Journal of Statistical Computation and Simulation
Vol. 74, No. 10, October 2004, pp. 741–764

Taylor & Francis
Taylor & Francis Group

EMPIRICAL BAYES APPROACH TO WAVELET REGRESSION USING ε-CONTAMINATED PRIORS

CLAUDIA ANGELINI[a] and THEOFANIS SAPATINAS[b,*]

[a]*Instituto per le Applicazioni del Calcolo 'Mauro Picone', Sezione di Napoli,*
Consiglio Nazionale delle Ricerche, Italy;
[b]*Department of Mathematics and Statistics, University of Cyprus,*
P.O. Box 20537, CY 1678 Nicosia, Cyprus

(Received 22 November 2002; In final form 31 October 2003)

We consider an empirical Bayes approach to standard nonparametric regression estimation using a nonlinear wavelet methodology. Instead of specifying a single prior distribution on the parameter space of wavelet coefficients, which is usually the case in the existing literature, we elicit the ε-contamination class of prior distributions that is particularly attractive to work with when one seeks robust priors in Bayesian analysis. The type II maximum likelihood approach to prior selection is used by maximizing the predictive distribution for the data in the wavelet domain over a suitable subclass of the ε-contamination class of prior distributions. For the prior selected, the posterior mean yields a thresholding procedure which depends on one free prior parameter and it is level- and amplitude-dependent, thus allowing better adaptation in function estimation. We consider an automatic choice of the free prior parameter, guided by considerations on an exact risk analysis and on the shape of the thresholding rule, enabling the resulting estimator to be fully automated in practice. We also compute pointwise Bayesian credible intervals for the resulting function estimate using a simulation-based approach. We use several simulated examples to illustrate the performance of the proposed empirical Bayes term-by-term wavelet scheme, and we make comparisons with other classical and empirical Bayes term-by-term wavelet schemes. As a practical illustration, we present an application to a real-life data set that was collected in an atomic force microscopy study.

Keywords: Atomic force microscopy; ε-Contaminated priors; Empirical Bayes; Exact risk analysis; Nonparametric regression; Type-II maximum likelihood priors; Wavelet shrinkage estimation; Wavelet transform

1 INTRODUCTION

Over the last decade, the nonparametric regression literature has been dominated by *nonlinear wavelet* methods. These methods are based on the idea of thresholding, which typically amounts to individual assessment of every empirical wavelet coefficient. If an empirical wavelet coefficient is sufficiently large in magnitude, that is if its magnitude exceeds a predetermined threshold, then the corresponding term in the empirical wavelet expansion is retained (or shrunk towards zero); otherwise it is omitted. The resulting term-by-term wavelet thresholding estimators are typically implemented through fast algorithms which makes them very appealing in practice (see, *e.g.*, Donoho and Johnstone, 1994; 1995; Donoho *et al.*, 1995).

Various empirical Bayes approaches for term-by-term wavelet shrinkage and wavelet thresholding estimators have also been proposed. (To introduce terminology, a *shrinkage* rule shrinks

[*] Corresponding author. E-mail: t.sapatinas@ucy.ac.cy

ISSN 0094-9655 print; ISSN 1563-5163 online
DOI: 10.1080/00949650310001643162

empirical wavelet coefficients to zero, whilst a *thresholding* rule shrinks and, in addition, sets to zero all empirical wavelet coefficients below a certain level.) These approaches impose a prior distribution on the wavelet coefficients of the unknown response function, designed to capture the sparseness of wavelet expansions common to most applications. A popular prior model for each wavelet coefficient is a mixture of two distributions, one mixture component corresponding to *negligible* wavelet coefficients, the other to *significant* wavelet coefficients. Usually, a mixture of two normal distributions or a mixture of one normal distribution and a point mass at zero is considered. Then the function is estimated by applying a suitable Bayes rule to the resulting posterior distribution of the wavelet coefficients. Different choices of loss function lead to different Bayes rules and hence to different (usually *level-dependent)* wavelet shrinkage and wavelet thresholding rules (see, *e.g.*, Chipman *et al.*, 1997; Abramovich *et al.*, 1998; Clyde *et al.*, 1998; Vidakovic, 1998; Clyde and George, 1999; 2000).

Extensive reviews and descriptions of various classical and empirical Bayes term-by-term wavelet schemes can be found in, for example, the book by Vidakovic (1999) and the review papers by Abramovich *et al.* (2000) and Antoniadis *et al.* (2001). The relative small sample performance of most of these wavelet schemes was also examined in an extensive simulation study by Antoniadis *et al.* (2001), using a variety of sample sizes, test functions, signal-to-noise ratios and wavelet filters. While empirical Bayes term-by-term wavelet shrinkage and wavelet thresholding methods have proven to be an effective tool in nonparametric function estimation, their computational cost and/or careful hand-tuning of their free prior parameters may be a handicap, when compared with some classical term-by-term wavelet thresholding methods. For example, the very best of them require computationally expensive iterative procedures (like the EM algorithm of Dempster *et al.*, 1977) to obtain estimates of the free prior parameters, while others require careful hand-tuning of the various free prior parameters to obtain overall good numerical performances.

In this article, instead of specifying a single prior distribution on the parameter space of wavelet coefficients, which is usually the case in the existing literature, we elicit a class of plausible prior distributions. We consider the *ε-contamination* class of prior distributions that is particularly attractive to work with when one seeks robust priors in Bayesian analysis (see *e.g.*, Berger, 1985, Chapter 4). The *type II maximum likelihood prior* (ML-II prior) approach of Berger and Berliner (1986) and Berger and Sellke (1987) is used by maximizing the predictive distribution for the data in the wavelet domain over a suitable subclass of the ε-contamination class of prior distributions. For the prior selected, the proposed empirical Bayes term-by-term wavelet methodology possess the following advantageous features: (i) its posterior mean (the Bayes rule under L^2-loss) yields a *bonafide* thresholding rule which is *level- and amplitude-dependent*, thus allowing better adaptation in function estimation (this is different from existing empirical Bayes term-by-term wavelet schemes that are usually only level-dependent); (ii) it only depends on *one* free prior parameter (it is therefore an almost-free method compared with existing empirical Bayes term-by-term wavelet schemes that usually depend on more than one free prior parameter); and (iii) its computational cost is *low* (it is much less computationally expensive than those empirical Bayes term-by-term wavelet schemes requiring iterative procedures to obtain estimates of their free prior parameters).

We note that a class of prior distributions similar to the ε-contaminated class used in this article has been also recently considered by Angelini and Vidakovic (2004) to study wavelet shrinkage in nonparametric regression via a Γ-minimax approach. However, their methodology is suitable when prior information about the energy of the signal of interest is available. Moreover, although their methodology is almost computationally inexpensive, the resulting posterior-based rule is a shrinkage rule, depends on two free prior parameters and is level-dependent; hence, it does not possess the appealing features of the proposed thresholding rule. Furthermore, by adapting the simulation-based approach of Barber (2001), we have obtained a computationally very fast and easy to apply scheme to compute pointwise Bayesian credible intervals for the resulting function estimate obtained from the proposed empirical Bayes term-by-term wavelet thresholding methodology.

The article is organized as follows. In Section 2 we describe the proposed empirical Bayes term-by-term wavelet thresholding methodology to the standard nonparametric regression estimation and present a simulation-based approach to compute pointwise Bayesian credible intervals for the resulting function estimate. In Section 3, we carry out numerically an exact risk analysis of the resulting thresholding rule and propose an automatic choice of its free prior parameter. In Section 4 we provide several simulated examples to illustrate the performance of the proposed empirical Bayes term-by-term wavelet thresholding methodology and the simulation-based pointwise Bayesian credible intervals for the resulting function estimates. Moreover, we compare the proposed empirical Bayes term-by-term wavelet thresholding methodology with various standard classical and empirical Bayes term-by-term wavelet schemes. As a practical illustration, we also present an application to a real-life data set that was collected in an atomic force microscopy study. Some concluding remarks are made in Section 5.

2 EMPIRICAL BAYES APPROACH TO WAVELET REGRESSION USING ε-CONTAMINATED PRIORS

In this section we first briefly describe the standard nonparametric regression setting in the data domain and the equivalent setting in the wavelet domain. We then explain in detail the steps of the proposed empirical Bayes term-by-term wavelet scheme to the standard nonparametric regression estimation and present a simulation-based approach to computing pointwise Bayesian credible intervals for the resulting function estimate.

2.1 The Standard Nonparametric Regression Setting

Consider the standard nonparametric regression setting

$$y_i = g(t_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $t_i = i/n, n = 2^J$ for some positive integer J, ε_i are independent and identically distributed $N(0, 1)$ random variables and the noise level σ may, or may not, be known. The problem is to estimate the underlying function g from the observations $\mathbf{y} = (y_1, \dots, y_n)'$ without assuming any particular parametric structure for its form.

The wavelet approach to this problem is easily described. Given a suitable wavelet basis and *primary* resolution level $j_0 \geq 0$, the discrete wavelet transform (DWT) of $\mathbf{y}$ gives rise to an n-dimensional vector $\hat{\mathbf{d}}$ consisting of what are known as the *empirical scaling* coefficients $\hat{c}_{j_0 k} (k = 0, \dots, 2^{j_0} - 1)$ and the *empirical wavelet* coefficients $\hat{d}_{jk} (j = j_0, \dots, J - 1; k = 0, \dots, 2^j - 1)$. In practice, the DWT (and its inverse (IDWT)) may be performed through a computationally fast algorithm developed by Mallat (1989) that requires only O(n) operations. Due to the orthogonality of the DWT, it follows from Eq. (1) that

$$\hat{c}_{j_0 k} = \hat{c}_{j_0 k} + \sigma \varepsilon_{j_0 k}, \quad k = 0, 1, \dots, 2^{j_0} - 1, \tag{2}$$

$$\hat{d}_{jk} = d_{jk} + \sigma \varepsilon_{jk}, \quad j = j_0, \dots, J - 1, \quad k = 0, \dots, 2^j - 1, \tag{3}$$

where the ε_{jk} are themselves independent and identically distributed $N(0, 1)$ random variables, and the c_{j_0k} and d_{jk} are, respectively, the true *scaling* and *wavelet* coefficients of the (unknown) vector of function values $\mathbf{g} = (g(t_1), \ldots, g(t_n))'$.

Throughout the article we assume that the noise level σ is unknown and it is robustly estimated by the median absolute deviation of the empirical wavelet coefficients of the data at the highest resolution level divided by 0.6745 (see Donoho and Johnstone, 1994; 1995), *i.e.*,

$$\hat{\sigma} = \frac{\text{median}(\{|\hat{d}_{J-1,k}|: k = 0, 1, \ldots, 2^{J-1} - 1\})}{0.6745} \tag{4}$$

Following asymptotic considerations (see Härdle *et al.*, 1998, Chapter 10), the primary resolution level j_0 that we have used throughout our examples was chosen to be

$$j_0 = [\log_2(\log(n))] + 1, \tag{5}$$

where $[x]$ denotes the integer part of x.

