^2 I(\sqrt{n}|d_{jk} - \theta_{jk}| > \varrho(n))] = O(n^{-4r/(2r+1)}).$$

PROOF.   Validity of statements (5.6) and (5.7) follows directly from the fact that [cf. (2.5)]

$$
(5.9) \qquad d_{jk} - \theta_{jk} \sim (1 - \lambda_j)\sqrt{n}(\sqrt{2\pi}\sigma)^{-1} \exp\{-nx^2/(2\sigma^2)\}
$$
$$
+ \lambda_j \sqrt{n}\mu(\sqrt{n}x). \qquad \square
$$

LEMMA 4.   *Let ϕ and ψ be boundary coiflets introduced in [18, 19] possessing $s > r$ vanishing moments and based on orthonormal coiflets supported in $[-S + 1, S]$, $s < S$. Assume that $p \geq 1$, $r > \max(1/2, 1/p, r_p)$, where r_p is defined in (3.1) and $L \geq \log_2(6S - 6)$. If $f \in B_{p,q}^r(A)$, then for some absolute positive constants A_1 and A_2*

$$
(5.10) \qquad \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} (\tilde{\theta}_{jk} - \theta_{jk})^2 \leq A_1 n^{-2r/(2r+1)},
$$

$$
(5.11) \qquad \sum_{k=0}^{2^j-1} \theta_{jk}^2 \leq A_2 2^{-2j[r-(1/p-1/2)_+]}.
$$

PROOF.   The proof is based on Proposition 5 of [18], which under conditions of Lemma 4 can be written as

$$
(5.12) \qquad \sum_{k=0}^{2^j-1} |\tilde{\theta}_{jk} - \theta_{jk}|^p \leq AC(r, p, \varphi, \psi)2^{-rJp+(1/p-1/2)jp}.
$$

To prove (5.10), consider cases $p \geq 2$ and $1 \leq p < 2$ separately. If $p \geq 2$, the Hölder inequality leads to $\sum_{k=0}^{2^j-1}(\tilde{\theta}_{jk} - \theta_{jk})^2 = O(2^{-2rJ+2(1/p-1/2)j}2^{j(1-2/p)}) = O(2^{-2rJ})$. Adding the terms together we obtain

$$
\sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} (\tilde{\theta}_{jk} - \theta_{jk})^2 = O\left(\sum_{j=L}^{J-1} 2^{-2rJ}\right) = O(n^{-1}\ln n) = O(n^{-2r/(2r+1)}).
$$

If $1 \leq p < 2$, then $\sum_{k=0}^{2^j-1}(\tilde{\theta}_{jk} - \theta_{jk})^2 \leq (\sum_{k=0}^{2^j-1} |\tilde{\theta}_{jk} - \theta_{jk}|^p)^{2/p} = O(2^{-2rJ+2(1/p-1/2)j})$. Summing the terms, we arrive at $\sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1}(\tilde{\theta}_{jk} - \theta_{jk})^2 = O(2^{-2J(r+1/2-1/p)}) = O(n^{-2r/(2r+1)})$ provided $r > r_p$.

To derive (5.11), note that $\tilde{\theta} \in B_{p\infty}^r(A)$, where with some abuse of notation we use $B_{p\infty}^r$ to denote Besov space of infinite sequences. Hence $\sum_{k=0}^{2^j-1} \tilde{\theta}_{jk}^2 = O(2^{2j(r-(1/p-1/2)_+)})$, which in combination with (5.10) yields (5.11). $\square$

PROOF OF THEOREM 1. Since the wavelet basis is orthonormal,

$$(5.13) \quad R_n\big(B^r_{p,q}(A), \hat{f}\big) = \sum_{k=0}^{2^L-1} E(\hat{\theta}_k - \tilde{\theta}_k)^2$$

$$+ \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \tilde{\theta}_{jk})^2 + \sum_{j=J}^{\infty} \sum_{k=0}^{2^j-1} \tilde{\theta}_{jk}^2.$$

Observe that the first term in (5.13) is bounded by $2\sum_{k=0}^{2^L-1}[\mathrm{Var}(\hat{\theta}_k) + (\theta_k - \tilde{\theta}_k)^2] = O(n^{-1})$, while the last term is bounded by $A2^{-2rJ} = O(n^{-2r})$ due to (5.11). By Lemma 4, the second term in (5.13) is dominated by

$$(5.14) \quad \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \tilde{\theta}_{jk})^2 \le 2 \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2 + A_2 n^{-2r},$$

that is, the main contribution to $R_n(B^r_{p,q}(A), \hat{f})$ is made by the first term in (5.14). Therefore, we need to construct an asymptotic upper bound for $R = \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2 = R_1 + R_2$ with

$$(5.15) \quad R_1 = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2, \qquad R_2 = \sum_{j=j_0+1}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2.$$

Let us examine each of the terms in turn. Denote

$$(5.16) \quad A_{jn}(d) = \beta_{jn} \sqrt{n} \eta_j(\sqrt{n}d)/I_0(d)$$

and note that $R_1 \le 2(R_{11} + R_{12})$ where

$$R_{11} = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E\left(\frac{I_1(d_{jk})}{I_0(d_{jk})} - \theta_{jk}\right)^2,$$

$$(5.17)$$

$$R_{12} = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E\left(\frac{I_1(d_{jk})/I_0(d_{jk})}{1 + A_{jn}(d_{jk})} - \frac{I_1(d_{jk})}{I_0(d_{jk})}\right)^2.$$

To establish an asymptotic upper bound for R_{11} observe that by combination of Lemma 2 and (3.8), for $j \le j_0$

$$E\big(I_1(d_{jk})/I_0(d_{jk}) - \theta_{jk}\big)^2$$

$$\le 2\big[E\big(I_1(d_{jk})/I_0(d_{jk}) - d_{jk}\big)^2 + E(d_{jk} - \theta_{jk})^2\big]$$

$$(5.18)$$

$$= O\big(E(v_j^4 d_{jk}^2/n^2) + v_j^2/n^2 + \sigma^2/n\big)$$

$$= O(v_j^4 \theta_{jk}^2/n^2 + v_j^2/n^2 + \sigma^2/n),$$

so that by (3.14)

$$(5.19) \quad R_{11} = O\left(\sum_{j=L}^{j_0} [2^{-2jr} v_j^4/n^2 + 2^j v_j^2/n^2 + 2^j/n]\right) = O(n^{-2r/(2r+1)}).$$

In the case of R_{12}, note that $R_{12} = R_{121} + R_{122}$ where

$$(5.20)$$
$$R_{121} = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E\left(\frac{I_1(d_{jk})}{I_0(d_{jk})} \frac{A_{jn}(d_{jk})}{1 + A_{jn}(d_{jk})} I(v_j|d_{jk}| > C_\delta)\right)^2,$$

$$R_{122} = \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E\left(\frac{I_1(d_{jk})}{I_0(d_{jk})} \frac{A_{jn}(d_{jk})}{1 + A_{jn}(d_{jk})} I(v_j|d_{jk}| \le C_\delta)\right)^2.$$

For R_{121}, recall that since $v_j \le \sqrt{n}$ and by assumptions (A3) and (A4)

$$(5.21) \quad \frac{\eta_j^2(\sqrt{n}d_{jk})}{\xi^2(v_jd_{jk})} \le C_{\xi,\eta}^2 \left[\frac{(\sqrt{n}d_{jk})^{2+\delta}\eta_j(\sqrt{n}d_{jk})}{(v_jd_{jk})^{2+\delta}\eta_j(v_jd_{jk})}\right]^2 \frac{v_j^{4+2\delta}}{n^{2+\delta}} \le C_{\xi,\eta}^2 \frac{v_j^{4+2\delta}}{n^{2+\delta}}.$$

