Biometrika (2010), **97**, 1, *pp*. 181–198 doi: 10.1093/biomet/asp080

Printed in Great Britain

On Bayesian testimation and its application to wavelet thresholding

By FELIX ABRAMOVICH

Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel

felix@post.tau.ac.il

VADIM GRINSHTEIN

Department of Mathematics, The Open University of Israel, Raanana 43107, Israel

vadimg@openu.ac.il

ATHANASIA PETSA and THEOFANIS SAPATINAS

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

map6pa2@ucy.ac.cy fanis@ucy.ac.cy

Summary

We consider the problem of estimating the unknown response function in the Gaussian white noise model. We first utilize the recently developed Bayesian maximum a posteriori testimation procedure of Abramovich et al. (2007) for recovering an unknown high-dimensional Gaussian mean vector. The existing results for its upper error bounds over various sparse l_p-balls are extended to more general cases. We show that, for a properly chosen prior on the number of nonzero entries of the mean vector, the corresponding adaptive estimator is asymptotically minimax in a wide range of sparse and dense l_p-balls. The proposed procedure is then applied in a wavelet context to derive adaptive global and level-wise wavelet estimators of the unknown response function in the Gaussian white noise model. These estimators are then proven to be, respectively, asymptotically near-minimax and minimax in a wide range of Besov balls. These results are also extended to the estimation of derivatives of the response function. Simulated examples are conducted to illustrate the performance of the proposed level-wise wavelet estimator in finite sample situations, and to compare it with several existing counterparts.

Some key words: Adaptive estimation; Besov space; Gaussian sequence model; Gaussian white noise model; l_p-ball; Multiple testing; Thresholding; Wavelet estimation.

1. Introduction

We consider the problem of estimating the unknown response function in the Gaussian white noise model, where one observes Gaussian processes $Y_n(t)$ governed by

$$dY_n(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}\, dW(t), \quad t \in [0, 1]. \tag{1}$$

The noise parameter $\sigma > 0$ is assumed to be known, W is a standard Wiener process and $f \in L^2[0, 1]$ is the unknown response function. Under some smoothness constraints on f, such a model is asymptotically equivalent in Le Cam sense to the standard nonparametric regression setting (Brown & Low, 1996).

In a consistent estimation theory, it is well known that f should possess some smoothness properties. We assume that f belongs to a Besov ball $B^s_{p,q}(M)$ of a radius $M > 0$, where $0 < p, q \leqslant \infty$ and $s > \max(0, 1/p - 1/2)$. The latter restriction ensures that the corresponding Besov spaces are embedded in $L^2[0, 1]$. The parameter s measures the degree of smoothness while p and q specify the type of norm used to measure the smoothness. Besov classes contain various traditional smoothness spaces such as Hölder and Sobolev spaces as special cases. However, they also include different types of spatially inhomogeneous functions (Meyer, 1992).

The fact that wavelet series constitute unconditional bases for Besov spaces has caused various wavelet-based estimation procedures to be widely used for estimating the unknown response $f \in B^s_{p,q}(M)$ in the Gaussian white noise model (1). The standard wavelet approach for the estimation of f is based on finding the empirical wavelet coefficients of the data and denoising them, usually by some type of thresholding rule. Transforming them back to the function space then yields the resulting estimate. The main statistical challenge in such an approach is a proper choice of a thresholding rule. A series of various wavelet thresholds originating from different ideas has been proposed in the literature during the last decade, e.g. the universal threshold (Donoho & Johnstone, 1994a), Stein's unbiased risk estimation threshold (Donoho & Johnstone, 1995), the false discovery rate threshold (Abramovich & Benjamini, 1996), the crossvalidation threshold (Nason, 1996), the Bayes threshold (Abramovich et al., 1998) and the empirical Bayes threshold (Johnstone & Silverman, 2005).

Abramovich & Benjamini (1996) demonstrated that thresholding can be viewed as a multiple hypothesis testing procedure, where one first simultaneously tests the wavelet coefficients of the unknown response function for significance. The coefficients concluded to be significant are then estimated by the corresponding empirical wavelet coefficients of the data, while the nonsignificant ones are discarded. Such a testimation procedure evidently mimics a hard thresholding rule. Various choices for adjustment to multiplicity on the testing step lead to different thresholds. In particular, the universal threshold of Donoho & Johnstone (1994a) and the false discovery rate threshold of Abramovich & Benjamini (1996) fall within such a framework corresponding to Bonferroni and false discovery rate multiplicity corrections, respectively.

In this paper, we proceed along the lines of this testimation approach, where we utilize the recently developed maximum a posteriori Bayesian multiple testing procedure of Abramovich & Angelini (2006). Their hierarchical prior model is based on imposing a prior distribution on the number of false null hypotheses. Abramovich et al. (2007) applied this approach to estimating a high-dimensional Gaussian mean vector and showed its minimax optimality where the unknown mean vector was assumed to be sparse.

We first extend the results of Abramovich et al. (2007) to more general settings. Consider the problem of estimating an unknown high-dimensional Gaussian mean vector, where one observes y_i governed by

$$y_i = \mu_i + \sigma_n z_i \quad (i = 1, 2, \ldots, n). \tag{2}$$

The variance $\sigma_n^2 > 0$, which may depend on n, is assumed to be known, z_i are independent $N(0, 1)$ random variables and the unknown mean vector $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is assumed to lie in a strong l_p-ball $l_p[\eta_n]$, $0 < p \leqslant \infty$, of a normalized radius η_n, that is, $\|\mu\|_p \leqslant C_n$, where $C_n = n^{1/p}\sigma_n\eta_n$. Abramovich et al. (2007) considered the Gaussian sequence model (2) with $\sigma_n^2 = \sigma^2$ and derived upper error bounds for the quadratic risk of an adaptive Bayesian maximum a posteriori estimator of μ in the sparse case, where $0 < p < 2$ and $\eta_n \to 0$ as $n \to \infty$. We extend their results for all combinations of p and η_n and for the variance in (2) that may depend on n. We show, in particular, that for a properly chosen prior distribution on the number of nonzero entries of μ, the corresponding estimator, up to a constant factor, is asymptotically minimax for almost all l_p-balls including both sparse and dense cases.

We then apply the proposed approach to the wavelet thresholding estimation in the Gaussian white noise model (1). We show that, under mild conditions on the prior distribution on the number of nonzero wavelet coefficients, the resulting global wavelet estimator of f, up to a logarithmic factor, attains the minimax convergence rates simultaneously over the entire range of Besov balls. Furthermore, we demonstrate that estimating wavelet coefficients at each resolution level separately, allows one to remove the extra logarithmic factor. Moreover, the procedure can also be extended to the estimation of derivatives of f. These results, in some sense, complement the adaptively minimax empirical Bayes estimators of Johnstone & Silverman (2005).

2. Estimation in the Gaussian sequence model

2·1. *The Bayesian maximum a posteriori estimation procedure*

We start by reviewing the Bayesian maximum a posteriori estimation procedure for the Gaussian sequence model (2) developed by Abramovich et al. (2007).

For this model, consider the multiple hypothesis testing problem, where we wish to simultaneously test

$$H_{0i} : \mu_i = 0 \quad \text{versus} \quad H_{1i} : \mu_i \neq 0 \quad (i = 1, 2, \ldots, n).$$

A configuration of true and false null hypotheses is uniquely defined by the indicator vector $x = (x_1, \ldots, x_n)^{\mathrm{T}}$, where $x_i = \mathbb{I}(\mu_i \neq 0)$ and $\mathbb{I}(A)$ denotes the indicator function of the set A. Let $\kappa = x_1 + \cdots + x_n = \|\mu\|_0$ be the number of nonzero μ_i, i.e. $\|\mu\|_0 = \#\{i : \mu_i \neq 0\}$. Assume some prior distribution π_n on κ with $\pi_n(\kappa) > 0$, $\kappa = 0, \ldots, n$. For a given κ, all the corresponding different vectors x are assumed to be equally likely a priori, that is, conditionally on κ,

$$\mathrm{pr}\left(x \,\middle|\, \sum_{i=1}^n x_i = \kappa\right) = \binom{n}{\kappa}^{-1}.$$

Naturally, $\mu_i \mid x_i = 0 \sim \delta_0$, where δ_0 is a probability atom at zero. To complete the prior specification, we assume that $\mu_i \mid x_i = 1 \sim N(0, \tau_n^2)$.