2.2 The $\boldsymbol{\varepsilon}$-Contamination Prior Model

It is advisable to keep the coefficients on the lower coarse resolution levels intact because they represent 'low-frequency' terms that usually contain important components of the function g. Thus, the scaling coefficients $\{c_{j_0k}: k = 0, \ldots, 2^{j_0} - 1\}$ are assumed to be *mutual independent* random variables and vague priors are placed on them

$$c_{j_0k} \sim N(0, \varepsilon), \quad \varepsilon \longrightarrow \infty. \tag{6}$$

The wavelet coefficients $\{d_{jk}: j = j_0, \ldots, J - 1; k = 0, \ldots, 2^j - 1\}$ are assumed to be *mutual independent* random variables and to belong to the ε-contaminated class of prior distributions

$$\Gamma = \{\pi(d_{jk}) = (1 - \varepsilon_j)\delta(0) + \varepsilon_j q(d_{jk}), q \in \mathcal{D}\}, \tag{7}$$

where $0 \leq \varepsilon_j \leq 1, \pi$ denotes the prior distribution of the wavelet coefficients d_{jk}, $\delta(0)$ is a point mass at zero (which models wavelet coefficients with *negligible* amplitudes) and $\mathcal{D}$ denotes a class of possible *spread* distributions (which models wavelet coefficients with *large* amplitudes). According to the ε-contamination prior model (7), at each resolution level $j = j_0, \ldots, J - 1$, each wavelet coefficient d_{jk} is either zero with probability $(1 - \varepsilon_j)$ or with probability ε_j is distributed with a probability distribution from the class $\mathcal{D}$ of plausible prior distributions. Note that we are using the same prior parameter ε_j at each resolution level $j = j_0, \ldots, J - 1$ and it should be specified appropriately (see Sec. 3.2).

For each $j = j_0, \ldots, J - 1; k = 0, \ldots, 2^j - 1$, the predictive distribution (denoted throughout the article by $m(\hat{d}_{jk}|\pi)$ to emphasize its dependence on the prior π, rather than using the standard notation $m(\hat{d}_{jk})$)

$$m(\hat{d}_{jk}|\pi) = \int f(\hat{d}_{jk}|d_{jk})\pi(d_{jk})\,\mathrm{d}d_{jk},$$

where $f(\hat{d}_{jk}|d_{jk})$ is the probability density function given in Eq. (3), reflects the plausibility of π. Therefore, a natural method of choosing π is to use the type II maximum likelihood approach, (ML-II approach) (see, *e.g.*, Berger, 1985, Sec. 3.5.4). In other words, the ML-II prior $\hat{\pi} \in \Gamma$ satisfies (for the observed empirical wavelet coefficients $\hat{d}_{jk}$)

$$m(\hat{d}_{jk}|\hat{\pi}) = \sup_{\pi \in \Gamma} m(\hat{d}_{jk}|\pi), \tag{8}$$

where Γ is given by Eq. (7).

The attractive richness and flexibility of the ε-contamination class of prior distributions given in Eq. (7) (through appropriate choice of $\mathcal{D}$ one can assume that Γ contains all plausible priors and no implausible ones) has substantial calculational advantages as well. Surprisingly, when Γ is *infinite* dimensional, the calculation of the ML-II prior given in Eq. (8) is not too involved. In what follows we exploit one such a case which is appropriate for the proposed empirical Bayes term-by-term wavelet methodology; the case $\mathcal{D} = \mathcal{D}_{\text{uni}}$ where

$$\mathcal{D}_{\text{uni}} = \{\text{the class of all densities of the form } q(|d_{jk}|),\ q \text{ nonincreasing}\}. \tag{9}$$

Note that in Eq. (9), only symmetric densities around zero are allowed, and only those which are unimodal. This seems to be in agreement with the observation that, at each resolution level, the wavelet coefficients of most noiseless signals or images encountered in practice possess a density function with a marked peak at zero and heavy tails (see, *e.g.*, Mallat, 1989). Various probability models have been used to model such behavior, including the mixtures of two normal distributions (see, *e.g.*, Chipman *et al.*, 1997), the mixture of a normal distribution and a point mass at zero (see, *e.g.*, Abramovich *et al.*, 1998; Clyde *et al.*, 1998; Clyde and George, 1999; 2000), and the generalized Gaussian distribution (see, *e.g.*, Moullin and Liu, 1999; Figueiredo and Nowak, 2001). However, probability models with flatter tails could give a better match between probability models and noiseless data in the wavelet domain. The ε-contamination class of prior models that we consider in Section 2.3 allows such models to be selected when the ML-II priors are chosen.

2.3 The ML-II Approach to Prior Selection

For any π in the ε-contamination class of prior distributions Γ given in Eq. (7), it is clear that

$$m(\hat{d}_{jk}|\pi) = (1 - \varepsilon_j) m_0(\hat{d}_{jk}) + \varepsilon_j m(\hat{d}_{jk}|q),$$

where

$$m_0(\hat{d}_{jk}) = \int f(\hat{d}_{jk}|d_{jk}) \delta(0)\, \mathrm{d}d_{jk} \tag{10}$$

and

$$m(\hat{d}_{jk}|q) = \int f(\hat{d}_{jk}|d_{jk}) q(d_{jk})\, \mathrm{d}d_{jk}. \tag{11}$$

Hence, the ML-II, prior for each $j = j_0, \ldots, J-1$; $k = 0, \ldots, 2^j - 1$ can be found by maximizing $m(\hat{d}_{jk}|\pi)$ over $\pi \in \Gamma$ given in Eq. (7), or equivalently by maximizing $m(\hat{d}_{jk}|q)$ over $q \in \mathcal{D}$.

By considering now $q \in \mathcal{D}_{\text{uni}}$, where $\mathcal{D}_{\text{uni}}$ is given in Eq. (9), for each $j = j_0, \ldots, J-1$; $k = 0, \ldots, 2^j - 1$, the predictive distribution $m(\hat{d}_{jk}|q)$ is maximized over $q \in \mathcal{D}_{\text{uni}}$ at $\hat{q}$, where

$\hat{q}$ follows a uniform distribution on $(-l_{jk}, l_{jk})$ (*i.e.*, $\hat{q} \sim \mathcal{U}(-l_{jk}, l_{jk})$), with l_{jk} being chosen to maximize (with a slight abuse of notation and assuming that the maximum is attained)

$$m(\hat{d}_{jk}|l_{jk}) = \int_{-l_{jk}}^{l_{jk}} \frac{1}{2l_{jk}} f(\hat{d}_{jk}|d_{jk})\,\mathrm{d}d_{jk}.$$

Indeed, following Berger and Berliner (1986) and Berger and Sellke (1987), for each $j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1$, any density $q \in \mathcal{D}_{\text{uni}}$ can be written as

$$q(|d_{jk}|) = \int_0^{\infty} \frac{1}{2l_{jk}} I_{[0,l_{jk})}(|d_{jk}|)\,\mathrm{d}F(l_{jk}),$$

where I_A is the indicator function of the set A and F is some distribution function on $[0, \infty)$. Therefore, it is not difficult to see that, for each $j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1$,

$$\begin{aligned} m(\hat{d}_{jk}|q) &= \int_{-\infty}^{\infty} f(\hat{d}_{jk}|d_{jk}) q(|d_{jk}|)\,\mathrm{d}d_{jk} \\ &= \int_0^{\infty} m(\hat{d}_{jk}|l_{jk})\,\mathrm{d}F(l_{jk}). \end{aligned}$$

Clearly, for each $j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1$, this is maximized by choosing F to be a point mass at a l_{jk} maximizing $m(\hat{d}_{jk}|l_{jk})$ (the corresponding value of q is indicated by $\hat{q}$).

Recall from Eq. (3) that $f(\hat{d}_{jk}|d_{jk})$ is a $N(d_{jk}, \sigma^2)$ probability density function. Then, it is easily seen that, for each $j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1$,

$$m(\hat{d}_{jk}|l_{jk}) = \frac{1}{2l_{jk}} \left\{ \Phi\left(\frac{l_{jk} - \hat{d}_{jk}}{\sigma}\right) - \Phi\left(\frac{-l_{jk} - \hat{d}_{jk}}{\sigma}\right) \right\}, \tag{12}$$

where Φ is the $N(0, 1)$ cumulative distribution function. Therefore, for each $j = j_0, \ldots, J-1$; $k = 0, \ldots, 2^j - 1$, the maximum of $m(\hat{d}_{jk}|l_{jk})$ is unique and can be found by differentiating Eq. (12) with respect to l_{jk} and setting equal to zero (unless $|\hat{d}_{jk}| \le \sigma$, in which case the maximum is attained at $l_{jk}^{\max} = 0$). After some simple algebra we found that, for each $j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1$,

$$\Phi(l_{jk}^* - z_{jk}) - \Phi(-(l_{jk}^* + z_{jk})) = l_{jk}^* \{\phi(l_{jk}^* - z_{jk}) + \phi(-(l_{jk}^* + z_{jk}))\},$$

where $z_{jk} = |\hat{d}_{jk}|/\sigma$, $l_{jk}^* = l_{jk}/\sigma$ and ϕ is the $N(0, 1)$ probability density function. A useful rewriting of this equation, especially for iterative calculations, is

$$l_{jk}^* = z_{jk} + \left\{ -2\log\left(\sqrt{2\pi} \left[\frac{1}{l_{jk}^*} (\Phi(l_{jk}^* - z_{jk}) - \Phi(-(l_{jk}^* + z_{jk}))) - \phi(-(l_{jk}^* + z_{jk})) \right] \right) \right\}^{1/2}. \tag{13}$$

For iterative calculations, we plug in a guess for l_{jk}^* on the right hand side of Eq. (13) (an initial – good guess – is by taking $l_{jk}^* = z_{jk}$), carry out the calculation in Eq. (13) to obtain a revised l_{jk}^*, and iterate until the value stabilizes. Figure 1(a) shows the values of l_{jk}^* as a function of z_{jk} (for fixed j and k).

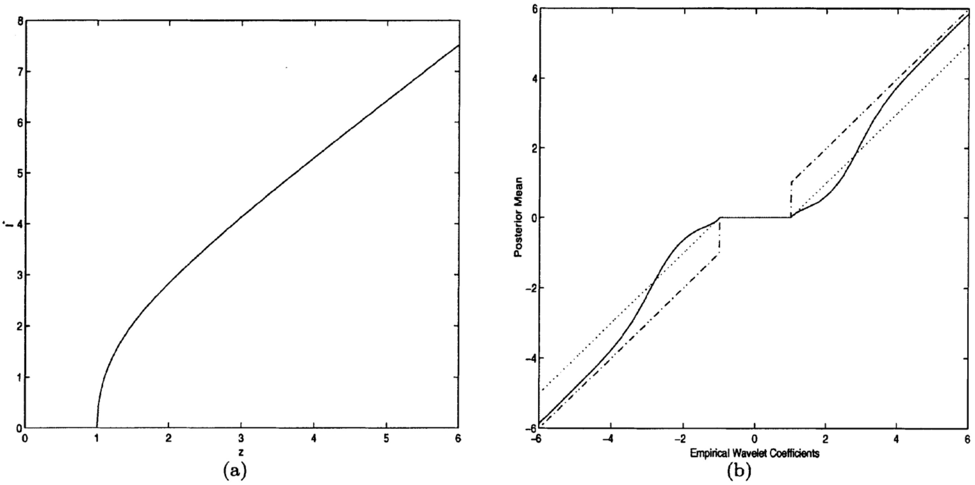

FIGURE 1 (a) Graph of l^*_{jk} as a function of z_{jk} (for fixed j and k); (b) ML-II thresholding rule (——) as a function of the empirical wavelet coefficients, $\hat{d}_{jk}$, with $\varepsilon_j = 0.4$ (for fixed j and k), superimposed with the hard (—··— ··— ··), soft (·····) thresholding rules with threshold value 1. For the three thresholding rules, $\sigma = 1$.

