Statistica Sinica 4(1994), 651-678

SMOOTHING NOISY DATA WITH COIFLETS

Anestis Antoniadis

Laboratoire de Statistique et Modélisation Stochastique

Abstract: This paper is concerned with an orthogonal wavelet series regression estimator of an unknown smooth regression function observed with noise on a bounded interval. The method is based on applying results of the recently developed theory of wavelets and uses the specific asymptotic interpolating properties of the wavelet approximation generated by a particular wavelet basis, Daubechie's coiflets. Conditions are given for the estimator to attain optimal convergence rates in the integrated mean square sense as the sample size increases to infinity. Moreover, the estimator is shown to be pointwise consistent and asymptotically normal. The numerical implementation of the estimation procedure relies on the discrete wavelet transform; and the algorithm for smoothing a noisy sample of size n requires order $\mathcal{O}(n)$ operations. The general theory is illustrated with simulated and real examples and a comparison with other nonparametric smoothers is made.

Key words and phrases: Nonparametric regression, curve smoothing, wavelets, multiresolution analysis, splines.

1. Introduction

The purpose of this paper is to contribute to the methodology available for estimating smooth regression functions from noisy data.

Let $Y_0, Y_1, \ldots, Y_{N-1}$ denote data generated by the fixed-design regression model

$$Y_i = g(x_i) + \epsilon_i, \quad 0 \leq i \leq N-1, \tag{1.1}$$

where the design points $x_0, x_1, \ldots, x_{N-1}$ all belong to an interval $[0,T]$ and are assumed to be equally spaced. The random variables (Y_i) are measurements of the unknown regression function $g : [0,T] \to \mathbb{R}$, contaminated with errors ϵ_i which are assumed to be i.i.d. random variables with mean 0 and variance σ^2.

There are two main approaches to estimate the regression function g: the parametric approach which assumes that g follows a parametric (linear or nonlinear) model and the nonparametric approach which only assumes that g is "smooth". In recent years several methods of nonparametric curve fitting have been proposed and investigated, the basic motivation for investigations of this type being the doubt concerning the usual assumptions in the classical finite parameter theory. Recent references with extensive bibliographies and interesting discussions on the applicability of such smoothing methods are Eubank (1988), Müller (1988) and Wahba (1990).

There are a number of good estimators for g in model (1.1). A popular choice is series estimation obtained by regression on functions ϕ_j selected from an orthogonal basis for

$$L^2[0,T] = \left\{f : \int_0^T f^2(x)dx < \infty\right\}.$$

Assuming that g can be expanded as

$$g(x) = \sum_{k=1}^{\infty} a_k\phi_k(x),$$

one uses estimators of the form

$$\hat{g}(x) = \sum_{k=1}^{m} \hat{a}_k\phi_k(x), \tag{1.2}$$

where m is some positive integer and $\hat{a}_k$ are appropriate estimators of the a_k. The parameter m in (1.2) governs the number of terms and, hence, the smoothness of the estimator. Problems of this nature have been studied by Rutkowski (1982), Rafajlowicz (1987), Cox (1988) and Eubank, Hart and Speckman (1990) to cite only a few. However such estimates have not been widely used in practice because their quality depends very much on the selected orthogonal system. Our aim is to suggest a new orthogonal series regression estimator which turns out to have not only theoretical, but also computational advantages over classical orthogonal series estimators, while its smoothing properties remain very close to the smoothing properties of kernel or penalized least-squares smoothing estimators.

Our estimator can be treated as a projection estimate. Roughly speaking, using an increasing sequence of closed subspaces V_m of the parameter space provided by an analysis of compactly supported wavelets in $L^2[0,T]$, the orthogonal projection of the unknown function g onto V_m is estimated. The appropriate choice of the wavelet basis, the coiflets, allows us to consider the data as an estimate of the projection of g on V_n, from which the estimated projections on lower resolution subspaces are easily obtained. By means of the properties of the wavelet basis we obtain precise rates of convergence. Moreover, a key aspect of our estimator is that its tuning parameter m ranges over a smaller set of values than those of other orthogonal series estimators.

The paper is organized as follows. Section 2 reviews some background on wavelet theory. The wavelet estimator for nonparametric regression 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. Section 4 discusses the numerical implementation of the estimator and contains some extensive simulations and a comparison of the estimator with other smoothers. A discussion on possible extensions and some conclusions are presented in Section 5. The proofs of the results are deferred to the end of the paper.

2. A Short Review of Wavelets

In this section we sketch an account of some relevant properties of orthogonal wavelets and multiresolution analysis, which will be used to derive our estimator. The following succinct review suffices for the understanding of this paper. For a more detailed exposition, examples and proofs the reader is referred to the sizable wavelet literature, especially Daubechies (1992), Meyer (1990) and Chui (1992).

Wavelets are functions generated from one basic function by dilatations and translations. A particularly interesting development is the recent discovery of orthonormal bases of wavelets. For particular functions $\psi \in L^2(\mathbb{R})$, the family

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j,k \in \mathbb{Z}, \tag{2.1}$$

constitutes an orthonormal basis of $L^2(\mathbb{R})$. A classical example of such a basis is the Haar basis; smoother choices of compactly supported ψ were constructed by Daubechies (1988).

In all the interesting examples, orthonormal wavelet bases can be constructed via a *multiresolution analysis*, a framework developed by Mallat (1989), in which the wavelet coefficients $< f, \psi_{j,k} >$ of a function f for a fixed j describe the difference between two approximations of f, one with resolution 2^j, and one with the coarser resolution 2^{j-1}.

A multiresolution analysis (or approximation) of L^2 consists of a nested sequence of closed subspaces $V_j, j \in \mathbb{Z}$, of $L^2(\mathbb{R})$,

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

such that their intersection is trivial and union is dense in $L^2(\mathbb{R})$,

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

they are dilates of one another,

$$f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1},$$

and there exists a *scaling* function $\phi \in V_0$ whose integer translates span V_0, the approximation space with resolution 1,

$$V_0 = \Big\{f \in L^2(\mathbb{R}) : f(x) = \sum_{k\in\mathbb{Z}} \alpha_k \phi(x-k)\Big\}.$$

One can always choose ϕ such that $\int_{\mathbb{R}} \phi(x)dx = 1$. An orthonormal basis of V_j, the approximation space with resolution 2^{-j} is then given by the family $\{\phi_{j,k} : k \in \mathbb{Z}\}$, where

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$$

is a dilated and translated version of ϕ. The orthogonal projection of a function $f \in L^2(\mathbb{R})$ into V_j is given by

$$P_n f = \sum_{k\in\mathbb{Z}} < f, \phi_{j,k} > \phi_{j,k},$$

and can be thought of as an approximation of f with resolution 2^{-j}.

The multiresolution analysis is said to be r-regular if ϕ is in the Hölder space $C^r(\mathbb{R})$, and if both ϕ and its derivatives have a fast decay,

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

Note that, since $\phi \in V_0 \subset V_1 = \overline{\mathrm{Span}\{\phi_{1,k} : k \in \mathbb{Z}\}}$, the function ϕ necessarily satisfies an equation of the type

$$\phi(x) = \sum_{k\in\mathbb{Z}} c_k \phi(2x-k), \tag{2.2}$$

for some coefficients c_k.

The multiresolution analysis leads directly to an orthogonal decomposition of $L^2(\mathbb{R})$. Defining W_j as the orthogonal complement of V_j in V_{j+1}, we get another sequence $\{W_j : j \in \mathbb{Z}\}$ of closed mutually orthogonal subspaces of L^2, such that each W_j is a dilate of W_0, and their direct sum is $L^2(\mathbb{R})$.

The space W_0 is spanned by integer translates of a single function ψ which can be defined by

$$\psi(x) = \sum_{k\in\mathbb{Z}} (-1)^k c_{-k+1} \phi(2x-k), \tag{2.3}$$

where the c_n are given by Eq. (2.2). Note that we have assumed that the c_k are real. The whole analysis carries through, modulo some complex conjugates, for complex c_k. Using this decomposition of $L^2(\mathbb{R})$ and Eq. (2.3), it follows that the family $\{\psi_{j,k} : j,k \in \mathbb{Z}\}$ given in Eq. (2.1) constitutes an orthonormal basis of $L^2(\mathbb{R})$.

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)

$$\int_{-\infty}^{+\infty} x^k \psi(x)dx = 0 \quad \text{for} \quad 0 \le k \le r. \tag{2.4}$$

The converse is generally false, and the number of vanishing moments is usually larger than the regularity of the multiresolution analysis. An important consequence of this property is that polynomials of degree less than or equal to r can be expressed as linear combination of the translates of ϕ, i.e. they belong to V_0.

Daubechies (1988) constructed a class $\{\phi^{(N)} : N \in \mathbb{N}\}$ of scaling functions, such that each $\phi^{(N)}$, in its orthogonalized version, has compact support and N vanishing moments, and such that $\phi^{(N)} \in C^{\mu N}(\mathbb{R})$ where $\mu \simeq 0.1936$. From the definition of the coefficients c_k in Eq. (2.2), it follows that only finitely many c_k are nonzero, so that ψ reduces to a finite linear combination of compactly supported and regular functions, and therefore has compact support and the same regularity. Constructing ψ from the c_k has the advantage of allowing better control of the supports of ϕ and ψ. If $c_k = 0$ for $k < N_1, k > N_2$, then the support of ϕ is included in the interval $[N_1, N_2]$. Note however that the size of the support increases linearly with the number of vanishing moments. 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.

$$\begin{aligned} &\int_{\mathbb{R}} \phi(x)dx = 1, \\ &\int_{\mathbb{R}} x^k \phi(x)dx = 0, \quad 1 \le k \le N-1, \\ &\int_{\mathbb{R}} x^k \psi(x)dx = 0, \quad 0 \le k \le N-1. \end{aligned} \tag{2.5}$$

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

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)).