For the proposed hierarchical prior, the posterior probability of a given vector x with κ nonzero entries is

$$\pi_n(x, \kappa \mid y) \propto \binom{n}{\kappa}^{-1} \pi_n(\kappa)\, \mathbb{I}\left(\sum_{i=1}^n x_i = \kappa\right) \prod_{i=1}^n \left(B_i^{-1}\right)^{x_i}, \tag{3}$$

where the Bayes factor B_i of H_{0i} is

$$B_i = \surd(1 + \gamma_n)\ \exp\left\{-\frac{y_i^2}{2\sigma_n^2(1 + 1/\gamma_n)}\right\} \tag{4}$$

and $\gamma_n = \tau_n^2/\sigma_n^2$ is the variance ratio (Abramovich & Angelini, 2006).

Given the posterior distribution $\pi_n(x, \kappa \mid y)$, we apply the maximum a posteriori rule to choose the most likely indicator vector. Generally, to find the posterior mode of $\pi_n(x, \kappa \mid y)$, one should look through all 2^n possible sequences of zeroes and ones. However, for the proposed model, the number of candidates for a mode is, in fact, reduced to $n+1$ only. Indeed, let $\hat{x}(\kappa)$ be a maximizer of (3) for a fixed κ that indicates the most plausible vector x with κ nonzero entries. From (3), it follows immediately that $\hat{x}_i(\kappa)=1$ at the κ entries corresponding to the smallest Bayes factors B_i and zeroes otherwise. Due to the monotonicity of B_i in $|y|_i$ in (4), it is equivalent to setting $\hat{x}_i(\kappa)=1$ for the κ largest $|y|_i$ and zeroes for others. The proposed Bayesian multiple testing procedure then leads to finding $\hat{\kappa}$ that maximizes

$$\log \pi_n\{\hat{x}(\kappa), \kappa \mid y\} = c + \sum_{i=1}^{\kappa} y_{(i)}^2 + 2\sigma_n^2(1+1/\gamma_n)\log\left\{\binom{n}{\kappa}^{-1}\pi_n(\kappa)(1+\gamma_n)^{-\kappa/2}\right\}$$

for some constant c or, equivalently, minimizes

$$\sum_{i=\kappa+1}^{n} y_{(i)}^2 + 2\sigma_n^2(1+1/\gamma_n)\log\left\{\binom{n}{\kappa}\pi_n^{-1}(\kappa)(1+\gamma_n)^{\kappa/2}\right\},$$

where $|y|_{(1)} \geqslant \cdots \geqslant |y|_{(n)}$. The $\hat{\kappa}$ null hypotheses corresponding to $|y|_{(1)}, \ldots, |y|_{(\hat{\kappa})}$ are rejected. The resulting Bayesian estimation yields a hard thresholding with a data-driven threshold $\hat{\lambda}_{\text{MAP}} = |y|_{(\hat{\kappa})}$, i.e.

$$\hat{\mu}_i = \begin{cases} y_i, & |y_i| \geqslant \hat{\lambda}_{\text{MAP}}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

If $\hat{\kappa}=0$, then all y_i $(i=1,2,\ldots,n)$, are thresholded and $\hat{\mu}\equiv 0$.

From a frequentist view, the above estimator $\hat{\mu}=(\hat{\mu}_1,\ldots,\hat{\mu}_n)^{\mathrm{T}}$ in (5) is evidently a penalized likelihood estimator with the complexity penalty

$$P_n(\kappa) = 2\sigma_n^2(1+1/\gamma_n)\log\left\{\binom{n}{\kappa}\pi_n^{-1}(\kappa)(1+\gamma_n)^{\kappa/2}\right\}. \tag{6}$$

In this sense, it can also be considered within the framework of Birgé & Massart (2001). In the following section, we will discuss these relations in more detail.

2·2. *Upper error bounds*

Abramovich et al. (2007, Theorem 6) obtained upper error bounds for the l^2-risk of (5) in the Gaussian sequence model (2) for sparse $l_p[\eta_n]$-balls, where $0<p<2$ and $\eta_n \to 0$ as $n\to\infty$. We now extend these results to more general settings.

Fix a prior distribution $\pi_n(\kappa)>0$ $(\kappa=0,\ldots,n)$ on the number of nonzero entries of μ, and let $\gamma_n=\tau_n^2/\sigma_n^2$ be the variance ratio.

PROPOSITION 1. *Let $\hat{\mu}$ be the estimator (5) of μ in the Gaussian sequence model (2), where $\mu\in l_p[\eta_n]$, $0<p\leqslant\infty$. Assume that there exist positive constants γ_- and γ_+ such that $\gamma_-\leqslant\gamma_n\leqslant\gamma_+$.*

(1) *Let $0<p\leqslant\infty$. Assume that $\pi_n(n)\geqslant e^{-c_0 n}$ for some $c_0>0$. Then, as $n\to\infty$,*

$$\sup_{\mu\in l_p[\eta_n]} E\left(\|\hat{\mu}-\mu\|_2^2\right) = O\left(n\sigma_n^2\right).$$

(2) *Let $2\leqslant p\leqslant\infty$. Assume that there exists $\beta\geqslant 0$ such that $\pi_n(0)\geqslant n^{-c_1 n^{-\beta}}$ for some $c_1>0$. Then, as $n\to\infty$,*

$$\sup_{\mu\in l_p[\eta_n]} E\left(\|\hat{\mu}-\mu\|_2^2\right) = O\left(\sigma_n^2 n\eta_n^2\right) + O\left(\sigma_n^2 n^{-\beta}\log n\right).$$

(3) *Let* $0 < p < 2$. *Assume* $\pi_n(\kappa) \geqslant (\kappa/n)^{c_2\kappa}$ *for all* $\kappa = 1, 2, \ldots, \alpha_n n$, *where* $n^{-1}(2\log n)^{p/2} \leqslant \alpha_n \leqslant \exp\{-c(\gamma_n)\}$, $c(\gamma_n) = 8(\gamma_n + 3/4)^2 > 9/2$, *and for some* $c_2 > 0$. *Then, as* $n \to \infty$,

$$\sup_{\mu \in l_p[\eta_n]} E\left(\|\hat{\mu} - \mu\|_2^2\right) = O\left\{\sigma_n^2 n \eta_n^p \left(2 \log \eta_n^{-p}\right)^{1-p/2}\right\}$$

for all $n^{-1}(2\log n)^{p/2} \leqslant \eta_n^p \leqslant \alpha_n$.

(4) *Let* $0 < p < 2$. *Assume that there exists* $\beta \geqslant 0$ *such that* $\pi_n(0) \geqslant n^{-c_1 n^{-\beta}}$ *for some* $c_1 > 0$. *Then, as* $n \to \infty$,

$$\sup_{\mu \in l_p[\eta_n]} E\left(\|\hat{\mu} - \mu\|_2^2\right) = O\left(\sigma_n^2 n^{2/p} \eta_n^2\right) + O\left(\sigma_n^2 n^{-\beta} \log n\right)$$

for all $\eta_n^p < n^{-1}(2\log n)^{p/2}$.

The proof of Proposition 1 is given in the Appendix. Similar to Abramovich et al. (2007), analogous results can be obtained for other types of balls, e.g. weak l_p-balls, $0 < p < \infty$, and l_0-balls, with necessary changes in the proofs; see Chapter 3 of a 2009 University of Cyprus PhD thesis by Petsa.

Since the prior assumptions in Proposition 1 do not depend on the parameters p and η_n of the l_p-ball, the estimator (5) is inherently adaptive. The condition on $\pi_n(n)$ guarantees that its risk is always bounded by an order of $n\sigma_n^2$, corresponding to the risk of the maximum likelihood estimator, $\hat{\mu}_i^{\text{MLE}} = y_i$, in the Gaussian sequence model (2).

The following corollary of Proposition 1 essentially defines dense and sparse zones for $2 \leqslant p \leqslant \infty$, and dense, sparse and super-sparse zones for $0 < p < 2$ of different behaviour for the quadratic risk of the proposed estimator (5). To evaluate its accuracy, we also compare the resulting risks with the corresponding minimax risks $R(l_p[\eta_n]) = \inf_{\tilde{\mu}} \sup_{\mu \in l_p[\eta_n]} E(\|\tilde{\mu} - \mu\|_2^2)$ that can be found, e.g. in Donoho & Johnstone (1994b). In what follows, $g_1(n) \asymp g_2(n)$ denotes that $0 < \liminf\{g_1(n)/g_2(n)\} \leqslant \limsup\{g_1(n)/g_2(n)\} < \infty$ as $n \to \infty$.