By (3.16), (5.16) and since $\delta + (2r+1)^{-1} > (2r+1)^{-1}(2+\delta) - 1$, the latter implies that

$$(5.22)$$
$$R_{121} = O\left(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} \beta_{jn}^2 v_j^{2+2\delta} n^{-(1+\delta)} [E(d_{jk} - \theta_{jk})^2 + \theta_{jk}^2]\right)$$
$$= O\left(n^{-2r/(2r+1)} \sum_{j=L}^{j_0} \beta_{jn}^2 \left(\frac{v_j^2}{n}\right)^{1/(2r+1)+\delta}\right) = O(n^{-2r/(2r+1)}).$$

For the sake of construction of an upper bound for R_{122}, note that since $v_j|d_{jk}| \le C_\delta$ we have $I_0(d_{jk}) \sim v_j\xi(v_jd_{jk}) \ge v_j\xi(C_{dd})$. Thus, by $x/(1+x) \asymp \min(x, 1)$, we derive that $A_{jn}(d_{jk}) = O(\min(1, \beta_{jn}^2 n v_j^{-2} \eta_j^2(\sqrt{n}d_{jk})))$. Let α_{jn} be a unique positive solution of $\beta_{jn}^2 n v_j^{-2} \eta_j^2(\alpha_{jn}) = 1$, that is,

$$(5.23) \qquad\qquad \alpha_{jn} = \eta_j^{-1}\left(\frac{v_j}{\sqrt{n}\beta_{jn}}\right).$$

Since $\eta_j(x)$ are decreasing as $x > 0$, $A_{jn}(d_{jk}) \le 1$ iff $|d_{jk}| \ge \alpha_{jn}/\sqrt{n}$. Hence, by condition (A4), $E[d_{jk}^2 \min(1, A_{jn}^2(d_{jk}))] = E[d_{jk}^2 I(\sqrt{n}|d_{jk}| \le \alpha_{jn})] + E[d_{jk}^2 \beta_{jn}^2 n v_j^{-2} \eta_j^2(\sqrt{n}d_{jk}) I(\sqrt{n}|d_{jk}| > \alpha_{jn})]$, so that

$$(5.24) \quad E[d_{jk}^2 \min(1, A_{jn}^2(d_{jk}))] \le \alpha_{jn}^2/n + \beta_{jn}^2 n v_j^{-2} \eta_j^2(\alpha_{jn})\alpha_{jn}^2/n = 2\alpha_{jn}^2/n,$$

since $x\eta_j(x)$ are nonincreasing functions of x for $x > 0$. Now, note that, as $v_j|x| \ge C_\delta$, by condition (A4) $|x|^{2+\delta}\eta_j(x) \le C_\delta^{2+\delta}\eta_j(C_\delta) \le C_\delta^{2+\delta}\eta_j(0) \le$

$C_\delta^{2+\delta} C_{\xi\eta} \xi(0) \equiv C_\eta^*$. Therefore, $\eta_j(x) \leq C_\eta^{(1)} |x|^{-(2+\delta)}$ for any x and $\eta_j^{-1}(z) \leq C_\eta^{(2)} z^{-1/(2+\delta)}$. Let $\beta_{jn} = (\sqrt{n}/v_j)^\alpha$. Then, from (5.16)

$$(5.25) \qquad \alpha_{jn} = \eta_j^{-1}([v_j/\sqrt{n}]^{1+\alpha}) = O([\sqrt{n}/v_j]^{(1+\alpha)/(2+\delta)}).$$

Combining (5.20), (5.24), (5.25) and Lemma 1, we derive

$$
\begin{aligned}
R_{122} &= O\left(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E[d_{jk}^2 \min(1, A_{jn}^2(d_{jk}))] \right) \\
&= O\left(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} n^{-1} [n/v_j^2]^{(1+\alpha)/(2+\delta)} \right) \\
&= O\left(\sum_{j=L}^{j_0} \frac{2^j}{n} \left[\frac{n}{v_j^2}\right]^{(1+\alpha)/(2+\delta)} \right) \\
&= n^{-2r/(2r+1)} O\left(\sum_{j=L}^{j_0} \left[\frac{2^{(2r+1)j}}{n}\right]^{[1/(2r+1)-(1+\alpha)/(2+\delta)]} \right) \\
&= O(n^{-2r/(2r+1)})
\end{aligned}
$$

(5.26)

since $(1+\alpha)(2r+1) < 2+\delta$. Combination of (5.19), (5.20), (5.22) and (5.26) ensures that $R_1 = O(n^{-2r/(2r+1)})$.

Now, consider R_2 given by (5.15). Since $|\hat{\theta}_{jk}| \leq |I_1(d_{jk})/I_0(d_{jk})|$, by combination of Lemma 2 and (3.10)

$$
\begin{aligned}
R_2 &= O\left(\sum_{j=j_0+1}^{J-1} \sum_{k=0}^{2^j-1} \left[E\left(\frac{I_1(d_{jk})}{I_0(d_{jk})}\right)^2 + \theta_{jk}^2 \right] \right) \\
&= O\left(\sum_{j=j_0+1}^{J-1} \sum_{k=0}^{2^j-1} \left[\frac{n}{v_j^4} + \frac{n^2 E d_{jk}^2}{v_j^4} + \theta_{jk}^2 \right] \right) \\
&= O\left(\sum_{j=j_0+1}^{J-1} [2^j n 2^{-j(4r+2)} + 2^{-2jr}] \right) \\
&= O(n^{-2r/(2r+1)}),
\end{aligned}
$$

(5.27)

which completes the proof. $\square$

PROOF OF COROLLARY 1. Proof follows directly from Theorem 1 by Lemma 2 whenever $\lambda_\xi = \lambda_\eta = 0$. If $\xi(x)$ and $\eta_j(x)$ are normal p.d.f.'s, validity of (3.8)–(3.10) can be verified by direct calculation of $I_0(d)$ and $I_1(d)$. $\square$

PROOF OF COROLLARY 1*. Using properties of the Fourier transform and formulae (8.432.5), (8.468), (3.944.5) and (3.944.6) of [15], one can show by direct calculation that $|I_1(d)/I_0(d) - d| = O(|d|v_j/\sqrt{n})$, $I_0(d) \sim v_j\xi(v_jd)$, if $v_j/\sqrt{n} \to 0$ and $|I_1(d)/I_0(d)| = O(\sqrt{n}|d|/v_j)$ if $v_j/\sqrt{n} \to \infty$, so Theorem 1 remains valid. $\square$