If f belongs to the Sobolev space $H^s(\mathbb{R})$ and the multiresolution analysis is r-regular, then

$$\|f - P_j f\|_{L^2} \le o(2^{-j \min(s,r)}) \quad \text{as} \quad j \to \infty. \tag{2.6}$$

If f belongs to some Hölder space $C^s(\mathbb{R})$ then

$$| < f, \psi_{j,k} > | \leq C 2^{-j(\min(s,r)+1/2)} \quad \text{for all} \quad j \geq 0, \ k \in \mathbb{Z}, \tag{2.7}$$

where C is a constant independent of j and k.

3. Nonparametric Curve Estimation

The setting considered here is the well known fixed design regression model where the ordinates of the data points in the scatter plot are regarded as deterministic values. These ordinate values are usually chosen by the experimenter, as in a designed experiment. Since in the absence of other information it is intuitively sensible to take such points equally spaced, this is frequently the case. Moreover such an assumption is very often made in theoretical investigations of nonparametric regression. See Wahba (1983), Rice (1984), Eubank (1988), and Hall and Titterington (1992) for ample precedent.

The fixed design model is given by

$$Y_i = g(t_i) + \epsilon_i, \quad 0 \leq i \leq N-1, \tag{3.1}$$

where the Y_i are noisy measurements of the "mean" or regression function $g(t)$ taken at equidistant nonrandom design points t_i within $[0,1]$ (without loss of generality), and the ϵ_i's are independent identically distributed random variables with mean 0 and finite variance σ^2. In the above model the function g is viewed as a member of a class $\mathcal{F}$ of possible regressions over $[0,1]$. The specification of $\mathcal{F}$ is an approximation of the "real" g whose explicit form can never be known. A smooth type condition of order $m > 0$ will be imposed on g. More precisely, if $m \notin \mathbb{N}$, we shall assume that $\mathcal{F}$ is the set of functions $[m]$ times continuously differentiable in $\mathbb{R}$, and such that their $[m]$th derivatives satisfy the Lipschitz condition of order $m-[m]$. If $m \in \mathbb{N}$, $\mathcal{F}$ will be the set of functions $m-1$ times continuously differentiable in $\mathbb{R}$, and such that their $(m-1)$th derivatives satisfy a Lipschitz condition of order 1. A function g in $\mathcal{F}$ will be called m-smooth. Considering the structure of the wavelet multiresolution, the design points t_i will be assumed to have the form $t_i = i\Delta$, with $\Delta = 2^{-n}$ for $i = 0, \ldots, N-1$, where $N = 2^n$.

The asymptotic results will be obtained as n goes to infinity. Our attention will be limited to wavelet estimates that are linear functions of the observations. If the errors are normal and if g belongs to a Sobolev space, this is no restriction if a minimax approach is adopted since in this case linear methods can achieve minimax rates of convergence for squared error loss. However, when $\mathcal{F}$ is an infinite dimensional class of partially smooth functions presenting some jumps (e.g.

Besov class) it is generally true that linear methods do not suffice for obtaining minimax mean square error and nonlinear methods are generally preferable. Donoho and Johnstone (1992) developed nonlinear wavelet based estimation procedures for noisy functions with gaussian errors using decision-theoretic criteria based on Stein's unbiased estimate of risk. There are however some limitations to the interesting results of Donoho and Johnstone in their paper on regression estimation by wavelet shrinkage. First, they implicitly suppose, as in Mallat (1989) or Cohen (1990), that it is the wavelet coefficients $< g, \phi_{n,k} >$ that are directly observed instead of the sampled values of the g. Secondly, the errors are taken to be $N(0, \sigma^2)$ and furthermore their minimax estimation procedure requires the knowledge of the noise level σ^2, though it is claimed that this can be relaxed.

All the one-dimensional wavelets we have reviewed so far lead to bases for $L^2(\mathbb{R})$. In the application we have in mind we are interested in only part of the real line, since the regression function will be observed within the bounded interval $[0,1]$. One could, of course, decide to use standard wavelet bases to analyze g, setting the data equal to zero outside the interval, but this introduces an artificial jump at the edges, reflected in the wavelet coefficients. Another way to deal with this problem is to "periodize" the data and use the usual wavelets to analyze the "periodized" version. Unless the regression function is already periodic this will again introduce jumps at the boundaries, which will be reflected by large fine scale wavelet coefficients near the boundaries. For this reason we follow Hall (1983) and use an integrated mean squared error (IMSE) of the form

$$R_n = R(\hat{g}) = \int_0^1 E[\hat{g}(t) - g(t)]^2 u(t) dt \tag{3.2}$$

to assess the performance of an estimator $\hat{g}$ of g on $[0,1]$. Here $u > 0$ denotes a continuously differentiable weight function on $\mathbb{R}$, supported in an interval $[\alpha, \beta] \subset (0,1)$ and included to reduce the impact of boundary effects. When viewed through the asymptotic behavior of this criterion, wavelet estimators merit serious consideration as competitors to second order kernel or cubic-spline estimators.

Hereafter, ϕ and ψ denote the scaling function and the wavelet associated with an r-regular multiresolution analysis of $L^2(\mathbb{R})$. We assume further that the scaling function ϕ is a coiflet of order L with $L > m + 1$.

Since g in Eq.(3.1) does not necessarily belong to $L^2(\mathbb{R})$, we replace g by a function f such that

$$f(t) = g(t) \quad \text{for all } t \text{ in } [0,1], \tag{3.3}$$

with f in $C^m(\mathbb{R})$, compactly supported in an interval $[-\epsilon, 1+\epsilon]$ for some small ϵ. This substitution will have little effect since the properties of the estimator are only evaluated on $[0,1]$, and (3.1) for the observations coincides with

$$Y_i = f(t_i) + \epsilon_i, \quad 0 \le i \le N-1. \tag{3.4}$$

It follows from Eq. (3.3) that $f \in H^m(\mathbb{R})$, where $H^m(\mathbb{R})$ denotes the Sobolev space of order m.

Since

$$L^2(\mathbb{R}) = V_n \oplus (\oplus_{j\ge n} W_j),$$

the function f admits the following expansion in H^m (and in L^2):

$$f(t) = \sum_{k\in\mathbb{Z}} < f, \phi_{n,k} > \phi_{n,k}(t) + \sum_{j\ge n}\sum_{\ell\in\mathbb{Z}} < f, \psi_{j,\ell} > \psi_{j,\ell}(t),$$

with

$$< f, \phi_{n,k} >= \int_{\mathbb{R}} f(t)\phi_{n,k}(t)dt \quad \text{and} \quad < f, \psi_{j,\ell} >= \int_{\mathbb{R}} f(t)\psi_{j,\ell}(t)dt.$$

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_n for which the coefficients $< f, \phi_{n,k} >$ are given. However, when a function is given in sampled form there is no general method for deriving the coefficients $< f, \phi_{n,k} >$. A first step towards our curve estimation method is to try to approximate the projection P_n by some operator Π_n in terms of the sampled values $f(\frac{k}{2^n})$ and to derive then a reasonable estimator of the approximation $\Pi_n f$. Such a device has also been used by Istas (1992) for the estimation of the covariance function of a Gaussian process. Since the coiflets have L vanishing moments, one can define such an estimator of $\Pi_n f$ by

$$\hat{f}_n(t) = \widehat{\Pi_n f(t)} = 2^{-n/2}\sum_{k\in\mathbb{Z}} Y_k\phi_{n,k}(t) = 2^{-n/2}\sum_{k=0}^{2^n-1} Y_k\phi_{n,k}(t). \tag{3.5}$$

This choice can be justified by the following lemma.

Lemma 3.1. *The set of non zero coefficients $f^{\{n\}}(k) = 2^{n/2} < f, \phi_{n,k} >$ has a cardinality equivalent to $\mathcal{O}(2^n)$. Moreover, with $L > [m]+1$, the following uniform (in $0 \le k \le 2^n - 1$) bound holds:*

$$|f^{\{n\}}(k) - g\Big(\frac{k}{2^n}\Big)| \le C_1 2^{-nm}, \tag{3.6}$$

where C_1 is a constant depending only on the coiflet ϕ.

By Lemma 3.1 one is therefore able to approximate the coefficients $\langle f, \phi_{n,k}\rangle$ with an error $\mathcal{O}(2^{-n/2}2^{-nm})$. It is therefore natural to approximate $P_n f$ by

$$(\Pi_n f)(t) = \sum_{k\in\mathbb{Z}} f\Big(\frac{k}{2^n}\Big)\phi(2^n t - k) = 2^{-n/2}\sum_{k\in\mathbb{Z}} f\Big(\frac{k}{2^n}\Big)\phi_{n,k}(t).$$

Using such an approximation and Lemma 3.1, we have:

$$\begin{aligned}\|P_n f - \Pi_n f\|_\infty &\le \sup_{t\in\mathbb{R}}\{2^{-n/2}\sum_{k\in\mathbb{Z}} |f^{\{n\}}(k) - f\Big(\frac{k}{2^n}\Big)|\,|\phi_{n,k}(t)|\} \\ &\le \mathcal{O}(2^{-nm})\sup_{t\in\mathbb{R}}\{\sum_{k\in\mathbb{Z}} |\phi(t-k)|\} \le \mathcal{O}(2^{-nm}). \qquad (3.7)\end{aligned}$$

Observing that $E(Y_k) = g(\frac{k}{2^n}) = f(\frac{k}{2^n})$ completely justifies our choice of the estimator $\hat{f}_n$.