Thus, for each $j = j_0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, the ML-II prior, $\hat{\pi}$, for this case is given by

$$\hat{\pi}(d_{jk}) = \begin{cases} \delta(0) & \text{if } l^{\max}_{jk} = 0 \\ (l - \varepsilon_j)\delta(0) + \varepsilon_j \dfrac{1}{2l^{\max}_{jk}} I_{(-l^{\max}_{jk}, l^{\max}_{jk})} & \text{if } l^{\max}_{jk} > 0, \end{cases}$$

where $l^{\max}_{jk} = l^*_{jk}\sigma$ with l^*_{jk} being the solution of Eq. (13) if $|\hat{d}_{jk}| > \sigma$, otherwise $l^*_{jk} = 0$. For each $j = j_0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, the ML-II prior, $\hat{\pi}$, is therefore either a point mass at zero or a mixture of two distributions; a point mass at zero and a uniform distribution on $(-l^{\max}_{jk}, l^{\max}_{jk})$ (with weights depending only on the particular resolution level). Note also that, for each $j = j_0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, the value of $l^{\max}_{jk}$ depends on the corresponding empirical wavelet coefficient $\hat{d}_{jk}$.

2.4 The Posterior Distribution and the Corresponding Bayes Rule for the L^2-Loss Function

Subject to the ε-contamination prior model (7), the posterior distribution is easily evaluated. More specifically, for each $j = j_0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, after some straightforward calculations, we have the following closed form for the resulting posterior distribution

$$\pi(d_{jk}|\hat{d}_{jk}) = (1 - \varepsilon^*_{jk})\delta(0) + \varepsilon^*_{jk} q(d_{jk}|\hat{d}_{jk}), \tag{14}$$

where

$$\varepsilon^*_{jk} = \frac{\varepsilon_j m(\hat{d}_{jk}|q)}{(1-\varepsilon_j)m_0(\hat{d}_{jk}) + \varepsilon_j m(\hat{d}_{jk}|q)}, \tag{15}$$

$$q(d_{jk}|\hat{d}_{jk}) = \frac{q(d_{jk}) f(\hat{d}_{jk}|d_{jk})}{m(\hat{d}_{jk}|q)},$$

and $m_0(\hat{d}_{jk})$, $m(\hat{d}_{jk}|q)$ are given respectively by Eqs. (10) and (11). Although Eq. (14) seems to express the posterior class as an ε-contamination class, this is not the case because ε_{jk}^* is not fixed; Eq. (15) clearly shows that ε_{jk}^* depends on $q(d_{jk})$.

For each $j = j_0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, using the results of the ML-II prior analysis considered in Section 2.3 for $q \in \mathcal{D}_{\text{uni}}$, where $\mathcal{D}_{\text{uni}}$ is given in Eq. (9), and after some simple algebra, we have the following closed form for the resulting posterior distribution

$$\hat{\pi}(d_{jk}|\hat{d}_{jk}) = \begin{cases} \delta(0) & \text{if } |\hat{d}_{jk}| \le \sigma \\ (1-\varepsilon_{jk}^*)\delta(0) + \varepsilon_{jk}^*\hat{q}(d_{jk}|\hat{d}_{jk}) & \text{otherwise,} \end{cases} \tag{16}$$

where

$$\hat{q}(d_{jk}|\hat{d}_{jk}) = \left\{ \Phi\left(\frac{l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) - \Phi\left(\frac{-l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) \right\}^{-1} \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left(-\frac{1}{2\sigma^2}(d_{jk} - \hat{d}_{jk})^2\right), \tag{17}$$

ε_{jk}^* is given by Eq. (15) with

$$m_0(\hat{d}_{jk}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\hat{d}_{jk}^2}{2\sigma^2}\right) \tag{18}$$

and

$$m(\hat{d}_{jk}|q) = \frac{1}{2l_{jk}^{\max}} \left\{ \Phi\left(\frac{l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) - \phi\left(\frac{-l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) \right\}, \tag{19}$$

and $l_{jk}^{\max} = l_{jk}^*\sigma$ with l_{jk}^* being the solution of Eq. (13) if $|\hat{d}_{jk}| > \sigma$, otherwise $l_{jk}^* = 0$.

For each $j = j_0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, the corresponding Bayes rule for the L^2-loss function is obviously obtained by the posterior mean. After some simple algebra, for each $j = j_0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, the posterior mean $\mathbb{E}(d_{jk}|\hat{d}_{jk})$ corresponding to the posterior distribution $\hat{\pi}(d_{jk}|\hat{d}_{jk})$ given in Eqs. (16) and (17) is given by

$$\mathbb{E}(d_{jk}|\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \le \sigma \\ \varepsilon_{jk}^*\{\hat{d}_{jk} - \sigma A_{jk} B_{jk}\} & \text{otherwise,} \end{cases} \tag{20}$$

where ε_{jk}^* is given by Eqs. (15), (18) and (19), and A_{jk} and B_{jk} are given respectively by

$$A_{jk} = \left\{ \Phi\left(\frac{l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) - \Phi\left(\frac{-l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) \right\}^{-1} \tag{21}$$

and

$$B_{jk} = \left\{ \phi\left(\frac{l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) - \phi\left(\frac{-l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right) \right\}, \tag{22}$$

where $l_{jk}^{\max} = l_{jk}^{*}\sigma$ with l_{jk}^{*} being the solution of Eq. (13) if $|\hat{d}_{jk}| > \sigma$, otherwise $l_{jk}^{*} = 0$. Clearly the posterior mean given in Eq. (20) yields a thresholding rule (which we call, *ML-II thresholding rule*) with threshold value σ. It depends on one free prior parameter (*i.e.*, it depends only on ε_j) and it is level- and amplitude-dependent, thus allowing better adaptation in function estimation. For a plot of the proposed ML-II thresholding rule for a particular case see Figure 1(b).

Finally, due to the vague priors (6) imposed on the scaling coefficients c_{j_0k} and the fact that from Eq. (2) we have $f(\hat{c}_{j_0k}|c_{j_0k})$ to be a $N(c_{j_0k}, \sigma^2)$ probability density function, the L^2-loss function results in estimating c_{j_0k} by their empirical counterparts $\hat{c}_{j_0k}$. Hence, the vector $\hat{\mathbf{g}}$ of the corresponding estimate of the unknown response function g at the observed data-points can be derived by simply performing the IDWT of $\{\hat{c}_{j_0k}\colon k = 0, 1, \ldots, 2^{j_0} - 1\}$ and $\{\mathbb{E}(d_{jk}|\hat{d}_{jk})\colon j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1\}$.

2.5 An Algorithm to Computing Pointwise Bayesian Credible Intervals for the Resulting Function Estimate

After some simple algebra, for each $j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1$, the cumulative distribution function of the posterior distribution given in Eq. (16) can be expressed as

$$F(d_{jk}|\hat{d}_{jk}) = \begin{cases} I_{[0,\infty)}(d_{jk}) & \text{if } |\hat{d}_{jk}| \le \sigma \\ (1-\varepsilon_{jk}^{*})I_{[0,\infty)}(d_{jk}) + \varepsilon_{jk}^{*}A_{jk}\Gamma_{jk} & \text{otherwise,} \end{cases} \tag{23}$$

where

$$\Gamma_{jk} = \left\{\Phi\left(\frac{d_{jk} - \hat{d}_{jk}}{\sigma}\right) - \Phi\left(\frac{-l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right)\right\},$$

ε_{jk}^{*} is given by Eq. (15), A_{jk} is given by Eq. (21) and $l_{jk}^{\max} = l_{jk}^{*}\sigma$ with l_{jk}^{*} being the solution of Eq. (13) if $|\hat{d}_{jk}| > \sigma$, otherwise $l_{jk}^{*} = 0$. Clearly, when $|\hat{d}_{jk}| > \sigma$, the cumulative distribution function F given in Eq. (23) is supported in $(-l_{jk}^{\max}, l_{jk}^{\max})$ and we have

$$\lim_{d_{jk}\to -l_{jk}^{\max}} F(d_{jk}|\hat{d}_{jk}) = 0 \quad \text{and} \quad \lim_{d_{jk}\to l_{jk}^{\max}} F(d_{jk}|\hat{d}_{jk}) = 1.$$

It also follows that $F(d_{jk}|\hat{d}_{jk})$ has a discontinuity only at $d_{jk} = 0$; it is strictly monotonic and invertible in each interval $(-l_{jk}^{\max}, 0)$ and $(0, l_{jk}^{\max})$, while it is not invertible on the whole real line. We now define

$$u' = \lim_{d_{jk}\uparrow 0} F(d_{jk}|\hat{d}_{jk}) = \varepsilon_{jk}^{*}A_{jk}\left\{\Phi\left(\frac{-\hat{d}_{jk}}{\sigma}\right) - \Phi\left(\frac{-l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right)\right\}$$

and

$$u'' = \lim_{d_{jk}\downarrow 0} F(d_{jk}|\hat{d}_{jk}) = (1-\varepsilon_{jk}^{*}) + \varepsilon_{jk}^{*}A_{jk}\left\{\Phi\left(\frac{-\hat{d}_{jk}}{\sigma}\right) - \Phi\left(\frac{-l_{jk}^{\max} - \hat{d}_{jk}}{\sigma}\right)\right\}.$$

The inverse cumulative distribution function, $F^{-1}(u|\hat{d}_{jk})$ for $u \in [0, 1]$, is defined as

$$F^{-1}(u|\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \le \sigma \\ G^{-1}(u|\hat{d}_{jk}) & \text{otherwise,} \end{cases}$$

where

$$G^{-1}(u|\hat{d}_{jk}) = \begin{cases} 0 & \text{if } u \in [u', u''] \\ \sigma\Phi^{-1}\left\{u\dfrac{1}{A_{jk}\varepsilon^*_{jk}} + \Phi\left(\dfrac{-l^{\max}_{jk} - \hat{d}_{jk}}{\sigma}\right)\right\} + \hat{d}_{jk} & \text{if } u \in [0, u') \\ \sigma\Phi^{-1}\left\{(u - 1 + \varepsilon^*_{jk})\dfrac{1}{A_{jk}\varepsilon^*_{jk}} + \Phi\left(\dfrac{-l^{\max}_{jk} - \hat{d}_{jk}}{\sigma}\right)\right\} + \hat{d}_{jk} & \text{if } u \in (u'', 1]. \end{cases}$$