PROOF OF THEOREM 2. It is easy to see that condition (A3) is used only for derivation of the R_{121} term in (5.20). Since $I(v_j|d_{jk}| > C_\delta) \leq I(\sqrt{n}|d_{jk} - \theta_{jk}| > a\sqrt{\ln n}) + I(v_j|d_{jk}| > C_\delta)I(v_j|d_{jk}| \leq v_j|\theta_{jk}| + av_j/\sqrt{\ln n}\sqrt{n})$, we have $R_{121} = R_{1211} + R_{1212}$. Here, by Lemma 3,

$$R_{1211} = O\left(\sum_{j=L}^{j_0}\sum_{k=0}^{2^j-1} E[d_{jk}^2 I(\sqrt{n}|d_{jk} - \theta_{jk}| > a\sqrt{\ln n})]\right)$$

$$= O\left(\sum_{j=L}^{j_0}\sum_{k=0}^{2^j-1} [E(d_{jk} - \theta_{jk})^2 + \theta_{jk}^2 P(\sqrt{n}|d_{jk} - \theta_{jk}| > a\sqrt{\ln n})]\right)$$

$$= O(n^{-2r/(2r+1)}) + O\left(\sum_{j=L}^{j_0} 2^{-2jr}n^{-a^2/2\sigma^2}\right) = O(n^{-2r/(2r+1)}),$$

provided $a^2 \geq 4r\sigma^2/(2r+1)$. For R_{1212}, comparing with (5.21) and (5.22), take into account that $v_j|\theta_{jk}| \leq B_1 2^{j/2}$ and $v_j\sqrt{\ln n}/\sqrt{n} = o(2^{j/2})$. Using (3.19), obtain

$$R_{1212} = O\left(\sum_{j=L}^{j_0}\sum_{k=0}^{2^j-1}\left[\frac{\beta_{jn}^2}{v_j^2}\left(\frac{v_j^2}{n}\right)^{1+\delta}\left(v_j|\theta_{jk}| + a\frac{v_j}{\sqrt{n}}\sqrt{\ln n}\right)^2\right.\right.$$

$$\left.\left.\times U^2\left(v_j|\theta_{jk}| + a\frac{v_j}{\sqrt{n}}\sqrt{\ln n}\right)\right]\right)$$

$$= O\left(\sum_{j=L}^{j_0}\frac{\beta_{jn}^2}{2^{2rj}}\left(\frac{v_j^2}{n}\right)^{1+\delta} 2^j U^2(2B_1 2^{j/2})\right) = O(n^{-2r/(2r+1)}). \qquad \square$$

PROOF OF THEOREM 3. Note that since condition (3.10) is no longer valid, we need to derive new upper bounds for R_2. Note that $|\hat{\theta}_{jk} - \theta_{jk}| \leq |\theta_{jk}| + |\hat{\theta}_{jk}|$ and $\sum_{j=j_0+1}^{J_0}\sum_{k=0}^{2^j-1} \theta_{jk}^2 = O(n^{-2r/(2r+1)})$, so we shall be concerned with the $|\hat{\theta}_{jk}|$ term only. Partition $\sum_j\sum_k \hat{\theta}_{jk}^2$ into the sum over $j_0 \leq j \leq J_0$ and $J_0 + 1 \leq j \leq J - 1$ and denote the respective sums by R_3 and R_4.

To analyze R_3 note that $|\hat{\theta}_{jk}| \leq |I_1(d_{jk})/I_0(d_{jk})|$ and that

$$(5.28) \quad \begin{aligned} 1 \leq{}& I(\sqrt{n}|\theta_{jk}| > a\sqrt{\ln n}) + I(\sqrt{n}|d_{jk} - \theta_{jk}| > a\sqrt{\ln n}) \\ &+ I(\sqrt{n}|d_{jk} - \theta_{jk}| \leq a\sqrt{\ln n})I(\sqrt{n}|\theta_{jk}| \leq a\sqrt{\ln n}). \end{aligned}$$

Hence $R_3 = R_{31} + R_{32} + R_{33}$, where by Lemmas 1 and 3

$$R_{31} = O\left(\sum_{j=j_0+1}^{J_0} \sum_{k=0}^{2^j-1} E[d_{jk}^2 I(\sqrt{n}|\theta_{jk}| > a\sqrt{\ln n})]\right)$$

$$(5.29) \qquad = O\left(\sum_{j=j_0+1}^{J_0} \sum_{k=0}^{2^j-1} \left[\theta_{jk}^2 + E(d_{jk} - \theta_{jk})^2 I\left(\frac{1}{n} < \frac{\theta_{jk}^2}{a^2 \ln n}\right)\right]\right)$$

$$= o(n^{-2r/(2r+1)}),$$

$$R_{32} = O\left(\sum_{j=j_0+1}^{J_0} \sum_{k=0}^{2^j-1} E[d_{jk}^2 I(\sqrt{n}|d_{jk} - \theta_{jk}| > a\sqrt{\ln n})]\right)$$

$$= O\left(\sum_{j=j_0+1}^{J_0} \sum_{k=0}^{2^j-1} \left[\sqrt{E(d_{jk} - \theta_{jk})^4}\right.\right.$$

$$(5.30) \qquad \left.\left. \times \sqrt{P(|d_{jk} - \theta_{jk}|\sqrt{n} > a\sqrt{\ln n})} + \theta_{jk}^2\right]\right)$$

$$= O\left(\sum_{j=j_0+1}^{J_0} \sum_{k=0}^{2^j-1} [n^{-1}n^{-2r/(2r+1)} + \theta_{jk}^2]\right)$$

$$= O(2^{J_0}n^{-(4r+1)/(2r+1)}) = o(n^{-2r/(2r+1)})$$

provided $a^2 \geq 8\sigma^2 r/(2r+1)$. To derive an asymptotic upper bound for R_{33}, note that

$$I(\sqrt{n}|d_{jk} - \theta_{jk}| \leq a\sqrt{\ln n})I(\sqrt{n}|\theta_{jk}| \leq a\sqrt{\ln n})$$

$$\leq I(\sqrt{n}|d_{jk}| \leq 2a\sqrt{\ln n})$$

$$(5.31) \qquad \leq I(\sqrt{n}|d_{jk}| \leq 2a\sqrt{\ln n})I\left(\frac{\sqrt{n}}{v_j}(\sqrt{\ln n})^{\lambda_\eta} \to 0\right)$$

$$+ I\left(\frac{v_j}{\sqrt{n}}(\sqrt{\ln n})^{-\lambda_\eta} = O(1)\right)$$

$$\leq I(v_j^{-1}\sqrt{n}(\sqrt{n}|d_{jk}|)^{\lambda_\eta} \to 0) + I(2^{j(2r+1)} = O(n(\ln n)^{\lambda_\eta})).$$

Note that Lemma 2 and $\sqrt{n}v_j^{-1}(\sqrt{n}|d_{jk}|)^{\lambda_\eta} \to 0$ imply that

$$E[[I_1(d_{jk})/I_0(d_{jk})]^2 I(v_j^{-1}\sqrt{n}(\sqrt{n}|d_{jk}|)^{\lambda_\eta} \to 0)]$$

$$= O(nv_j^{-4}(\sqrt{n}|d_{jk}|)^{2\lambda_\eta} I(v_j^{-1}\sqrt{n}(\sqrt{n}|d_{jk}|)^{\lambda_\eta} \to 0))$$

$$= O(v_j^{-2}) = O(2^{-(2r+1)j}).$$

Therefore, by calculations similar to (5.27), the portion of R_{33} corresponding to the first term in (5.31) is $O(n^{-2r/(2r+1)})$.

By Lemma 1, $E[I_1(d_{jk})/I_0(d_{jk})]^2 = O(E[d_{jk} - \theta_{jk}]^2 + \theta_{jk}^2)$, so the second term in the portion is

$$O\left(\sum_{j=j_0+1}^{J_0} \sum_{k=0}^{2^j-1} [n^{-1} + \theta_{jk}^2] I[2^j = O(n^{1/(2r+1)}(\ln n)^{\lambda_\eta/(2r+1)})]\right)$$

$$= O(n^{-2r/(2r+1)}(\ln n)^{\lambda_\eta/(2r+1)}).$$

Consequently,

$$(5.32) \qquad\qquad R_{33} = O(n^{-2r/(2r+1)}(\ln n)^{\lambda_\eta/(2r+1)}),$$

and (5.29), (5.30) and (5.32) imply that $R_{321} = O(n^{-2r/(2r+1)}(\ln n)^{\lambda_\eta/(2r+1)})$.