The above calculations suggest that given our observed sampled values of g, the observations are equivalent to the estimator $\hat{f}_n$. This estimator, while presenting a very small bias, leads to a highly oscillatory solution that perfectly fits the data. In order to smooth the data, we associate with each sample size $N = 2^n$ a resolution $j(n)$, and estimate the unknown function g by the orthogonal projection of $\hat{f}_n$ onto $V_{j(n)}$. The parameter $j(n)$ governs the smoothness of our estimator. For the purpose of this paper, it will be treated as being deterministic rather than depending on the data. It is however important to choose it judiciously because it controls the trade-off between fidelity to the data and the smoothness of the resulting solution. Too small a value of $j(n)$ leads to an oversmoothed, biased solution. From a theoretical viewpoint, in the derivation of asymptotic results, the smoothing parameter must tend to infinity at the correct rate as the amount of information in the data grows to infinity. The following theorem addresses the appropriate decay rate of the integrated mean squared error defined in Eq.(3.2).

Theorem 3.1. *Under the smoothness assumptions imposed on g and the weight function u in this section, the IMSE satisfies*

$$R_N \le O(2^{-2j(n)\min([m],r)}) + O(2^{j(n)-n}) + O(2^{j(n)}2^{-2(n-j(n))(\alpha+1/2)}),$$

where α denotes the Hölder exponent of the coiflet ϕ.

In Ibragimov and Hasminskii (1982), Stone (1982) and Nussbaum (1985) it has been proved that the best global convergence rate, in the IMSE sense, of any nonparametric estimator of g in the class $\mathcal{F}$ of m-smooth function, that we considered is $\mathcal{O}(n^{-s})$ with $s = \frac{2[m]}{2[m]+1}$. If one takes $j(n) = n/(1+2[m]) = \log_2(N)/(1+2[m])$, it is clear that, with enough regularity of the coiflets, our estimator attains the best possible convergence rate.

We devote the rest of this section to a discussion of the consistency of our estimator. Regarding the pointwise consistency of the coiflet estimator, we have:

Theorem 3.2. *If the errors ϵ_i, $0 \leq i \leq N-1$, are independent identically distributed with zero mean and finite variance σ^2, then, for any dyadic $t \in]0,1[$,*
(a) $\hat{g}_n(t) \rightarrow g(t)$ *almost surely.*
(b) $\sqrt{N}(\hat{g}(t) - g(t))$ *converges in distribution to a zero mean Gaussian random variable with variance $\sigma^2 w^2(t)$ where*

$$w^2(t) = \lim_{n\to\infty} \frac{1}{2^n} \sum_{k=0}^{2^n-1} \phi_{n,k}^2(t).$$

Let us remark that, using the multivariate version of the central limit theorem of Jennrich (1969), it is straightforward to extend Theorem 3.2 to the multivariate case.

4. Numerical Aspects, Simulations and Examples

This section is devoted to the numerical application of the estimation procedure. For the convenience of the reader, a short description of the fast wavelet transform is first presented, followed by the analysis of some simulated and some real examples.

4.1. The discrete wavelet transform

The L^2-setting for wavelet decompositions used in the previous sections was appropriate for theoretical investigations. We review here a class of numerical algorithms, originally developed by Beylkin, Coifman and Rokhlin (1991) that are suitable for calculating our wavelet estimator. An excellent pedagogical review, as well as some fortran procedures of the discrete wavelet transform (DWT) that we are going to use, are given in Press (1991). Like the fast Fourier transform (FFT), the discrete wavelet transform (DWT) 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.

We start with $V_n = \mathbb{R}^N$, $N = 2^n$. The multiresolution approximation defined in Section 2, is now replaced by a finite chain

$$V_0 \subset V_1 \subset \cdots \subset V_n,$$

$$V_n = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_{n-1},$$

where $\dim(V_j) = \dim(W_j) = 2^j$. A function f in V_n

$$f = \sum_j s_j^n \phi_{n,j}$$

can be written as the sum of its components in V_{n-1}, W_{n-1}

$$f = \sum_k s_k^{n-1} \phi_{n-1,k} + \sum_k d_k^{n-1} \psi_{n-1,k}.$$

The coefficients of either of these representations can be calculated from the other by the formulas

$$s_k^{n-1} = \sum_j c_{j-2k} s_j^n, \tag{4.1}$$

$$d_k^{n-1} = \sum_j c_{2k-j+1} s_j^n, \tag{4.2}$$

and

$$s_j^n = \sum_k [c_{j-2k} s_k^{n-1} + c_{2k-j+1} d_k^{n-1}], \tag{4.3}$$

where the recursion coefficients c_n are the coefficients defining the original compactly supported scaling function ϕ in Eq. (2.2).

Decomposition of a vector $\nu \in V_n$ into its subspace components and subsequent reconstruction are achieved by repeated application of (4.1), (4.2) and (4.3) if all subscripts are taken modulo 2^n. What is interesting is that the functions ϕ or ψ do not explicitly enter the picture. The algorithm continues to work on a purely algebraic level.

Application of (4.1), (4.2) and (4.3) can be considered as matrix multiplication which is useful in visualizing the effect on a vector. Each step of the orthogonal wavelet decomposition of $\nu \in V_n$, as well as any partial decomposition, corresponds to multiplying ν by an orthogonal matrix W. Reconstruction corresponds to multiplication by W^T, and recovers the original $v = W^T(W\nu)$.

The DWT consists of applying a wavelet coefficient matrix like W *hierarchically*, first to the full data vector of length N, then to the "smooth" vector of length $N/2$, then to the "smooth-smooth" vector of length $N/4$, and so on until only a trivial number of "smooth-...-smooth" components (usually 2) remain. The procedure is sometimes called a *pyramidal algorithm* (see Mallat (1989)), for obvious reasons. The output of the DWT consists of these remaining components and all the "detail" components that were accumulated along the way. The following configuration borrowed from Press (1991) should make the procedure clear:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \end{pmatrix} \xrightarrow{W} \begin{pmatrix} s_1 \\ d_1 \\ s_2 \\ d_2 \\ s_3 \\ d_3 \\ s_4 \\ d_4 \\ s_5 \\ d_5 \\ s_6 \\ d_6 \\ s_7 \\ d_7 \\ s_8 \\ d_8 \end{pmatrix} \xrightarrow{\text{permute}} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \\ \overline{d_1} \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{pmatrix} \xrightarrow{\text{W}} \begin{pmatrix} S_1 \\ D_1 \\ S_2 \\ D_2 \\ S_3 \\ D_3 \\ S_4 \\ D_4 \\ \overline{d_1} \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{pmatrix} \xrightarrow{\text{permute}} \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ \overline{D_1} \\ D_2 \\ D_3 \\ D_4 \\ \overline{d_1} \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{pmatrix} \xrightarrow{W} \begin{pmatrix} \mathcal{S}_1 \\ \mathcal{S}_2 \\ \overline{\mathcal{D}_1} \\ \mathcal{D}_2 \\ \overline{D_1} \\ D_2 \\ D_3 \\ D_4 \\ \overline{d_1} \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{pmatrix}. \tag{4.4}$$

If the length of the data vector were a higher power of two, there would be more stages of applying W (or any other wavelet coefficients) and permuting. The endpoint will always be a vector with two $\mathcal{S}$'s and a hierarchy of $\mathcal{D}$'s, D's, d's, etc. Notice that once d's are generated, they simply propagate through to all subsequent stages.

To invert the DWT, one simply reverses the procedure, starting with the smallest level of the hierarchy and working in (4.4) from right to left. The inverse matrix W^T is of course used instead of the matrix W. Applying the DWT to an arbitrary vector is therefore an order $\mathcal{O}(N)$ numerical procedure.

This is the algorithm we used for the numerical applications to follow in the next section. The wavelet transform W 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).

4.2. Simulated and real examples

To examine the performance of our coiflet estimator and to compare it with penalized least-squares smoothers we designed a small simulation. We did not compare our estimator with kernel smoothers, mainly because they are costly to compute.

In this simulation all data is of the form $Y_i = g(x_i) + \epsilon_i$ with $\{\epsilon_i\}$ i.i.d. $N(0, \sigma^2)$.

The sample sizes selected were 128 (intermediate sample size) and 256 (large). For each sample size we used two different functions. These are the following (see Fig. 4.1)

$$\begin{aligned} g_1(x) &= x(x-0.3)(x-0.5)(x-1)+0.2, \\ g_2(x) &= \sin(2\pi x)\sin(2\pi(x-0.3))\sin(2\pi(x-0.5)). \end{aligned}$$

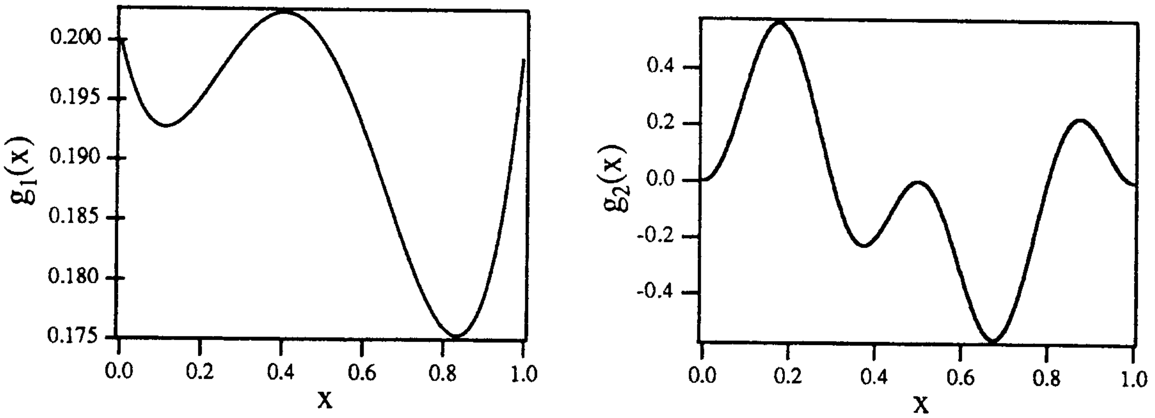

Figure 4.1. The two functions used in the simulations.