It is easily seen now that, if $-l^{\max}_{jk} < d_{jk} < 0$ and $|\hat{d}_{jk}| > \sigma$,

$$d_{jk} = \sigma\Phi^{-1}\left\{\frac{u}{A_{jk}\varepsilon^*_{jk}} + \Phi\left(\frac{-l^{\max}_{jk} - \hat{d}_{jk}}{\sigma}\right)\right\} + \hat{d}_{jk}$$

while, if $0 < d_{jk} < l^{\max}_{jk}$ and $|\hat{d}_{jk}| > \sigma$,

$$d_{jk} = \sigma\Phi^{-1}\left\{\frac{u - (1 - \varepsilon^*_{jk})}{A_{jk}\varepsilon^*_{jk}} + \Phi\left(\frac{-l^{\max}_{jk} - \hat{d}_{jk}}{\sigma}\right)\right\} + \hat{d}_{jk}.$$

Following the idea of Barber (2001), we can now generate a sample value for each wavelet coefficient, d^s_{jk}, from the posterior distribution given in Eq. (16) using the following algorithm

- If $|\hat{d}_{jk}| < \sigma$, set $d^s_{jk} = 0$.
- Else, generate $u \sim U[0, 1]$.
  - if $u \in [u', u'']$, set $d^s_{jk} = 0$.
  - if $u < u'$, set
  $$d^s_{jk} = \sigma\Phi^{-1}\left\{\frac{u}{A_{jk}\varepsilon^*_{jk}} + \Phi\left(\frac{-l^{\max}_{jk} - \hat{d}_{jk}}{\sigma}\right)\right\} + \hat{d}_{jk}.$$
  - if $u > u''$, set
  $$d^s_{jk} = \sigma\Phi^{-1}\left\{\frac{u - 1 + \varepsilon^*_{jk}}{A_{jk}\varepsilon^*_{jk}} + \Phi\left(\frac{-l^{\max}_{jk} - \hat{d}_{jk}}{\sigma}\right)\right\} + \hat{d}_{jk}.$$

We can now use the above algorithm to resample a value for each wavelet coefficient, d^s_{jk}, generating thus a full set of sampled wavelet coefficients $\mathbf{d}^s = \{d^s_{jk}\colon j = j_0, \ldots, J-1; k = 0, \ldots, 2^j - 1\}$. This set of coefficients can be transformed via the IDWT (see Sec. 2.1), to construct a sample from the posterior distribution of the vector of function values $\mathbf{g}$, given the observations $\mathbf{y}$. By sampling a total, say S, of such sets of values, $(1-\alpha)100\%$ pointwise Bayesian credible intervals can be found for each function point, $g(t_i)$, $i = 1, \ldots, n$, by ordering the sampled values and taking the central $(1-\alpha)100\%$ as the resulting pointwise Bayesian credible interval.

3 EXACT RISK ANALYSIS OF THE ML-II THRESHOLDING RULE AND ELICITATION OF THE FREE PRIOR PARAMETER

In this section we first carry out numerically an exact risk analysis of the ML-II thresholding rule (20) to explore robustness in risk, bias and variance when the free prior parameter ε_j $(j = j_0, \dots, J-1)$ change. We then propose an automatic choice of this level-dependent free prior parameter that is guided by considerations on the exact risk analysis and on the shape of the ML-II thresholding rule, enabling the resulting estimator to be fully automated in practice.

3.1 Exact Risk Analysis of the ML-II Thresholding Rule

Exact risk analysis of any proposed wavelet shrinkage or wavelet thresholding rule has received considerable attention since it allows for comparison of different classical and Bayesian term-by-term wavelet schemes in nonparametric regression estimation. When the corresponding rule is given in a simple form, then the exact risk analysis can be carried out explicitly. For instance, Donoho and Johnstone (1994) and Bruce and Gao (1996) provide exact risk analysis for *hard* and *soft* thresholding rule under squared error loss. Gao and Bruce (1997) give the rationale for introducing the *firm* (or *semi-soft*) thresholding rule utilizing exact risk analysis. However, the form of the ML-II thresholding rule (20) is more complex and the exact risk analysis has to be carried numerically using a suitable quadrature formula. The aim of our analysis is to explore the robustness of the ML-II thresholding rule in terms of risk (under squared loss), squared bias, and variance when the free prior parameter ε_j ranges in [0, 1]. The analysis has been carried out in MATLAB using Gauss–Lobatto quadrature formula to evaluate the integrals. In the following, we briefly describe the numerical findings.

The ML-II thresholding rule, for any choice of the free prior parameter ε_j, has a threshold value σ. As depicted in Figure 2(a), it sets small values of empirical wavelet coefficients to zero (*i.e.*, when $|\hat{d}_{j,k}| \leq \sigma$) and shrinks (nonlinearly) large values of empirical wavelet coefficients (*i.e.*, when $|\hat{d}_{j,k}| > \sigma$). However, it does not overpenalize large values of empirical wavelet coefficients and hence does not create excessive biases when the wavelet coefficients have large values. In fact, for large values of empirical wavelet coefficients (when $|\hat{d}_{jk}| \to \infty$), the proposed ML-II thresholding rule approaches the hard thresholding rule. Apart from the fixed threshold value σ, the amount of shrinkage of the ML-II thresholding rule essentially depends on the choice of ε_j that represents the probability of a wavelet coefficient being significant; the smaller the value of ε_j the heavier the amount of shrinkage.

The risks (under squared losses) of the ML-II thresholding rules given in Figure 2(a) are presented in Figure 2(b). Additionally, the risks of the hard and soft thresholding rules are computed and superimposed in the figure as a reference. We notice an obvious trade-off in the risk performance of the ML-II thresholding rule for small, medium and large values of wavelet coefficients, respectively. When ε_j is small, the risk remains very close to 0, for small values of $|d_{j,k}|$; the risk rapidly increases for medium values of $|d_{jk}|$ (the rate and the amplitude of such increase is larger when ε_j is smaller). Finally, for large values of $|d_{jk}|$ the risks of the various ML-II thresholding rules are approaching the hard thresholding risk (mimic the fact that the shape of the rules approaches the hard thresholding rule). For this range, the risks are always lower than the soft thresholding risks.

The squared-biases of the ML-II thresholding rules given in Figure 2(a) are depicted in Figure 2(c). We observe that they are not much influenced from the free prior parameter ε_j for small values of $|d_{jk}|$, remaining always very close to zero. Analogously, for large values of $|d_{jk}|$, the squared-bias of the ML-II thresholding rule tends to the squared-bias of the hard thresholding rule, remaining lower than the corresponding function for the soft thresholding

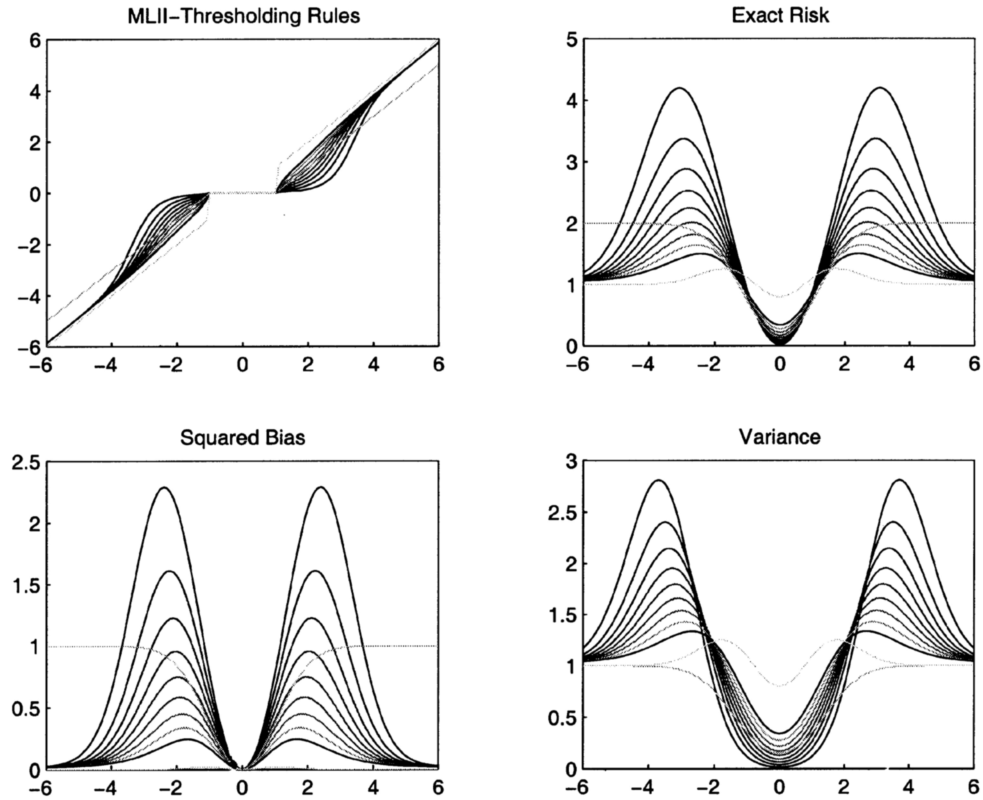

FIGURE 2 (a) ML-II thresholding rules for ε_j (for fixed j) ranging from 0.1 (upper envelope function) to 0.9 (lower envelop function). (b) Exact risks of the ML-II thresholding rules given in (a); (c) Squared-biases of the ML-II thresholding rules given in (a); (d) Variances of the ML-II thresholding rules given in (a). (Both hard and soft thresholding rules are superimposed in all figures with threshold value 1. For all figures, $\sigma = 1$.)

rule. For medium values of $|d_{jk}|$ and for large values of ε_j, the squared-bias of the ML-II thresholding rule increases very fast.

The behavior of the variance functions of the ML-II thresholding rules given in Figure 2(a) are illustrated in Figure 2(d). We observe that, for small values of $|d_{j,k}|$, the variance functions of the ML-II thresholding rules are lower than the variance of the hard thresholding rule and decrease with ε_j. For large values of $|d_{j,k}|$, the variance functions rapidly tend to the variance of the hard and soft thresholding rules remaining, however, a little larger. For medium values of $|d_{j,k}|$, the smaller the value of ε_j the larger the variance function observed.

3.2 Elicitation of the Free Prior Parameter

The elicitation of prior parameters is one of the major issues in Bayesian analysis. In order to have an effective ML-II thresholding rule, the free prior parameter ε_j should be carefully elicited at each resolution level. The elicitation is guided by considerations of the exact risk analysis and of the shape of the ML-II thresholding rule presented in Section 3.1.

It has been shown that, at each resolution level, ε_j regulates the amount of shrinkage at zero. In fact, it should be close to 0 (large shrinkage) at the finest level of detail, where most of the coefficients are zeros; and close to 1 (small shrinkage) at coarse levels, where most of the coefficients are significant. However, the exact risk analysis presented in Section 3.1 shows that the ML-II thresholding rule is robust with respect to the choice of ε_j for large values of $|d_{j,k}|$, it is relatively robust for small values of $|d_{j,k}|$, while the influence of ε_j is visible for medium values of $|d_{j,k}|$. Hence the capability of the resulting estimator will depend on the elicitation of ε_j at medium values of $|d_{j,k}|$.

For practical purposes, and taking into account the exact risk analysis and the shape of the ML-II thresholding rule discussed above, we consider an automatic choice of ε_j according to the suggestions by Vidakovic and Ruggeri (2001). The level-dependent values of ε_j are defined as

$$\varepsilon_j = \frac{1}{(j - j_0 + 1)^{\gamma}}, \quad j_0 \leq j \leq J - 1, \tag{24}$$

where j_0 represents the primary resolution level (see Sec. 2.1) and γ is empirically chosen. An automatic choice for γ that was found to work well in our examples and the sensitivity of the results on a range of γ values is discussed in Section 4.1.

