Wavelets in Statistics: A Review

A. Antoniadis*
Laboratoire IMAG-LMC,
University Joseph Fourier
BP 53
38041 Grenoble Cedex 9
France

February 1997

Abstract

The field of nonparametric function estimation has broadened its appeal in recent years with an array of new tools for statistical analysis. In particular, theoretical and applied research on the field of wavelets has had noticeable influence on statistical topics such as nonparametric regression, nonpararametric density estimation, nonparametric discrimination and many other related topics. This is a survey article that attempts to synthetize a broad variety of work on wavelets in statistics and includes some recent developments in nonparametric curve estimation that have been omitted from review articles and books on the subject. After a short introduction to wavelet theory, wavelets are treated in the familiar context of estimation of "smooth" functions. Both "linear" and "nonlinear" wavelet estimation methods are discussed and cross-validation methods for choosing the smoothing parameters are addressed. Finally, some areas of related research are mentioned, such as hypothesis testing, model selection, hazard rate estimation for censored data, and nonparametric change-point problems. The closing section formulates some promising research directions relating to wavelets in statistics.

Keywords and phrases: Wavelets, multiresolution analysis, nonparametric curve estimation, density estimation, regression, model selection, orthogonal series, thresholding, cross-validation, shrinkage, denoising.

* Research supported by the IDOPT project (CNRS-INRIA-UJF-INPG).

1 Introduction

Wavelet theory has provided statisticians with powerful new techniques for nonparametric inference by combining recent advances in approximation theory with insights gained from applied signal analysis. When faced with the problem of recovering a 'piecewise' smooth function when only noise measurements are available, wavelet smoothing methods provide a natural and flexible approach to the estimation of the true function due their outstanding ability and efficiency to respond to local variations without allowing pathological behavior.

This article surveys recent developments and applications of wavelets in nonparametric curve estimation, as well as topics that were omitted from previous review articles and books. Both "linear" and "nonlinear" wavelet estimation methods are presented and the relative advantages and disadvantages of each method are discussed. Our exposition assumes no prior knowledge of the theory of wavelets, and we briefly develop all the necessary tools, under minimal conditions. As in almost all nonparametric smoothing methods, there are some smoothing parameters which determine how much the data are smoothed to produce the estimate. Automatic choices of these parameters by cross-validation methods are addressed.

Our present discussion is organized as follows: Section 2 deals with the fundamentals of wavelet theory. It contains a short overview of the basic definitions and the main properties of wavelets that will be used throughout this article. The next section focus on linear wavelet estimators in univariate regression and density estimation while Section 4 is devoted to nonlinear wavelet smoothers. Cross-validation methods for selection of the smoothing parameters are also discussed. In Section 5 we discuss the use of wavelets in a variety of other statistical problems such as model selection, hazard rate estimation for censored data, and nonparametric change-point problems as well as their use in time series analysis. In the concluding section that closes this article we try to identify a number of challenging open problems and some promising research directions. Note that the references, though numerous, should not be regarded as exhaustive.

2 Some background on wavelets

In this section we give a brief exposition of the relevant aspects of wavelet theory that will be used in the sequel and try to explain why wavelets are desirable in nonparametric curve smoothing. For precise mathematical statements, clear definitions and detailed expositions we refer the reader to Meyer [62], Mallat [58], Daubechies [27], Chui [24], Wickerhauser [92], Cohen and Ryan [26] and Holschneider [52].

2.1 Wavelet analysis

Wavelet analysis requires a description of two basic functions, the *scaling function* $\varphi(x)$ and the *wavelet* $\psi(x)$. The function $\varphi(x)$ is a solution of a two-scale difference equation

$$\varphi(x) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_k \varphi(2x-k) \tag{1}$$

with normalization $\int_{\mathbb{R}} \varphi(x)\,dx = 1$. The function $\psi(x)$ is defined by

$$\psi(x) = \sqrt{2}\sum_{k\in\mathbb{Z}} (-1)^k h_{1-k} \varphi(2x-k). \tag{2}$$

The coefficients h_k are called the *filter coefficients*, and it is through careful choice of these that wavelet functions with desirable properties can be constructed.