The choice of the first function represents a polynomial and was used to see how well our wavelet smoother does on such regular data (this function was used by Hall and Koch (1992) in the context of image analysis). The second function also used in the above mentioned paper represent a class of functions with slowly changing curvature and some inflection points.

Each function g_i, $i = 1, 2$, was discretized to 128 and 256 equally spaced points in the interval $[-0.2, 1.2]$, in such a way that $x_0 = -0.2$ and $x_N = 1.2$. The choice of the design interval $[-0.2, 1.2]$ was made to avoid boundary effects at the points 0 and 1. The values of σ^2 were determined in the same way as in Breiman and Peters (1992) who made an extensive simulation study designed to compare some classical smoothers). For a given design set $\{x_i\}$, and a given function g, the empirical standard deviation $sd(g)$ of the set $\{g(x_i)\}$ was computed and two values of the signal/noise ratio $sd(g)/\sigma$ were taken, determined by

$$sd(f)/\sigma = 1,\ 2.0.$$

In terms of percent of variance explained by the function, these values correspond to 50% and 80%.

To summarize the behavior of a smooth estimate, we used a partial mean squared error criterion PMSE. That is, if $\hat{g}(x_i)$ is the estimated function value at x_i, and M the number of points within the interval $[0, 1]$, then

$$\text{PMSE} = \frac{1}{M} \sum_{x_i \in [0,1]} (g(x_i) - \hat{g}(x_i))^2.$$

In each run all factors were held constant, except the $\{\epsilon_i\}$ which were regenerated. All the computations were performed on a Macintosh IIcx personal computer. The penalized least-squares splines were computed with Hutchinson's "generalized cross-validated" cubic spline smoother cubgcv.

The function g_1 was discretized to 128 points and 256 points. For each setting, the discretized function was degraded by a Gaussian additive noise of mean zero and standard deviation $\sigma = 0.294$ or $\sigma = 0.294/2$. The observations are displayed in Fig. 4.2 by dots.

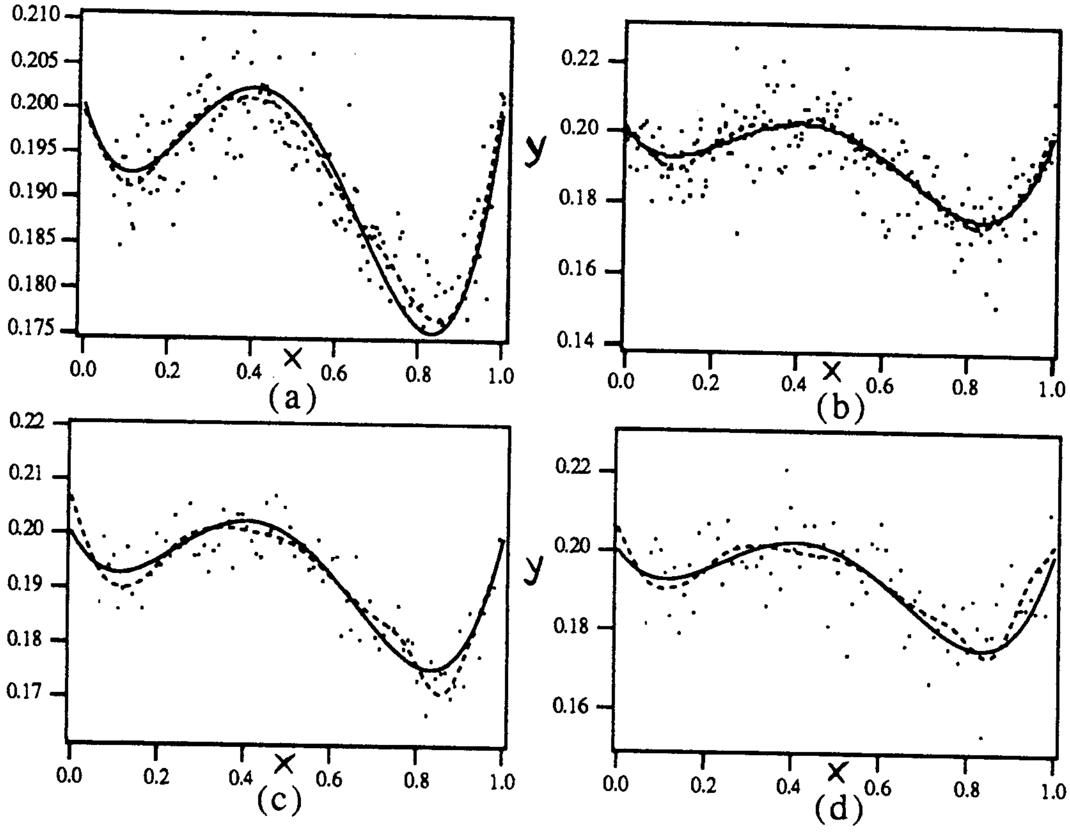

Figure 4.2. The true function $g_1(x)$ (solid line) and the wavelet reconstruction (dashed line) using coiflets of order 6. The dots represent the noisy data obtained from g_1. The graphs (a) and (b) correspond to a 256-point discretization with a signal/noise ratio $sd(g_1)/\sigma = 1$ for (a) and $sd(g_1)/\sigma = 2$ for (b). The graphs (c) and (d) correspond to 128-point discretization with similar signal/noise ratio.

For the 128-discretization, the estimator was calculated at the 128 points by means of the discrete wavelet transform, using the projection onto the space V_3, according to the conclusion of Theorem 3.1. Note that the choice of $j(n) = 3$ is automatic and by no means suggested by the data. The PMSE was calculated for each estimation. For each estimate plotted as a dashed line in Fig. 4.2, we found the following values of PMSE: 0.0439 for the case (a), 0.0487 for (b), 0.0581 for (c) and 0.0914 for (d).

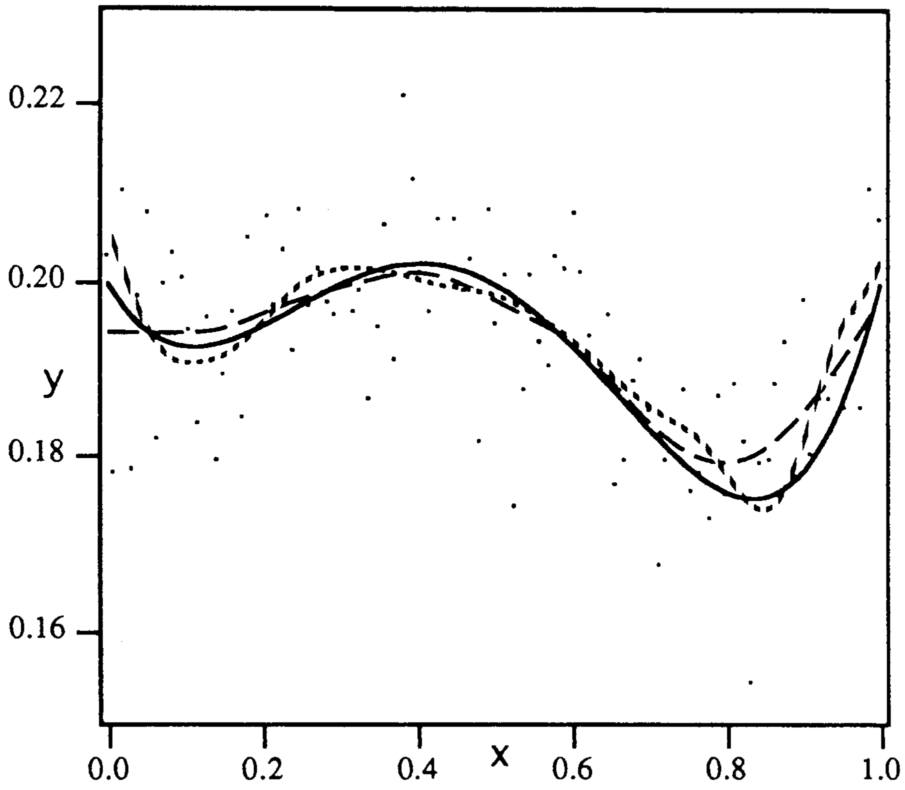

Figure 4.3. The true function $g_1(x)$ (solid line), the wavelet reconstruction (gray line) and the cross-validated smoothing spline (dashed line) for the data displayed in Fig. 4.2 (d).

To examine further the performance of our estimator, for the intermediate sample size and for the larger value of σ we calculated the corresponding smoothing spline. The corresponding fit is displayed in Fig. 4.3. The smoothing spline estimate, using an optimal smoothing parameter $\lambda = 0.00009$ obtained by cross-validation, presents a PMSE = 0.009. This shows that the coiflet estimator does not perform as well (as it is expected from asymptotic theory) when the data is very regular. However, for large sample sizes, the difference was less pronounced.

For the second curve g_2 the wavelet estimator performs much better than the spline smoother. As before, the estimator was calculated at the 256 points by means of the discrete wavelet transform, using the projection onto the space V_4 this time. For each estimate plotted as a dashed line in Fig. 4.4, we found the following values of PMSE: 0.0014 for the case (a), 0.0015 for (b), 0.0021 for (c) and 0.0028 for (d). As one can see, the wavelet smoother is especially suitable here.