4 NUMERICAL RESULTS AND COMPARISONS

The purpose of this section is to provide several examples to illustrate the performance of the proposed empirical Bayes term-by-term wavelet thresholding methodology. First, we carry out an extensive simulation study to investigate the finite sample performance of this methodology. We also give comparisons with various standard classical and empirical Bayes term-by-term wavelet schemes. Then, we apply the proposed methodology to a real-life data set that was collected in an atomic force microscopy study.

The computational algorithms related to wavelet analysis were performed using Version 8 of the WaveLab toolbox (see Buckheit *et al.*, 1995) for MATLAB that is freely avallable from http://www-stat.stanford.edu/software/software.html and the GaussianWaveDen toolbox (see Antoniadis *et al.*, 2001) for MATLAB that is freely available from http://www.jstatsoft.org/v06/i06. The entire study was carried out using the MATLAB programming environment.

4.1 Simulation Study

The results of the simulation study will now be presented, with the remainder of this section devoted to the discussion of these results. We compare the proposed empirical Bayes term-by-term wavelet scheme (which we call *ML-IIThresh*) with three classical term-by-term wavelet schemes (the *VisuShrink* method of Donoho and Johnstone, 1994, the hybrid version of the *SureShrink* method of Donoho and Johnstone, 1995, and the ‘leave-out-half’ version of the *Cross Validation* method of Nason, 1996) and three empirical Bayes term-by-term wavelet schemes (the *PostMean* method of Clyde and George, 1999, the *PostMedian* method of Abramovich *et al.*, 1998, and the *ABE* method of Figueiredo and Nowak, 2001). For excellent numerical performances, we consider the VisuShrink and the ‘leave-out-half’ version of the Cross Validation methods with hard thresholding, while the versions of PostMean and PostMedian methods that we consider 10 use the EM-algorithm for estimating their free prior parameters. Note that the hybrid version of the *SureShrink* method is only defined for soft thresholding, while the *ABE* method *does not* contain any free prior parameters.

In this simulation study, we evaluate the various classical and empirical Bayes term-by-term wavelet schemes by estimating the noise level σ according to Eq. (4) and by choosing the primary resolution level j_0 according to Eq. (5). For the ML-Thresh method, ε_j was chosen according to Eq. (24). We have considered the following nine test functions: *Wave*, *Blip*, *HeaviSine*, *Doppler*, *Angles*, *Parabolas*, *Time Shifted Sine*, *Spikes* and *Corner*; these functions are supposed to caricature spatially variable signals arising in a number of scientific fields. For all test functions, Daubechies' nearly symmetric wavelets of order 8, *Symmlet 8*, were used.

For each test function, $M = 200$ samples were generated by adding independent random noise $\varepsilon \sim N(0, \sigma^2)$ to $n = 256$ (small sample size), 512 (moderate sample size) and 1024 (large sample size) equally spaced points on [0,1]. The value of σ was taken to correspond to the values 3 (high noise level), 5 (moderate noise level) and 7 (low noise level) for the root signal-to-noise ratio (RSNR)

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

The nine test functions based on $n = 1024$ design points with the addition of independent normally distributed noise with mean zero and RSNR = 5, giving a visual impression of the large sample size and moderate noise level used in the simulation study, are shown in Figure 3.

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

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

its average mean absolute deviation (AMAD) from the M simulations, defined as

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

and its average maximal absolute deviation (AMXD) from the M simulations, defined as

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

In order to examine the effect of the parameter γ (the only free parameter through ε_j according to Eq. (24)) on the numerical performance of the ML-IIThresh method, for each test function, sample size and RSNR we have computed the AMSE, AMAD and AMXD for a range of γ values. For brevity, we only report here in detail the results for AMSE and $n = 1024$. Different combinations of goodness-of-fit measures and sample sizes yield basically similar results. We have found that for moderate or high RSNR, the performance of the estimator in terms of AMSE is quite robust with respect to the choice of γ. Larger values of γ would provide an almost free noise reconstruction at the price of oversmoothing the singularities. When RSNR decreases, AMSE exhibits a significant influence with respect to γ, showing preference to relatively large values of γ in most cases. Figure 4 shows the behavior of AMSE, as γ ranges in [1,3], for the nine test functions based on $n = 1024$ design points. The value of $\gamma = 1.8$ has been selected as a default value to compromise the bias and the variance in the reconstruction, when no additional information on the true signal is available. Although not reproduced here, larger values of γ can often improve the estimator, especially for very large sample sizes. As an illustration of the visual appearance of the ML-IIThresh method with the default value $\gamma = 1.8$, Figure 5 shows the estimates obtained using the noisy samples of the nine test functions given in Figure 3.

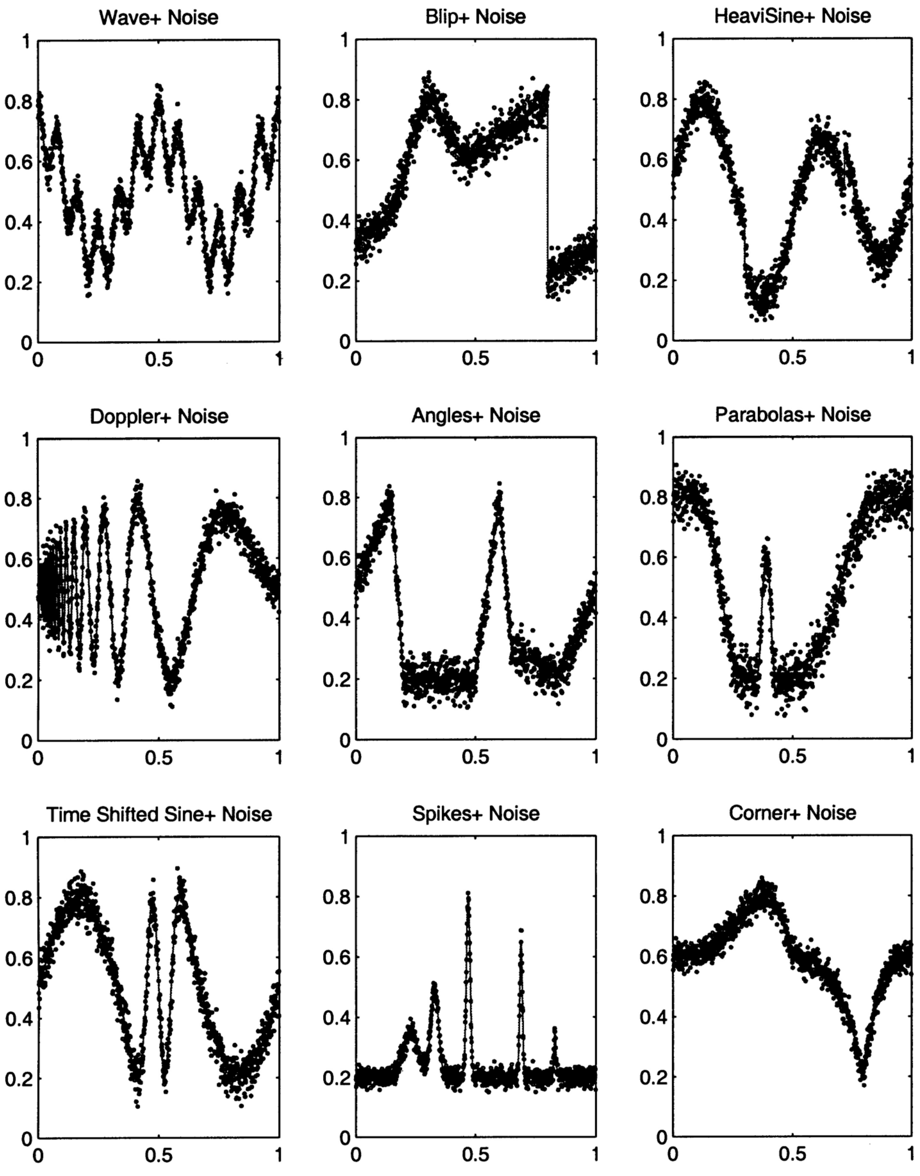

FIGURE 3 The nine test functions (——) used in the simulation study, based on $n = 1024$ design points, with the addition of independent normally distributed noise with mean zero and RSNR = 5 (∘), giving a visual impression of the large sample size and moderate noise level used in the simulation study.

The various goodness-of-fit measures, AMSE, AMAD and AMXD, have been computed for the ML-IIThresh method using the default value $\gamma = 1.8$ and compared with the corresponding indices computed for the various classical and empirical Bayes term-by-term wavelet schemes used in this simulation study. The results across the various combinations of test functions, sample sizes and RSNR show that the ML-IIThresh method outperforms the VisuShrink method with hard thresholding, the hybrid version of the SureShrink method, the 'leave-out-half' version of the CrossValidation method with hard thresholding and the ABE method, and performs as well as (sometimes even better than) PostMean and PostMedian methods. Moreover, it is seen that the VisuShrink method with hard thresholding and the hybrid version of

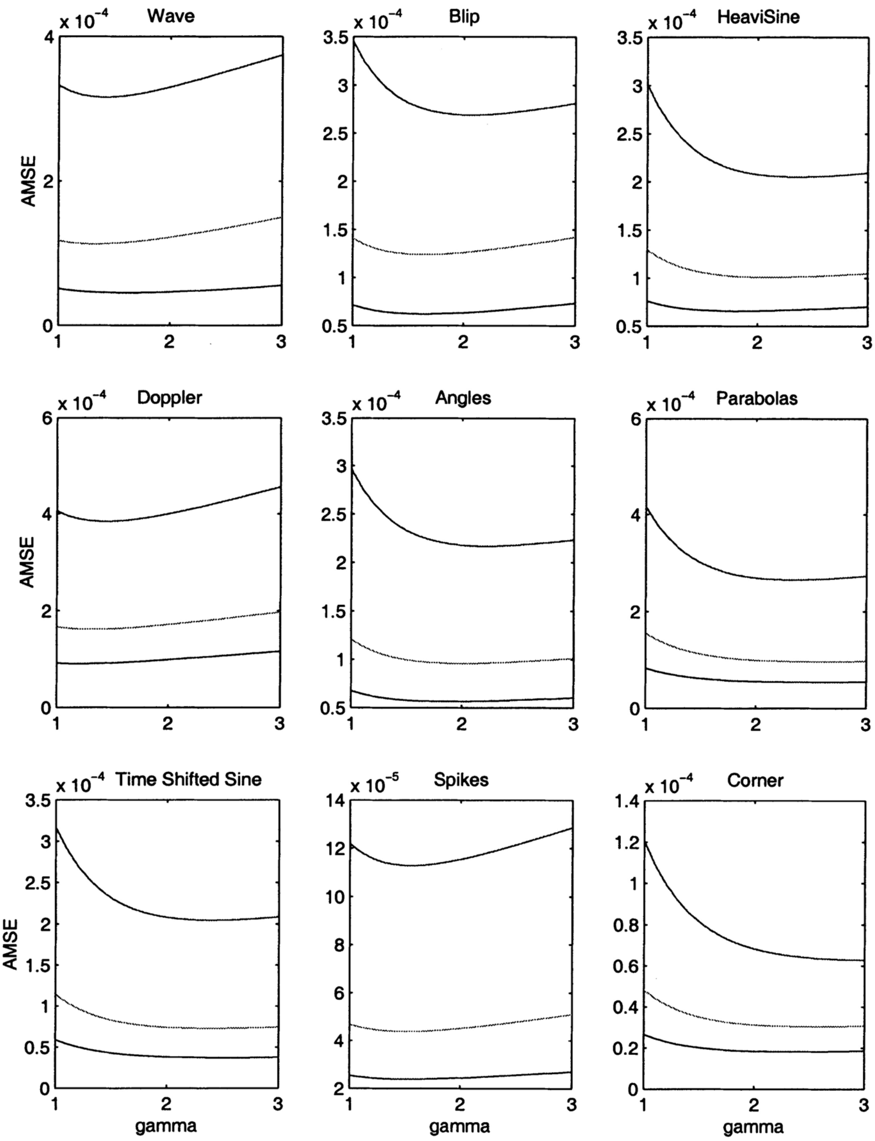

FIGURE 4 The behavior of AMSE for the ML-IIThresh method as γ ranges in [1, 3], for the nine test functions based on $n = 1024$ design points. The top line corresponds to RSNR = 3, the middle line corresponds to RSNR = 5, and the bottom line corresponds RSNR = 7.

