Smoothing Noisy Data with Tapered Coiflets Series.

Anestis Antoniadis†

ABSTRACT

This paper is concerned with an orthogonal wavelet series estimator of an unknown smooth regression function observed with noise on a bounded interval. A penalized least squares approach is adopted and our method uses the specific asymptotic interpolating properties of the wavelet approximation generated by a particular wavelet basis, Daubechie's coiflets. A simple procedure is described to estimate the smoothing parameter of the penalizing functional and conditions are given for the estimator to attain optimal convergence rates in the integrated mean square sense as the sample size increases to infinity. The results are illustrated with simulated and real examples and a comparison with other nonparametric smoothers is made.

Keywords and phrases: Nonparametric regression, curve smoothing, wavelets, multiresolution analysis, splines, penalized least-squares.

1. Introduction and summary.

Consider the nonparametric regression problem (x_i, Y_i), $i = 0, \ldots, m-1$, $Y_i = g(x_i)+\epsilon_i$, where the ϵ_i's are uncorrelated, with zero mean and constant variance σ^2. For simplicity we consider the case of a fixed design where the design points $x_0, x_1, \ldots, x_{m-1}$ all belong to the interval $[0, 1]$ and are assumed to be equally spaced. The x values should be doubly indexed, because x_i depends on m, but this dependence will not be carried in the notation. Our goal is to estimate the regression function $g(x)$ and we will judge the quality of an estimator $\hat{g}(x)$ by its integrated mean squared error (MISE):

$$MISE\,(\hat{g}) = \mathbb{E} \int_0^1 \left[\hat{g}\,(x) - g\,(x)\right]^2 \, dx. \tag{1.1}$$

Smoothing splines are one popular method for such a nonparametric regression problem. Let ν be a positive integer. An νth order smoothing spline $g_\lambda(x)$, is defined to be

† Research granted by CNRS–INRIA, Program 6 (IDOPT project).

that function with square integrable νth derivative which minimizes over the Sobolev space $H_2^\nu = H_2^\nu[0,1]$ the "discrete" functional:

$$\frac{1}{m}\sum_{i=0}^{m-1}\left(y_i - f\left(x_i\right)\right)^2 + \lambda \int_0^1 \left(f^{(\nu)}\left(x\right)\right)^2 dx,$$

where $f^{(\nu)}$ indicates the νth derivative of f. The asymptotic properties of smoothing splines have been studied by a number of authors, including Wahba (1975, 1990), Rice and Rosenblatt (1981), Rice (1984), Speckman (1985), Nussbaum (1985) and Cox (1988), to cite only a few.

The continuous analogue of spline smoothing is called Tikhonov regularization. The estimates of g to be considered in this paper are obtained by such a method of regularization and are based on wavelet decompositions on the interval. Our purpose is to show that wavelets can be used within the framework of regularization methods, thus augmenting rather than supplanting the many popular univariate curving methods.

In regularization, a target function g is approximated by a smoother function f. To be precise, we must define what we mean by "approximate" and "smoother". To this end, as in Devore and Lucier (1992), we introduce the error function space $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{X}}$ and the "smoothness" function space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$. The smoothness of a function f is measured by $\|f\|_{\mathcal{H}}$ and the size of the error between g and f is given by $\|g-f\|_{\mathcal{X}}$ (it is usually assumed that the space $\mathcal{H}$ is continuously embedded in $\mathcal{X}$). Under such a setting a Tikhonov regularizer of g is g_λ, a minimizer of

$$\inf_{f\in\mathcal{H}} \left\{\|g-f\|_{\mathcal{X}} + \lambda\|f\|_{\mathcal{H}}\right\}. \tag{1.2}$$

The parameter λ determines the relative importance of the smoothness of f and the approximation error of g.

In the following we shall assume that $\mathcal{X}$ is the space $L^2([0,1])$, the space of square integrable functions. A family of spaces useful for measuring the smoothness of f is the family of Sobolev spaces $H^\delta\left([0,1]\right)$, $\delta > 1/2$, which consist of all functions with δ-order "derivatives" in $L^2([0,1])$. When δ is an integer, one can define

$$\|f\|^2_{H^\delta([0,1])} = \sum_{0\le k\le\delta} \|D^k f\|^2_{L^2([0,1])}.$$

When δ is not an integer, the fractional Sobolev spaces $H^\delta\left([0,1]\right)$ can be defined in terms of the Besov spaces $B^\delta_{2,2}([0,1])$. This choice for the pair $(X,\mathcal{H})$ and the use of wavelet decompositions of f and g, allows us to find nearly optimal solutions to the variational problem given in (1.2). This is possible because the norms $\|f\|_{L^2([0,1])}$ and $\|f\|^2_{H^\delta([0,1])}$ can be determined simultaneously by examining the wavelet coefficients of f. A specific family of wavelets, the coiflets, will allow us to relate and approximate the discrete problem by a continuous one and to obtain an asymptotically valid estimate of the unknown regression function g.

The paper is organized as follows. Section 2 reviews some background on wavelet theory on the interval. The wavelet estimator for our nonparametric regression problem is introduced in section 3, and conditions are given for the estimator to attain optimal convergence rates in the integrated mean square sense as the sample size increases to infinity. This section also discusses the choice of the regularization parameter λ. Section 4 concerns the numerical implementation of the estimator and contains some simulations and a comparison of the estimator with adaptative kernel and cubic spline smoothers.

2. Orthogonal wavelets on $[0, 1]$.

A general introduction to the theory of wavelets can be found in Daubechies (1988). An introduction that considers the statistical uses of wavelets can be found in Nason and Silverman (1994).

In this section we sketch an account of the relevant properties of orthogonal wavelets and multiresolution analysis on closed subsets of the real line, that will be used to derive our estimator.

The construction of orthonormal wavelet bases for $L^2(\mathbb{R})$ is now well understood. There are many families of wavelets. For the construction of orthonormal bases of compactly supported wavelets for $L^2(\mathbb{R})$, in particular, one starts with a couple of special, compactly supported functions known as the scaling function ϕ and the wavelet ψ. The functions $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, $j, k \in \mathbb{Z}$, then constitute an orthonormal basis for $L^2(\mathbb{R})$. For fixed $j \in \mathbb{Z}$, the $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$, $k \in \mathbb{Z}$ are an orthonormal basis for a subspace $V_j \subset L^2(\mathbb{R})$; the spaces V_j constitute a multiresolution analysis, meaning in particular that

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots$$

with

$$\cap_j V_j = \{0\}, \qquad \overline{\cup_j V_j} = L^2(\mathbb{R}),$$

and

$$P_{j+1}f = P_j f + \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k},$$

where P_j denotes the orthogonal projector onto the closed subspace V_j of $L^2(\mathbb{R})$.
The multiresolution analysis is said to be r-regular ($r \geq 0$) if ϕ is an element of the Hölder space C^r, and if both ϕ and its derivatives have a fast decay,

$$|\partial^\alpha \phi\,(x)\,| \leq C_m\,(1+|x|)^{-m}\,, \quad \forall m \in \mathbb{N}, \quad 0 \leq \alpha \leq r.$$

One can prove that if a multiresolution analysis is r-regular, the wavelet ψ is also C^r and has vanishing moments up to the order r (see e.g. Daubechies (1992), Corollary 5.2)

$$\int_{-\infty}^{+\infty} x^k \psi\,(x)\,dx = 0 \quad \text{for} \quad 0 \leq k \leq r. \tag{2.1}$$

The converse is generally false, and the number of vanishing moments is usually larger than the regularity of the multiresolution analysis. The smoother wavelets provide not only orthonormal bases for $L^2(\mathbb{R})$, but also unconditional bases for function spaces consisting of more regular functions.