Corollary 1. *Let* $\hat{\mu}$ *be the estimator* (5) *of* μ *in the Gaussian sequence model* (2), *where* $\mu \in l_p[\eta_n]$, $0 < p \leqslant \infty$. *Assume that there exist positive constants* γ_- *and* γ_+ *such that* $\gamma_- \leqslant \gamma_n \leqslant \gamma_+$. *Define* $c(\gamma_n) = 8(\gamma_n + 3/4)^2 > 9/2$ *and let the prior* π_n *satisfy the following conditions*:

(1) $\pi_n(0) \geqslant n^{-c_1 n^{-\beta}}$ *for some* $\beta \geqslant 0$ *and* $c_1 > 0$;
(2) $\pi_n(\kappa) \geqslant (\kappa/n)^{c_2\kappa}$ *for all* $\kappa = 1, 2, \ldots, \alpha n$, *where* $\alpha = \exp(-9/2)$ *or* $\alpha = \exp\{-c(\gamma_-)\}$ *if* γ_- *is known, and for some* $c_2 > 0$; *and*
(3) $\pi_n(n) \geqslant e^{-c_0 n}$ *for some* $c_0 > 0$.

Then, as $n \to \infty$, *depending on* p *and* η_n, *one has the following.*

Case 1. Let $0 < p \leqslant \infty$, $\eta_n^p > \alpha$. Then,

$$\sup_{\mu \in l_p[\eta_n]} E\left(\|\hat{\mu} - \mu\|_2^2\right) = O\left(n\sigma_n^2\right), \quad R(l_p[\eta_n]) \asymp n\sigma_n^2.$$

Case 2. Let $2 \leqslant p \leqslant \infty$, $\eta_n^p \leqslant \alpha$. Then,

$$\sup_{\mu \in l_p[\eta_n]} E\left(\|\hat{\mu} - \mu\|_2^2\right) = O\left(\sigma_n^2 n \eta_n^2\right) + O\left(\sigma_n^2 n^{-\beta} \log n\right), \quad R(l_p[\eta_n]) \asymp \sigma_n^2 n \eta_n^2.$$

Case 3. Let $0 < p < 2,\ n^{-1}(2\log n)^{p/2} \leqslant \eta_n^p \leqslant \alpha$. Then,

$$\sup_{\mu \in l_p[\eta_n]} E\left(\|\hat{\mu} - \mu\|_2^2\right) = O\{\sigma_n^2 n \eta_n^p (2\log \eta_n^{-p})^{1-p/2}\}, \quad R(l_p[\eta_n]) \asymp \sigma_n^2 n \eta_n^p (2\log \eta_n^{-p})^{1-p/2}.$$

Case 4. Let $0 < p < 2,\ \eta_n^p < n^{-1}(2\log n)^{p/2}$. Then,

$$\sup_{\mu \in l_p[\eta_n]} E\left(\|\hat{\mu} - \mu\|_2^2\right) = O\left(\sigma_n^2 n^{2/p} \eta_n^2\right) + O\left(\sigma_n^2 n^{-\beta} \log n\right), \quad R(l_p[\eta_n]) \asymp \sigma_n^2 n^{2/p} \eta_n^2.$$

For $\beta = 0$ one can easily verify that all three conditions of Corollary 1 are satisfied, for example, for the truncated geometric prior TrGeom$(1-q)$ $(0 < q < 1)$, where $\pi_n(\kappa) = (1-q)q^\kappa/(1-q^{n+1})$ $(\kappa = 0, \ldots, n)$. On the other hand, for any β, no binomial prior Bin(n, p_n) can kill three birds with one stone. The requirement $\pi_n(0) = (1-p_n)^n \geqslant n^{-c_1 n^{-\beta}}$ necessarily implies $p_n \to 0$ as $n \to \infty$. However, to satisfy $\pi_n(n) = p_n^n \geqslant e^{-c_0 n}$, one needs $p_n \geqslant e^{-c_0}$.

The impact of Corollary 1 is that, up to a constant multiplier, the proposed estimator (5) is adaptively minimax for almost all l_p-balls $(0 < p \leqslant \infty)$, except those with very small normalized radiuses, where $\eta_n^2 = o(n^{-\{\beta+2/\min(p,2)\}} \log n)$. Hence, while the optimality of most existing threshold estimators, e.g. universal, Stein's unbiased risk, false discovery rate, has been established only over various sparse settings, the Bayesian estimator (5) is appropriate for both sparse and dense cases. To the best of our knowledge, such a wide adaptivity range can be compared only with the penalized likelihood estimators of Birgé & Massart (2001) and the empirical Bayes threshold estimators of Johnstone & Silverman (2004b, 2005); see the PhD thesis by Petsa for more details.

There are interesting asymptotic relationships between the Bayesian estimator (5) and the penalized likelihood estimator of Birgé & Massart (2001) that may explain their similar behaviour. For estimating the normal mean vector in (2) within l_p-balls, Birgé & Massart (2001) considered a penalized likelihood estimator with a specific complexity penalty

$$\tilde{P}_n(\kappa) = C\sigma_n^2 \kappa\{1 + \surd(2L_\kappa)\}^2, \tag{7}$$

where $L_\kappa = \log(n/\kappa) + (1+\theta)(1+\log(n)/\kappa)$ for fixed $C > 1$ and $\theta > 0$ (Birgé & Massart, 2001, § 6.3). For large n and $\kappa < n/e$, this penalty is approximately of the following form:

$$\tilde{P}_n(\kappa) \sim 2\sigma_n^2 c\kappa L_\kappa \sim 2\sigma_n^2 \tilde{c}_1 \left\{\log \binom{n}{\kappa} + \tilde{c}_2 \kappa\right\} \tag{8}$$

for some positive constants $c,\ \tilde{c}_1,\ \tilde{c}_2 > 1$; see also Lemma A1 in the Appendix. Thus, within this range, $\tilde{P}_n$ in (7)–(8) behaves in a way similar to a particular case of the penalty P_n in (6) corresponding to the geometric type prior $\pi_n(\kappa) \propto (1/\tilde{c}_2)^\kappa$. This prior satisfies the second condition on π_n of Corollary 1. Such a Bayesian interpretation can also be helpful in providing some intuition behind the penalty $\tilde{P}_n$ motivated in Birgé & Massart (2001) mostly due to technical reasons. In addition, under the conditions of Corollary 1, $P_n(n) \sim \tilde{P}_n(n) \sim cn$.

Furthermore, for sparse cases, where $\kappa \ll n$, under the conditions on the prior π_n of Corollary 1, both penalties P_n and $\tilde{P}_n$ are of the same so-called $2\kappa\log(n/\kappa)$-type penalties of the form $2\sigma_n^2 \zeta\kappa\{\log(n/\kappa) + c_{\kappa,n}\}$, where $\zeta > 1$ and $c_{\kappa,n}$ is negligible relative to $\log(n/\kappa)$. Such types of penalties have appeared within different frameworks in a series of recent works on estimation and model selection (Foster & Stine, 1999; George & Foster, 2000; Birgé & Massart, 2001; Abramovich et al., 2006, 2007).

3. Estimation in the Gaussian white noise model

3·1. *General algorithm*

In this section we apply the results of § 2 on estimation in the Gaussian sequence model (2) to wavelet estimation of the unknown response function f in the Gaussian white noise model (1).

Given a compactly supported scaling function ϕ of regularity $r > s$ and the corresponding mother wavelet ψ, one can generate an orthonormal wavelet basis on the unit interval from a finite number C_{j_0} of scaling functions $\phi_{j_0 k}$ at a primary resolution level j_0 and wavelets ψ_{jk} at resolution levels $j \geqslant j_0$ and scales $k = 0, \ldots, 2^j - 1$ (Cohen et al., 1993; Johnstone & Silverman, 2004a). For clarity of exposition, we use the same notation for interior and edge wavelets, and in what follows denote $\phi_{j_0 k}$ by $\psi_{j_0-1,k}$.

Then, f is expanded in the orthonormal wavelet series on [0, 1] as

$$f(t) = \sum_{j=j_0-1}^{\infty} \sum_{k=0}^{2^j-1} \theta_{jk}\psi_{jk}(t),$$

where $\theta_{jk} = \int_0^1 f(t)\psi_{jk}(t)dt$. In the wavelet domain, the Gaussian white noise model (1) becomes

$$Y_{jk} = \theta_{jk} + \epsilon_{jk} \quad (j \geqslant j_0 - 1,\ k = 0, \ldots, 2^j - 1),$$

where the empirical wavelet coefficients Y_{jk} are given by $Y_{jk} = \int_0^1 \psi_{jk}(t)dY(t)$ and ϵ_{jk} are independent $N(0, \sigma^2/n)$ random variables.

Define $J = \log_2 n$. Estimate wavelet coefficients θ_{jk} at different resolution levels j by the following scheme:

(1) set $\hat{\theta}_{j_0-1,k} = Y_{j_0-1,k}$;
(2) apply the Bayesian estimation procedure of Abramovich et al. (2007) described in § 2 to estimate θ_{jk} at resolution levels $j_0 \leqslant j < J$ by the corresponding $\hat{\theta}_{j,k}$;
(3) set $\hat{\theta}_{jk} = 0,\ \ j \geqslant J$.