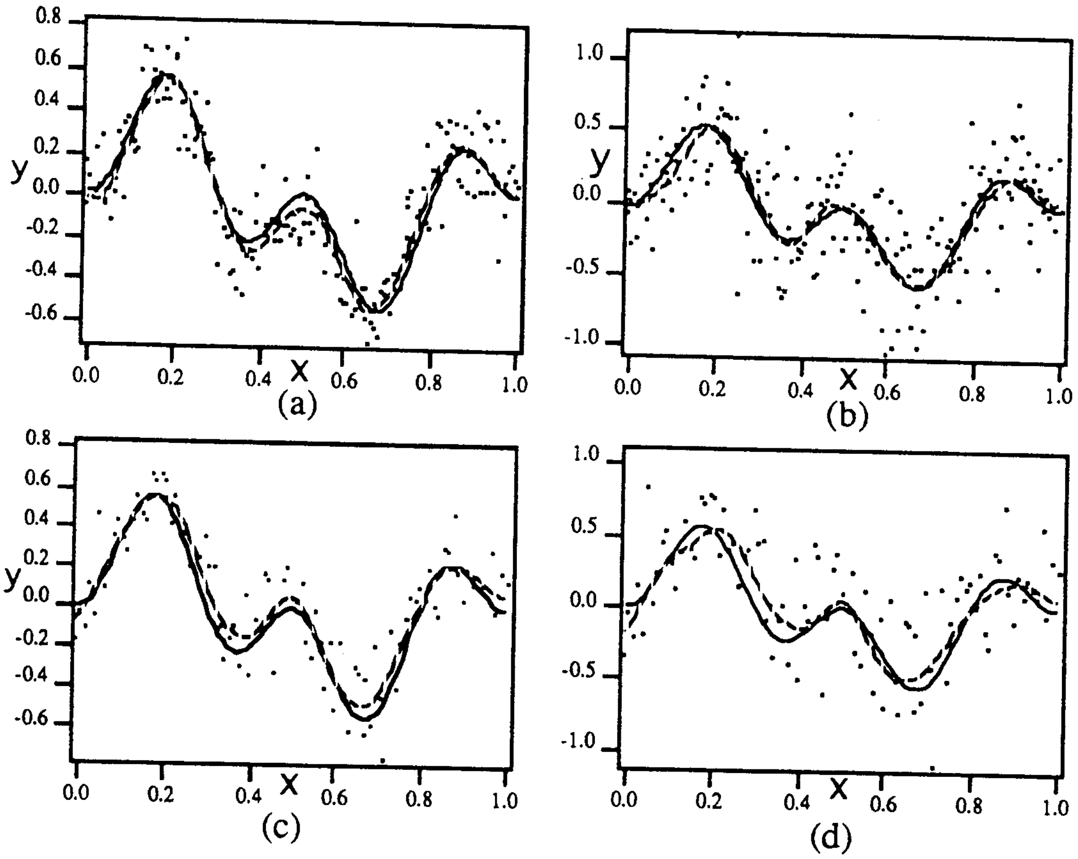

Figure 4.4. The true function $g_2(x)$ (solid line) and the wavelet reconstruction (dashed line) using coiflets of order 6. The dots represent the noisy data obtained from g_2. The graphs (a) and (b) correspond to a 256-point discretization with a signal/noise ratio $sd(g_1)/\sigma = 2$ for (a) and $sd(g_1)/\sigma = 1$ for (b). The graphs (c) and (d) correspond to 128-point discretization with similar signal/noise ratios.

Fig. 4.5 illustrates a further comparison between the wavelet estimator and penalized least-squares splines. A smoothing spline was fitted to the data corresponding to case (d) in Fig. 4.4. The optimal data driven smoothing parameter chosen by cross-validation was $\lambda = 0.089$, leading to a fit with a corresponding PMSE equal to 0.0739. As one can see, the spline approach suffers from a local lack of fit, a relatively large bias occuring at peaks.

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 (see next section) and the interpolated values on an equispaced grid of 128 points was fitted. The resulting coiflet smoother obtained by projection onto the space V_3 is displayed

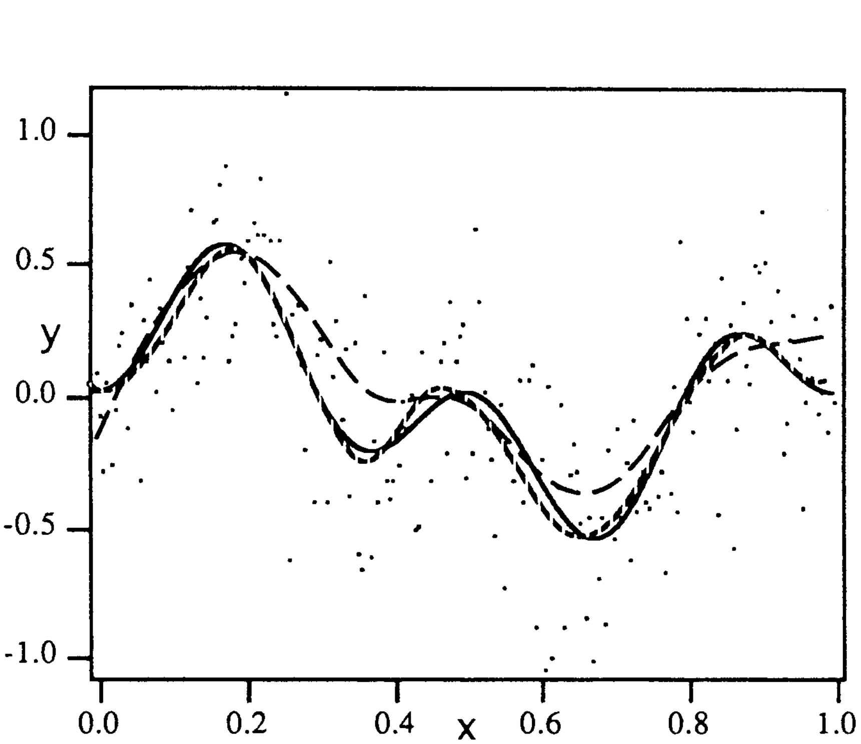

Figure 4.5. The true function $g_2(x)$ (solid line), the wavelet reconstruction (gray line) and the cross-validated smoothing spline (dashed line) for the data displayed in Fig. 4.4 (b).

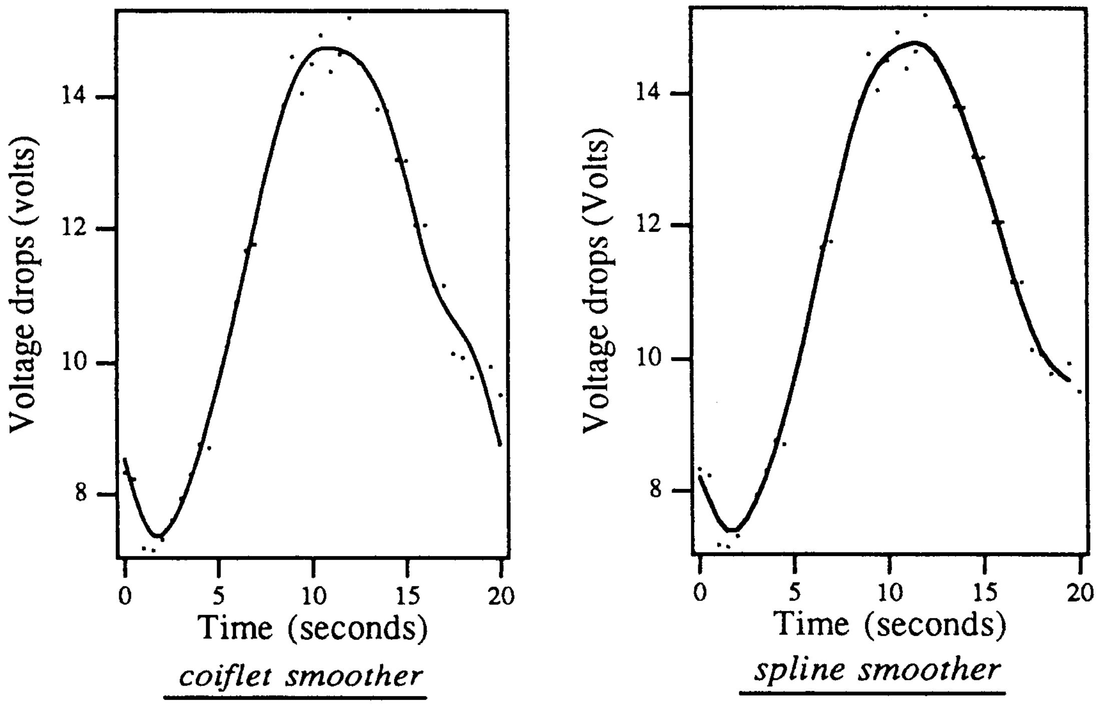

Figure 4.6. Plot of the voltage drop data. The wavelet estimator is displayed at the left figure, while the right graph displays the cubic spline fitted by cross-validation.

at the left plot of Fig. 4.6. For comparison purposes, the right plot displays the cross-validatory spline smoother fitted to the original data with a cross-validation parameter $\lambda = 0.094$. Considering that the wavelet estimator chose a tunning parameter from among only two different values ($j(n) = 2$ or 3), it gives an outstanding result.

As an overall conclusion, the coiflet smoother gives good performance, does not rely on a data driven method for choosing its smoothing parameters and is very fast to compute.

5. Discussion and Possible Extensions

We have assumed that the data are equally spaced, an assumption that is very often made in theoretical investigations of nonparametric regression. This made possible the use of the discrete wavelet transform to obtain a nice computational algorithm for $\hat{g}$ and to derive its asymptotic properties. However, an analysis for the more general situation would have to take account of unequally spaced data and of the boundaries and would proceed along different lines.