A wavelet system is the infinite collection of translated and scaled versions of φ and ψ defined by:

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

Some additional conditions on the filter coefficients imply that $\{\psi_{j,k},\ j,k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$, and $\{\varphi_{j,k},\ k \in \mathbb{Z}\}$ is an orthonormal system in $L^2(\mathbb{R})$ for each $j \in \mathbb{Z}$;

A key observation of Daubechies ([27]) is that it is possible to construct finite-length sequences of filter coefficients satisfying all of these conditions, resulting in compactly supported φ and ψ that have space-frequency localization (this localization allows parsimonious representation for a wide set of different functions in wavelet series). The derived wavelet basis is well localized in space since the total energy of a wavelet is restricted to a finite interval. Frequency localization simply means that the Fourier transform of a wavelet is localized, i.e., a wavelet mostly contains frequencies from a certain frequency band. The Heisenberg uncertainty principle puts a lower bound on the product of space and frequency variance. The decay towards high frequencies corresponds to the smoothness of the function. The smoother the function, the faster the decay. If the decay is exponential, the function is infinitely many times differentiable. The decay towards low frequencies corresponds to the number of vanishing moments of the wavelet. A wavelet ψ has N vanishing moments in case

$$\int_{\mathbb{R}} x^p \psi(x)\,dx = 0,$$

for $0 \le p < N$. Thinking of "frequency localization" in terms of smoothness and vanishing moments, allows a generalization of this notion to settings where no Fourier transform is available.

In technical terms, a scaling function φ is said to be r-regular ($r \in \mathbb{N}$) if for any $\ell \leq r$ and for any integer k one has

$$\left| \frac{d^\ell \varphi}{dx^\ell} \right| \leq C_k (1 + |x|)^{-k}$$

where C_k is a generic constant depending only on k.

We assume throughout that φ is r-regular for some $q \in \mathbb{N}$. Of course the primary wavelet inherits the regularity of the scaling function. Moreover if ψ is regular enough, the resulting wavelet orthonormal basis provides unconditional bases for a wide set of function spaces, such as Besov or Triebel spaces, see Meyer [62].

The wavelet representation of a function $g \in L^2(\mathbb{R})$ is

$$g = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} w_{j,k} \psi_{j,k} \tag{3}$$

where the wavelet coefficients $w_{j,k}$ are given by $w_{j,k} = \int_{\mathbb{R}} g(t) \psi_{j,k}(t) dt$.

Typically we want algorithms with linear or linear-logarithmic complexity to pass between a function g and its wavelet coefficients $\mathbf{w}$. Such algorithms are referred to as a *fast wavelet transform*. Fast wavelet transforms are often obtained through multiresolution analysis, a a framework developed by Mallat [58], in which the wavelet coefficients $< g, \psi_{j,k} >$ of a function g for a fixed j describe the difference between two approximations of g, one with resolution 2^j, and one with the coarser resolution 2^{j-1}.

A multiresolution analysis (or approximation) of $L^2(\mathbb{R})$ 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 they have intersection that is trivial and union that 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,

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

An orthonormal basis of V_j, the approximation space with resolution 2^{-j} is then given by the family $\{\phi_{j,k} : \; k \in \mathbb{Z}\}$. The orthogonal projection of a function $f \in L^2(\mathbb{R})$ into V_j is given by

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

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 r-regular.

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 $L^2(\mathbb{R})$, 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 the wavelet ψ that is associated to φ through equation (2).

According to the above, if g represents a square integrable function, it can be represented by

$$g = \sum_{k\in\mathbb{Z}} c_{j_0,k}\varphi_{j_0,k} + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k} \tag{4}$$

where j_0 represents a “coarse” level of approximation. The first part of (4) is the projection $P_{j_0}g$ of g onto the coarse approximating space V_{j_0}, and the second part represents the details.