The resulting wavelet estimator $\hat{f}_n$ of f is then defined as

$$\hat{f}_n(t) = \sum_{k=0}^{C_{j_0}-1} Y_{j_0-1,k}\psi_{j_0-1,k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \hat{\theta}_{jk}\psi_{jk}(t). \tag{9}$$

Theorem 1 below shows that, under mild conditions on the prior π_n, the resulting global wavelet estimator (9) of f, where the estimation procedure is applied to the entire set of wavelet coefficients at all resolution levels $j_0 \leqslant j < J$, up to a logarithmic factor, attains the minimax convergence rates over the whole range of Besov classes. Furthermore, Theorem 2 demonstrates that performing the estimation procedure at each resolution level separately allows one to remove the extra logarithmic factor. Moreover, a level-wise version of (9) allows one to estimate the derivatives of f at optimal convergence rates as well.

3·2. *Global wavelet estimator*

The number of wavelet coefficients at all resolution levels up to J is $\tilde{n} = 2^J - 2^{j_0} \sim n$ for large n. Let $\pi_n(\kappa) > 0$ $(\kappa = 0, \ldots, \tilde{n})$, be a prior distribution on the number of nonzero wavelet coefficients of f at all resolution levels $j_0 \leqslant j < J$, and let the prior variance of nonzero coefficients at the jth resolution level be τ_j^2/n; the corresponding level-wise variance ratios are $\gamma_j = \tau_j^2/\sigma^2$.

It is well known (Donoho & Johnstone, 1998) that, as $n \to \infty$, the minimax convergence rate for the L^2-risk of estimating the unknown response function f in the model (1) over Besov balls $B^s_{p,q}(M)$, where $0 < p, q \leqslant \infty$, $s > \max(0, 1/p - 1/2)$ and $M > 0$, is given by

$$\inf_{\tilde{f}_n} \sup_{f \in B^s_{p,q}(M)} E\left(\|\tilde{f}_n - f\|_2^2\right) \asymp n^{-2s/(2s+1)}.$$

THEOREM 1. *Let ψ be a mother wavelet of regularity r and let $\hat{f}_n$ be the corresponding global wavelet estimator* (9) *of f in the Gaussian white noise model* (1)*, where $f \in B^s_{p,q}(M)$, $0 < p, q \leqslant \infty$, $1/p < s < r$ and $M > 0$. Assume that there exist positive constants γ_- and γ_+ such that $\gamma_- \leqslant \gamma_j \leqslant \gamma_+$ for all $j = j_0, \ldots, J-1$. Let the prior π_n satisfy $\pi_n(\kappa) \geqslant (\kappa/n)^{c\kappa}$ for all $\kappa = 1, 2, \ldots, \exp(-9/2)n$ or, for a shorter range $\kappa = 1, 2, \ldots, \exp\{-c(\gamma_-)\}n$ if γ_- is known. Then, as $n \to \infty$,*

$$\sup_{f \in B^s_{p,q}(M)} E\left(\|\hat{f}_n - f\|_2^2\right) = O\left\{\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+1}}\right\}. \quad (10)$$

The proof of Theorem 1 is based on the relationship between the smoothness conditions on functions within Besov spaces and the conditions on their wavelet coefficients. Namely, if $f \in B^s_{p,q}(M)$, then the sequence of its wavelet coefficients $\{\theta_{jk},\ k = 0, \ldots, 2^j - 1,\ j = j_0, \ldots, J-1\}$ belongs to a weak $l_{2/(2s+1)}$-ball of a radius aM, where the constant a depends only on a chosen wavelet basis (Donoho, 1993, Lemma 2). One can then apply the corresponding results of Abramovich et al. (2007) for estimation over weak l_p-balls. Details of the proof of Theorem 1 are given in the Appendix.

The resulting global wavelet estimator does not rely on the knowledge of the parameters s, p, q and M of a specific Besov ball and it is, therefore, inherently adaptive. Theorem 1 establishes the upper bound for its L^2-risk and shows that the resulting adaptive global wavelet estimator is asymptotically near-optimal within the entire range of Besov balls. In fact, the additional logarithmic factor in (10) is the unavoidable minimal price for adaptivity for any global wavelet threshold estimator (Donoho et al., 1995; Cai, 1999), and in this sense, the upper bound for the convergence rates in (10) is sharp. To remove this logarithmic factor one should consider level-wise thresholding.

3·3. *Level-wise wavelet estimator*

Consider now the level-wise version of the wavelet estimator (9), where estimation is applied separately at each resolution level j. The number of wavelet coefficients at the jth resolution level is $n_j = 2^j$. Let $\pi_j(\kappa) > 0$ ($\kappa = 0, \ldots, 2^j$), be the prior distribution on the number of nonzero wavelet coefficients, and let τ_j^2/n be their level-wise prior variance, $j_0 \leqslant j < J$; the corresponding level-wise variance ratios are $\gamma_j = \tau_j^2/\sigma^2$.

THEOREM 2. *Let ψ be a mother wavelet of regularity r and let $\hat{f}_n(\cdot)$ be the corresponding level-wise wavelet estimator* (9) *of f in the Gaussian white noise model* (1)*, where $f \in B^s_{p,q}(M)$, $0 < p, q \leqslant \infty$, $1/p < s < r$ and $M > 0$. Assume that there exist positive constants γ_- and γ_+ such that $\gamma_- \leqslant \gamma_j \leqslant \gamma_+$ for all $j = j_0, \ldots, J-1$. Let the priors π_j satisfy the following conditions for all $j = j_0, \ldots, J-1$:*

(1) *$\pi_j(0) \geqslant 2^{-c_1 j}$ for some $c_1 > 0$;*

(2) $\pi_j(\kappa) \geqslant (\kappa 2^{-j})^{c_2\kappa}$ *for all* $\kappa = 1, \ldots, \alpha_j 2^j$, *where* $c_2 > 0$ *and* $0 < c_\alpha \leqslant \alpha_j \leqslant \exp\{-c(\gamma_j)\}$ *for some constant* $c_\alpha > 0$, *and the function* $c(\gamma_j) = 8(\gamma_j + 3/4)^2$ *was defined in Proposition* 1; *and*

(3) $\pi_j(2^j) \geqslant e^{-c_0 2^j}$ *for some* $c_0 > 0$.

Then, as $n \to \infty$, $\sup_{f \in B^s_{p,q}(M)} E(\|\hat{f}_n - f\|_2^2) = O(n^{-\frac{2s}{2s+1}})$.

For $f \in B^s_{p,q}(M)$, the sequence of its wavelet coefficients at the jth resolution level belongs to $l_p[\eta_j]$, where $\eta_j = C_0 n^{1/2} 2^{-j(s+1/2)}$ for some $C_0 > 0$ (Meyer, 1992, § 6.10). The conditions on the prior in Theorem 2 ensure that all the four statements of Proposition 1 simultaneously hold at all resolution levels $j_0 \leqslant j < J$ with $\beta = 0$, and one can exploit any of them at each resolution level. It is necessary for adaptivity of the resulting level-wise wavelet estimator (9).

As mentioned in § 2·2, all three conditions of Theorem 2 hold, for example, for the truncated geometric prior TrGeom$(1 - q_j)$, where q_j are bounded away from zero and one.

It turns out that requiring a slightly more stringent condition on $\pi_j(0)$ allows one also to estimate derivatives of f by the corresponding derivatives of its level-wise wavelet estimator $\hat{f}_n$ at the optimal convergence rates. Such a plug-in estimation of $f^{(m)}$ by $\hat{f}_n^{(m)}$ is, in fact, along the lines of the vaguelette-wavelet decomposition approach of Abramovich & Silverman (1998).

Recall that, as $n \to \infty$, the minimax convergence rate for the L^2-risk of estimating an mth derivative of the unknown response function f in the model (1) over Besov balls $B^s_{p,q}(M)$, where $0 \leqslant m < \min\{s, (s + 1/2 - 1/p)p/2\}$, $0 < p, q \leqslant \infty$ and $M > 0$, is given by (Donoho et al., 1997; Johnstone and Silverman, 2005)

$$\inf_{\tilde{f}_n^{(m)}} \sup_{f \in B^s_{p,q}(M)} E\left(\|\tilde{f}_n^{(m)} - f^{(m)}\|_2^2\right) \asymp n^{-2(s-m)/(2s+1)}.$$

The following Theorem 3 is a generalization of Theorem 2 for simultaneous level-wise wavelet estimation of a function and its derivatives.