An advantage of having a high number of vanishing moments for ψ is that the fine scale wavelet coefficients of a function are essentially zero where the function is smooth. Since $\int \phi(x)dx = 1$, the same thing can never happen for the $< f, \phi_{j,k} >$, but it is possible to construct compactly supported orthonormal wavelets such that the scaling function ϕ has $N-1$ vanishing moments, i.e.

$$\begin{aligned} \int_{\mathbb{R}} \phi\,(x)\;dx &= 1\,, \\ \int_{\mathbb{R}} x^k \phi\,(x)\;dx &= 0 \quad 1 \leq k \leq N-1\,, \\ \int_{\mathbb{R}} x^k \psi\,(x)\;dx &= 0 \quad 0 \leq k \leq N-1\,. \end{aligned} \tag{2.2}$$

Such wavelets were constructed by Daubechies (1990) and were named *coiflets* after Ronald Coifman who asked for their construction.

All the above concern bases for $L^2(\mathbb{R})$. Let us consider now orthogonal wavelets on the interval $[0,1]$. The starting point of this construction is any one of the compactly supported bases in Daubechies (1988), with N-vanishing moments, regularity r and support included in $[-2K, 2K+1]$.

To summarize, for parameters $N > 0$ and j_0 with $2^{j_0} \geq 2N$, the construction furnishes a finite set of 2^{j_0} scaling functions $\phi_{j_0,k}$, and for each $j \geq j_0$, 2^j functions $\psi_{j,k}$, such that the collection of these functions forms a complete orthonormal system of $L^2([0,1])$. With this notation, the $L^2([0,1])$ reconstruction formula is

$$f = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k} \phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \beta_{j,k} \psi_{j,k}$$

All the polynomials on $[0,1]$ of degree $\leq N-1$ can be written as linear combinations of the scaling functions at any fixed scale. It follows that all the corresponding wavelets, at the edges as well as in the interior, have N vanishing moments. Therefore by selecting the parameter N large and using the most regular wavelets, one gets unconditional bases for various functional spaces.

Finally, we state a generic theorem that characterizes the Besov spaces $B^{\delta}_{p,p}([0,1])$ in terms of wavelet expansions in a bounded closed interval. Heuristically, these are spaces of functions that have δ bounded "derivatives" in $L^p([0,1])$. There are several equivalent definitions of Besov spaces. One of them is as follows. Let $1 \leq p \leq \infty$, $1 \leq q \leq \infty$. For $0 < s < 1$ let us denote

$$\gamma^s_{p,q}(f) = \left(\int_{\mathbb{R}} \left(\frac{\|\tau_h f - f\|_p}{|h|^s} \right)^q \frac{dh}{|h|} \right)^{\frac{1}{q}},$$

where $\tau_h f(x) = f(x-h)$, and for $s=1$ let us denote

$$\gamma^1_{p,q}(f) = \left(\int_{\mathbb{R}} \left(\frac{\|\tau_h f + \tau_{-h}(f) - 2f\|_p}{|h|} \right)^q \frac{dh}{|h|} \right)^{\frac{1}{q}}$$

with the appropriate modification for $q = \infty$. With these notations, a complete definition of Besov spaces, in terms of the modulus of continuity, is the following one:

Definition 2.1. For $0 < s \leq 1$, $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, the Besov space $B^s_{p,q}$ is the normed space

$$B^s_{p,q} = \{ f \in L_p |\, \gamma^s_{p,q}(f) < \infty \},$$

equipped with the norm

$$\|f\|_{B^s_{p,q}} = \|f\|_{L_p} + \gamma^s_{p,q}(f).$$

For $s > 1$, writing $s = n + s_0$ with $n \in \mathbb{N}$ and $0 < s_0 \leq 1$, $B^s_{p,q}$ is defined by

$$B^s_{p,q} = \{ f | f^{(m)} \in B^{s_0}_{p,q},\ \forall m \leq n \}$$

with the norm

$$\|f\|_{B^s_{p,q}} = \|f\|_p + \sum_{m \leq n} \gamma^{s_0}_{p,q}\left(f^{(m)}\right).$$

The following characterization is especially convenient for the purposes of this paper. Its proof can be found in Devore and Lucier (1992).

Theorem 2.1. *If the wavelet ψ corresponding to a multiresolution analysis of $L^2([0,1]$ is r-regular, then a function f in $L^p([0,1])$, $1 < p \leq \infty$, $\max\{(1/p-1), 0\} < \delta < r$, is in the Besov space $B^{\delta}_{p,p}([0,1])$ if and only if it has the wavelet decomposition*

$$f = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}$$

and

$$\|f\|^p_{B^\delta_{p,p}([0,1])} \simeq \| \left(\alpha_{j_0,k}\right)_k \|^p_{\ell_p} + \sum_{j\geq j_0} 2^{jp\delta} \sum_{k=0}^{2^j-1} |\beta_{j,k}|^p,$$

with constants of equivalency that do not depend on f, but which may depend on p, j_0 and the wavelet ψ.

To end this preliminary section let us mention some approximation properties of regular wavelets that will be used in the sequel of this paper (for a detailed account see Meyer (1990), (1991)).

If f belongs to the Sobolev space $H^\delta([0,1])$ and the multiresolution analysis is r-regular, then

$$\|f - P_j f\|_{L^2} \leq o\left(2^{-j\min(\delta,r)}\right) \quad \text{as } j \to \infty\,. \tag{2.3}$$

If f belongs to some Hölder space $C^\delta([0,1])$ then

$$|< f, \psi_{j,k} >| \leq C 2^{-j(\min(\delta,r)+1/2)} \quad \text{for all} \quad j \geq j_0,\ k = 0, \ldots, 2^j - 1, \tag{2.4}$$

where C is a constant independent of j and k.

3. Nonparametric regression.

The fixed design regression model considered in the introduction was given by

$$Y_i = g\left(x_i\right) + \epsilon_i, \quad 0 \leq i \leq m - 1, \tag{3.1}$$

where the Y_i are noisy measurements of the "mean" or regression function $g(x)$ taken at equidistant nonrandom design points x_i within $[0,1]$ and the ϵ_i's are independent identically distributed random variables with mean 0 and finite variance σ^2. The function g is viewed as a member of a class $\mathcal{F}$ of possible regressions over $[0,1]$. Hereafter, we shall assume that $\mathcal{F}$ is a (possibly fractional) Hölder space of exponent δ, $\delta > 0$. Considering the structure of the wavelet multiresolution, the design points x_i will be of the form $x_i = i\Delta$, with $\Delta = 2^{-n}$ for $i = 0, \ldots, m-1$, where $m = 2^n$. The asymptotic results will be obtained as n goes to infinity.

We consider the simple model (3.1) only for convenience. The restriction of working with equidistant fixed design points x_i can be dropped, and one can work with the situation where the design points x_i are either nonrandom and unequally spaced or random and satisfying some regularity conditions. Recent work on this aspect is provided in Antoniadis, Grégoire and McKeague (1994). See also Antoniadis and Grégoire (1994). As a referee has pointed out, of interest would also be to consider an error term of the form $\sigma^2(x_i)\epsilon_i$, with $\mathrm{var}(\epsilon_i) = 1$ allowing for heteroscedasticity. Antoniadis and Lavergne (1994) deal with this situation and estimate the variance function by wavelet techniques. However, the context is different and a proper analysis of this problem is beyond the scope of this paper. We also assume that the sample size is of power of two, i.e. $m = 2^n$. Here again this restriction is assumed for convenience of presentation and can be relaxed. See for example McGill and Taswell (1993) and Herley (1995).