Derivation of an asymptotic upper bound for R_4 is, by and large, similar to that for R_3. First, assume that condition (3.22) holds. Rewrite (5.28) and (5.31) with $a\sqrt{\ln n}$ replaced by $\varrho(n)$ and observe that under assumption (3.22) the second indicator $I(2^{j(2r+1)} = O(n(\ln n)^{\lambda_\eta}))$ in (5.31) vanishes. Now, same as before, $R_{41} = o(n^{-2r/(2r+1)})$ and $R_{42} = O(n^{-2r/(2r+1)})$ by (5.8). To ensure that $R_{43} = O(n^{-2r/(2r+1)})$ repeat calculations for the R_{33} corresponding to the first term in (5.31).

If assumption (3.22) is violated, the only term in the proof which is affected is R_{43}. Namely, the second indicator in (5.31) does not vanish. Recall that $R_{43} = O(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E[\hat{\theta}_{jk}^2 I(\sqrt{n}|d_{jk}| \le 2\varrho(n))])$. Since $|\hat{\theta}_{jk}| \le |I_1(d_{jk})|/[\beta_{jn}\sqrt{n}\eta_j(\sqrt{n}d_{jk})]$ and $|I_1(d)| \le \eta(0)\sqrt{n}\nu_j^{-1} \int_{-\infty}^{\infty} |z|\xi(z)\,dz$, we derive that, by (3.23), $R_{43} = O(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E[\beta_{jn}\nu_j\eta_j(2\varrho(n))]^{-2})$. To complete the proof note that the assumption that μ has at least four finite moments implies that $\varrho(n) \le \sqrt{C_\mu} n^{r/(2r+1)}$. Consequently, $R_{43} = O(n^{-2r/(2r+1)})$. $\square$

PROOF OF COROLLARY 2.   The proof follows directly from Lemma 2. $\square$

PROOF OF THEOREM 4.   The fact that condition (3.25) replaces (3.8) affects the term R_{11} only. Denote $\Delta_{jk} = I_1(d_{jk})/I_0(d_{jk}) - I_1(\theta_{jk})/I_0(\theta_{jk})$ and observe that

$$(5.33) \quad R_{11} \le 2 \sum_{j=L}^{J-1} \sum_{k=0}^{2^j-1} \left[E(I_1(\theta_{jk})/I_0(\theta_{jk}) - \theta_{jk})^2 + E\Delta_{jk}^2\right] \equiv 2(R_{111} + R_{112}).$$

For an upper bound on R_{111} note that (5.11) implies that $|\theta_{jk}| \le \sqrt{A}2^{-jr}$, so that $(\nu_j|\theta_{jk}|)^{\lambda_\xi}\nu_j/\sqrt{n} = O(2^{j(r+\lambda_\xi/2+1/2)}/\sqrt{n})$. The last expression does not turn to zero only if $2^j > Cn^{1/(2r+1+\lambda_\xi)}$ for some $C > 0$. Also, it follows from Lemma 2 that

$$(5.34) \quad (\nu_j|\theta_{jk}|)^{\lambda_\xi}\nu_j/\sqrt{n} \to 0 \quad \Longrightarrow \quad |I_1(\theta_{jk})/I_0(\theta_{jk}) - \theta_{jk}| = O(1/\sqrt{n}).$$

Hence,

$$R_{11} = O\left(\sum_{j=L}^{j_0} 2^j/n\right) + O\left(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} \theta_{jk}^2 I[2^j > Cn^{1/(2r+1+\lambda_\xi)}]\right)$$

$$= O(n^{-2r/(2r+\lambda_\xi+1)}).$$

To examine R_{112} consider three separate cases. If both $(v_j|\theta_{jk}|)^{\lambda_\xi} v_j/\sqrt{n}$ and $(v_j|d_{jk}|)^{\lambda_\xi} v_j/\sqrt{n}$ tend to zero, then $\Delta_{jk} \leq |I_1(d_{jk})/I_0(d_{jk}) - d_{jk}| + |d_{jk} - \theta_{jk}| + |I_1(\theta_{jk})/I_0(\theta_{jk}) - \theta_{jk}|$ and, by Lemma 2 and (5.34) we have $E\Delta_{jk}^2 = O(1/n)$. Consequently, the respective portion of R_{112} is $O(n^{-2r/(2r+1)})$. If $(v_j|\theta_{jk}|)^{\lambda_\xi} v_j/\sqrt{n}$ does not tend to zero, then, as was mentioned earlier, $2^j > Cn^{1/(2r+1+\lambda_\xi)}$ for some $C > 0$. Hence the respective portion of R_{112} is $O(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} [\theta_{jk}^2 I(2^j > Cn^{1/(2r+1+\lambda_\xi)}) + E(d_{jk} - \theta_{jk})^2]) = O(n^{-2r/(2r+\lambda_\xi+1)})$. The third case occurs if $(v_j|\theta_{jk}|)^{\lambda_\xi} v_j/\sqrt{n} \to 0$ but $(v_j|d_{jk}|)^{\lambda_\xi} v_j/\sqrt{n}$ does not tend to zero. Then $\Delta_{jk}^2 = O(1/n) + O(|d_{jk} - \theta_{jk}|^2) + O(d_{jk}^2)$ and $\sqrt{n}|d_{jk} - \theta_{jk}| \geq C^*(n/v_j^2)^{(\lambda_\xi+1)/2\lambda_\xi}$. Considering cases $n/v_j^2 > \ln n$ and $n/v_j^2 < \ln n$ separately and applying Lemma 3, we obtain $R_{112} = O(n^{-2r/(2r+\lambda_\xi+1)})$. $\square$

The proof of Corollary 5 is based on the following lemmas.

LEMMA 5. *Let $\xi(\cdot)$ and $\eta_j(\cdot) \equiv \eta(\cdot)$ be such that $|I_1(d)/I_0(d) - d| \geq |d|/2$ for $d = \sqrt{A/2}\, 2^{-i_0 r}$ with $i_0 < j_0$. Then $R(n, H^r(A)) \geq (A/8)2^{-2i_0 r}$.*

PROOF. Consider $f(x) = \theta_{i_0,0}^* \psi_{i_0,0}(x)$ with $\theta_{i_0,0}^* = \sqrt{A/2}\, 2^{-i_0 r}$, $i_0 \geq L$. It is easy to check that $\theta_{i_0,0} = \theta_{i_0,0}^*$ and $\theta_{jk} = 0$ otherwise. Thus, the coefficients θ_{jk} satisfy condition (5.11) with $A_2 = A$ and $p = 2$. For this function f, under the condition of Lemma 5, the bias term exceeds $R_1 \geq C_2^2 \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} \theta_{jk}^2 = \theta_{i_0,0}^2 \geq (A/8)2^{-2i_0 r}$.

$\square$

LEMMA 6. *If $\xi(x)$ is a p.d.f. of the normal distribution and $\eta(x)$ is a p.d.f. of the double-exponential or the Student t distribution, then Lemma 5 holds with $i_0 = (2r + 2)^{-1}\log_2 n$.*

PROOF. Let $v_j/\sqrt{n} \to 0$, and $d = \sqrt{A/2}\, 2^{-jr}$. If $\eta(x)$ is the Student t distribution, then direct calculations show that $|I_1(d)/I_0(d) - d| \sim |d(C_v\sigma^2/(v_j d)^2 - 1)|$ where C_v is a constant depending on the degrees of freedom of the t distribution; hence $|I_1(d)/I_0(d) - d| \geq |d|/2$ for $j \geq i_0$ with $2^{i_0} \geq n^{1/(2r+2)}$.