THEOREM 3. *Let* ψ *be a mother wavelet of regularity* r *and let* $\hat{f}_n$ *be the level-wise wavelet estimator* (9) *of* f *in the Gaussian white noise model* (1), *where* $f \in B^s_{p,q}(M)$, $0 < p, q \leqslant \infty$, $1/p < s < r$ *and* $M > 0$. *Assume that there exist positive constants* γ_- *and* γ_+ *such that* $\gamma_- \leqslant \gamma_j \leqslant \gamma_+$ *for all* $j = j_0, \ldots, J - 1$. *Let the priors* π_j *satisfy the following conditions for all* $j = j_0, \ldots, J - 1$:

(1) $\pi_j(0) \geqslant 2^{-c_1 j 2^{-\beta j}}$ *for some* $\beta \geqslant 0$ *and* $c_1 > 0$;

(2) $\pi_j(\kappa) \geqslant (\kappa 2^{-j})^{c_2\kappa}$ *for all* $\kappa = 1, \ldots, \alpha_j 2^j$, *where* $c_2 > 0$ *and* $0 < c_\alpha \leqslant \alpha_j \leqslant \exp\{-c(\gamma_j)\}$ *for some constant* $c_\alpha > 0$, *and the function* $c(\gamma_j) = 8(\gamma_j + 3/4)^2$ *was defined in Proposition* 1;

(3) $\pi_j(2^j) \geqslant e^{-c_0 2^j}$ *for some* $c_0 > 0$.

Then, for all mth derivatives $f^{(m)}$ *of* f, *where* $0 \leqslant m \leqslant \beta/2$ *and* $m < \min\{s, (s + 1/2 - 1/p)p/2\}$, *as* $n \to \infty$,

$$\sup_{f \in B^s_{p,q}(M)} E\left(\|\hat{f}_n^{(m)} - f^{(m)}\|_2^2\right) = O\left(n^{-\frac{2(s-m)}{2s+1}}\right).$$

Theorem 2 is evidently a particular case of Theorem 3 corresponding to the case $m = 0$, for $\beta = 0$ in the condition on $\pi_j(0)$. Theorem 3 shows that the same proposed adaptive level-wise wavelet estimator (9) is simultaneously optimal for estimating a function and an entire range of its derivatives. This range is the same as that for the empirical Bayes shrinkage and threshold estimators appearing in Theorem 1 of Johnstone & Silverman (2005). The proof of Theorem 3 is given in the Appendix.

4. NUMERICAL STUDY

4·1. *Preamble*

In this section, we present a simulation study to illustrate the performance of the developed level-wise wavelet estimator (9) and compare it with three empirical Bayes wavelet estimators: the posterior mean and the posterior median of Johnstone & Silverman (2005), and the Bayes Factor of Pensky & Sapatinas (2007); and two other estimators: the block wavelet estimator NeighBlock of Cai & Silverman (2001) and the complex-valued wavelet hard thresholding estimator of Barber & Nason (2004). All the above Bayesian estimators and the block wavelet estimator are asymptotically minimax in a wide range of Besov balls. Although no such theoretical results have been established so far for the complex-valued wavelet estimator, it has performed well in simulations (Barber & Nason, 2004).

In practice, one typically deals with discrete data of a sample size n and the sampled data analogue of the Gaussian white noise model (1) is the standard nonparametric regression model

$$Y_i = f(i/n) + \epsilon_i \quad (i = 1, \ldots, n),$$

where ϵ_i are independent $N(0, \sigma^2)$ random variables. The corresponding global and level-wise Bayesian maximum a posteriori wavelet estimation procedures then use the empirical wavelet coefficients obtained by the discrete wavelet transforms of the data. However, utilizing the machinery of Johnstone & Silverman (2004a, 2005) for development of appropriate boundary-corrected wavelet bases, one can show that discretization does not affect the order of magnitude of the accuracy of the resulting wavelet estimates (Johnstone & Silverman, 2004a, 2005; PhD thesis at University of Cyprus of Petsa).

The computational algorithms were performed using the WaveLab and EbayesThresh software. The entire study was carried out using the Matlab programming environment.

4·2. *Estimation of parameters*

To apply the proposed level-wise wavelet estimator (9) one should specify the priors π_j, the noise variance σ^2 and the prior variances τ_j^2 or, equivalently, the variance ratios $\gamma_j = \tau_j^2/\sigma^2$. We used the truncated geometric priors TrGeom$(1 - q_j)$ discussed in § 3·3. Since the parameters σ^2, q_j and γ_j are rarely known a priori in practice, they should be estimated from the data in the spirit of empirical Bayes.

The unknown σ was robustly estimated by the median of the absolute deviation of the empirical wavelet coefficients at the finest resolution level $J - 1$, divided by 0·6745 as suggested by Donoho & Johnstone (1994a), and usually applied in practice. For a given σ, we then estimate q_j and γ_j by the conditional likelihood approach of Clyde & George (1999).

Consider the prior model described in § 2·1. The corresponding marginal likelihood of the observed empirical wavelet coefficients, say Y_{jk}, at the jth resolution level is then given by

$$L(q_j, \gamma_j; Y_j) \propto \sum_{\kappa=0}^{2^j} \pi_j(\kappa) \binom{2^j}{\kappa}^{-1} (1+\gamma_j)^{-\kappa/2} \sum_{x_i : \sum_k x_{ik} = \kappa} \exp\left\{ \frac{\gamma_j \sum_k x_{ik} Y_{jk}^2}{2\sigma^2(1+\gamma_j)} \right\},$$

where $\pi_j(\kappa) = (1 - q_j)q_j^{\kappa}/(1 - q_j^{2^j+1})$ and x_i are indicator vectors. Instead of direct maximization of $L(q_j, \gamma_j; Y_j)$ with respect to q_j and γ_j, regard the indicator vector x as a latent variable and consider the corresponding loglikelihood for the augmented data (Y_j, x), i.e.

$$l(q_j, \gamma_j; Y_j, x) = c + \log \pi_j(\kappa) - \log \binom{2^j}{\kappa} - \frac{\kappa}{2}\log(1+\gamma_j) + \frac{\gamma_j \sum_k x_k Y_{jk}^2}{2\sigma^2(1+\gamma_j)}, \tag{11}$$

where c is a constant. The EM-algorithm iteratively alternates between computation of the expectation of $l(q_j, \gamma_j; Y_j, x)$ in (11) with respect to the distribution of x given Y_j evaluated using the current estimates for the parameters' values at the E-step, and then updating the parameters by maximizing it with respect to q_j and γ_j at the M-step. However, for a general prior distribution π_n and for the truncated geometric prior, in particular, the EM-algorithm does not allow one to achieve analytic expressions on the E-step. Instead, we apply the conditional likelihood estimation approach originated by George & Foster (2000) and adapted to the wavelet estimation context by Clyde & George (1999). The approach is based on evaluating the augmented loglikelihood (11) at the mode for the indicator vector x at the E-step rather than using the mean as in the original EM-algorithm (Abramovich & Angelini, 2006).

For a fixed number κ of its nonzero entries, it is evident from (11) that the most likely vector $\hat{x}(\kappa)$ is $\hat{x}_k(\kappa) = 1$ for the κ largest $|Y_{jk}|$ and zero otherwise. For the given κ, maximizing (11) with respect to γ_j after some algebra yields $\hat{\gamma}_j(\kappa) = \max\{0, \sum_{k=1}^{\kappa} Y_{(k)}^2/(\kappa\sigma^2) - 1\}$. To simplify maximization with respect to q_j, approximate the truncated geometric distribution π_j in (11) by a non-truncated one. This approximation does not strongly affect the results, especially at sufficiently high resolution levels, and allows one to obtain analytic solutions for $\hat{q}_j$, i.e. $\hat{q}_j(\kappa) = \kappa/(\kappa+1)$. It is now straightforward to find $\hat{\kappa}$ that maximizes (11) together with the corresponding $\hat{\gamma}_j(\hat{\kappa})$ and $\hat{q}_j(\hat{\kappa})$. The above conditional likelihood approach results, therefore, in rapidly computable estimates for γ_j and q_j in closed forms.

4·3. *Simulation study*

We now present and discuss the results of the simulation study. For all three empirical Bayes wavelet estimators, we used the double exponential prior, where the corresponding prior parameters were estimated level-by-level by marginal likelihood maximization, as described in Johnstone & Silverman (2005). The prior parameters for the proposed level-wise wavelet estimator (9) were estimated by conditional likelihood maximization procedure described in § 4·2 above. For the block wavelet estimator, the lengths of the blocks and the thresholds were selected as suggested by Cai & Silverman (2001). Finally, for all competing methods, σ was estimated by the median of the absolute value of the empirical wavelet coefficients at the finest resolution level divided by 0·6745 as discussed in § 4·2.