From now on, ϕ and ψ will denote the scaling function and the wavelet on $[0,1]$ associated to an r-regular multiresolution analysis of $L^2(\mathbb{R})$. We will assume further that the scaling function ϕ is a coiflet of order $L = 2K$ with $L > [\delta] + 1$.

Since

$$L^2\left([0,1]\right) = V_{j_0} \oplus \left(\oplus_{j \geq j_0} W_j\right),$$

the function g admits the following expansion in H^δ (and in L^2):

$$g\left(t\right) = \sum_{k=0}^{2^{j_0}-1} < g, \phi_{j_0,k} > \ \phi_{j_0,k}\left(t\right) + \sum_{j \geq j_0} \sum_{\ell=0}^{2^j-1} < g, \psi_{j,\ell} > \ \psi_{j,\ell}\left(t\right),$$

with

$$< g, \phi_{j_0,k} >= \int_0^1 g\left(t\right) \phi_{j_0,k}\left(t\right)\, dt \quad \text{and} \quad < g, \psi_{j,\ell} >= \int_0^1 g\left(t\right) \psi_{j,\ell}\left(t\right)\, dt.$$

One advantage of the nested structure of a multiresolution analysis is that it leads to an efficient tree-structured algorithm for the decomposition of functions in V_ℓ for which the fine-scale theoretical wavelet coefficients $< g, \phi_{\ell,k} >$ are given. However, when a function is given in sampled form, one does not have access to the fine scale integrals $< g, \phi_{\ell,k} >$.

In Antoniadis (1994) we described a way, given a smooth function g, to determine, with a certain accuracy and in a computational favorable way, the coefficients $< g, \phi_{n,k} >$ of g in the space V_n which are needed to start the fast wavelet transform.

A first step towards this is to try to approximate the projection $P_n g$ of g on V_n by $\Pi_n g = 2^{-n/2} \sum_{k=0}^{2^n-1} g(\frac{k}{2^n})\phi_{n,k}(t)$ to derive then a reasonable estimator of the approximation $\Pi_n g$. Since the coiflets have $L = 2K$ vanishing moments, one can define such an estimator by

$$\hat{g}_n\left(t\right) = 2^{-n/2} \sum_{k=0}^{2^n-1} Y_k \phi_{n,k}\left(t\right). \tag{3.2}$$

This choice can be justified by the following lemma, whose proof may be found in Antoniadis (1994).

Lemma 3.1. *Let $g^{\{n\}}(k) = 2^{n/2} < g, \phi_{n,k} >$. With $L > [\delta] + 1$, the following uniform (in k; $0 \le k \le 2^n$) bound holds:*

$$|g^{\{n\}}(k) - g\left(\tfrac{k}{2^n}\right)| \le C_1 2^{-n\delta}, \tag{3.3}$$

where C_1 is a constant only depending on the coiflet ϕ and its support length.

By Lemma 3.1 one is therefore able to approximate the coefficients $< g, \phi_{n,k} >$, $0 \le k \le 2^n - 1$, with an error $\mathcal{O}(2^{-n/2}2^{-n\delta})$. Using such an approximation, and Lemma 3.1 we have:

$$\begin{aligned}\sup_{t\in[0,1]} \{2^{-n/2} \sum_{k=0}^{2^n-1} |g^{\{n\}}(k) - g\left(\tfrac{k}{2^n}\right)| \, |\phi_{n,k}(t)|\} \qquad\qquad \qquad\qquad (3.4)\\ \le \mathcal{O}\left(2^{-n\delta}\right) \sup_{t\in[0,1]} \{\sum_{k=0}^{2^n-1} |\phi(t-k)|\} \le \mathcal{O}\left(2^{-n\delta}\right).\end{aligned}$$

Observing that $\mathbb{E}(Y_k) = g(\frac{k}{2^n})$ justifies completely our choice of the estimator $\hat{g}_n$. Other sampling procedures, such as quasi-interpolation and quadrature formulae were proposed by Shensa (1992).

The above calculations suggest that given our observed sampled values y_i of Y_i, the observations are equivalent to the "raw" estimator $\hat{g}_n$, which lies now in $H^\delta([0,1])$. This estimator, while presenting a very small bias on $[0,1]$, leads to a highly oscillatory solution being to fidel to the data. The orthogonal discrete wavelet transform on the interval, say W, applied on $\hat{g}_n$, identified by the vector $\mathbf{y}$ of observed values, gives a new sequence of numbers $\{\left(c_{j_0,k}\right)_{k=0,\ldots,2^{j_0}-1}, \left(d_{j,k}\right)_{j=j_0,\ldots,n;k=0,\ldots,2^{j_0}-1}\}$, say $\left((c_{j,k}), (d_{j,k})\right)$ for short, that are interpreted as the coefficients of the expansion of the function $\hat{g}_n$. Under the noise model underlying (3.1), noise contaminates all wavelet coefficients of $\Pi_n g$ equally. Consequently, the empirical wavelet coefficients $\left((c_{j,k}), (d_{j,k})\right)$, obtained by applying the discrete orthogonal transform on the data vector $2^{-n/2}\mathbf{y}$, can be considered as a version of the coefficients $\left((\alpha_{j,k}), (\beta_{j,k})\right)$ of $\Pi_n g$, contaminated by a similar white noise $\epsilon_{j,k}$ of variance $\sigma^2/2^n$. This is the heuristic of our method.

Following Devore and Lucier (1992) we can now express the continuous regularization problem (1.2) in terms of wavelet coefficients. For any $g \in H^\delta$, $\Pi_n g$ can be expanded as

$$\Pi_n g = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{n} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}.$$

Furthermore, according to our remarks above, we have

$$\hat{g}_n = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{n} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}.$$

Replacing the regularization functional used in expression (1.2) by

$$\inf_{f\in\mathcal{H}}\left\{\|g-\Pi_n f\|_{\mathcal{X}}+\lambda\|\Pi_n f\|_{\mathcal{H}}\right\}. \tag{3.5}$$

and using the equivalent sequence norms and these expressions, minimizing the penalized functional given above is equivalent to minimize the expression

$$\sum_{k=0}^{2^{j_0}-1}\left[c_{j_0,k}-\alpha_{j_0,k}\right]^2+\sum_{j\geq j_0}^{n}\sum_{k=0}^{2^j-1}\left[\left(d_{j,k}-\beta_{j,k}\right)^2+\lambda 2^{2\delta j}\beta_{j,k}^2\right], \tag{3.6}$$

where $d_{j,k}=0$ for $j>n$ and $k=0,\ldots,2^j-1$. Clearly, one minimizes the entire expression by minimizing each term separately, and it is easy to see that the solution is the function $g_{n,\delta,\lambda}=g_\lambda$ with wavelet coefficients

$$\hat{\alpha}_{j_0,k}=c_{j_0,k},\quad k=0,\ldots,2^{j_0}-1$$

and

$$\hat{\beta}_{j,k}=\tfrac{d_{j,k}}{1+\lambda 2^{2\delta j}},\quad j\geq j_0, k=0,\ldots,2^j-1$$

that is

$$g_\lambda=\sum_{k=0}^{2^{j_0}-1}c_{j_0,k}\phi_{j_0,k}+\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\hat{\beta}_{j,k}\psi_{j,k},$$

which appears to be a tapered wavelet series estimator of the regression function, i.e. g_λ may be viewed as the result of passing the "raw" wavelet series estimate $\hat{g}_n$ through a low pass filter controlled by the parameters λ and δ.