One can associate to the projector P_j onto V_j its integral kernel defined by:

$$g \to P_j(g) = \int_{\mathbb{R}} E_j(\cdot, t)g(t)\,dt = \text{projection of } g \text{ onto } V_j,$$

where $E_j(x,y) = 2^j \sum_{k\in\mathbb{Z}} \phi_{j,k}(x)\phi_{j,k}(y)$. It is easy to see that $E_j(x,y) = 2^j E_0(2^j x, 2^j y)$ and that $E_0(x+k, y+k) = E_0(x,y)$ for $k \in \mathbb{Z}$. Obviously, E_0 is not a convolution kernel, but the regularity of φ and ψ implies that it is bounded above by a convolution kernel, that is $|E_0(x,y)| \leq K(x-y)$ where K is some positive, bounded, integrable function satisfying moment conditions, see Meyer ([62], p. 33).

The main properties and the key to applications of wavelets and multiresolution analyses are their powerful approximating qualities, i.e., they provide accurate approximations of a function using only a relatively small fraction of the coefficients. Given that the norm of a function g usually depends only on the absolute value of its wavelet coefficients, one can show (see Devore *et al.* [31]) that the best approximation of a function g with M coefficients, is obtained by

$$g_M = \sum_{j,k\in\Lambda_M} w_{j,k}\psi_{j,k}$$

where Λ_M contains the indexes of the M largest in absolute value coefficients. Note that this approximation is nonlinear. The speed of convergence of this approximation as we add more terms quantifies the approximation properties of g_M. This is given by the largest positive exponent α for which

$$\|g - g_M\| = \mathcal{O}(M^{-\alpha}). \tag{5}$$

The question on how to find α has been extensively studied in the area of nonlinear approximation and smoothness spaces. The main result (see Devore *et al.* [31]) says that if the normed space to which g belongs is a Besov space of smoothness index α, then

(5) holds. We will give a precise mathematical definition of Besov spaces in the next subsection. But to gain some intuition on why they are important, note that functions that are only *piecewise* smooth still belong to Besov spaces with high smoothness index. For such functions it is known that Fourier-based methods give very slow convergence rates ($\alpha = 1$). Therefore, wavelets are optimal bases for compressing and recovering functions in such spaces.

In many practical situations, the functions involved are only defined on a compact interval, such as the interval $[0, 1]$, and to apply wavelets requires some modifications. Cohen *et al.* [25] have obtained the necessary boundary corrections to retain orthonormality, and their wavelets on $[0, 1]$ also constitute unconditional bases for the Besov spaces on the interval with an associated multiresolution analysis structure. For the phenomena that we wish to present in the following sections one may work with such wavelets without altering the results.

2.2 Besov spaces on the interval

In this subsection we shall only mention the minimum aspects of the Besov spaces on the interval to be explicitly invoked in the sequel. For a more detailed study we refer to Triebel [81].

We restrict the consideration to the range of parameters $1 \leq p, q \leq \infty$, $s > 0$ and denote the respective Besov space by $B^s_{pq} = B^s_{pq}(R)$. If the wavelet ψ has regularity $r > s$ (more precisely, if $\psi \in B^r_{1\infty} \cap B^r_{\infty\infty}$), then $\{\varphi_{0,k}, \psi_{j,k}; k \in \mathbb{Z}, j \geq 0\}$ is a Riesz basis simultaneously for all $B^s_{pq}(R)$, $1 \leq p, q \leq \infty$, $0 < s < r$, so that for $g \in B^s_{pq}(R)$

$$g(t) = \sum_{k \in \mathbb{Z}} \alpha_{0k} \varphi_{0,k}(t) + \sum_{j=0}^{\infty} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{j,k}(t)$$