A possible approach for unequally spaced data is to first interpolate the data vector $\boldsymbol{y}$ by a piecewise polynomial. For instance, suppose that the design points $x_0, x_1, \ldots, x_{N-1}$ form a regular design sequence in the sense that

$$\max_{1 \leq i \leq N-1} \mid x_i - x_{i-1} \mid = \mathcal{O}(N^{-1}).$$

Following De Boor (1978), let $p_{k+1,i}(x)$ be the polynomial of order $k+1$ which interpolates the noisy function at the points x_{i+j}, $j = 0, \ldots, k$, and let $[x_i, \ldots, x_{i+k}]\boldsymbol{y}$ be its leading coefficient (the kth divided difference). Using the properties of polynomial interpolation it is not difficult to show that for an m-smooth function g and for each $x \in [x_i, x_{i+1}]$,

$$E(\mid g(x) - \hat{p}(x) \mid) \leq \mathcal{O}(\mid x_{[m]+1} - x_0 \mid^m),$$

where the interpolating polynomial $\hat{p}$ is defined by

$$\hat{p}(x) = p_{[m],i}(x) - (x - x_i) \cdots (x - x_{i+[m]})[x_i, \ldots, x_{i+[m]}]\boldsymbol{y}$$

for $x \in [x_i, x_{i+1}]$. Using the upper bound above, one obtains a uniform upper bound on $[0, 1]$:

$$E(\mid g(x) - \hat{p}(x) \mid) \leq \mathcal{O}(N^{-m}).$$

It then seems natural to estimate the coefficients $g_{n,k} = 2^n \int g(x)\phi(2^n x - k)dx$ at resolution n by

$$\hat{g}_{n,k} = 2^n \int \hat{p}(x)\phi(2^n x - k)dx.$$

We conjecture that our results extend to the estimator based on $\hat{g}_{n,k}$ and should be comparable to the equispaced case. However, a deeper analysis of this estimator and several obvious modifications of it go beyond the intent of this paper, but provide interesting topics for future research.

Concerning the boundary effects, one could use, instead of wavelets defined on the whole real line or their periodized version, wavelet bases on the unit interval with preassigned boundary conditions. Such bases have been constructed recently (see Auscher (1992), Meyer (1991) and Cohen, Daubechies and Vial (1992)) and provide a wavelet analysis for the Sobolev spaces $H^m([0,1])$ or subspaces of $H^m([0,1])$ whose members satisfy specific boundary conditions. However, for the moment, these constructions are too cumbersome to be efficiently used for numerical purposes.

To conclude, let us say that while DWT smoothers lack the flexibility of more sophisticated techniques such as penalized least-squares smoothing, they seem to give satisfactory results in many cases, are capable of attaining mean squared error rates that are optimal, get high marks for their implementational simplicity and do not rely on choosing bandwidths or weights in loss functions. Smoothers with such qualities are probably the kind of estimators to be used in more complicated situations, for example, when fitting generalized additive models or using wavelet estimators to test the adequacy of linear models.

6. Proofs of the Results

In this section the proofs are given of the lemmas and theorems stated in Section 3. The notation and assumptions in Section 2 are in force throughout.

Proof of Lemma 3.1. The first assertion follows from the fact that both ϕ and f are compactly supported. For the second assertion, one has:

$$\begin{aligned} f^{\{n\}}(k) - g\Big(\frac{k}{2^n}\Big) &= f^{\{n\}}(k) - f\Big(\frac{k}{2^n}\Big) \\ &= \int_{\mathbb{R}} 2^{n/2} f(t)\phi_{n,k}(t)dt - f\Big(\frac{k}{2^n}\Big) \\ &= \int_{\mathbb{R}} 2^n f(t)\phi(2^n t - k)dt - f\Big(\frac{k}{2^n}\Big) \\ &= \int_{\mathbb{R}} \Big(f\Big(\frac{v+k}{2^n}\Big) - f\Big(\frac{k}{2^n}\Big)\Big)\phi(v)dv, \end{aligned}$$

since $\int_{\mathbb{R}} \phi(v)dv = 1$.

A Taylor expansion at the point $k/2^n$ up to the order $[m]$, and the vanishing moments ϕ lead to:

$$\begin{aligned}&f\Big(\frac{v+k}{2^n}\Big)-f\Big(\frac{k}{2^n}\Big)\\&=\sum_{\ell=1}^{[m]-1}f^{(\ell)}\Big(\frac{k}{2^n}\Big)\Big(\frac{v}{2^n}\Big)^{\ell}+\Big(\frac{v}{2^n}\Big)^{[m]}\int_0^1\frac{(1-x)^{[m]-1}}{([m]-1)!}f^{([m])}\Big(\frac{k}{2^n}+x\frac{v}{2^n}\Big)dx.\end{aligned}$$

Now,

$$\begin{aligned}\int_0^1\frac{(1-x)^{[m]-1}}{([m]-1)!}f^{([m])}\Big(\frac{k}{2^n}+x\frac{v}{2^n}\Big)\Big(\frac{v}{2^n}\Big)^{[m]}dx &= \frac{(\frac{v}{2^n})^{[m]}f^{([m])}(\frac{k}{2^n})}{([m]+1)!}\\&+\Big(\frac{v}{2^n}\Big)^{[m]}\int_0^1\frac{(1-x)^{[m]-1}}{([m]-1)!}\Big[f^{([m])}\Big(\frac{k}{2^n}+x\frac{v}{2^n}\Big)-f^{([m])}\Big(\frac{k}{2^n}\Big)\Big]dx,\end{aligned}$$

since $\int_0^1\frac{(1-x)^j}{j!}dx=\frac{1}{(j+1)!}$.

It follows, that

$$\begin{aligned}&f^{\{n\}}(k)-f\Big(\frac{k}{2^n}\Big)\\&=\int_{\mathbb{R}}\Big(\frac{v}{2^n}\Big)^{[m]}\phi(v)\int_0^1\frac{(1-x)^{[m]-1}}{([m]-1)!}\Big[f^{([m])}\Big(\frac{k}{2^n}+x\frac{v}{2^n}\Big)-f^{([m])}\Big(\frac{k}{2^n}\Big)\Big]dxdv\\&=2^{-n[m]}\int_a^b v^{[m]}\phi(v)\int_0^1\frac{(1-x)^{[m]-1}}{([m]-1)!}\Big[f^{([m])}\Big(\frac{k}{2^n}+x\frac{v}{2^n}\Big)-f^{([m])}\Big(\frac{k}{2^n}\Big)\Big]dxdv,\end{aligned}$$

since the support of ϕ is within $[a,b]$.

To end the proof, note that

$$\int_0^1\frac{(1-x)^{[m]-1}}{([m]-1)!}\mid\Big[f^{([m])}\Big(\frac{k}{2^n}+x\frac{v}{2^n}\Big)-f^{([m])}\Big(\frac{k}{2^n}\Big)\Big]dx\mid\le C\max\Big(\frac{|a|,|b|}{2^n}\Big)^{m-[m]},$$

by the Hölder continuity of f.

Proof of Theorem 3.1. Since the weight function u has support within the interval $[0,1]$ and g coincides with f on $[0,1]$, the integrated squared error R_n can be written as

$$R_n=R(\hat{g})=E\Big(\int_{\mathbb{R}}[\hat{g}(t)-g(t)]^2u(t)dt\Big)=E\Big(\int_{\mathbb{R}}[\hat{g}(t)-f(t)]^2u(t)dt\Big).$$

Applying the usual variance-bias decomposition to the integrated mean squared error yields

$$\begin{aligned}R(\hat{g})&=E\Big(\int_{\mathbb{R}}[\hat{g}(t)-E(\hat{g}(t))]^2u(t)dt\Big)+\Big(\int_{\mathbb{R}}(f(t)-E(\hat{g}(t)))^2u(t)dt\Big)\\&=V_n+B_n^2.\end{aligned}$$

We shall concentrate first on bounding the squared bias B_n^2 of $\hat{g}$.

Recall from G. Beylkin and al. (1991) that given a function h in V_n of the form

$$h(x) = \sum_{k=0}^{2^n-1} s_k^0 \phi_{n,k}(t)$$

it can be expressed in the form

$$h(x) = \sum_{k=0}^{2^{(n-f(n))}-1} s_k^{j(n)} \phi_{n-j(n),k}(t) + \sum_{k=0}^{2^{(n-j(n))}-1} d_k^{j(n)} \psi_{n-j(n),k}(t),$$

where the $s_k^{j(n)}$ denote the coefficients of the projection of h onto $V_{j(n)}$. It follows that

$$\hat{g}(t) = \left(P_{V_{j(n)}} \hat{f}\right)(t) = \sum_{k=0}^{2^{(n-j(n))}-1} \hat{s}_k^{j(n)} \phi_{n-j(n),k}(t).$$

Note also that since

$$\begin{aligned} f &= P_{V_n} f + (I - P_{V_n}) f = \Pi_n f + P_{V_n} f - \Pi_n f + (I - P_{V_n}) f \\ &= P_{V_{j(n)}} \Pi_n f + \Pi_n f - P_{V_{j(n)}} \Pi_n f + (P_{V_n} f - \Pi_n f) + (I - P_{V_n}) f, \end{aligned}$$

we have

$$\begin{aligned} f(t) = \sum_{k=0}^{2^{(n-j(n))}-1} s_k^{j(n)} \phi_{n-j(n),k}(t) + \sum_{\ell=j(n)}^{n} \sum_{k=0}^{2^{(n-\ell)}-1} d_k^{\ell} \psi_{n-\ell,k}(t) \\ + (P_{V_n} f - \Pi_n f)(t) + ((I - P_{V_n}) f)(t). \end{aligned}$$

Using the above expressions, for any $t \in [0,1]$, the bias term can be decomposed as follows

$$E[\hat{g}(t)] - f(t) = B_1(t) + B_2(t) + B_3(t),$$

where

$$B_1(t) = \sum_{\ell=j(n)}^{n} \sum_{k=0}^{2^{(n-\ell)}-1} d_k^{\ell} \psi_{n-\ell,k}(t), \tag{6.1}$$

$$B_2(t) = (P_{V_n} f - \Pi_n f)(t), \tag{6.2}$$

and

$$B_3(t) = \left(\left(I - P_{V_n}\right) f\right)(t). \tag{6.3}$$

The first two terms are due to the fact that

$$E\left[(P_{V_{j(n)}} \hat{g})(t)\right] \neq (P_{V_{j(n)}} f)(t),$$

while the third term $B_3(t)$ is the usual term that appears in the bias whenever a projection method is used, and represents the error of approximating f by its projection on V_n.