the SureShrink method both outperform the ABE method. If one instead uses the VisuShrink with soft thresholding and SureShrink methods then, although not reproduced here, the ABE method is superior to these two latter methods; a point that has been highlighted as one of the main features in the simulation study of Figueiredo and Nowak (2001). Obviously, our simulation study reveals features that affect the conclusions drawn by Figueiredo and Nowak (2001) about the relative performance of their free parameter empirical Bayes term-by-term wavelet scheme with standard classical term-by-term wavelet schemes. For brevity, we only show the results obtained for AMSE. Figure 6 shows the boxplots of the AMSE computed for the 9 test functions based on $n = 1024$ design points and RSNR = 5 given in Figure 3.

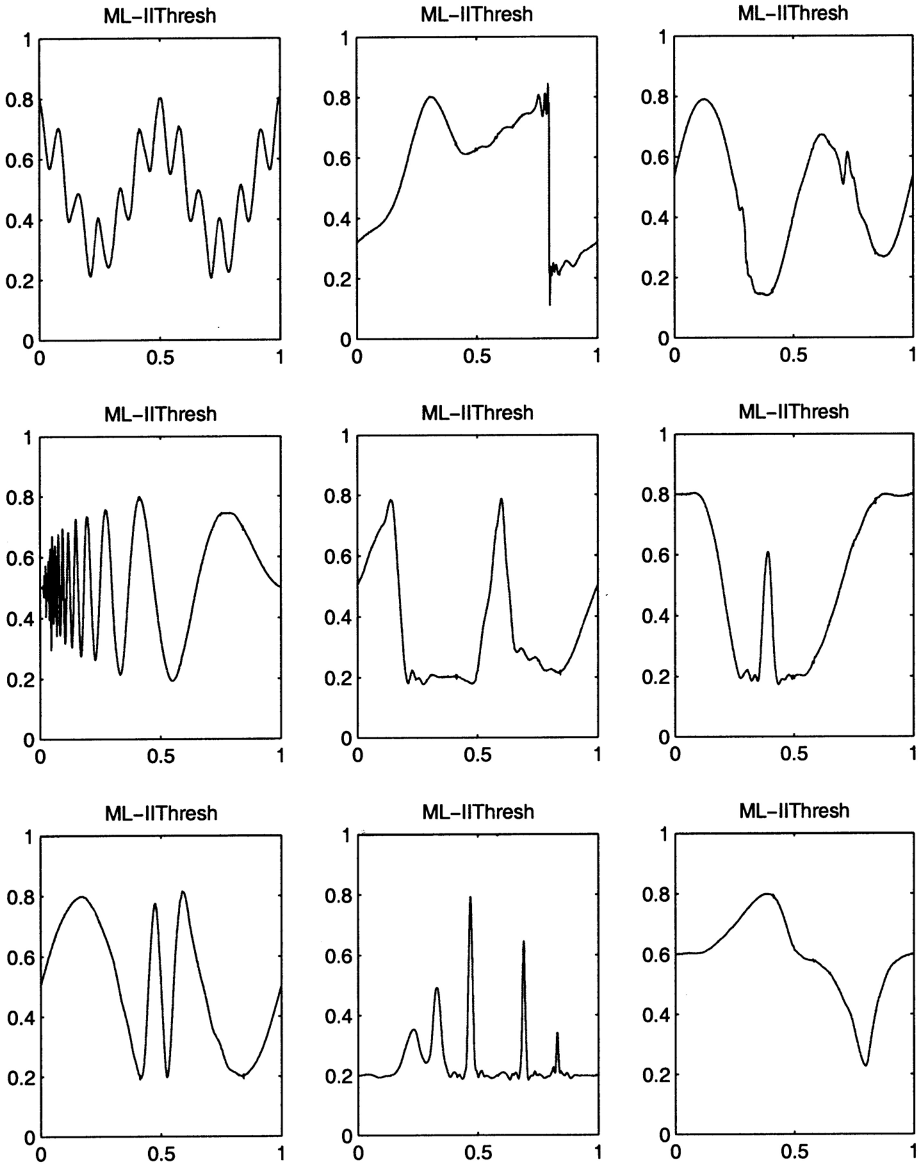

FIGURE 5 Estimates obtained using the ML-IIThresh method with the default value $\gamma = 1.8$ for the noisy samples of the nine test functions given in Figure 3.

The ML-IIThresh method using the default value $\gamma = 1.8$ has also been compared with the classical and empirical Bayes term-by-term wavelet schemes in terms of the average CPU time. The results show that obviously the ML-IIThresh method has a larger computational cost with respect to the VisuShrink with soft thresholding, hybrid version of the SureShrink and ABE methods (that are almost computationally inexpensive procedures). However its computational cost is much smaller than the corresponding computational cost of the 'leave-out-half' version of the CrossValidation with hard thresholding, PostMean and PostMedian methods. We also

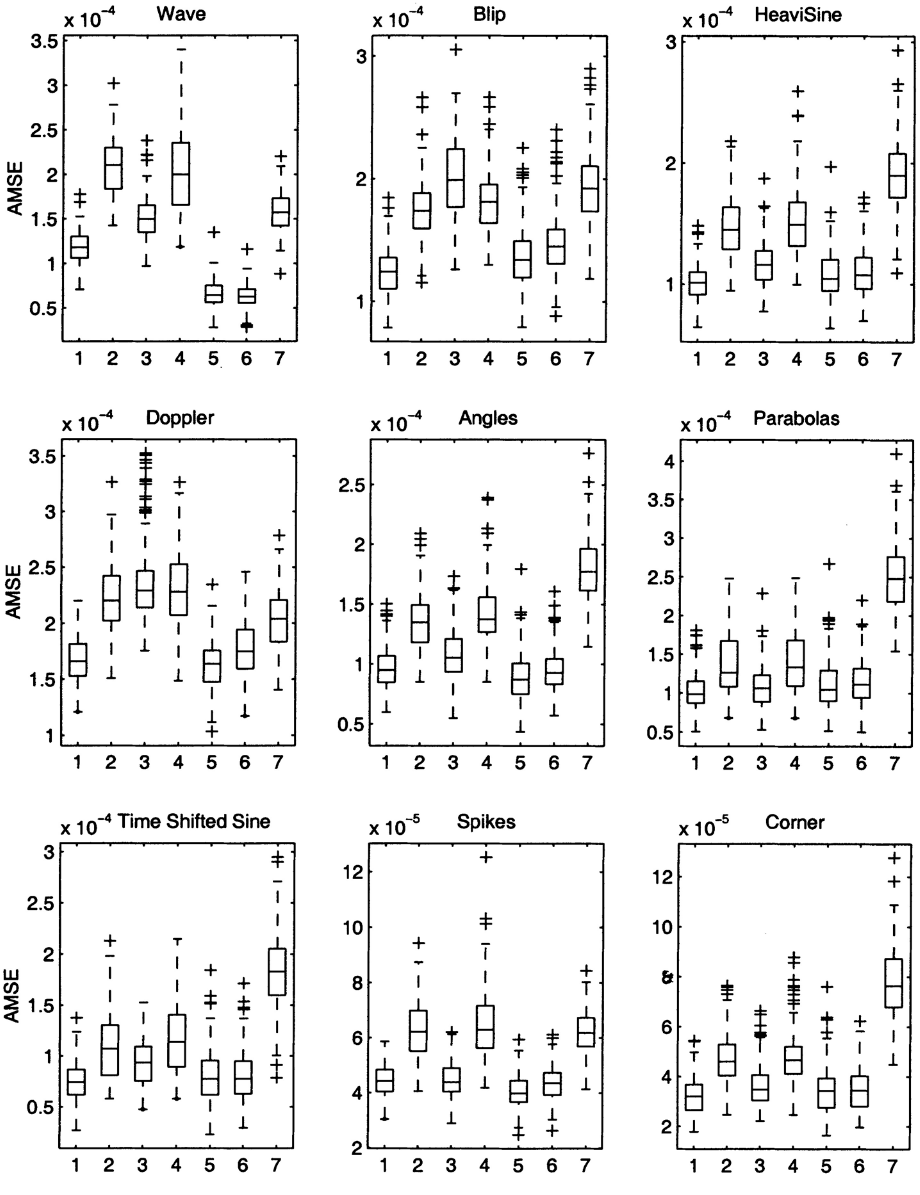

FIGURE 6 Boxplots of the AMSE for the various methods (1) ML-IIThresh ($\gamma = 1.8$), (2) VisuShrink with hard thresholding, (3) hybrid version of SureShrink, (4) 'leave-out-half' version of CrossValidation with hard thresholding, (5) PostMean, (6) PostMedian, and (7) ABE, computed for the nine test functions based on $n = 1024$ design points and RSNR $= 5$ given in Figure 3.

note that the large computational cost of the PostMean and PostMedian methods is mainly due to the EM-algorithm (used for estimating their free prior parameters) and its average value is accomplished with a large standard deviation. The relatively high cost of the ML-IIThresh method that can be observed in some cases is mainly due to the large number of iterations needed to compute l^*_{jk} when $|\hat{d}_{jk}| \downarrow \sigma$, as expected observing Figure 1(a), while for $|\hat{d}_{jk}| > 1.2\sigma$ the convergence is reached within 3–4 iterations at most. Figure 7 shows the barplots of the CPU time computed for the nine test functions based on $n = 1024$ design points and RSNR $= 5$ given in Figure 3.