In the case where $\eta(x)$ is the double-exponential distribution, $|I_1(d)/I_0(d) - d| \sim |d - \sigma^2 \sqrt{n} v_j^{-2}| \geq |d|/2$ for $j \geq i_0$ with i_0 satisfying $2^{i_0} > (8n\sigma^4 A^{-1})^{1/(2r+1)}$. □

PROOF OF COROLLARY 3. Corollary 3 follows directly from Lemmas 5 and 6. □

The proof of Theorem 5 is based on the following lemma.

LEMMA 7. *If $f \in B_{p,q}^r(A)$, then for $b > 0$*

$$\sum_j \sum_{k=0}^{2^j-1} \theta_{jk}^2 I(\sqrt{n}|\theta_{jk}| \leq (\ln n)^b) = O(n^{-2r/(2r+1)}[\ln n]^{4br/(2r+1)}),$$

$$\sum_j \sum_{k=0}^{2^j-1} n^{-1} I(\sqrt{n}|\theta_{jk}| > (\ln n)^b) = O(n^{-2r/(2r+1)}[\ln n]^{-pb}).$$

PROOF. The proof is an adaptation of the proof in Donoho, Johnstone, Kerkyacharian and Picard [13]. □

PROOF OF THEOREM 5. Similarly to (5.15), partition R as $R = R_1 + R_5 + R_2$ where R_1, R_5 and R_2 correspond to $j \leq j_0$, $j_0 < j \leq j_1$ and $j > j_1$, respectively, and consider each term separately. Note also that in condition (A4), δ can be made as large as one desires if $\gamma > 0$ and $\delta \leq \nu - 2$ if $\gamma = 0$.

Low resolution levels: $j \leq j_0$. Note that v_j is chosen so that $v_j/\sqrt{n} = o(1)$ and $\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} (n^{-2} v_j^2 + n^{-2} v_j^4 \theta_{jk}^2) = O(n^{-2r/(2r+1)})$, so that the upper bound (5.19) for R_{11} is still valid. Partition $R_{12} = R_{121} + R_{122}$ as in (5.20). Consider the cases $\gamma > 0$ and $\gamma = 0$ separately. If $\gamma > 0$, then taking into account that $v_j|d_{jk}| > C_\delta$,

$$(5.35) \qquad \frac{\eta_j^2(\sqrt{n}d_{jk})}{\xi^2(v_j d_{jk})} \leq \exp\left\{-\lambda\left[\left(\frac{n}{v_j^2}\right)^{\gamma/2} - 1\right]C_\delta^\gamma\right\} = O\left(\left[\frac{v_j}{\sqrt{n}}\right]^\upsilon\right)$$

for any $\upsilon > 0$. In the case $\gamma = 0$,

$$(5.36) \qquad \frac{\eta_j^2(\sqrt{n}d_{jk})}{\xi^2(v_j d_{jk})} \leq C_{\xi,\eta} \frac{(1 + v_j^2 d_{jk}^2)^{\nu/2}}{(1 + n d_{jk}^2)^{\nu/2}} = O\left(\left[\frac{v_j}{\sqrt{n}}\right]^\nu\right).$$

Hence, following (5.22) and denoting $u = \upsilon$ for $\gamma > 0$ and $u = \nu$ for $\gamma = 0$, we obtain

$$(5.37) \qquad R_{121} = O\left(\sum_{j=L}^{j_0} n^{-(u-1-\alpha_1)} 2^{2j[m_1(u-1-\alpha_1)-(r+1/2-1/p)]}\right)$$

$$= O(n^{-2r/(2r+1)})$$

provided α_1 satisfies (3.32) if $\gamma > 0$ (we have no restrictions on α if $\gamma > 0$). The last term, R_{122}, we partition into $R_{122} = R_{1221} + R_{1222}$ depending on the values of $\sqrt{n}|\theta_{jk}|$. Here

$$R_{1221} = O\left(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E\left[d_{jk}^2 \min(1, A_{jn}^2(d_{jk})) I(\sqrt{n}|\theta_{jk}| \leq (\ln n)^b)\right]\right)$$

(5.38)
$$= O\left(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} \left[n^{-1} + \theta_{jk}^2 I(\sqrt{n}|\theta_{jk}| \leq (\ln n)^b)\right]\right)$$

$$= O\left([n(\ln n)^{-2b}]^{-2r/(2r+1)}\right)$$

by Lemma 7. For R_{1222}, repeating (5.24) and (5.25), we derive

$$(5.39) \quad R_{1222} = O\left(\sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} n^{-1}[\eta_j^{-1}([\nu_j/\sqrt{n}]^{1+\alpha_1}) I(\sqrt{n}|\theta_{jk}| > (\ln n)^b)]^2\right).$$

By assumption on the decay of η_j we derive that

$$(5.40) \quad \eta_j^{-1}([\nu_j/\sqrt{n}]^{1+\alpha_1}) \leq \begin{cases} C[\ln(n/\nu_j^2)]^{1/\gamma}, & \text{if } \gamma > 0, \\ C(n/\nu_j^2)^{(1+\alpha_1)/(2\nu)}, & \text{if } \gamma = 0. \end{cases}$$

Consider the two cases separately. If $\gamma > 0$, then by (5.40) and Lemma 7,

$$(5.41) \quad R_{1222} = O\left(\frac{(\ln n)^{2/\gamma}}{n} \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} I(\sqrt{n}|\theta_{jk}| > (\ln n)^b)\right)$$

$$= O\left(n^{-2r/(2r+1)}[\ln n]^{2/\gamma - pb}\right).$$

Now, to obtain (3.34), choose b minimizing $\max(4br/(2r+1), 2/\gamma - pb)$. If $\gamma = 0$, then using (3.32) and (5.40), we derive by direct calculation that

$$(5.42) \quad R_{1222} = O\left(\sum_{j=L}^{j_0} n^{-1} 2^j (n/\nu_j^2)^{(1+\alpha_1)/\nu}\right)$$

$$= O\left(n^{-2r/(2r+1)+(1+\alpha_1)/(\nu(2r+1))(1/p-1/2)}\right),$$

which agrees with (3.34).

High resolution levels: $j_1 + 1 \leq j \leq J - 1$. Repeat (5.27) with j_0 replaced by j_1. Then $R_2 = O(n^{-2r/(2r+1)})$ follows from (3.29) and (5.27) and from the fact that $\sum_{j=j_1+1}^{J-1} \sum_{k=0}^{2^j-1} n/\nu_j^4 = O(n^{-2r/(2r+1)})$.