is always convergent in the norm topology of the space and

$$J_{spq}(\alpha, \beta) = \|\alpha_{0\cdot}\|_p + \left(\sum_{j=0}^{\infty} (2^{j(s+(1/2)-(1/p))} \|\beta_{j\cdot}\|_p)^q \right)^{1/q} \tag{6}$$

is an equivalent norm in B^s_{pq}. Here the notation

$$\|\alpha_{0\cdot}\|_p = \left(\sum_{k \in \mathbb{Z}} |\alpha_{0k}|^p \right)^{1/p} \qquad \|\beta_{j\cdot}\|_p = \left(\sum_{k \in \mathbb{Z}} |\beta_{jk}|^p \right)^{1/p}$$

has been used. Hence, for the above range of the parameters, the Besov space $B^s_{pq}(R)$ can be defined as $B^s_{pq}(R) = \{g \in L^p(\mathbb{R});\ J_{spq}(g) < \infty\}$.

An important fact is that Besov spaces can also be defined on the interval $[0, 1]$ (see Triebel [81]). For the considered range of parameters p, q, s all Besov spaces over $\mathbb{R}$ and

$[0,1]$ are continuously embedded in $L_{1,loc}$, hence, consist of regular distributions only, and elements of $B^s_{pq}([0,1])$ are obtained by taking the usual pointwise restriction on $[0,1]$ of a function defined Lebesgue-a.e. on $\mathbb{R}$.

When restricted to an interval, the Besov norm for a function $g \in B^s_{pq}([0,1])$ is related to the sequence space norm of the wavelet coefficients as follows

$$J_{spq}(\alpha, \beta) = \|\alpha_{j_0 \cdot}\|_p + \left(\sum_{j=0}^{\infty} (2^{j(s+(1/2)-(1/p))} \|\beta_{j \cdot}\|_p)^q \right)^{1/q} \tag{7}$$

where $\alpha_{j_0 \cdot} = \{\alpha_{j_0 k}; k = 0, \ldots, 2^{j_0} - 1\}$ are the coarse scale coefficients obtained from the wavelet transform on the interval.

The Besov scales include, in particular, the well-known Sobolev and Hölder scales of smooth functions H^m and C^s (B^m_{22} and $B^s_{\infty\infty}$ respectively), but in addition less traditional spaces, like the space of functions of bounded variation, sandwiched between B^1_{11} and $B^1_{1\infty}$. The latter functions are of statistical interest because they allow for better models of spatial inhomogeneity (e.g. Meyer [62], Donoho & Johnstone [35]).

2.3 Computational algorithms and the discrete wavelet transform

An algorithm described in Daubechies and Lagarias ([28], p. 17) (the cascade algorithm) allows the construction of orthogonal compactly supported wavelets as limits of step functions which are finer and finer scale approximations of φ. The algorithm is easy to implement on a computer and converges quite rapidly. Given a finite sequence of filter coefficients, $h_0, \ldots, h_N$, define the linear operator A by

$$(Aa)_n = \sum_{k \in \mathbb{Z}} h_{n-2k} a_k, \qquad a = (a_k)_{k \in \mathbb{Z}}$$

where it is understood that $h_k \equiv 0$ if $k < 0$ or $k > N$. Define $a^j = A^j a^0$, where $(a^0)_0 = 1$ and $(a^0)_k = 0$ for $k \neq 0$. Set

$$\varphi_j(x) = \sum_{k \in \mathbb{Z}} a^j_k \chi(2^j x - k), \tag{8}$$

where χ is the indicator function of the interval $[-\frac{1}{2}, \frac{1}{2}[$. Under certain conditions (see Daubechies [27]), the sequence of functions φ_j converges pointwise to a limit function φ that satisfies the two-scale difference equation (1). Note that the projection integral kernel E_j can be written as

$$E_j(t,s) = 2^j \sum_{k \in \mathbb{Z}} \varphi(2^j t - k) \varphi(2^j s - k).$$

When φ has compact support then this is a finite sum, each term of which can be evaluated by the cascade algorithm.