In the simulation study, we evaluated the above six wavelet estimators for a series of test functions. We present here the results for the now-standard Bumps, Blocks, Doppler and Heavisine functions of Donoho & Johnstone (1994a), and the Wave (Marron et al., 1998; Antoniadis et al., 2001) and Peak (Angelini et al., 2003) functions defined, respectively, as

$$f(t) = 0{\cdot}5 + 0{\cdot}2\cos 4\pi t + 0{\cdot}1\cos 24\pi t \quad (0 \leqslant t \leqslant 1)$$

and

$$f(t) = \exp\{-|t - 0{\cdot}5|\} \quad (0 \leqslant t \leqslant 1).$$

See Fig. 1 for Wave and Peak test functions.

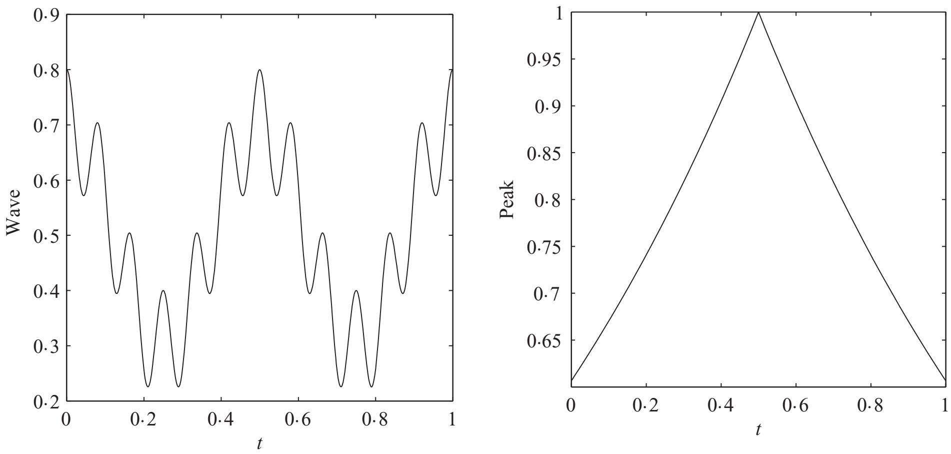

Fig. 1. Wave (left) and Peak (right) test functions.

For each test function, $M = 100$ samples were generated by adding independent Gaussian noise $\varepsilon \sim N(0, \sigma^2)$ to $n = 256$, 512 and 1024 equally spaced points on [0,1]. The value of the root signal-to-noise ratio was taken to be 3, 5 and 7 corresponding respectively to high, moderate and low noise levels. The goodness-of-fit for an estimator $\hat{f}$ of f in a single replication was measured by its mean squared error.

For brevity, we report the results only for $n = 1024$ using the compactly supported mother wavelet Coiflet 3 (Daubechies, 1992, p. 258) and the Lawton mother wavelet (Lawton, 1993) for the complex-valued wavelet estimator. The primary resolution level was $j_0 = 4$. Different choices of sample sizes and wavelet functions yielded similar results in magnitude.

The sample distributions of mean squared errors over replications for different wavelet estimators in the conducted simulation study were typically asymmetrical and affected by outliers. Therefore, we preferred the sampled medians of mean squared errors rather than means to gauge the estimators' goodness-of-fit. Thus, for each wavelet estimator, test function and noise level, we calculated the sample median of mean squared errors over all 100 replications. To quantify the comparison between the competing wavelet estimators over various test functions and noise levels, for each model we found the best wavelet estimator among the six, i.e. the one achieving the minimum median mean squared error. We then evaluated the relative median mean squared error of each estimator defined as the ratio between the minimum and the estimator's median mean squared errors; see Table 1.

As expected, Table 1 shows that there is no uniformly best wavelet estimator. Each one has its own favourite and challenging cases, and its relative performance strongly depends on the specific test function. Thus, the complex-valued estimator indeed demonstrates excellent results for Donoho & Johnstone's functions as has been reported in Barber & Nason (2004), but is much less successful for Peak and Wave. The block estimator is the best for the Peak and Doppler but the worst for Blocks and Bumps. The proposed Bayesian estimator (9) outperforms others for Wave but is less efficient for Donoho & Johnstone's (1994a) examples. Interestingly, the relative performance of the estimators is much less sensitive to the noise level. For each of the test functions, the corresponding best estimator is essentially the same for all noise levels.

The minimal relative median of mean squared errors of an estimator over all cases reflects its inefficiency at the most challenging combination of a test function and noise level, and can be viewed as a natural measure of its robustness. In this sense, the posterior mean estimator is the most robust although it is not the winner in any particular case.

Table 1. *Relative median mean squared errors,* MSE, *for various test functions, levels of the root signal-to-noise ratio,* RSNR, *and different wavelet estimators*

Signal	RSNR	MAP	BF	Postmed	Postmean	Block	CW
Peak	3	0·8697	0·1763	0·8279	0·6589	1	0·5795
	5	0·7772	0·1497	0·7864	0·6525	1	0·6234
	7	0·8033	0·186	0·8501	0·6958	1	0·6979
Wave	3	1	0·5614	0·9841	0·9103	0·4570	0·9189
	5	0·9841	0·4603	1	0·9165	0·6072	0·8265
	7	1	0·6241	0·9900	0·9303	0·7498	0·7793
Bumps	3	0·5968	0·6254	0·6814	0·7569	0·4769	1
	5	0·5221	0·5641	0·5893	0·6671	0·4788	1
	7	0·5132	0·5537	0·5707	0·6420	0·5202	1
Blocks	3	0·6595	0·6807	0·8815	0·9500	0·5606	1
	5	0·6875	0·727	0·8541	0·9065	0·4416	1
	7	0·6921	0·7134	0·7806	0·8535	0·4288	1
Doppler	3	0·7214	0·611	0·8277	0·8709	0·9878	1
	5	0·6962	0·6739	0·8116	0·8583	1	0·9119
	7	0·7655	0·7122	0·8236	0·883	1	0·9382
Heavisine	3	0·7523	0·3566	0·9333	0·9154	0·8406	1
	5	0·6640	0·3764	0·8622	0·8427	0·5796	1
	7	0·6931	0·3505	0·8298	0·8424	0·5028	1

MAP, the proposed Bayesian estimator; BF, Bayes factor; Postmed, posterior median; Postmean, posterior mean; Block, block; and CW, complex-valued hard thresholding.

We also compared different thresholding estimators in terms of sparsity measured by the average percentage of nonzero wavelet coefficients that remained after thresholding; see Table A1. The posterior mean estimator was not included in this comparison since it is a nonlinear shrinkage but not a thresholding estimator. Similar to the previous results on the goodness-of-fit, the relative sparsity strongly depends on the test function. However, except for the Doppler example, the estimator (9) is consistently the most sparse among Bayesian estimators.

Apart from providing a theoretical justification, our numerical results show that our estimator demonstrates good performance in finite sample settings and can, therefore, be viewed as a contribution to the list of useful wavelet-based function estimation tools.

Acknowledgement

Felix Abramovich and Vadim Grinshtein were supported by the Israel Science Foundation. The authors would like to thank the editor, the associate editor and three anonymous referees for helpful comments on improvements to this paper. Valuable remarks of Bernice Oberman are gratefully acknowledged.

Appendix

Throughout the proofs we use C to denote a generic positive constant, not necessarily the same each time it is used, even within a single equation.

Proof of Proposition 1. We start the proof of the proposition with the following lemma that establishes the bounds for binomial coefficients.

Table A1. *Average percentages of remaining coefficients for various test functions, levels of the root signal-to-noise ratio and different wavelet thresholding estimators*

Signal	RSNR	MAP	BF	Postmed	Block	CW
Peak	3	7·57	14·87	8·92	1·64	1·74
	5	5·62	15·89	8·39	1·61	1·93
	7	4·26	12·93	7·81	1·63	2·05
Wave	3	11·39	18·81	12·79	5·21	5·52
	5	11·51	20·19	12·52	6·30	6·28
	7	10·38	20·84	13·07	6·30	7·01
Bumps	3	10·63	12·13	10·86	16·63	12·23
	5	11·17	12·60	12·45	21·13	14·52
	7	12·65	13·90	13·80	23·70	16·03
Blocks	3	17·10	15·32	11·62	12·27	8·39
	5	10·39	12·20	11·72	18·03	12·07
	7	11·12	12·87	12·63	22·47	14·20
Doppler	3	11·46	14·73	8·58	5·69	5·13
	5	7·24	9·23	6·52	6·63	6·60
	7	6·42	9·58	6·57	7·27	7·86
Heavisine	3	6·35	19·27	10·75	1·98	2·17
	5	8·11	19·73	10·62	3·17	2·69
	7	10·87	18·52	12·55	4·00	3·39

LEMMA A1. *For all $n \geqslant 2$ and $\kappa = 1, 2, \ldots, n-1$,*

$$\left(\frac{n}{\kappa}\right)^{\kappa} \leqslant \binom{n}{\kappa} < \left(\frac{ne}{\kappa}\right)^{\kappa}. \tag{A1}$$

In particular, for $\kappa \leqslant n/e$,

$$\binom{n}{\kappa} < \left(\frac{n}{\kappa}\right)^{2\kappa}. \tag{A2}$$

This lemma generalizes Lemma A.1 of Abramovich et al. (2007), where the upper bound similar to that in (A1) was obtained for $\kappa = o(n)$.