Medium resolution levels: $j_0 < j \le j_1$.   Partition $R_5 = R_6 + R_7$ where

$$(5.43) \qquad R_6 = \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2 I(\sqrt{n}|\theta_{jk}| > (\ln n)^b),$$

and R_7 has $\le$ instead of $>$ inside the indicator function; then further partition R_6 into R_{61} and R_{62} in a manner similar to (5.17). Repeat (5.18) and note that m_2 is chosen so that

$$(5.44) \qquad \sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} [v_j^2 n^{-2} + \theta_{jk}^2 v_j^4 n^{-2}] = O(n^{-2r/(2r+1)}).$$

Then, by Lemma 7,

$$
\begin{aligned}
(5.45) \quad R_{61} &= O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} \left[\frac{v_j^2}{n^2} + \frac{\theta_{jk}^2 v_j^4}{n^2} + \frac{\sigma^2 I(\sqrt{n}|\theta_{jk}| > (\ln n)^b)}{n} \right] \right) \\
&= O(n^{-2r/(2r+1)}).
\end{aligned}
$$

Partitioning R_{62} further into R_{621} and R_{622} in a manner similar to R_{121} and R_{122} in (5.20), we derive

$$
\begin{aligned}
R_{621} &= O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} E\left[\frac{I_1(d_{jk})}{I_0(d_{jk})} \frac{A_{jn}(d_{jk})I(v_j|d_{jk}| \le C_\delta)}{1 + A_{jn}(d_{jk})} \right]^2 \right. \\
(5.46) & \hspace{5cm} \left. \times\, I(\sqrt{n}|\theta_{jk}| > (\ln n)^b) \right) \\
&= O(n^{-2r/(2r+1)}) + O\left(\sum_{j=j_0+1}^{j_1} \beta_{jn}^2 v_j^{2+2\delta} n^{-(1+\delta)} 2^{-2j(r+1/2-1/p)} \right) \\
&= O(n^{-2r/(2r+1)})
\end{aligned}
$$

by Lemmas 1 and 7, (5.21), condition (3.32) on α_2 and the choice of m_2. For R_{622} we obtain an upper bound similar to (5.39):

$$(5.47) \quad R_{622} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} n^{-1} [\eta_j^{-1}([v_j/\sqrt{n}]^{1+\alpha})]^2 I(\sqrt{n}|\theta_{jk}| > (\ln n)^b) \right).$$

In the case where $\eta_j(x)$ has exponential decay ($\gamma > 0$) we just repeat (5.41) to obtain (3.34). In the case $\gamma = 0$ denote $(1 + \alpha_2)/v = h$ and observe that

$$R_{622} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} n^{-1}(n/v_j^2)^h I(\sqrt{n}|\theta_{jk}| > (\ln n)^b) \right)$$

(5.48)
$$= O\left(\sum_{j=j_0+1}^{j_1} \frac{n^{p/2-1+h}}{(\ln n)^{pb}} \frac{v_j^{-2h} \|\theta_{j\cdot}\|_{rp\infty}^p}{2^{2j(r+1/2-1/p)}} \right)$$

$$= O\left(n^{-2r/(2r+1)+h(1/p-1/2)(r+1)/(r(r+1/2))} \right),$$

which agrees with (3.34). Now, to complete the proof we need to consider the term R_7 given by (5.43). Since

$$|\hat{\theta}_{jk} - \theta_{jk}| \leq \frac{|I_1(d_{jk})/I_0(d_{jk}) - d_{jk}|}{1 + A_{jn}(d_{jk})} + \frac{|d_{jk} - \theta_{jk}|}{1 + A_{jn}(d_{jk})}$$

$$+ \frac{|\theta_{jk}|}{1 + A_{jn}(d_{jk})} + |\theta_{jk}|$$

(5.49)
$$= O\left(\frac{v_j}{n} + \frac{v_j^2 |d_{jk}| n^{-1}}{1 + A_{jn}(d_{jk})} + \frac{|d_{jk} - \theta_{jk}|}{1 + A_{jn}(d_{jk})} + |\theta_{jk}| \right)$$

$$= O\left(|\theta_{jk}| + \frac{v_j}{n} + \frac{|d_{jk} - \theta_{jk}|}{1 + A_{jn}(d_{jk})} \right),$$

we partition R_7 into R_{71}, R_{72} and R_{73}. Here, by Lemma 7 and (5.44)

$$R_{71} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} \theta_{jk}^2 I(\sqrt{n}|\theta_{jk}| \leq (\ln n)^b) \right)$$

(5.50)
$$= O\left([n(\ln n)^{-2b}]^{-2r/(2r+1)} \right),$$

(5.51)
$$R_{72} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} v_j^2/n^2 \right) = (n^{-2r/(2r+1)}).$$

The third term, R_{73}, we partition into R_{731} and R_{732} where

$$R_{731} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} E\left[\frac{(d_{jk} - \theta_{jk})^2}{(1 + A_{jn}(d_{jk}))^2} \right] I(1 \leq \sqrt{n}|\theta_{jk}| \leq (\ln n)^b) \right),$$

(5.52)
$$R_{732} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} E\left[\frac{(d_{jk} - \theta_{jk})^2}{(1 + A_{jn}(d_{jk}))^2} \right] I(\sqrt{n}|\theta_{jk}| \leq 1) \right).$$

By Lemma 7,

$$(5.53) \quad R_{731} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} n^{-1} I(n^{-1} < \theta_{jk}^2) I(\sqrt{n}|\theta_{jk}| \le (\ln n)^b)\right)$$

$$= O(n^{-2r/(2r+1)}[\ln n]^{4br/(2r+1)}).$$

For an upper bound for R_{732} note that since $\xi(x)$ is bounded

$$E[(d_{jk} - \theta_{jk})^2 (1 + A_{jn}(d_{jk}))^{-2}] I(\sqrt{n}|\theta_{jk}| \le 1)$$

$$= O\left(\int_{-\infty}^{\infty} \frac{(x - \theta_{jk})^2}{(1 + \beta_{jn}\sqrt{n}v_j^{-1}\eta_j(\sqrt{n}x))^2}\right.$$

$$\left. \times \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-n(x-\theta_{jk})^2/2\sigma^2} dx \, I(\sqrt{n}|\theta_{jk}| \le 1)\right)$$

$$= O\left(v_j^2 \beta_{jn}^{-2} n^{-2} \int_{-\infty}^{\infty} z^2 \exp(-z^2/2\sigma^2)[\eta_j(1 + |z|)]^{-2} dz\right)$$

$$= O(v_j^2 \beta_{jn}^{-2} n^{-2})$$

by condition (A5). Hence, whenever $\alpha_2 \ge 0$, by (5.44) we have

$$(5.54) \quad R_{732} = O\left(\sum_{j=j_0+1}^{j_1} \sum_{k=0}^{2^j-1} v_j^2 n^{-2}\right) = O(n^{-2r/(2r+1)}).$$

Combination of (5.37), (5.38), (5.41), (5.42) and (5.45)–(5.54) completes the proof. □

PROOF OF THEOREM 6.   Note that for $j \le j_1$, the derivation of the error is identical to that for Theorem 5. When $j > j_1$, the proof is very similar to the proof of Theorem 3 with j_0 replaced by j_1 and $\lambda_\eta = 1$. Again, since $|\hat{\theta}_{jk} - \theta_{jk}| \le |\theta_{jk}| + |\hat{\theta}_{jk}|$ and due to (3.29), we shall be concerned with the $|\hat{\theta}_{jk}|$ term only. Noting that $|\hat{\theta}_{jk}| \le |I_1(d_{jk})/I_0(d_{jk})|$ and using the inequality (5.28), we partition the error and derive upper bounds identical to those in (5.29), (5.30) and (5.32), which leads to $\sum_{j_1+1}^{J_0} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2 = O(n^{-2r/(2r+1)}(\ln n)^{1/(2r+1)})$. Derivation of the upper bound for $\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2$ repeats the derivation of R_4 in Theorem 3, and similarly, we obtain $\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2 = O(n^{-2r/(2r+1)})$. □