The following weights will be useful for some of the estimators that we are going to consider later. If A_i denotes a bounded interval to evaluate the weights $\int_{A_i} E_j(t,s)\,ds$ one can employ an integrated version of (8):

$$\int_u^v \varphi_j(x)\,dx = \sum_{k\in\mathbb{Z}} a_k^j \int_u^v \chi(2^j x - k)\,dx.$$

The sequence $\int_u^v \varphi_j(x)\,dx$ converges to $\int_u^v \varphi(x)\,dx$ for each $u < v$.

If the projection of a square integrable function g onto a fine multiresolution space V_n is known, it can be written as

$$P_{V_n} g = \sum_{k\in\mathbb{Z}} c_{n,k}\varphi_{n,k}$$

Given a lower resolution $J_0 < n$, the projection $P_{V_n} g$ can be decomposed as

$$P_{V_n} g = \sum_{k\in\mathbb{Z}} c_{J_0,k}\varphi_{J_0,k} + \sum_{j=J_0}^{n} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}.$$

Due to the multiresolution analysis structure, given the V_N coefficients $c_{n,k}$, we find $c_{J_0,k}$ and $d_{j,k}$ by the following recursive formulas:

$$c_{j-1,k} = \sum_m h_{m-2k} c_{j,m}, \quad d_{j,k} \sum_m g_{m-2k} c_{j,m}, \quad j = n, \dots J_0 + 1,$$

where $g_k = (-1)^k h_{1-k}$. The above computations can be summarized as follows: Let $\mathbf{f} = (f_1, \dots, f_n)$ be an element of the Hilbert space $\ell_2(n)$ of all square summable sequences of length n. The discrete wavelet transform of $\mathbf{f}$ is an $\ell_2(n)$ sequence

$$\beta = (G\mathbf{f}, GH\mathbf{f}, \dots, GH^{J_0-1}\mathbf{f}, H^J\mathbf{f})$$

where H and G are operators from $\ell_2(2M)$ to $\ell_2(M)$ (M is the length of the filter sequence h) defined coordinate-wise via

$$\text{for } a \in \ell_2(2M), \quad (Ha)_k = \sum_m h_{m-2k} a_m, \quad (Ga)_k = \sum_m g_{m-2k} a_m.$$

The discrete wavelet transformation described above is linear and orthonormal and can be represented in matrix form. Given the lowpass filter coefficients $\{h_k\}$ one can write the DWT-transformation matrix $\mathcal{W}_{n,J_0}$ explicitly. In terms of $\mathcal{W}_{n,J_0}$, $\beta = \mathcal{W}_{n,J_0}\mathbf{f}$ and $\mathbf{f} = \mathcal{W}_{n,J_0}^T\beta$. If $n = 2^J$ for some positive J, both DWT and inverse DWT are performed by Mallat's [58] fast algorithm that requires only $\mathcal{O}(n)$ operations and is available in several standard implementations, for example in the S-plus packages `WaveThresh` (Nason & Silverman [67]) or `S+Wavelets` (Bruce & Gao [18]) or in the Matlab package `WaveLab` (Buckheit *et al.* [21]).

3 Linear wavelet methods for curve estimation

Among the first to consider (linear) wavelet methods in statistics are Doukhan and Léon [40], Antoniadis and Carmona [6], Kerkyacharian and Picard [56] and Walter [87] for density estimation and Doukhan and Léon [40], Antoniadis, Grégoire and McKeague [7] for nonparametric regression. In the following subsection we will address first the performance of such wavelet estimators in the case of a single model for nonparametric regression in close analogy with the classical theory of curve estimation. However the case of nonparametric density estimation is important and is addressed in its own right in subsections of Sections 3 and 4.