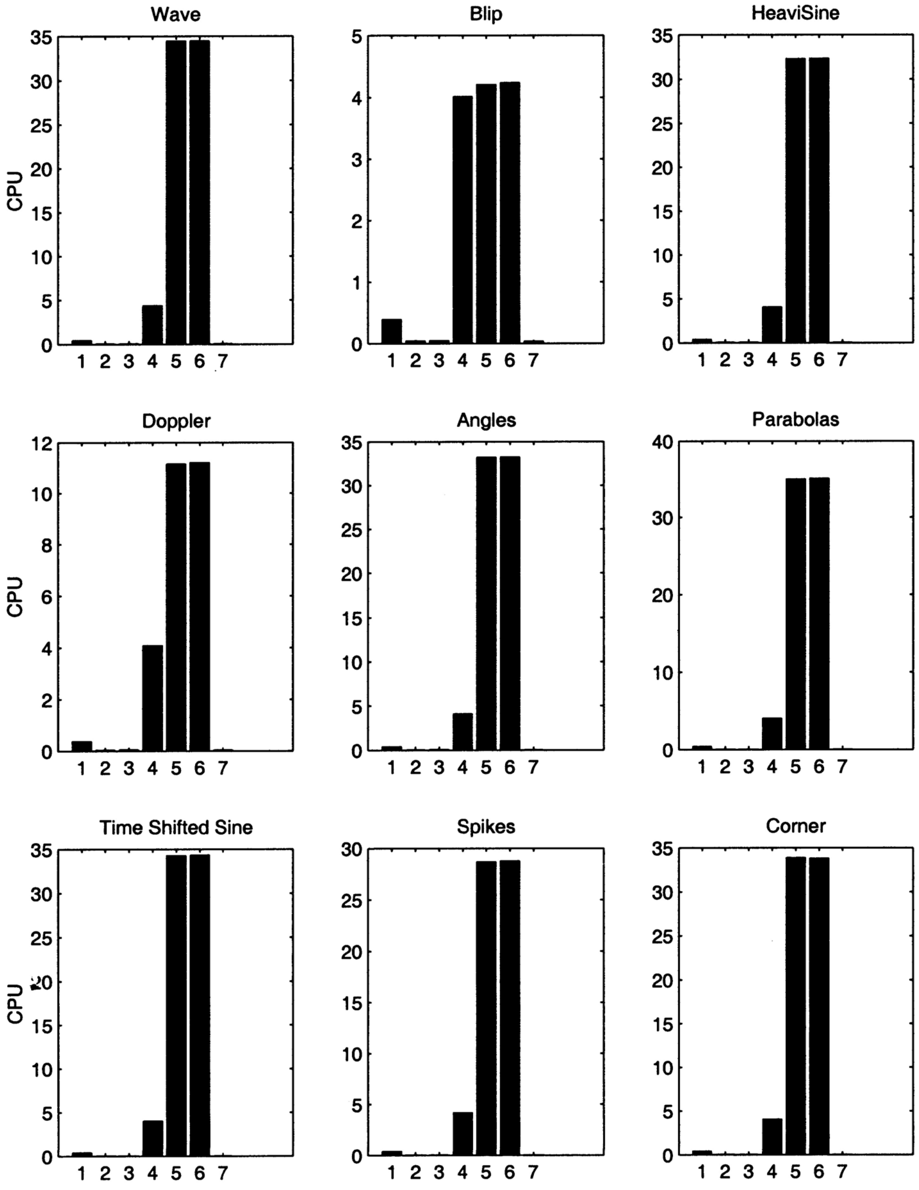

FIGURE 7 Barplots of the CPU time for the various methods (1) ML-IIThresh ($\gamma = 1.8$), (2) VisuShrink with hard thresholding, (3) hybrid version of SureShrink, (4) 'leave-out-half' version of CrossValidation with hard thresholding, (5) PostMean, (6) PostMedian, and (7) ABE, computed for the nine test functions based on $n = 1024$ design points and RSNR $= 5$ given in Figure 3.

Pointwise Bayesian credible intervals have been computed according to the simulation-based procedure described in Section 2.5 for the standard nominal coverage probabilities 0.90, 0.95 and 0.99. The empirical coverage rates and interval widths have been analyzed for each test function, sample size and RSNR. The corresponding pointwise Bayesian credible intervals have been generated by resampling $S = 200$ runs from the posterior distribution, and $M = 200$ samples have been considered in order to compute the empirical coverage rates and the interval widths. The resulting pointwise Bayesian credible intervals computed using the default value

$\gamma = 1.8$ present noisy spikes that are recognized to be typical of wavelet regression methods (see Barber, 2001; Barber *et al.*, 2002). However, although the maximum interval widths are nonnegligible due to the presence of the spikes, the average interval widths are quite tight and the pointwise Bayesian credible intervals have good average empirical coverage rates. Obviously, the empirical coverage rates varies greatly across each test function and, unsurprisingly, it becomes much better where the test function is smoother and less variable. Moreover, the performance of the proposed pointwise Bayesian credible intervals improves as the nominal coverage probability increases. As pointed out by Barber *et al.* (2002), this

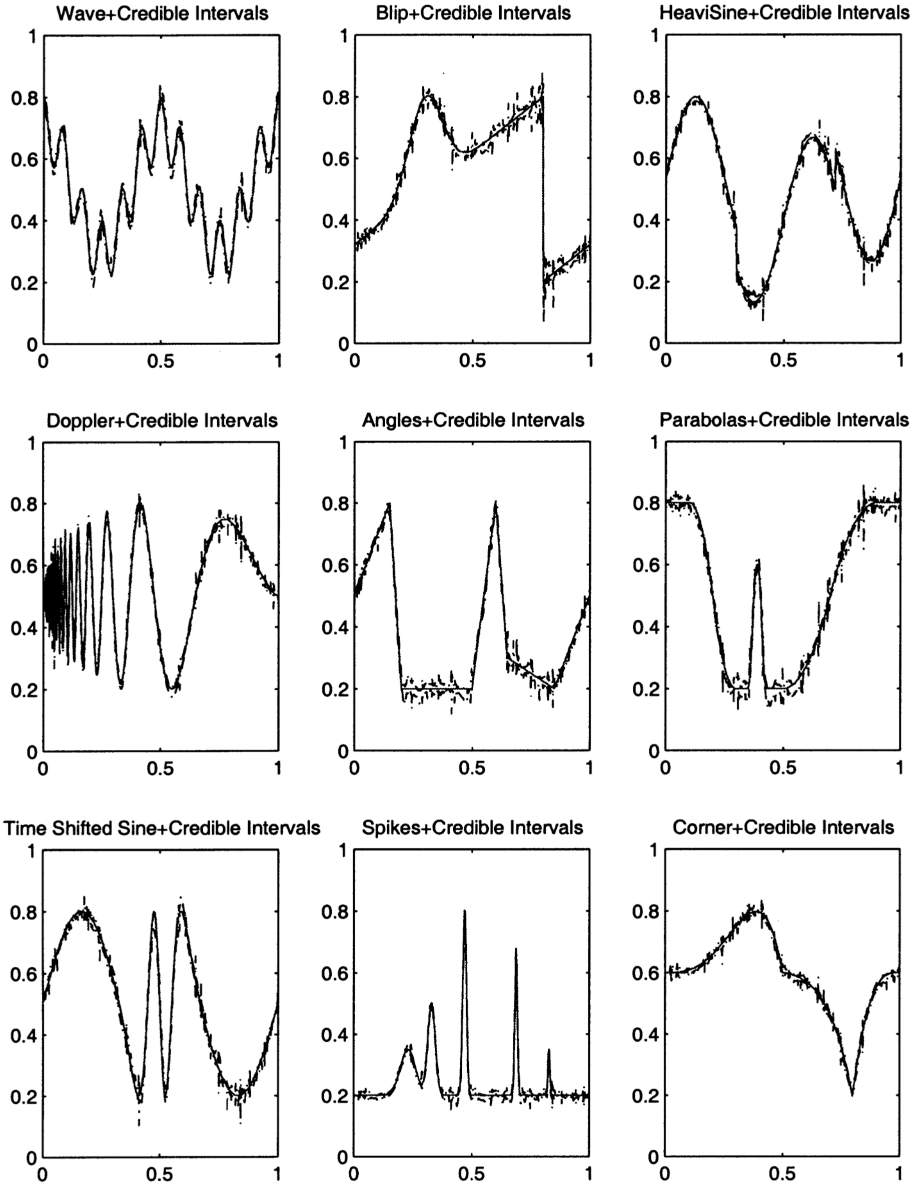

FIGURE 8 The 95% pointwise Bayesian credible intervals computed for the nine test functions based on $n = 1024$ design points and RSNR = 5, using the simulation-based procedure described in Section 2.5 for the ML-II Thresh method with $\gamma = 1.8$.

can be, in part, attributed to the kurtosis of the posterior distribution. As the nominal coverage increases, the limits of the pointwise Bayesian credible intervals move out into the tails of the posterior distributions. Therefore, with heavy tails distributions, like the posterior distribution given in Eq. (16), small increases in the nominal coverage rates can produce substantially wider pointwise Bayesian credible intervals. The brief spikes which occur in the interval bands can be smoothed out by increasing the parameter γ. However, this risks oversmoothing the data and it is usually payed by a decrease in the empirical coverage rate. For brevity, we only present some of the numerical findings. Figure 8 shows the 95% pointwise Bayesian credible intervals computed for the nine test functions based on $n = 1024$ design points and RSNR $= 5$ for the choice $\gamma = 1.8$. Figure 9 shows the boxplots of the empirical coverage rates and interval widths for the nominal 90%, 95% and 99% pointwise Bayesian credible intervals for the nine test functions based on $n = 1024$ design points and RSNR $= 5$ for the choice $\gamma = 1.8$.

4.2 Atomic Force Microscopy

The atomic force microscopy is a type of scanned proximity probe microscopy that can measure the adhesion strength between two materials at the nanonewton scale (see Binning *et al.*, 1986). In atomic force microscopy, a cantalevar beam is adjusted until it bonds with the surface of a sample, and then the force required to separate the beam and the sample is measured from beam deflection. Beam vibration can be caused by various factors, such as thermal energy of the surrounding air or the footsteps of someone outside the laboratory, and it acts as noise on the deflection signal. The atomic force microscopy signal sampled at $n = 2048$ data points from

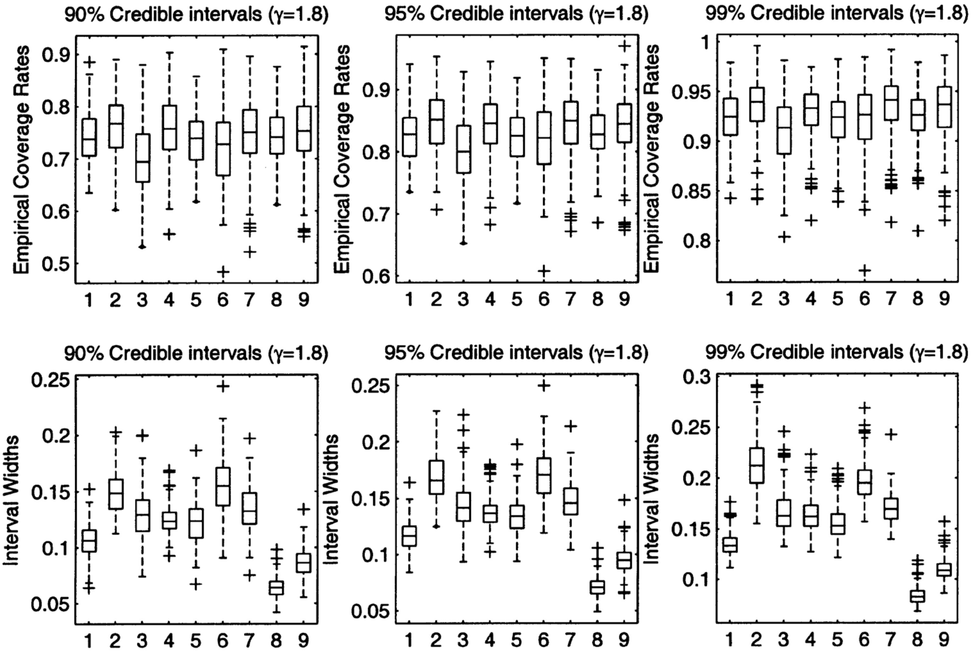

FIGURE 9 Boxplots of the empirical coverage rates (top panel) and interval widths (bottom panel) computed for the nominal (left panel) 90%, (middle panel) 95% and (right panel) 99% pointwise Bayesian credible intervals using the simulation-based procedure described in Section 2.5 for the ML-II Thresh method with $\gamma = 1.8$, for each of the nine test functions (1) Wave, (2) Blip, (3) HeaviSine, (4) Doppler, (5) Angles, (6) Parabolas, (7) Time Shifted Sine, (8) Spikes and (9) Corner, based on $n = 1024$ design points and RSNR $= 5$.

the adhesion measurements between carbohydrate and the cell adhesion molecule E-selectin is plotted in the top-right panel of Figure 10. The technical descriptions of this data set are provided in Marshall *et al.* (2001). In order for the data to be useful for subsequent analysis, the noise that arose from beam vibration must be removed.