PROOF OF THEOREM 7.   The validity of Theorem 6 follows from Theorems 1, 3, 5 and 6. The only portion of the error which needs additional consideration is

the one corresponding to $j > J_0$. Recall that

$$(5.55) \qquad \hat{\theta}_{jk} = \frac{\lambda_j^* I_1(d_{jk}) + (1 - \lambda_j^*) I_1^*(d_{jk})}{\lambda_j^* I_0(d_{jk}) + (1 - \lambda_j^*) I_0^*(d_{jk}) + \beta_{jn} \sqrt{n} \eta_j(\sqrt{n} d_{jk})},$$

where $\eta_j(x)$ and λ_j^* are given by (3.36),

$$I_i(d) = \int_{-\infty}^{\infty} x^i \sqrt{n} (\sqrt{2\pi} \sigma_0)^{-1} \exp(-n(d - x)^2 / (2\sigma_0^2)) v_j \xi(v_j x) \, dx,$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 0, 1,$$

$$I_i^*(d) = \int_{-\infty}^{\infty} x^i \sqrt{n} \zeta(\sqrt{n}(d - x)) v_j \xi(v_j x) \, dx, \qquad\qquad i = 0, 1.$$

Consider the cases when ξ is a normal and a heavy-tailed prior separately. If ξ is a normal prior, then by Lemma 2

$$R_8 = \sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2$$

$$(5.56) \qquad = O\left(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E\left[\left(\frac{I_1(d_{jk})}{I_0(d_{jk})}\right)^2 + \left(\frac{I_1^*(d_{jk})}{I_0^*(d_{jk})}\right)^2 + \theta_{jk}^2 \right] \right)$$

$$= O\left(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} (n^2 v_j^{-4} E d_{jk}^2 + n v_j^{-4} + \theta_{jk}^2) \right) = O(n^{-2r/(2r+1)}).$$

If $\lambda_\xi = 0$, then partition R_8 as $R_8 = R_{81} + R_{82}$ where

$$R_{81} = O\left(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} \left[E\left\{ \left(\frac{I_1(d_{jk})}{I_0(d_{jk})}\right)^2 I\left(\frac{n|d_{jk}|}{v_j} \to 0\right) \right\} + E\left(\frac{I_1^*(d_{jk})}{I_0^*(d_{jk})}\right)^2 + \theta_{jk}^2 \right] \right),$$

so that $R_{81} = O(n^{-2r/(2r+1)})$ by Lemma 2 and calculation similar to (5.56), and

$$R_{82} = O\left(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E\left[\left(\frac{I_1(d_{jk})}{I_0(d_{jk}) + (\lambda_j^*)^{-1} \beta_{jn} \sqrt{n} \zeta(\sqrt{n} d_{jk})}\right)^2 \right. \right.$$

$$\left. \left. \times I\left(\frac{n|d_{jk}|}{v_j} \geq M\right) \right] \right).$$

Let $\zeta_1(x)$ be a heavy-tailed p.d.f. such that

$$(5.57) \qquad C_{\zeta,1} (\sigma_0 \sqrt{2\pi})^{-1} \exp\{-x^2/2\sigma_0^2\} \leq \zeta_1(x) \leq C_{\zeta,2} (x^2 + 1)^{-1} \zeta(x)$$

for some positive $C_{\zeta,1}$ and $C_{\zeta,2}$. Then, it is easy to see that $I_0(d) \leq C_{\zeta,1}^{-1} \int \sqrt{n} \zeta_1 \times (\sqrt{n}(d - x)) v_j \xi(v_j x) \, dx \sim \sqrt{n} \zeta_1(\sqrt{n} d)$. Thus, by Lemma 2 $\sqrt{n} \zeta(\sqrt{n} d_{jk})/$

$I_0(d_{jk}) \geq C_{\zeta,1}^{-1} \zeta(\sqrt{n} d_{jk})/\zeta_1(\sqrt{n} d_{jk})$ and

$$R_{82} = O\left(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E \frac{[I_1(d_{jk})/I_0(d_{jk})]^2 I(v_j^{-1} n |d_{jk}| \geq M)}{[\beta_{jn}(\lambda_j^*)^{-1} \zeta(\sqrt{n} d_{jk})/\zeta_1(\sqrt{n} d_{jk})]^2} \right)$$

(5.58)
$$= O\left(\sum_{j=J_0+1}^{J-1} \sum_{k=0}^{2^j-1} E[(n d_{jk}^2)^{-1} d_{jk}^2 I(v_j^{-1} n |d_{jk}| \geq M)] \right)$$

$$= O(n^{-2r/(2r+1)}),$$

which completes the proof. $\square$

PROOF OF THEOREM 8. The proof is based on the following obvious corollary of Kolmogorov's three series theorem: if $Z_1, Z_2, \ldots, Z_n$ are independent variables such that $\sum_n E|Z_n| < \infty$, then the series $\sum_n Z_n$ converges with probability 1 (see, e.g., [22], Section 4.2). Since we impose priors on each coefficient independently, it follows from the above that $f \in B_{pq}^r$ if

$$(5.59) \qquad S_1 = \sum_{j=L}^{\infty} 2^{j(r+1/2-1/p)q} E\left[\sum_{k=0}^{2^j-1} |\theta_{jk}|^p \right]^{q/p} < \infty$$

where the expectation is taken over the prior p.d.f. $\pi_{jn} v_j \xi(v_j x) + (1 - \pi_{jn}) \delta(0)$. Note that for $a > 0$ we have $E|\theta_{jk}|^a = C_a \pi_{jn} v_j^{-a}$ where $C_a = \int |x|^a \xi(x) \, dx$ provided the last integral is convergent. If $q/p \leq 1$, then using Jensen's inequality $E(|Y|^{q/p}) \leq (E|Y|)^{q/p}$ and the expression for $E|\theta_{jk}|^q$ we obtain

$$(5.60) \quad S_1 \leq \sum_{j=L}^{\infty} 2^{j(r+1/2-1/p)q} \left[\sum_{k=0}^{2^j-1} E|\theta_{jk}|^p \right]^{q/p} \leq \sum_{j=L}^{\infty} 2^{j(r+1/2)q} v_j^{-q} \pi_{jn}^{q/p}.$$

Taking into account that $\beta_{jn} = (1 + \pi_{jn})^{-1}$, we derive (3.41).

If $q > p$, note that if $f \in B_{p,p}^r$, then $f \in B_{p,q}^r$, so it is enough to consider $q = p$. Repeating the previous calculation, we arrive at (3.41) again. $\square$

PROOF OF COROLLARY 4. Plug the values of v_j and β_{jn} into (3.41) and check that the sum (3.41) is uniformly bounded by a constant independent of n.
$\square$

PROOF OF COROLLARY 5. Plugging the values of v_j given by (3.28) into (3.41), we derive that (3.41) holds whenever the sum $S^* = \sum_{j=L}^{\infty} 2^{jv} \times \beta_{jn}^{-\min(q/p,1)}$ is uniformly bounded, where

$$(5.61) \qquad v = \begin{cases} (1/p - 1/2)/2, & L \leq j \leq j_0, \\ (1/p - 1/2)(1 + 1/r), & j_0 < j \leq j_1, \\ 0, & j > j_1. \end{cases}$$

Now to complete the proof, check that S^* is uniformly bounded whenever (3.43) is valid. $\square$

The proof of Theorem 9 is based on the following lemma.