Let $T_{n,\delta}(\lambda)$ be the integrated squared difference between g and the regularizer $g_{n,\delta,\lambda}$, denoted by g_λ for convenience. One has

$$\begin{aligned}
T_{n,\delta}(\lambda)&=\|g-g_\lambda\|^2_{L^2([0,1])}=\|g-P_ng+P_ng-g_\lambda\|^2_{L^2([0,1])}\\
&=\|(I-P_n)g\|^2_{L^2([0,1])}+\|P_ng-g_\lambda\|^2_{L^2([0,1])}\\
&\leq\|(I-P_n)g\|^2_{L^2([0,1])}+2\|P_ng-\Pi_ng\|^2_{L^2([0,1])}+2\|\Pi_ng-g_\lambda\|^2_{L^2([0,1])}
\end{aligned}$$

For the last term in the right hand side of this inequality we have

$$\begin{aligned}
\|\Pi_ng-g_\lambda\|^2_{L^2([0,1])}&=\sum_{k=0}^{2^{j_0}-1}\left[\hat{\alpha}_{j_0,k}-\alpha_{j_0,k}\right]^2+\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\left(\hat{\beta}_{j,k}-\beta_{j,k}\right)^2\\
&=\sum_{k=0}^{2^{j_0}-1}\left[\hat{\alpha}_{j_0,k}-\alpha_{j_0,k}\right]^2+\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\left[\left(\tfrac{\lambda_j}{\lambda_j+\lambda}-1\right)\beta_{j,k}+\tfrac{\lambda_j}{\lambda_j+\lambda}2^{-n/2}\epsilon_{j,k}\right]^2,
\end{aligned} \tag{3.7}$$

where $\lambda_j=2^{-2\delta j}$.

We can now state the first result concerning the convergence properties of our estimate.

Theorem 3.1. *Under the smoothness assumptions imposed on g and on the weight function u in this section, and for any $\delta > 1/4$, the MISE $R_{n,\delta}(\lambda) = \mathbb{E}\left(T_{n,\delta}(\lambda)\right)$ satisfies*

$$R_{n,\delta}(\lambda) \leq o\left(2^{-2n\min(\delta,r)}\right) + \mathcal{O}\left(2^{-2n\delta}\right) + \mathcal{O}(\lambda) + \mathcal{O}\left(2^{j_0-n}\right) + \mathcal{O}\left(2^{-n}\lambda^{-1/(2\delta)}\right).$$

Moreover, if $\delta > 1/2$ and if $\lambda = \mathcal{O}\left(m^{-2\delta/(2\delta+1)}\right)$ then

$$R_{n,\delta}(\lambda) = \mathcal{O}\left(m^{-2\delta/(2\delta+1)}\right).$$

Proof of Theorem 3.1. Since $R_{n,\delta}(\lambda) = \mathbb{E}\left(T_{n,\delta}(\lambda)\right)$, the decomposition of $T_{n,\delta}(\lambda)$ leads to the following decomposition for $R_{n,\delta}(\lambda)$:

$$\begin{aligned} R_{n,\delta}(\lambda) &= \mathbb{E}\left[\|g - g_\lambda\|^2_{L^2([0,1])}\right] \\ &\leq \|(I - P_n)g\|^2_{L^2([0,1])} + 2\|P_n g - \Pi_n g\|^2_{L^2([0,1])} + 2\mathbb{E}\left[\|\Pi_n g - g_\lambda\|^2_{L^2([0,1])}\right] \\ &= B_1 + 2B_2 + 2B_3 \end{aligned} \tag{3.8}$$

The first term B_1 represent the error of approximating g by its projection on V_n. Since $g \in H^\delta([0,1])$ and the multiresolution analysis used is r-regular, we can use the upper bound given by Eq. (2.3) to obtain

$$\|(I - P_n)g\|^2_{L^2([0,1])} = o\left(2^{-2n\min(\delta,r)}\right). \tag{3.9}$$

By using the inequality (3.4) the second term B_2 is bounded above by

$$\|P_n g - \Pi_n g\|^2_{L^2([0,1])} \leq \mathcal{O}\left(2^{-2n\delta}\right). \tag{3.10}$$

It remains to control the third term B_3. The definition of our estimator gives easily

$$\mathbb{E}\left\{\sum_{k=0}^{2^{j_0}-1}\left[\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}\right]^2\right\} = \frac{\sigma^2}{2^n}2^{j_0}, \tag{3.11}$$

while

$$\begin{aligned} &\mathbb{E}\left\{\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\left[\left(\frac{\lambda_j}{\lambda_j+\lambda} - 1\right)\beta_{j,k} + \frac{\lambda_j}{\lambda_j+\lambda}2^{-n/2}\epsilon_{j,k}\right]^2\right\} \\ &= \sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\left(\frac{\lambda_j}{\lambda_j+\lambda} - 1\right)^2\beta^2_{j,k} + \frac{\sigma^2}{2^n}\sum_{j=j_0}^{n}2^j\frac{\lambda_j^2}{(\lambda_j+\lambda)^2} \end{aligned} \tag{3.12}$$

Note that, for $\delta > 1/4$,

$$\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\frac{\lambda^2}{(\lambda_j+\lambda)^2}\beta^2_{j,k} \leq \lambda\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}2^{2\delta j}\beta^2_{j,k} = \lambda\|g\|^2_{H^\delta([0,1])}$$

and, also that

$$\frac{\sigma^2}{2^n}\sum_{j=j_0}^{n} 2^j \frac{\lambda_j^2}{\left(\lambda_j+\lambda\right)^2} \le \frac{\sigma^2}{2^n \log 2}\int_1^\infty \frac{1}{\left(1+\lambda y^{2\delta}\right)^2}\, dy \le \mathcal{O}\left(2^{-n}\lambda^{-1/(2\delta)}\right). \tag{3.13}$$

The theorem now follows upon using Eqs. (3.7) through (3.11). ∎

Note that all constants involved in the upperbounds $\mathcal{O}$ of the MISE in Theorem 3.1 are known. The first one is C_1 (given by Lemma 3.1), the second one is σ^2, the third one is $\|g\|^2_{H^\delta([0,1])}$ and the last one is given by the integral in expression (3.13). Note also that according to the theorem, an optimal choice for j_0 is $\log(m)/(2\beta+1)$. In Ibragimov and Khasminski (1980), Stone (1982) and Nussbaum (1985) it has been proved that the best global convergence rate in the MISE sense of any nonparametric estimator of g in the class $\mathcal{F}$ of δ-smooth functions ($\delta > 1/2$) that we considered, is $\mathcal{O}(n^{-s})$ with $s = \frac{2\delta}{2\delta+1}$ and similar constants. It is clear that, with regular enough coiflets, our estimator attains the best possible convergence rate.

In practice, for finite sample size m, the above property, although important, can be viewed as a side issue. A major problem in nonparametric regression is to choose the smoothing parameter(s), objectively from the data. Here the major smoothing parameter is λ, and δ is a secondary "shape" parameter. The next contribution of this section is to provide a simple objective method for estimating from the data, the λ and δ which minimize weighted integrated mean square error.

Note that the only term involving λ in the upper bound of the risk $R_{n,\delta}(\lambda)$ is the expectation of $\|\Pi_n g - g_\lambda\|^2_{L^2([0,1])}$. Thus, an objective determination of λ (and δ) can be based on an appropriate estimate of this expectation.