Using the bound given by Eq. (2.6) we obtain

$$\int_{\mathbb{R}} B_3^2(t)dt = \|f - P_n f\|^2_{L^2(\mathbb{R})} \leq o\left(2^{-n \min([m],r)}\right) \tag{6.4}$$

as $n \to \infty$.

The term $B_1(t)$ can be written as

$$B_1(t) = (\Pi_n f)(t) - (P_{V_n} f)(t) + (P_{V_{j(n)}}(P_{V_n} f - \Pi_n f))(t) + ((P_{V_n} - P_{V_{j(n)}})f)(t).$$

By Eq. (3.7) we have

$$\|\Pi_n f - P_{V_n} f\|^2_\infty \leq \mathcal{O}\left(2^{-2nm}\right), \tag{6.5}$$

and since $P_{V_{j(n)}}$ is an orthogonal projector the same bound holds for

$$\|P_{V_{j(n)}}(\Pi_n f - P_{V_n} f)\|^2 \leq \mathcal{O}(2^{-2n \min(r,[m])}). \tag{6.6}$$

It remains to control the remaining term

$$((P_{V_n} - P_{V_{j(n)}})f)(t) = \sum_{j \geq j(n)}^{n} \sum_{k \in \mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}(t)$$

of $B_1(t)$. Considering precisely the supports of f, u and $\psi_{j,k}$ and using the assumed regularity of f and the uniform upper bound for $< f, \psi_{j,k} >$ given by Eq. (2.7), we have

$$\begin{aligned}\int_{\mathbb{R}} ((P_{V_n} - P_{V_{j(n)}})f)^2(t)u(t)dt &\leq \sum_{j \geq j(n)}^{n} \sum_{-b \leq k \leq 2^j - a} |< f, \psi_{j,k} >|^2 \\ &\leq C \sum_{j \geq j(n)} 2^{-2j \min(r,[m])} \leq \mathcal{O}(2^{-2j(n) \min(r,[m])}).\end{aligned} \tag{6.7}$$

Combining (6.4), (6.5), (6.6) and (6.7), we obtain

$$B_n^2 \leq \mathcal{O}(2^{-2j(n) \min(r,[m])}). \tag{6.8}$$

Now consider the variance term V_n. We have

$$\begin{aligned} V_n &= E\Big(\int_{\mathbb{R}} [\hat{g}(t) - E(\hat{g}(t))]^2 u(t)dt\Big) \\ &= E\Big(\int_{\mathbb{R}} [(P_{V_{j(n)}} \hat{f})(t) - (P_{V_{j(n)}} \Pi_n f)(t)]^2 u(t)dt\Big) \\ &= E\Big(\int_{\mathbb{R}} [P_{V_{j(n)}} (\hat{f} - (\Pi_n f))(t)]^2 u(t)dt\Big). \end{aligned}$$

Using the Schwarz inequality, we obtain

$$
\begin{aligned}
V_n &\le C\,E\Big(\int_{\mathbb{R}}\{P_{V_{j(n)}}\Big[\sum_{k\in\mathbb{Z}} 2^{-n/2}\Big(Y_k - f\Big(\frac{k}{2^n}\Big)\Big)\phi_{n,k}\Big](t)\}^2 u(t)dt\Big)\\
&= E\Big(2^{-n}\sum_{a\le\ell\le 2^{j(n)}+b}\ \sum_{k=0}^{2^n-1}\Big(Y_k - f\Big(\frac{k}{2^n}\Big)\Big)^2 |< \phi_{j(n),\ell}, \phi_{n,k} >|^2\Big)\\
&= C\,2^{-n}\sigma^2 \sum_{a\le\ell\le 2^{j(n)}+b}\ \sum_{k=0}^{2^n-1} |< \phi_{j(n),\ell}, \phi_{n,k} >|^2\,. \qquad (6.9)
\end{aligned}
$$

Noting that

$$< \phi_{j(n),\ell}, \phi_{n,k} > \; = \; < \phi, \phi_{n-j(n),k-2^{n-j(n)}\ell} >$$

and using the inequality (6.5.3, p.204) in Daubechies (1992), we have

$$|< \phi_{j(n),\ell}, \phi_{n,k} >|^2 = 2^{j(n)-n}\phi^2\Big(\frac{k}{2^{n-j(n)}} - \ell\Big) + \mathcal{O}(2^{-2(n-j(n))(\alpha+1/2)}),$$

where α is the exponent of the Hölder continuous wavelet ϕ, and the constant in $\mathcal{O}$ does not depend on n, $j(n)$ or k. Hence Eq. (6.9) becomes

$$
\begin{aligned}
V_n &\le C\,2^{-n}\sigma^2 \sum_{a\le\ell\le 2^{j(n)}+b}\ \sum_{k=0}^{2^n-1} 2^{j(n)-n}\phi^2\Big(\frac{k}{2^{n-j(n)}} - \ell\Big)\\
&\quad + \mathcal{O}(2^{j(n)}2^{-2(n-j(n))(\alpha+1/2)}). \qquad (6.10)
\end{aligned}
$$

Now,

$$2^{-n}\sum_{a\le\ell\le 2^{j(n)}+b}\ \sum_{k=0}^{2^n-1} \phi^2\Big(\frac{k}{2^{n-j(n)}} - \ell\Big) \to 1 \qquad (6.11)$$

as $n \to \infty$, and therefore Inequality (6.10) can be rewritten as

$$V_n \le \mathcal{O}(2^{j(n)-n}) + \mathcal{O}(2^{j(n)}2^{-2(n-j(n))(\alpha+1/2)}). \qquad (6.12)$$

which completes the proof.

The lemma to follow, a variant of which may also be found in Genon-Catalot et al. (1990) shows that expression (6.11) used in the proof of the Theorem 3.1 holds.

Lemma 6.1. *With the notation of this paper, if w is in $C^1[0,1]$, then*

$$2^{-n}\sum_{a\le\ell\le 2^{j(n)}+b}\ \sum_{k=0}^{2^n-1} \phi^2\Big(\frac{k}{2^{n-j(n)}} - \ell\Big)w\Big(\frac{k}{2^n}\Big) \to \int_0^1 w(s)ds,$$

as $n \to \infty$.

Proof. Let $q_k = 2^{j(n)}(\frac{k}{2^n} - r_\ell)$, $r_\ell = \ell/2^{j(n)}$ and set $T_{k,\ell} = [q_k, q_{k+1}] \times [r_\ell, r_{\ell+1}]$. The rectangles $T_{k,\ell}$ have an area equal to 2^{-n} and cover the domain

$$R_{j(n)} = \{(u,s) \in \mathbb{R}^2 : \frac{a}{2^{j(n)}} \leq s \leq 1 + \frac{b}{2^{j(n)}}, 0 \leq \frac{u}{2^{j(n)}} + s \leq 1\}.$$

Note that, by the change of variables $t = s + u/2^{j(n)}$, $v = u$, and since ϕ is normalized in $L^2(\mathbb{R})$, we have

$$I_n = \int_{R_{j(n)}} \phi^2(u) w\Big(s + u/2^{j(n)}\Big) du ds = \int_0^1 w(s) ds = I.$$

Letting $S_n(w)$ denote the expression of the lemma, we have

$$S_n(w) = S_1 + S_2 + I,$$

where

$$S_1 = \sum_{k,\ell} \int_{T_{k,\ell}} \Big[\phi^2(q_k) w\Big(s_\ell + 2^{-j(n)} q_k\Big) - \phi^2(q) w\Big(s + 2^{-j(n)} q\Big)\Big] dq ds$$

and

$$S_2 = -\int_{R_{j(n)} \setminus \cap_{k,\ell} T_{k,\ell}} \phi^2(q) w\Big(s + 2^{-j(n)} q\Big) dq ds.$$

It is easy to see that S_2 is bounded by $\mathcal{O}(2^{j(n)-n})$. Moreover, a Taylor expansion up to the order 1 of

$$\phi^2(q_k) w\Big(s_\ell + 2^{-j(n)} q_k\Big) - \phi^2(q) w\Big(s + 2^{-j(n)} q\Big)$$

yields the upper bound

$$\begin{aligned}|S_1 - I| \leq\ & 2^{-j(n)} C \sum_k (q_{k+1} - q_k) \phi^2(q_k) \\ & + D\, 2^{j(n)-n} \sum_k (q_{k+1} - q_k)\Big(\phi^2(q_k) + (\phi^2)'(q_k)\Big) + o(1),\end{aligned}$$

where C and D are generic constants. The desired result follows now from the fact that both ϕ_2 and $(\phi_2)'$ are integrable.

The study of the asymptotic distribution is facilitated with the following notation.