3.1 Nonparametric regression

Consider the following standard nonparametric regression model involving an unknown regression function g:

$$Y_i = g(X_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{9}$$

Two versions of this model are distinguished in the literature:

(i) the *fixed design* model in which the X_i's are nonrandom design points (in this case the X_i's are denoted t_i and taken to be ordered $0 \leq t_1 \leq \ldots \leq t_n \leq 1$), with the observation errors ϵ_i i.i.d. with mean zero and variance σ^2;

(ii) the *random design* model in which the (X_i, Y_i)'s are independent and distributed as (X, Y), with $g(x) = \mathbb{E}(Y|X = x)$ and $\epsilon_i = Y_i - g(X_i)$ (in this case let f denote the design density of the X_i's supposed to be bounded away from 0 and ∞).

In each case the problem is to estimate the regression function $g(t)$ for $0 < t < 1$.

In the context of non-uniform stochastic design there is a variety of ways to construct a wavelet estimator of the unknown mean function g. In this case, the basic wavelet estimator considered in Antoniadis *et al.* [7] is of the product of fg, which is then corrected by dividing by an estimator of the design density f which is constructed by a simple wavelet estimator or a kernel estimator. To simplify the exposition we will only review here the case of the fixed design model.

For the fixed design model, Antoniadis *et al.* [7] propose the estimator:

$$\hat{g}(t) = \sum_{i=1}^{n} Y_i \int_{A_i} E_J(t, s)\, ds, \tag{10}$$

where the $A_i = [s_{i-1}, s_i[$ are intervals that partition $[0, 1]$ with $t_i \in A_i$. This is a wavelet version of Gasser and Müller's [45] (convolution) kernel estimator or of Härdle's ([51], p.

51) orthogonal series estimator. It can be seen clearly that the kernel $E_J(t,s)$ is variable, since its form depends on t. This changing kernel allows the wavelet-based estimator to adapt itself automatically to local features of the data. The resolution level J acts as a tuning parameter, much as the bandwidth does for standard kernel smoothers. A key aspect of wavelet estimators is that the tuning parameter ranges over a much more limited set of values than is common with other nonparametric regression techniques. In practice, only a small number of values of J (say three or four) need to be considered. The problem of automatically selecting J is rather easier than the bandwidth selection problem for kernel estimators, since the bandwidth is essentially reduced to the form 2^{-J} where $J < \frac{1}{2}\log_2 n$. The selection rule used in Antoniadis *et al.* [7] is to choose J as the minimizer of the cross validation function

$$CV(J) = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{g}_{(i)}(t_i))^2,$$

where $\hat{g}_{(i)}(t)$ is the leave-one-out estimator obtained by evaluating $\hat{g}$ (as a function of J and t) with the ith data point removed. This gives reasonable results when applied to real and simulated data. In practice, for sample sizes between 100 and 200, they have found that it suffices to examine only $J = 3$, 4 and 5.

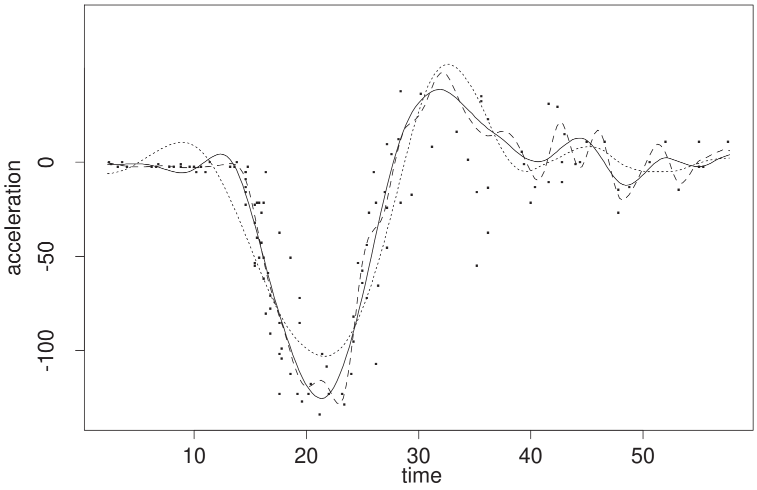

Figure 1: Plot of the motorcycle impact data together with the wavelet regression estimates $\hat{g}$ based on the Daubechies scaling function with filter of order 8 for $J = 3$ (dotted line), $J = 4$ (solid line) and $J = 5$ (dashed line). Cross validation selected the curve $J = 4$ as giving the best fit.