Consider first the case of σ^2 known. One could then form an unbiased estimate $\hat{\Lambda}_n(\delta,\lambda)$ of

$$\Lambda_n\left(\delta,\lambda\right) = \mathbb{E}\left\{\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\left[\left(\frac{\lambda_j}{\lambda_j+\lambda}-1\right)\beta_{j,k} + \frac{\lambda_j}{\lambda_j+\lambda}2^{-n/2}\epsilon_{j,k}\right]^2\right\},$$

and choose λ and δ to minimize $\hat{\Lambda}_n(\delta,\lambda)$. This sort of scheme, when δ is fixed, is classical in the context of choosing a smoothing parameter for a smoothing spline.

It very easy to see that, for $\beta > 1/2$ and $g \in H^\delta$, $\hat{\Lambda}_n(\delta,\lambda)$ defined by

$$\hat{\Lambda}_n\left(\delta,\lambda\right) = \frac{\sigma^2}{2^n}\left(\sum_{j=j_0}^{n} 2^j\left[\frac{\lambda_j}{\lambda_j+\lambda}\right]^2 - 2^j\left[\frac{\lambda}{\lambda_j+\lambda}\right]^2\right)$$
$$+\sum_{j=j_0}^{n}\sum_{k=0}^{2^j-1}\left[\left(\frac{\lambda}{\lambda_j+\lambda}\right)^2 d_{j,k}^2\right],$$

is an unbiased estimate of $\Lambda_n(\delta, \lambda)$.

In the following we consider the estimates λ_n and δ_n which minimize $\hat{\Lambda}_n(\delta, \lambda)$. We will first show that $\Lambda_n(\delta_n, \lambda_n) \to_P 0$ as $n \to \infty$.

Let $\xi_{j,k} = d_{j,k}^2 - \mathbb{E}\{d_{j,k}^2\}$. Since $\hat{\Lambda}_n(\delta, \lambda)$ is an unbiased estimator of $\Lambda_n(\delta, \lambda)$, we have:

$$\hat{\Lambda}_n\left(\delta, \lambda\right) = \Lambda_n\left(\delta, \lambda\right) + \Delta_n\left(\lambda, \delta\right),$$

where

$$\Delta_n\left(\lambda, \delta\right) = \sum_{j=j_0}^{n} \sum_{k=0}^{2^j-1} \left(\tfrac{\lambda}{\lambda+\lambda_j}\right)^2 \xi_{j,k},$$

and $\mathbb{E}\{\Delta_n(\lambda, \delta)\} = 0$.

Since $d_{j,k} = \beta_{j,k} + 2^{-n/2}\epsilon_{j,k}$, we clearly have:

$$\mathbb{E}\{d_{j,k}^2\} = \beta_{j,k}^2 + 2^{-n}\sigma^2. \tag{3.14}$$

We also have:

Lemma 3.2. *If the $\epsilon_{j,k}$'s have fourth order cumulant κ_4, then for $j, j' \geq j_0$, for $k = 0, \ldots, 2^j - 1$, $k' = 0, \ldots, 2^{j'} - 1$,*

$$Cov\left(d_{j,k}^2, d_{j',k'}^2\right) = \left(2^{-2n}\kappa_4 + 4\ 2^{-3n/2}\beta_{j,k}\mathbb{E}\{\epsilon_{j,k}^3\} + 4\ 2^{-n}\sigma^2\beta_{j,k}^2\right)\delta_{(j,k),(j',k')}. \tag{3.15}$$

The proof is straightforward but rather laborious, so details will be omitted. First, express the desired covariance in terms of the cumulants of the $d_{j,k}$'s and then use the fact that the $\epsilon_{j,k}$'s are uncorrelated.

We can now state the following theorem.

Theorem 3.2. *Assume that $\delta > 1/4$ and that the $\epsilon_{j,k}$'s have fourth moments. Then, for any constant $\epsilon > 0$,*

$$P\{\sup_{0<\lambda<1, 1/4<\delta<\infty} |\Delta_n\left(\lambda, \delta\right)| \geq \epsilon\} \leq \left(c/\epsilon^2\right) 2^{-n} \left(\log 2^n\right)^3,$$

where c does not depend on n.

Proof of Theorem 3.2. $\Delta_n(\lambda, \delta)$ can be expressed as

$$\Delta_n\left(\lambda, \delta\right) = \sum_{j=j_0}^{n} \left(\tfrac{\lambda}{\lambda+\lambda_j}\right)^2 \sum_{k=0}^{2^j-1} \xi_{j,k}.$$

Therefore,

$$\sup_{0<\lambda<1, 1/4<\delta<\infty} |\Delta_n\left(\lambda, \delta\right)| \leq \sup_{j_0 \leq j \leq n} |\sum_{k=0}^{2^j-1} \xi_{j,k}| \sup_{0<\lambda<1, 1/4<\delta<\infty} \sum_{j=j_0}^{n} \left(\tfrac{\lambda}{\lambda+\lambda_j}\right)^2.$$

The second sup is less than $n = \log m$ since $(\lambda/(\lambda + \lambda_j))^2 < 1$ for any $\delta > 0$ and $\lambda > 0$. For the first sup we can appeal to Lemma 4.1 of Chapter IV of Doob (1953). As stated, Doob's lemma applies to mean zero uncorrelated random variables Z_i with variances σ_i^2 and states that

$$\mathbb{E}\{\max_{j\leq n}|\sum_{i=1}^{j} Z_i|^2\} \leq (\log(4\times n)\,/\log 2)^2 \sum_{i=1}^{n} \sigma_i^2.$$

In our case, according to Lemma 3.2 the $\xi_{j,k}$'s are uncorrelated and we have:

$$\mathbb{E}\{\max_{j_0\leq j\leq n}|\sum_{k=0}^{2^j-1} \xi_{j,k}|^2\} \leq (\log(4\times 2^n)\,/\log 2)^2 \sum_{k=0}^{2^n-1} \mathrm{Var}\left(\xi_{n,k}\right).$$

Using the inequality (2.4) and Lemma 3.2, one can easily see that $\sum_{k=0}^{2^n-1}\mathrm{Var}(\xi_{n,k})$ behaves like $\mathcal{O}(2^{-n})$. Chebyshev's inequality completes now the proof of the theorem. ∎

Corollary 3.1. *Let λ_n and δ_n be the minimizers of $\hat\Lambda_n(\delta,\lambda)$. Then under the assumptions of the previous theorem, $\Lambda_n(\delta_n,\lambda_n)\to_P 0$.*

Proof. There exists a deterministic sequence $\bar\delta_n, \bar\lambda_n$ such that $\Lambda_n(\bar\delta_n,\bar\lambda_n)\to 0$. Now

$$\begin{aligned}\Lambda_n(\delta_n,\lambda_n) &= \hat\Lambda_n(\delta_n,\lambda_n) - \Delta_n(\delta_n,\lambda_n) \leq \hat\Lambda_n\left(\bar\delta_n,\bar\lambda_n\right) - \Delta_n(\delta_n,\lambda_n)\\ &= \Lambda_n\left(\bar\delta_n,\bar\lambda_n\right) + \Delta_n\left(\bar\delta_n,\bar\lambda_n\right) - \Delta_n(\delta_n,\lambda_n)\\ &\leq \Lambda_n\left(\bar\delta_n,\bar\lambda_n\right) + 2\sup_{\delta,\lambda}|\Delta_n(\delta,\lambda)| \to_P 0\end{aligned}$$