Let $\boldsymbol{x} = (x_j)$ and $\boldsymbol{y} = (y_j)$ be two sequences of real numbers, and let $[\boldsymbol{x}, \boldsymbol{y}]_n = n^{-1} \sum_{j=1}^n x_j y_j$. If $(\boldsymbol{x}, \boldsymbol{y})_n$ converges to a real number its limit $[\boldsymbol{x}, \boldsymbol{y}]$ will be called the tail product of $\boldsymbol{x}$ and $\boldsymbol{y}$. Let h and w be two sequence valued functions on a compact subset T of a Euclidian space. If $[h(\alpha), w(\beta)]_n$ converges to $[h(\alpha), w(\beta)]$ uniformly for all α and β in T, $[h, w]$ will denote the function on $T \times T$ which takes (α,β) into $[h(\alpha),w(\beta)]$. This function will be called the tail cross product of h and w. Note that if in addition the components of h and w are continuous then $[h,w]$, as a uniform limit of continuous functions, is also continuous.

Proof of Theorem 3.2. Recall that $\hat{g}_n = P_{V_{j(n)}}\hat{f}_n$ and $N = 2^n$. We first study the convergence of the sequence $\hat{f}_n$.

By definition,

$$\hat{f}_n - E(\hat{f}_n) = 2^{-n/2}\sum_{k=0}^{2^n-1}\epsilon_k\phi_{n,k}(t).$$

Now, for any $t \in [0,1]$ and any $n > 0$, we have

$$[\phi_{n,\cdot}(t),\phi_{n,\cdot}(t)]_n = \frac{1}{2^n}\sum_{k=0}^{2^n-1}\phi^2_{n,k}(t) = \sum_{k=0}^{2^n-1}\phi^2(2^nt-k) \leq \sum_{k=0}^{2^n-1}\frac{1}{(1+|2^nt-k|)^2}.$$

The last inequality follows by the assumed r-regularity of the multiresolution analysis. Setting $a = 2^nt$ in the previous expression yields

$$\begin{aligned}[\phi_{n,k}(t),\phi_{n,k}(t)]_n &\leq \sum_{k\in\mathbb{Z}}(1+|a-k|)^{-2} \leq \sup_{0\leq a'\leq 1}\sum_{k\in\mathbb{Z}}(1+|a'-k|)^{-2}\\ &\leq \sum_{\ell=0}^{\infty}(1+\ell)^{-2} < \infty.\end{aligned}$$

Moreover, for any dyadic $t \in [0,1]$, the series $\sum_{k=0}^{2^n-1}\phi^2(2^nt-k)$ converges. It follows that the tail cross product of the sequence valued function $\phi_{\cdot,\cdot}$ with itself exist. We shall denote this limit by $[\phi(t),\phi(t)]$. By a strong law of large numbers (see e.g. Theorem 4 of Jennrich (1969, p.636)), it follows that $\hat{f}_n(t) \to E(\hat{f}_n)(t)$ almost surely, for all $t \in [0,1]$. Moreover, by the form of the central limit theorem given in Theorem 5 of Jennrich (1969), we also know that, for any dyadic t in $[0,1]$, the sequence of random variables $\sqrt{N}(\hat{f}_n(t) - E(\hat{f}_n)(t))$ converges in distribution to a centered Gaussian variable with variance $\sigma^2w^2(t)$ where $w^2(t) = [\phi(t),\phi(t)]$. Since $j(n)$ goes to infinity as $n \to \infty$ and since, by Theorem 4.1 of Walter (1992, p.337) the pointwise error of approximation of any function in $H^m(\mathbb{R})$ by its orthogonal projection on $V_{j(n)}$ is uniformly bounded in the sup-norm by $\mathcal{O}(2^{-j(n)(m-1/2)})$, for every t in $]0,1[$, the sequence $\hat{g}_n(t)$ inherits the asymptotic properties of $\hat{f}_n(t)$. Therefore, $\hat{g}_n(t) \to E(\hat{g}_n)(t)$ almost surely as $n \to \infty$ and $\hat{g}_n(t)$ is asymptotically normal, with the same asymptotic distribution as that of $\hat{f}_n(t)$.

To end the proof it suffices now to control the local bias of the estimator. Note that, for any $t \in]0,1[$,

$$\begin{aligned}|E\{\hat{g}_n(t)\} - g(t)| &= |(P_{V_{j(n)}}\Pi_nf)(t) - f(t)|\\ &\leq |(P_{V_{j(n)}}\Pi_nf)(t) - (P_{V_{j(n)}}P_nf)(t)| + |((I - P_{V_{j(n)}})f)(t)|.\end{aligned}$$

With similar arguments as for the proof of Theorem 3.1, it is easy to see that the first term of the right-hand side is uniformly bounded by an upper bound of order $\mathcal{O}(2^{-j(n)\min([m],r)})$. The second term is equal to

$$((I - P_{V_{j(n)}})f)(t) = \sum_{j\geq j(n)} \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}(t).$$

The m-smoothness of f, together with the regularity of ψ, implies that the uniform upper bound for $< f, \psi_{j,k} >$ given by Eq. (2.7) holds. Moreover, since ψ is also compactly supported, the number of nonzero $\psi_{j,k}(t)$'s for $t \in]0,1[$ is of the order $\mathcal{O}(2^j)$. Thus,

$$\begin{aligned} |\sum_{j\geq j(n)} \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}(t)| &\leq C \sum_{j\geq j(n)} 2^{-j(n)\min([m],r)+1/2} 2^j \\ &= \mathcal{O}(2^{-j(n)\min([m],r)-1/2}), \end{aligned}$$

which concludes the proof.

Acknowledgments

We wish to thank I. Daubechies, P. Auscher and A. Cohen for providing us with a preprint of their paper prior to publication. We also owe thanks to the referee whose valuable comments and recommendations helped to improve this article. Special thanks are due to W. Press for providing the scientific community the fast wavelet transform subroutines.

References

Auscher, P. (1992). Wavelets with boundary conditions on the interval. In *Wavelets: A tutorial in theory and applications* (Edited by C. K. Chui), 217-236.

Beylkin, G., Coifman, R. and Rokhlin, V. (1991). Fast wavelet transforms and numerical algorithms. *Comm. Pure Appl. Math.* **44**, 141-183.

Breiman, L. and Peters, S. (1992). Comparing automatic smoothers. *Internat. Statist. Rev.* **60**, 271-290.

Chui, C. K. (1992). *An Introduction to Wavelets.* Academic Press, New York.

Cohen, A. (1990). *Ondelettes, Analyses Multirésolutions et Traitement Numérique du Signal.* Thesis, Université Paris IX.

Cohen, A. and Daubechies, I. (1991). Orthonormal bases of compactly supported wavelets. Better frequency resolution. AT & T Bell Laboratories, preprint.

Cohen, A., Daubechies, I. and Vial, P. (1992). Wavelets and fast wavelet transform on the interval. AT & T Bell Laboratories, preprint.

Cox, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16**, 713-732.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41**, 909-996.

Daubechies, I. (1990). Orthonormal bases of compactly supported wavelets II. Variation on a theme. Preprint, submitted to *SIAM Journal Math. Anal.*

Daubechies, I. (1992). *Ten Lectures on Wavelets.* CBMS-NSF regional conferences series in applied mathematics, SIAM, Philadelphia, Pennsylvania.

De Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

Donoho, D. L. and Johnstone, I. M. (1992). Minimax estimation via Wavelet Shrinkage. Technical Report, Stanford University.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

Eubank, R. L., Hart, J. D. and Speckman, P. (1990). Trigonometric series regression estimators with an application to partially linear models. *J. Multivariate Anal.* **32**, 70-83.

Genon-Catalot, Laredo C. and Picard, D. (1990). Estimation non paramétrique de la variance d'une diffusion par méthodes d'ondelettes. *C. R. Acad. Sci. Paris* **311**, 379-382.

Hall, P. (1983). Measuring the efficiency of trigonometric series estimates of a density. *J. Multivariate Anal.* **13**, 234-256.

Hall, P. and Koch, I. (1992). On the feasibility of cross-validation in image analysis. *SIAM J. Appl. Math.* **52**, 292-313.

Hall, P. and Titterington, D. M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34**, 429-440.

Ibragimov, I. A. and Khas'minskii R. Z. (1982). Bounds for the risks of non-parametric regression estimates. *Theory Probab. Appl.* **27**, 84-99.

Istas, J. (1992). *Statistique des Processus Gaussiens Stationnaires Continus par Méthodes D'ondelettes.* Thesis, Université Paris VII, Paris.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40**, 633-643.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence* **11**, 674-693.

Meyer, Y. (1990). *Ondelettes et Opérateurs.* Hermann, Paris.

Meyer, Y. (1991). Ondelettes sur l'intervalle. *Rev. Math. Iberoamer.* **7**, 115-133.

Müller, H. G. (1988). *Nonparametric regression analysis of longitudinal data.* Lecture Notes in Statist. **46**, Springer-Verlag, Berlin.

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13**, 984-997.

Press, H. W. (1991). Wavelet transforms. Technical Report 3184, Harvard-Smithsonian center for Astrophysics, Cambridge.

Rafajlowicz, E. (1987). Nonparametric orthogonal series estimators of regression: A class attaining the optimal convergence rate in L_2. *Statist. Probab. Lett.* **5**, 219-224.

Rutkowski, L. (1982). On system identification by nonparametric function fitting. *IEEE Trans. Automat. Control* **27**, 225-227.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040-1053.

Wahba, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45**, 133-150.

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS-NSF regional conferences series in applied mathematics, SIAM, Philadelphia, Pennsylvania.

Walter, G. G. (1992). Approximation of the delta function by wavelets. *J. Approx. Theory* **71**, 329-343.

Laboratoire de Statistique et Modélisation Stochastique, IMAG-LMC, B.P. 53 X, 38042 Grenoble Cedex, France.

(Received January 1993; accepted December 1993)