Proof of Lemma A1. The obvious lower bound for the binomial coefficient in (A1) has been shown in Lemma A.1 of Abramovich et al. (2007). To prove the upper bound in (A1), using Stirling's formula one has

$$\binom{n}{\kappa} \leqslant \left(\frac{n}{e}\right)^{n} \left(\frac{e}{n-\kappa}\right)^{n-\kappa} \left(\frac{e}{\kappa}\right)^{\kappa} = \left(\frac{n}{\kappa}\right)^{\kappa} \left(\frac{n}{n-\kappa}\right)^{n-\kappa} \tag{A3}$$

for all $n \geqslant 2$ and $\kappa = 1, 2, \ldots, n-1$.

Note that $\log\{x/(x-1)\} < 1/(x-1)$ for all $x > 1$. In particular, for $x = n/\kappa$ it implies $\log\{n/(n-\kappa)\} < \kappa/(n-\kappa)$ and, therefore, $(\frac{n}{n-\kappa})^{n-\kappa} < \exp(\kappa)$ that together with (A3) completes the proof of (A1).

The second statement (A2) of the lemma is an immediate consequence of (A1) for $\kappa \leqslant n/e$. This completes the proof of Lemma A1. □

We now return to the proof of Proposition 1 and consider separately all the four cases covered by the proposition. The proof will exploit the general results of Abramovich et al. (2007) on the upper error bounds for the l^2-risk of the estimator (5), adapting them also for the case where the variance in the Gaussian sequence model (2) may depend on n.

Case 1. Under the condition $\pi_n \geqslant e^{-c_0 n}$, $c_0 > 0$, the definition (5) of $\hat{\mu}$ and (6) immediately imply $\|y - \hat{\mu}\|^2 \leqslant \|y - \hat{\mu}\|^2 + P_n(\hat{k}) \leqslant P_n(n) = O(n\sigma_n^2)$. Thus,

$$E(\|\hat{\mu} - \mu\|^2) \leqslant 2\{E(\|y - \hat{\mu}\|^2) + E(\|y - \mu\|^2)\} = O\big(n\sigma_n^2\big).$$

Case 2. Applying Corollary 1 of Abramovich et al. (2007) for $\kappa = 0$ yields

$$E\left(\|\hat{\mu} - \mu\|_2^2\right) \leqslant c_0(\gamma_n)\left\{\sum_{i=1}^{n} \mu_i^2 + 2\sigma_n^2(1 + 1/\gamma_n)\log \pi_n^{-1}(0)\right\} + c_1(\gamma_n)\{1 - \pi_n(0)\}\sigma_n^2,$$

where the exact expressions for $c_0(\gamma_n)$ and $c_1(\gamma_n)$ are given in Theorem 2 of Birgé & Massart (2001) with their $K = 1 + 1/(2\gamma_n)$; see the proof of Theorem 1 of Abramovich et al. (2007). In particular, under the assumptions of the proposition on the boundness of γ_n, the functions $c_0(\gamma_n)$ and $c_1(\gamma_n)$ are also bounded from above. For $2 \leqslant p \leqslant \infty$, the least favorable sequence μ_0 that maximizes $\sum_{i=1}^n \mu_i^2$ over $l_p[\eta_n]$ is $\mu_{01} = \cdots = \mu_{0n} = C_n n^{-1/p} = \eta_n\sigma_n$. As $n \to \infty$, one then has

$$\begin{aligned} E\left(\|\hat{\mu} - \mu\|_2^2\right) &\leqslant c_0(\gamma_n)\left\{\sigma_n^2\eta_n^2 n + 2\sigma_n^2(1 + 1/\gamma_n)\log \pi_n^{-1}(0)\right\} + c_1(\gamma_n)\{1 - \pi_n(0)\}\sigma_n^2 \\ &= O\left(\sigma_n^2 n\eta_n^2\right) + O\left(\sigma_n^2 n^{-\beta}\log n\right). \end{aligned}$$

Case 3. This is essentially a sparse case considered in Abramovich et al. (2007) and its proof is a direct consequence of their Theorem 6.

Case 4. The proof for this case is similar to that of Case 2 except that for $0 < p < 2$, the least favourable sequences μ_0 that maximize $\sum_{i=1}^n \mu_i^2$ over $\mu \in l_p[\eta_n]$ are permutations of the spike $(C_n, 0, \ldots, 0)$ and therefore $\sum_{i=1}^n \mu_{0i}^2 \leqslant \sigma_n^2 n^{2/p}\eta_n^2$. Repeating the arguments used in the proof of Case 2 for $\kappa = 0$, under the requirements of the proposition on boundedness of γ_n, we then get as $n \to \infty$,

$$\begin{aligned} E\left(\|\hat{\mu} - \mu\|_2^2\right) &\leqslant c_0(\gamma_n)\left\{\sigma_n^2 n^{2/p}\eta_n^2 + 2(1 + 1/\gamma_n)\sigma_n^2\log \pi_n^{-1}(0)\right\} + c_1(\gamma_n)\{1 - \pi_n(0)\}\sigma_n^2 \\ &= O\left(\sigma_n^2 n^{2/p}\eta_n^2\right) + O\left(\sigma_n^2 n^{-\beta}\log n\right). \end{aligned}$$

for all $\eta_n^p < n^{-1}(2\log n)^{p/2}$. This completes the proof of Proposition 1. □

Proof of Theorem 1. Let $R_j = \sum_{k=0}^{2^j-1} E\{(\hat{\theta}_{jk} - \theta_{jk})^2\}$, $j \geqslant j_0 - 1$, be the L^2-risk of the global wavelet estimator (9) at the jth resolution level. Due to the Parseval relation, $E(\|\hat{f}_n - f\|^2) = \sum_{j \geqslant j_0 - 1} R_j$. Scaling coefficients are not thresholded and therefore $R_{j_0-1} = C_{j_0}\sigma^2 n^{-1} = o(n^{-2s/(2s+1)})$ as $n \to \infty$. At very high resolution levels, where $j \geqslant J$, all wavelet coefficients $\hat{\theta}_{jk}$ are set to zero and, therefore, as $n \to \infty$,

$$\sum_{j=J}^{\infty} R_j = \sum_{j=J}^{\infty}\sum_{k=0}^{2^j-1} \theta_{jk}^2 = O(n^{-2s'}) = o\left(n^{-2s/(2s+1)}\right),$$

where $s' = s + 1/2 - 1/\min(p, 2)$ (Johnstone & Silverman, 2005).

Consider now $\sum_{j=j_0}^{J-1} R_j$. The set of wavelet coefficients $\{\theta_i\}$ of a function $f \in B_{p,q}^s(M)$ lies within a weak l_r-ball of a radius aM with $r = 2/(2s + 1)$, where the constant a depends only on a chosen wavelet basis: $m_r[\eta_n] = \{\theta : |\theta|_{(i)} \leqslant (aM)i^{-1/r}\}$ (Donoho, 1993, Lemma 2). The corresponding normalized radius $\eta_n = (\sigma/\surd n)^{-1}\tilde{n}^{-1/r}aM = O(n^{-s})$, where $\tilde{n} = n - 2^{j_0} \sim n$ for large n.

Under the conditions of the theorem, one can then apply Theorem 6 of Abramovich et al. (2007) for $m_r[\eta_n]$ to get

$$\sum_{j=j_0}^{J-1} R_j \leqslant \sup_{\theta \in m_r[\eta_n]} E\left(\|\hat{\theta} - \theta\|_2^2\right) = O\left\{\eta_n^r\left(2\log \eta_n^{-r}\right)^{1-r/2}\right\} = O\left\{\left(\frac{\log n}{n}\right)^{2s/(2s+1)}\right\}$$

as $n \to \infty$. This completes the proof of Theorem 1. □

Proof of Theorem 3. Let $R_j = \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \theta_{jk})^2,\ j \geqslant j_0 - 1$, be now the L^2-risk of the level-wise version of the wavelet estimator (9) at the jth resolution level. Johnstone & Silverman (2005, § 5.6) showed that $E(\|\hat{f}_n^{(m)} - f^{(m)}\|^2) \asymp \sum_{j \geqslant j_0-1} 2^{2mj} R_j$.

For any $f \in B^s_{p,q}(M)$, the sequence of its wavelet coefficients at the jth resolution level belongs to a strong l_p-ball of a normalized radius $\eta_j = C_0 n^{1/2} 2^{-j(s+1/2)}$ for some $C_0 > 0$ (Meyer, 1992, § 6.10).