∎

We would like now to show that λ_n and δ_n the minimizers of $\hat\Lambda_n(\delta,\lambda)$ are consistent estimates of the minimizers λ_n^* and δ_n^* of $\Lambda_n(\delta,\lambda)$, but we have not been able to do this in full generality. We will, however, show, that for fixed $\delta > 1/4$, and for any constants a and b with $(0 < a < b)$ there is a consistent local minimizer $0 \leq \lambda_n \leq 1$. In order to do this it is necessary to choose the optimal approximation level $j_0 = \log m/(2\delta+1)$. We then have:

Corollary 3.2. *Let $\delta > 1/4$ fixed and let λ_n and λ_n^* be the minimizers of $\hat\Lambda_n(\delta,\lambda)$ and $\Lambda_n(\delta,\lambda)$, respectively. Then under the assumptions of Theorem 3.1, and for $j_0 = \log m/(2\delta+1)$, one has $2^{2n\delta/(2\delta+1)}(\lambda_n - \lambda_n^*)\to_P 0$.*

The proof may be established by arguments similar to those in Rice (1984) for periodic smoothing splines. In order to apply his results with similar reasoning one must insure that

$$\sup_{0<\lambda<1} 2^{2n\delta/(2\delta+1)}|\Delta_n(\lambda,\delta)| \to_P 0$$

A closer look at the proof of Theorem 3.2 shows that this is the case.

In the previous paragraphs we have considered results which entails knowing the error variance σ^2. Since σ^2 is typically unknown, it is tempting to use in the previous expressions an estimate of σ^2. In the context of nonparametric regression with independent errors, various proposals as to how σ^2 might be estimated have been made. A general and rather natural class of such estimators are those which are quadratic in the data. Here we will consider in some detail the estimator suggested by Müller and Stadtmüller (1987), based on squared weighted sums of squares of the observations. The proposed class of variance estimates is given by

$$\hat{\sigma}^2 = \tfrac{1}{2^n-2} \sum_{j=1}^{2^n-2} \left(\sum_{k=-1}^{1} \omega_k Y_{j+k} \right)^2$$

where $\omega_{-1} = 1/\sqrt{6}$, $\omega_0 = -2/\sqrt{6}$ and $\omega_1 = 1/\sqrt{6}$. When g is assumed to γ-Hölder continuous ($\gamma > 0$) and the errors ϵ_i are stochastically independent, the following proposition holds (for a proof see Lemma 2.1 of Müller and Stadtmüller (1987)):

Lemma 3.3. *Assume that the error variables are independent, with zero mean and finite variance σ^2, and that $g \in C^\gamma([0,1]$ for some $\gamma > 0$. Then*

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \sigma^2 + \mathcal{O}\left(2^{-2n\gamma}\right)$$

By the Sobolev embedding theorem, $H^\delta([0,1]) \subset C^\gamma([0,1])$ for any $\delta > \gamma + 1/2$. This allows us to introduce our modified estimate of $\Lambda_n(\delta, \lambda)$ as follows:

Theorem 3.3. *If $g \in H^\delta([0,1])$ for $\delta > \gamma + 1/2$ and some $\gamma > 0$, then $\tilde{\Lambda}_n(\delta, \lambda)$, defined by*

$$\tilde{\Lambda}_n(\delta, \lambda) = \tfrac{\hat{\sigma}^2}{2^n} \left(\sum_{j=j_0}^{n} 2^j \left[\tfrac{\lambda_j}{\lambda_j+\lambda}\right]^2 - 2^j \left[\tfrac{\lambda}{\lambda_j+\lambda}\right]^2 \right)$$
$$+ \sum_{j=j_0}^{n} \sum_{k=0}^{2^j-1} \left[\left(\tfrac{\lambda}{\lambda_j+\lambda}\right)^2 d_{j,k}^2 \right],$$

is such that

$$\mathbb{E}\left(\tilde{\Lambda}_{n,\delta}(\lambda)\right) = \Lambda_{n,\delta}(\lambda) + \mathcal{O}\left(n 2^{-n(2\gamma+1)}\right),$$

and, for fixed δ, $a > 0$ and $b > 0$

$$\sup_{am^{-2\delta/(2\delta+1)} \leq \lambda \leq bm^{-2\delta/(2\delta+1)}} \{m^{2\delta/(2\delta+1)} |\tilde{\Lambda}_n(\delta, \lambda) - \hat{\Lambda}_n(\delta, \lambda)|\} \to 0$$

Proof. Note that

$$\tilde{\Lambda}_n(\delta, \lambda) - \hat{\Lambda}_n(\delta, \lambda) = \tfrac{\hat{\sigma^2} - \sigma^2}{2^n} \left(\sum_{j=j_0}^{n} 2^j \left[\tfrac{\lambda_j}{\lambda_j+\lambda}\right]^2 - 2^j \left[\tfrac{\lambda}{\lambda_j+\lambda}\right]^2 \right).$$

The first part follows easily from Lemma 3.3. The second part is an easy consequence of the above equality, Lemma 3.3 and the results of Theorem 3.1. ∎

The previous results suggest the following procedure to determine the data dependent λ and δ: compute first $\tilde{\Lambda}_{n,\delta}(\lambda)$, and then choose λ and δ as its minimizers.

In spline smoothing, another method for providing, via a further approximation, an objective estimate for the minimizers of $R_n(\delta, \lambda)$ is cross-validation (see Wahba (1990)). The basic idea in cross-validation is to delete one observation at a time from the data set, and endeavor to predict it from the model as fitted to the remaining observations. In the context of wavelet smoothing such a method would require a two-fold cross-validation, since the discrete wavelet transform necessitates sample sizes of power of two. Some Monte-Carlo experiments with an empirical cross-validation procedure suggest that the results obtained by the application of Theorem 3.3 should be as good if not better.

To end this section, let us remark that the wavelet estimator we introduced appears as a particular diagonal linear shrinker (see Donoho and Johnstone (1991)). For each resolution $j \geq j_0$, the wavelet coefficients $d_{j,k}$ are shrinked by a factor $1/(1 + \lambda 2^{2\delta j})$ which is level dependent. It is therefore not surprising that we obtain asymptotic rates not far from the optimal ones. This is due to the choice of our loss function (L^2 loss) and our functional parameter class (a Sobolev space). If however, the estimation problem was stated with other loss functions (e.g. L^p loss, $1 < p < 2$) and in different Besov spaces, Donoho et al (1995) showed that no linear method can achieve the optimal minimax rate. In order to achieve near optimal rates they proposed wavelet shrinkage methods, a desarmingly simple and efficient way of de-noising gaussian data. We believe that the regularization framework provided in our paper, if stated with non-Hilbertian regularization penalties, can lead naturally to thresholding methods with optimal properties for quite general distributed data. It is hoped that this question will be investigated in a future work.

4. Computations and examples.

To investigate the practical implications of the asymptotic results presented in the previous section, a small scale empirical study was carried out. This section reports the computational issues and the results of simulations made to compare our method with some of the automatic smoothing techniques found in the literature : cross-validated smoothing splines (Wahba (1990), code contributed by Hutchinson), data driven variable bandwidth for local polynomial fitting (Fan and Gijbels (1995), code contributed by Fan and Gijbels) and wavelet thresholding (Donoho et al. (1995), code contributed by Donoho and Johnstone).