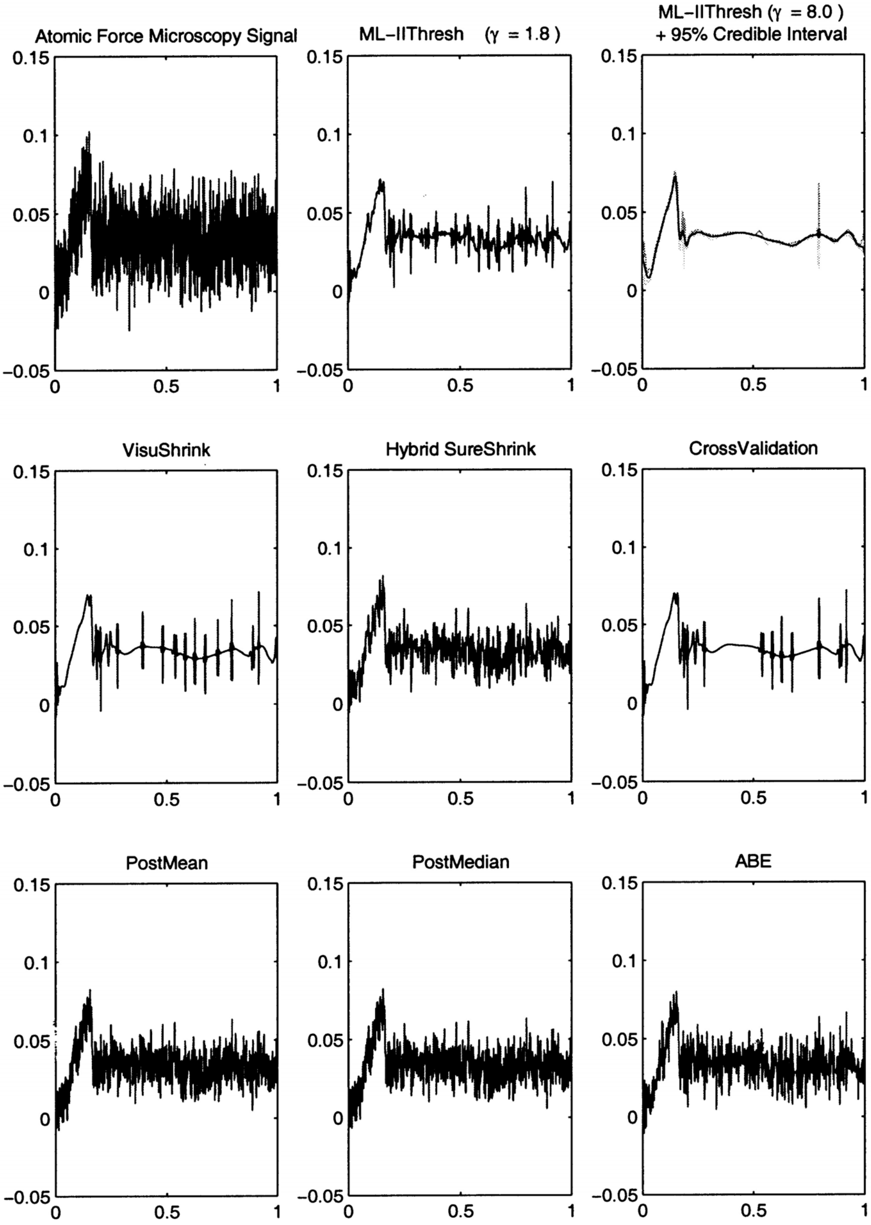

FIGURE 10 (Top panel) The atomic force microscopy signal sampled at $n = 2048$ data points (left); Estimate based on the ML-IIThresh method with $\gamma = 1.8$ (middle); Estimate based on the ML-II Thresh method with $\gamma = 8$ and corresponding 95% pointwise Bayesian credible intervals, using the simulation-based procedure described in Section 2.5 (right). (Middle panel) Estimate based on the VisuShrink method with hard thresholding (left); Estimate based on the Hybrid version of the SureShrink method (middle); Estimate based on the 'leave-out-half' version of the CrossValidation method with hard thresholding (right). (Bottom panel) Estimate based on the PostMean method (left); Estimate based the PostMedian method (middle); Estimate based on the ABE method (right).

The ML-IIThresh method with Daubechies's nearly symmetric wavelets of order 8, *Symmlet 8*, and the default value $\gamma = 1.8$ have been applied to the atomic force microscopy signal shown in the top-right panel of Figure 10, and the resulting estimate is shown in the top-middle panel of Figure 10. We can observe that the ML-IIThresh method with $\gamma = 1.8$ still exhibits a noisy shape; however, as Figure 10 shows, analogous estimates (and in most cases even noisier) have been obtained for the various classical and empirical Bayes term-by-term wavelet schemes used in the simulation study presented in Section 4.1. Guided from the consideration that high noise levels, as well as large sample sizes, would require a larger value of γ, the ML-IIThresh method has been applied to the atomic force microscopy data with the choice $\gamma = 8$. The resulting estimate, shown in the top-right panel of Figure 10, exhibits a smooth behavior, especially in the long-middle term without oversmoothing the bump at the beginning of the signal. Superimposed are shown the corresponding 95% pointwise Bayesian credible intervals, using the simulation-based procedure described in Section 2.5. Note that the largest widths of the resulting pointwise Bayesian credible intervals that appear at the boundaries of the atomic force microscopy signal are due to the fact that we have used periodic wavelets in our analysis.

5 CONCLUDING REMARKS

We have proposed an empirical Bayes approach to standard nonparametric regression estimation using a nonlinear wavelet methodology. This approach results in a thresholding procedure which depends on one free prior parameter and it is level- and amplitude-dependent, thus allowing better adaptation in function estimation. We have considered an automatic choice of the free prior parameter, guided by considerations on an exact risk analysis and on the shape of the thresholding rule, enabling the resulting estimator to be fully automated in practice. Pointwise Bayesian credible intervals for the resulting function estimate have also been considered using a simulation-based approach. It has been demonstrated that the proposed empirical Bayes term-by-term wavelet scheme outperforms standard classical term-by-term wavelet thresholding schemes and performs nearly as well as (sometimes even better than) much more computationally expensive empirical Bayes term-by-term wavelet shrinkage and wavelet thresholding schemes in finite sample situations. A simulation study also shows that the proposed pointwise Bayesian credible intervals for the resulting function estimates have good empirical coverage rates for standard nominal coverage probabilities. The proposed empirical Bayes term-by-term wavelet scheme and the resulting simulation-based procedure to computing pointwise Bayesian credible intervals could have been presented as a possibly useful addition to the growing range of nonlinear wavelet-based function estimation tools.

Acknowledgements

The work of Claudia Angelini was supported by the Agenzia Spaziale Italiana (Italian Space Agency). The work of Theofanis Sapatinas was partly supported by a 'University of Cyprus – 2002 Research Grant'. Helpful suggestions made by an anonymous reviewer are gratefully acknowledged.

References

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

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

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

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

Barber, S. (2001). Simulating from the posterior density of Bayesian wavelet regression estimates, *Tech. Rep. 01:29*, Department of Mathematics, University of Bristol, United Kingdom.

Barber, S., Nason, G. P. and Silverman, B. W. (2002). Posterior probability intervals for wavelet thresholding. *J. R. Statist. Soc.*, *Series B*, **64**, 189–206.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York.

Berger, J. O. and Berliner, L. M. (1986). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Ann. Statist.*, **14**, 461–486.

Berger, J. O. and Sellke, T. (1987). Testing of a point null hypothesis: The irreconcilability of significance levels and evidence (with discussion). *J. Am. Statist. Assoc.*, **82**, 112–139.

Binning, G. Quate, C. F. and Gerber, Ch. (1986). Atomic force microscope. *Phys. Rev. Lett.*, **56**, 930–933.

Bruce, A. G. and Gao, H. -Y. (1996). Understanding waveshrink: Variance and bias estimation. *Biometrika*, **83**, 727–745.

Buckheit, J. B., Chen, S., Donoho, D. L., Johnstone, I. M. and Scargle, J. (1995). About WaveLab, *Technical Report*, Department of Statistics, Stanford University, USA.

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

Clyde, M. and George, E. I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In: Müller, P. and Vidakovic, B. (Eds.), *Bayesian Inference in Wavelet Based Models, Lect. Notes Statist.*, Vol. 141, Springer-Verlag, New York, pp. 309–322.

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

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.*, *Series B*, **39**, 1–38.

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

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

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *J. R. Statist. Soc.*, *Series B*, **57**, 301–337.

Figueidero, M. A. T. and Nowak, R. D. (2001). Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior. *IEEE Trans. Image Process.*, **10**, 1322–1331.

Gao, H.-Y. and Bruce, A. G. (1997). WaveShrink with firm shrinkage. *Statist. Sinica*, **7**, 855–874.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*, Lecture Notes in Statistics 129. Springer-Verlag, New York.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pat. Anal. Mach. Intel.*, **11**, 674–693.

Marshall, B., McEver, R. and Zhu, C. (2001). Kinetic rates and their force dependence on the P-Selectin/PSGL-1 interaction measured by atomic force microscopy. In *Proceedings of AMSE 2001*, Bioengineering Conference, BED, Vol. 50.

Moullin, P. and Liu, J. (1999). Analysis of multiresolution image denoising schemes using generalised Gaussian and complexity priors. *IEEE Trans. Inform. Theory*, *45*, 909–919.

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

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

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

Vidakovic, B. and Ruggeri, F. (2001). BAMS method: Theory and simulations. *Sankhyā*, *Series B*, **63**, 234–249.