Define

$$j_1 = \frac{1}{2s+1} \log_2 \left(\frac{nC_0^2}{c_\alpha^{2/p}} \right) \sim \frac{1}{2s+1} \log_2 n.$$

For sufficiently large n, $j_1 > j_0$. Note that $\eta_j^p \geqslant c_\alpha$ for $j \leqslant j_1$ and $\eta_j^p < c_\alpha$ for $j > j_1$ with obvious modifications for $p = \infty$. Consider the following cases:

(1) Scaling coefficients: $j = j_0 - 1$. Similarly to the global wavelet estimator, for a fixed primary resolution level j_0, $2^{2m(j_0-1)} R_{j_0-1} = O(n^{-1}) = o(n^{-2(s-m)/(2s+1)})$ as $n \to \infty$.

(2) Coarse resolution levels: $j_0 \leqslant j \leqslant j_1$. Applying the first statement of Proposition 1 for each level, one has, as $n \to \infty$,

$$\sum_{j=j_0}^{j_1} 2^{2mj} R_j \leqslant C \sum_{j=j_0}^{j_1} 2^{2mj} n^{-1} \sigma^2 n_j \leqslant C n^{-1} \sum_{j=j_0}^{j_1} 2^{(2m+1)j} = O\left(n^{-2(s-m)/(2s+1)}\right).$$

(3) Middle and high resolution levels: $j_1 < j < J$. Consider separately the cases (a) $2 \leqslant p \leqslant \infty$ and (b) $0 < p < 2$.

(a) $2 \leqslant p \leqslant \infty$. Under the conditions of the theorem, the second statement of Proposition 1 at the jth resolution level yields

$$R_j \leqslant C n^{-1} \left(n_j \eta_j^2 + n_j^{-\beta} \log n_j \right) \leqslant C(2^{-2js} + n^{-1} 2^{-\beta j} j)$$

and, hence, as $n \to \infty$,

$$\begin{aligned} \sum_{j=j_1+1}^{J-1} 2^{2mj} R_j &\leqslant C\left(2^{-2j_1(s-m)} + n^{-1} J^2\right) \leqslant C\left(n^{-2(s-m)/(2s+1)} + n^{-1} \log_2^2 n\right) \\ &= O\left(n^{-2(s-m)/(2s+1)}\right). \end{aligned}$$

(b) $0 < p < 2$. Let j_2 be the largest integer for which $\eta_j^p \geqslant n_j^{-1} (2 \log n_j)^{p/2}$. One can easily verify that $j_1 < j_2 < J$.

Using the monotonicity arguments, $\eta_j^p \geqslant n_j^{-1} (2 \log n_j)^{p/2}$ for all middle resolution levels $j_1 < j \leqslant j_2$. One can then apply the third statement of Proposition 1, and after some algebra, to get, for $m < (s + 1/2 - 1/p)p/2$,

$$\begin{aligned} \sum_{j=j_1+1}^{j_2} 2^{2mj} R_j &\leqslant C n^{-1} \sum_{j=j_1+1}^{j_2} 2^{(2m+1)j} n^{p/2} 2^{-jp(s+1/2)} \left\{ \log(n^{-p/2} 2^{jp(s+1/2)}) \right\}^{1-p/2} \\ &\leqslant C n^{-(1-p/2)} 2^{-j_1 p(s+1/2-(2m+1)/p)} \log\left(n^{-p/2} 2^{j_1 p(s+1/2)}\right) \\ &= O\left(n^{-2(s-m)/(2s+1)}\right) \end{aligned}$$

as $n \to \infty$.

At high resolution levels $j_2 < j < J$, $\eta_j^p < n_j^{-1} (2 \log n_j)^{p/2}$, and the fourth statement of Proposition 1 implies

$$R_j \leqslant C\left(2^{-2j(s+1/2-1/p)} + n^{-1} 2^{-j\beta} j\right).$$

Hence, for $0 \leqslant m \leqslant \beta/2$ and $m < \min\{s, (s + 1/2 - 1/p)p/2\}$, one has

$$\sum_{j=j_2+1}^{J-1} 2^{2mj} R_j \leqslant C\big(2^{-2(j_2+1)(s+1/2-1/p-m)} + n^{-1}J^2\big) = S_1 + S_2,$$

where evidently $S_2 = O(n^{-1}\log_2^2 n) = o(n^{-2(s-m)/(2s+1)})$ as $n \to \infty$. From the definition of j_2, $2^{(j_2+1)(s+1/2-1/p)} > \surd\{nC/(j_2+1)\} > \surd(nC/\log_2 n)$, which after some algebra yields $S_1 = o(n^{-2(s-m)/(2s+1)})$ as $n \to \infty$.

(4) Very high resolution levels: $j \geqslant J$. Using the results of Johnstone & Silverman (2005), as $n \to \infty$, the tailed sum

$$\sum_{j \geqslant J} 2^{2mj} R_j = O\big(n^{-2(s'-m)}\big) = o\big(n^{-2(s-m)/(2s+1)}\big),$$

where $s' = s + 1/2 - 1/\min(p, 2)$. Summarizing,

$$\sum_{j \geqslant j_0-1} 2^{2mj} R_j = O\big(n^{-2(s-m)/(2s+1)}\big)$$

as $n \to \infty$. This completes the proof of Theorem 3. □

References

Abramovich, F. & Angelini, C. (2006). Bayesian maximum a posteriori multiple testing procedure. *Sankhya* **68**, 436–60.

Abramovich, F. & Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comp. Statist. Data Anal.* **22**, 351–61.

Abramovich, F., Benjamini, Y., Donoho, D. L. & Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653.

Abramovich, F., Grinshtein, V. & Pensky, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* **35**, 2261–86.

Abramovich, F. & Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115–29.

Abramovich, F., Sapatinas, T. & Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc.* B **60**, 725–49.

Angelini, C., De Canditiis, D. & Leblanc, F. (2003). Wavelet regression estimation in nonparametric mixed effects models. *J. Mult. Anal.* **85**, 267–91.

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

Barber, S. & Nason, G. P. (2004). Real nonparametric regression using complex wavelets. *J. R. Statist. Soc.* B **66**, 927–39.

Birgé, L. & Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203–68.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384–98.

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

Cai, T. T. & Silverman, B. W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhyā* **63**, 127–48.

Clyde, M. A. & George, E. I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet-based Models*, Ed. P. Müller and B. Vidakovic, Lecture Notes in Statistics 141, pp. 309–22. New York: Springer.

Cohen, A., Daubechies, I. & Vial, P. (1993). Wavelet bases on the interval and fast algorithms. *Appl. Comp. Harmon. Anal.* **1**, 54–81.

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

Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comp. Harmon. Anal.* **1**, 100–15.

Donoho, D. L. & Johnstone, I. M. (1994a). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–55.

Donoho, D. L. & Johnstone, I. M. (1994b). Minimax risk over l_p-balls for l_q-errors. *Prob. Theory Rel. Fields* **99**, 277–303.

DONOHO, D. L. & JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–24.

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

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. & PICARD, D. (1995). Wavelet shrinkage: asymptotia? (with discussion). *J. R. Statist. Soc.* B **57**, 301–69.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. & PICARD, D. (1997). Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam*, Ed. D. Pollard, E. Torgersen and G. L. Yang, pp. 183 – 218. Berlin: Springer.

FOSTER, D. P. & STINE, R. (1999). Local asymptotic coding. *IEEE Trans. Info. Theory* **45**, 1289–93.

GEORGE, E. I. & FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–47.

JOHNSTONE, I. M. & SILVERMAN, B. W. (2004a). Boundary coiflets for wavelet shrinkage in function estimation. *J. Appl. Prob.* **41A**, 81–98.

JOHNSTONE, I. M. & SILVERMAN, B. W. (2004b). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**, 1594–1649.

JOHNSTONE, I. M. & SILVERMAN, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33**, 1700–52.

LAWTON, W. (1993). Applications of complex valued wavelet transforms to subband decomposition. *IEEE Trans. Sig. Proces.* **41**, 3566–68.

MARRON, J. S., ADAK, S., JOHNSTONE, I. M., NEUMANN, M. H. & PATIL, P. (1998). Exact risk analysis of wavelet regression. *J. Comp. Graph. Statist.* **7**, 278–309.

MEYER, Y. (1992). *Wavelets and Operators*. Cambridge: Cambridge University Press.

NASON, G. P. (1996). Wavelet shrinkage using cross-validation. *J. R. Statist. Soc.* B **58**, 463–79.

PENSKY, M. & SAPATINAS, T. (2007). Frequentist optimality of Bayes factor estimators in wavelet regression models. *Statist. Sinica* **17**, 599–633.

[*Received July* 2008. *Revised August* 2009]