The L^2-setting for wavelet decompositions used in the previous sections was appropriate for theoretical investigations. A class of numerical algorithms that are suitable for calculating our wavelet estimator on the interval is described in Cohen, Daubechies and Vial (1993). Given $m = 2^n$ integrals $\alpha_{n,k} = \int_0^1 \phi_{n,k}(x)g(x)dx$, $k = 0, \ldots, 2^n - 1$, one iteratively applies a sequence of decimating high pass and low pass filter operators H_j, L_j: $\mathbb{R}^{2^j} \to \mathbb{R}^{2^{j-1}}$ via

$$(\alpha_{j-1,\cdot}) = L_j \circ (\alpha_{j,\cdot}) \quad (\beta_{j-1,\cdot}) = H_j \circ (\alpha_{j,\cdot})$$

for $j = n, n-1, \ldots, j_0 + 1$, producing a sequence of $m = 2^n$ coefficients

$$((\alpha_{j_0,\cdot}), (\beta_{j_0,\cdot}), (\beta_{j_0+1,\cdot}), \ldots, (\beta_{n-1,\cdot}))$$

Like the fast Fourier transform (FFT), the discrete wavelet transform (DWT) $W_{j_0,n}$ mapping $(\alpha_{n,\cdot})$ into this sequence is a fast, linear operation that operates on a data vector whose length is an integer power of two, transforming it into a numerically different vector of the same length. Also like the FFT, the wavelet transform is invertible and in fact orthogonal – the inverse transform, when viewed as a big matrix, is simply the transpose of the transform.

For empirical work, one does not have access to the fine-scale coefficients $\alpha_{n,k}$, but according to the Lemma 3.1 of the previous section, by using regular coiflets one perfectly controls the error made when applying the same algorithm to the normalized sampled values $2^{-n/2}g(k/2^n)$. There is however an important difference between wavelets on the line and wavelets on the interval, which results in the necessity to precondition the data by a prefiltering prior to their wavelet decomposition, in order to avoid boundary problems due to discretization. The details of this scheme can be found in Cohen, Daubechies and Vial (1993). This is the algorithm we used for the numerical applications to follow.

The wavelet transform $W_{j_0,n}$ is based on coiflets of order $L = 6$. These are defined through a set of 18 nonzero coefficients whose numerical values may be found in Daubechies (1992, Table 8.1, p. 261).

To examine the performance of our coiflet estimator and to compare it with other smoothers we designed a small simulation.

In this simulation all data is of the form $Y_i = g(x_i) + \epsilon_i$ with $\{\epsilon_i\}$ i.i.d. either Student's t noise with 3 d.f. or Gaussian noise. The sample sizes selected were $m = 128$ (intermediate

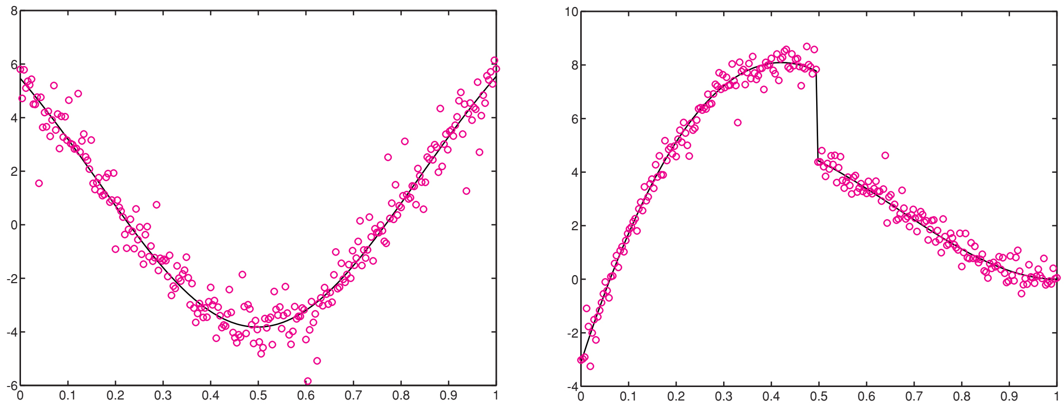

Figure 4.1: The two regression functions (g_1 at the left, g_2 at the right) used for the simulations (with similar additive noise structure).

sample size) and $m = 256$ (large). For each sample size, we used two different functions to ascertain the performance of the four smoothing methods. These are the following (see Fig. 4.1):

$$g_1(x) = \frac{3}{2}\cos 2\pi x + \frac{8}{3}\pi^2\left(x^2 - x + 1/6\right)$$

and

$$g_2(x) = \begin{cases} 11x^3 - 66x^2 + 52x - 3 & \text{if} \quad x \in [0, 0.5[\\ 32x(x-1)^2 & \text{if} \quad x \in [0.5, 1] \end{cases}$$

The first function is infinitely differentiable on $(0, 1)$ and satisfies $g_1(0) = g_1(1)$ but $g_1'(0) \neq g_1'(1)$. The second function is a piecewise polynomial with a discontinuity. It is similar to the function that appeared in Nason and Silverman (1994), to illustrate the adaptative properties of non linear wavelet shrinkers.

Each function g_i, $i = 1, 2$, was discretized to 128 and 256 equally spaced points in the interval $[0, 1]$, in such a way that $x_0 = 0$ and $x_m = 1$. The values of σ^2 were determined appropriately to achieve a similar noise structure for each signal (signal/noise ratio = 3) as in Breiman and Peters (1992) who made an extensive simulation study designed to compare some classical smoothers.

There is no one criterion that can adequately summarize the behavior of a smooth. The usually accepted criterion is MSE, that is if $\hat{g}(x_i)$ is the estimated function value at x_i, and m the sample size, then

$$MSE = \frac{1}{m}\sum_i \left(g(x_i) - \hat{g}(x_i)\right)^2$$

However, a smooth can have low MSE and appear noisy to the eye. To capture another aspect we also used the maximum deviation criterion (MXDV) defined in the Breiman and Peter's study. Nevertheless, as pointed out by a referee, MSE and MXDV could both be inappropriate for g_2: MSE does not sufficiently penalize the misestimation of the discontinuity, while MXDV does (e. g. if $g_2(x)$ was estimated by $\hat{g}(x) = g_2(x - 0.0001)$, MXDV would be large, in spite of the fact that $\hat{g}$ is an almost perfect "estimator"). Following the referee's suggestion we also used as a criterion the average of the ℓ largest absolute deviations (AMXDV), with $\ell = 10\%$ of the sample size. Each run was repeated 400 times, holding all factors constant, except the noise that was regenerated. The retained criteria were computed as follows:

a) RMSE (Root Mean Squared Error). The MSE was computed for each run and averaged over the 400 runs. Then its square root was taken.

b) MXDV (Maximum Deviation). This is the average over the 400 runs of

$$\max_i |g(x_i) - \bar{g}(x_i)|$$

c) AMXDV (Average Maximum Deviation). This is the average over the 400 runs of the ℓ largest absolute deviations ($\ell = 10\%\ m$).

All the computations were performed on a Macintosh IIcx personal computer. The simulation compares the following smoothers : GCVS (generalized cross-validation cubic splines), WAVREG (the wavelet regularized smoother developed in this article), VISUSHRINK (the universal wavelet hard thresholding estimate of Donoho and Johnstone), SURETHRESH (the SURE wavelet thresholding estimator of Donoho and Johnstone), both with boundary adjusted filters, and VARBAND (the variable bandwidth local polynomial smoother of Fan and Gijbels). The author is indebted to Donoho and Johnstone for providing him with the Matlab code for VISUSHRINK and SURESHRINK, and to Fan and Gijbels for supplying the VARBAND code.