LEMMA 8. *Under the conditions of Theorem 9, for some positive absolute constant C*

$$
(5.62) \qquad R_n\big(B^r_{p,q}(A), \hat{f}\big) \geq C \sum_{j=L}^{j_0} 2^{-2rj} I\bigg(\beta_{jn} > \frac{v_j}{\sqrt{n}} \eta_j\bigg(\frac{C_\delta \sqrt{n}}{v_j}\bigg)\bigg)
$$
$$
- O\big(n^{-2r/(2r+1)}\big).
$$

PROOF. Repeating the beginning of the proof of Theorem 1, we derive that $R_n(B^r_{p,q}(A), \hat{f}) \geq R \geq R_1$ where R_1 is defined in (5.15). Since for any random variables X and Y $E(X + Y)^2 \geq 0.5EX^2 - EY^2$, we obtain that $R_1 \geq 0.5R_{12} - R_{11}$ where, just as in (5.19), $R_{11} = O(n^{-2r/(2r+1)})$. Now, by (5.20) $R_{12} \geq R_{121}$. Consequently,

$$
(5.63) \qquad R_n\big(B^r_{p,q}(A), \hat{f}\big) \geq 0.5R_{121} - O\big(n^{-2r/(2r+1)}\big)
$$

where R_{121} is defined in (5.20). Recall (5.23) and by direct calculation verify that $\alpha_{jn} \geq C_\delta \sqrt{n} v_j^{-1}$ if and only if

$$
(5.64) \qquad \beta_{jn} \geq v_j\big[\sqrt{n}\eta(C_\delta \sqrt{n} v_j^{-1})\big]^{-1}.
$$

For the values of j for which (5.64) holds, $A_{jn}^2(d_{jk}) \geq 1$, so that $[A_{jn}(d_{jk})/(1 + A_{jn}(d_{jk}))]^2 \geq 0.25$. Hence

$$
(5.65) \qquad R_{121} \geq 0.25 \sum_{j=L}^{j_0} \sum_{k=0}^{2^j-1} E\big[d_{jk}^2 I(C_\delta v_j^{-1} < |d_{jk}| \leq \alpha_{jn}/\sqrt{n})\big]
$$
$$
\geq 0.25 C_\delta^2 \sum_{j=L}^{j_0} 2^j v_j^{-2} I\big(\beta_{jn} \geq v_j[\sqrt{n}\eta(C_\delta \sqrt{n} v_j^{-1})]^{-1}\big).
$$

Plug the value of v_j into (5.65). The combination of (5.63)–(5.65) completes the proof. Note that in Theorem 1 the choice of β_{jn} makes the inequality $\alpha_{jn} \geq C_\delta \sqrt{n} v_j^{-1}$ impossible. $\square$

PROOF OF THEOREM 9. Plugging $\beta_{jn} = 2^{bj}$ into (5.62) we obtain that

$$
\beta_{jn} > v_j\big[\sqrt{n}\eta(C_\delta \sqrt{n}/v_j)\big]^{-1} \iff v_j \geq C_\delta \sqrt{n}\big[\eta^{-1}\big(2^{j(r+1/2-b)}n^{-1/2}\big)\big]^{-1}.
$$

Recalling that $j \leq j_0$, we note that $2^{j(r+1/2-b)}n^{-1/2} \leq n^{-b/(2r+1)}$. Hence, by Lemma 5.62,

$$R_n\big(B_{p,q}^r(A), \hat{f}\big)$$

$$\geq C \sum_{j=L}^{j_0} 2^{-2rj} I\big(\beta_{jn} > \nu_j \big[\sqrt{n}\eta\big(C_\delta\sqrt{n}/\nu_j\big)\big]^{-1}\big)$$

$$\geq C \sum_{j=L}^{j_0} 2^{-2rj} I\big(2^j \geq \big[C_\delta\sqrt{n}\big]^{1/(r+1/2)}\big[\eta^{-1}\big(n^{-b/(2r+1)}\big)\big]^{-1/(r+1/2)}\big)$$

$$= C_1 n^{-2r/(2r+1)} \big[\eta^{-1}\big(n^{-b/(2r+1)}\big)\big]^{2r/(r+1/2)}$$

for some positive absolute constants C and C_1, which proves (3.45). To finish the proof, consider various cases for $\eta(x)$. $\quad\square$

REFERENCES

[1] ABRAMOVICH, F., AMATO, U. and ANGELINI, C. (2004). On optimality of Bayesian wavelet estimators. *Scand. J. Statist.* **31** 217–234. MR2066250

[2] ABRAMOVICH, F., BESBEAS, P. and SAPATINAS, T. (2002). Empirical Bayes approach to block wavelet function estimation. *Comput. Statist. Data Anal.* **39** 435–451. MR1910021

[3] ABRAMOVICH, F., SAPATINAS, T. and SILVERMAN, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 725–749. MR1649547

[4] ANGELINI, C. and SAPATINAS, T. (2004). Empirical Bayes approach to wavelet regression using ε-contaminated priors. *J. Stat. Comput. Simul.* **74** 741–764. MR2075964

[5] AUTIN, F., PICARD, D. and RIVOIRARD, V. (2004). Maxiset comparisons of procedures, application to choosing priors in a Bayesian nonparametric setting. Unpublished manuscript. Available at www.proba.jussieu.fr/mathdoc/textes/PMA-931.pdf.

[6] BARBER, S., NASON, G. P. and SILVERMAN, B. W. (2002). Posterior probability intervals for wavelet thresholding. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 189–205. MR1904700

[7] CHIPMAN, H. A., KOLACZYK, E. D. and MCCULLOCH, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92** 1413–1421.

[8] CLYDE, M. and GEORGE, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 681–698. MR1796285

[9] CLYDE, M., PARMIGIANI, G. and VIDAKOVIC, B. (1996). Bayesian strategies for wavelet analysis. *Statist. Comput. Graph. Newsletter* **7** 2.

[10] CLYDE, M., PARMIGIANI, G. and VIDAKOVIC, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85** 391–401. MR1649120

[11] DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia. MR1162107

[12] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

[13] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974

[14] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414

[15] GRADSHTEYN, I. S. and RYZHIK, I. M. (1980). *Tables of Integrals, Series, and Products.* Academic Press, New York. MR0582453

[16] JOHNSTONE, I. M. (2004). *Function Estimation and Gaussian Sequence Models.* Unpublished monograph. Available at www-stat.stanford.edu/~imj/.

[17] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Finding needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

[18] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Boundary coiflets for wavelet shrinkage in function estimation. *J. Appl. Probab.* **41A** 81–98. MR2057567

[19] JOHNSTONE, I. M. and SILVERMAN, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. MR2166560

[20] MALLAT, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Analysis Machine Intelligence* **11** 674–693.

[21] NEUMANN, M. H. and VON SACHS, R. (1995). Wavelet thresholding: Beyond the Gaussian i.i.d. situation. *Wavelets and Statistics. Lecture Notes in Statist.* **103** 301–329. Springer, New York. MR1364677

[22] SHIRYAEV, A. N. (1996). *Probability*, 2nd ed. Springer, New York. MR1368405

[23] SIMONCELLI, E. P. (1997). Statistical models for images: Compression, restoration and synthesis. In *Proc. 31st Asilomar Conference on Signals, Systems and Computers* 673–678. IEEE Comput. Soc., Pacific Grove, CA.

[24] VIDAKOVIC, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* **93** 173–179. MR1614620

[25] VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets.* Wiley, New York. MR1681904

[26] VIDAKOVIC, B. and RUGGERI, F. (2001). BAMS method: Theory and simulations. *Sankhyā Ser. B* **63** 234–249. MR1895791

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CENTRAL FLORIDA
ORLANDO, FLORIDA 32816-1364
USA
E-MAIL: mpensky@pegasus.cc.ucf.edu