The following tables summarize the simulation results. Table 4.1 gives the RMSE's, MXDV's and AMXDV's for the two functions g_1 and g_2, for a sample size 128, and for each smoothing method, while Table 4.2 corresponds to similar results for a sample size of 256. On average the standard error of the RMSE was about 0.03×RMSE.

RMSE over 400 runs (sample size 128, Student noise)					
function	GCVS	REGWAV	VISUSHRINK	SURESHRINK	VARBAND
$g_1(x)$	2.2960e-01	1.8831e-01	4.0886e-01	4.1094e-01	2.7276e-01
$g_2(x)$	4.1917e-01	3.7432e-01	4.6813e-01	4.1305e-01	2.6999e-01
MXDV over 400 runs (sample size 128, Student noise)					
$g_1(x)$	6.7980e-01	5.8214e-01	3.2311e+00	3.2764e+00	9.0556e-01
$g_2(x)$	1.8518e+00	1.7325e+00	1.9530e+00	1.8624e+00	1.3882e+00
AMXDV over 400 runs (sample size 128, Student noise)					
$g_1(x)$	4.2131e-01	3.8387e-01	8.1806e-01	8.2201e-01	5.6755e-01
$g_2(x)$	1.1287e+00	9.9237e-01	1.1932e+00	1.1260e+00	6.1110e-01

Table 4.1: Simulation results for each smoothing method (sample size 128, Student noise).

RMSE over 400 runs (sample size 256, Student noise)					
function	GCVS	REGWAV	VISUSHRINK	SURESHRINK	VARBAND
$g_1(x)$	1.7514e-01	1.6928e-01	3.9024e-01	3.4983e-01	2.4383e-01
$g_2(x)$	3.5924e-01	3.4098e-01	3.8176e-01	3.5569e-01	2.1315e-01
MXDV over 400 runs (sample size 256, Student noise)					
$g_1(x)$	5.1628e-01	4.8847e-01	3.9866e+00	3.8291e+00	8.3193e-01
$g_2(x)$	1.8548e+00	1.9784e+00	2.0879e+00	2.0583e+00	1.4641e+00
AMXDV over 400 runs (sample size 256, Student noise)					
$g_1(x)$	3.6224e-01	3.4733e-01	6.4045e-01	5.8726e-01	4.7142e-01
$g_2(x)$	9.8167e-01	9.0648e-01	9.7397e-01	9.2315e-01	4.7370e-01

Table 4.2: Simulation results for each smoothing method (sample size 256, Student noise).

WAVREG is markedly better on the smooth function g_1. We suspect, by looking at the MXDV criterion, that the cross-validation procedures used both in GCVS and VARBAND lead to a selection of window sizes on the small side, producing estimated curves with too much variability. VISUSHRINK and SURESHRINK performances are more adversely affected by the Student distribution of the noise and by the relatively small sample sizes. Indeed these smoothers are designed for gaussian data and for large sample sizes. In order to see how sensitive and robust are the estimators to a Gaussian noise assumption, we repeated the whole experiment with normal errors. The following tables summarize the results.

RMSE over 400 runs (sample size 128, Gaussian noise)					
function	GCVS	REGWAV	VISUSHRINK	SURESHRINK	VARBAND
$g_1(x)$	1.4314e-01	1.4297e-01	3.6663e-01	3.7266e-01	1.5583e-01
$g_2(x)$	4.1434e-01	3.5020e-01	4.6600e-01	4.0888e-01	1.9898e-01
MXDV over 400 runs (sample size 128, Gaussian noise)					
$g_1(x)$	5.6381e-01	5.5678e-01	3.2052e+00	3.3205e+00	4.3878e-01
$g_2(x)$	1.8536e+00	1.6369e+00	1.9102e+00	1.8333e+00	1.1667e+00
AMXDV over 400 runs (sample size 128, Gaussian noise)					
$g_1(x)$	3.0286e-01	3.0110e-01	7.0449e-01	7.0783e-01	3.0999e-01
$g_2(x)$	1.1363e+00	9.5809e-01	1.2048e+00	1.1546e+00	4.2968e-01

Table 4.3: Simulation results for each smoothing method (sample size 128, Gaussian noise).

RMSE over 400 runs (sample size 256, Gaussian noise)					
function	GCVS	REGWAV	VISUSHRINK	SURESHRINK	VARBAND
$g_1(x)$	1.0792e-01	1.0378e-01	2.8095e-01	2.8040e-01	1.1456e-01
$g_2(x)$	3.5023e-01	3.3031e-01	3.8172e-01	3.5604e-01	1.5912e-01
MXDV over 400 runs (sample size 256, Gaussian noise)					
$g_1(x)$	3.4808e-01	3.2721e-01	3.3312e+00	3.3974e+00	4.2561e-01
$g_2(x)$	1.8150e+00	2.0166e+00	2.1185e+00	2.1185e+00	1.3258e+00
AMXDV over 400 runs (sample size 256, Gaussian noise)					
$g_1(x)$	2.1650e-01	2.0744e-01	5.0514e-01	4.9500e-01	2.3557e-01
$g_2(x)$	9.8180e-01	8.9616e-01	9.7677e-01	9.5653e-01	3.2476e-01

Table 4.4: Simulation results for each smoothing method (sample size 256, Gaussian noise).

As expected, VISUSHRINK and SURESHRINK are getting relatively more efficient. Note also that as sample size increases all estimators became more efficient. Whatever the distribution of the noise is, larger values of MXDV occur for the second function g_2, due to the presence of the discontinuity point. The function hardest to fit is the discontinuous function g_2. In such a case, quite surprisingly, the VARBAND smoother does much better than wavelet methods. We believe that for recovering smooth functions a local polynomial fit with a constant bandwidth would be more advantageous than VARBAND and similarly, in that situation, REGWAV is more advantageous than the nonlinear wavelet estimators. For recovering irregular curves, the nonlinear wavelet estimators or other variable smoothing procedures as VARBAND could be employed.

As for computational efficiency, the ranking of the methods from the fastest to the slowest is : VISUSHRINK, SURESHRINK, WAVREG, GCVS and finally VARBAND.

To end this section, we now consider a real example. The data set is discussed in Example 3.4.5 of Eubank (1988, p. 82) and represents the voltage drop in the battery of a guided missile motor during flight. In this example the assumptions of an equispaced fixed design model are reasonable. In order to use the wavelet transform, the data was interpolated by a piecewise polynomial and the interpolated values on an equispaced grid of 128 points was fitted. The resulting tapered coiflet smoother obtained by our method with $j_0 = 3$ is displayed as a solid line on the plot of figure 4.2. For comparison purposes, the dashed line displays the cross-validated spline smoother fitted to the original data with a cross-validation parameter $\lambda = 0.094$.

As an overall conclusion, no smoother uniformly dominates in all aspects. However, the tapered coiflet smoother, introduced as a solution of a simple variational problem, is a complement to Donoho and Johnstone's nonlinear wavelet smoothers. When the underlying curve is relatively smooth it presents good performance, and appears to be more robust to the nature of the noise than other wavelet smoothers.

Acknowledgments

I wish to thank J. Fan and I. Gijbels for providing me with a preprint of their 1995 paper prior to publication and for providing me the c-code for the VARBAND smoother. D. Donoho and I. Johnstone are gratefully acknowledged for providing the WaveLab statistical package. I am very grateful to the referees for their numerous stimulating comments and detailed suggestions which have considerably improved this paper.