The wavelet estimator (10) has the advantage that the optimal asymptotic rates of mean square convergence hold for weaker conditions on the underlying function g ($g \in B^s_{22}$ for $s > 1/2$) than must be assumed in obtaining similar results for other types of smoothing. A disadvantage is that one can derive an asymptotic normality of the estimator at dyadic points only. At non-dyadic points, the asymptotic variance of the estimator, while remaining bounded, oscillates and asymptotic normality cannot be obtained. This phenomenon of erratic oscillations in the variance was also observed by Hall & Patil [48].

Figure 1 displays the motor-cycle impact data given in Härdle [51]. The observations consist of accelerometer readings taken through time in an experiment on the efficacy of crash helmets. For several reasons the time points are not regularly spaced. The cascade algorithm described in subsection 2.3 of Section 2 was used to compute the weights defining the estimator. The computational complexity of the algorithm for a general design is of the order $\mathcal{O}(n^2)$ and does not really take advantage of the fast discrete wavelet transform.

To obtain a faster computational algorithm in the fixed design model with equidistant nonrandom design points t_i within $[0, 1]$, Antoniadis [7] used another linear method that takes advantage of the DWT transform. Assuming that g is a function $[s]$-times continuously differentiable in $\mathbb{R}$, and such that its $[s]$th derivative satisfies a Lipschitz condition of order $s - [s]$, when $s > 1$ and $s \notin \mathbb{N}$, or that g is a function $s - 1$ times continuously differentiable in $\mathbb{R}$, and such that its $(s - 1)$th derivative satisfies a Lipschitz condition of order 1 when $s \in \mathbb{N}$, and taking the design points to be of the form $t_i = i\Delta$, with $\Delta = 2^{-N}$ for $i = 0, \ldots, n - 1$, where $n = 2^N$, he obtains a linear wavelet estimator that attains again the optimal asymptotic mean squared error rates and that is asymptotically normal at dyadic points. The approach is as follows.

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 $< g, \varphi_{n,k} >$ are given. However, when a function is given in sampled form there is no general method for deriving the coefficients $< g, \varphi_{n,k} >$. A first step towards the curve estimation method is to approximate the projection P_{V_n} by some operator Π_n in terms of the sampled values $g(\frac{k}{2^N})$ and to then derive a reasonable estimator of the approximation $\Pi_n g$. Using coiflets (see Daubechies [27]) that have L vanishing moments with $L > [s]$, such an estimator of $\Pi_n g$ is obtained by

$$\hat{g}_n(t) = \widehat{\Pi}_n g(t) = 2^{-N/2} \sum_{k \in \mathbb{Z}} Y_k \phi_{n,k}(t) = 2^{-N/2} \sum_{k=0}^{2^N - 1} Y_k \phi_{n,k}(t)$$

where the use of coiflets (wavelets for which the scaling function has integral 1 but admits zero moments) allows to approximate the coefficients $< g, \phi_{n,k} >$ by $2^{-N/2} g(\frac{k}{2^N})$ with an error $\mathcal{O}(2^{-N/2} 2^{-ns})$. In order to smooth correctly the data, to each each sample size $n = 2^N$ one then associates a resolution $j(n) = \log_2(n)/(1 + 2[s])$, and estimates the unknown function g by the orthogonal projection of $\hat{g}_n$ onto $V_{j(n)}$. Once again the parameter $j(n)$ governs the smoothness of the estimator.

Another class of linear wavelet estimators that used in literature is derived within the framework of regularization methods. Such estimators appear in Devore & Lucier [30], in a 1993 technical report of Antoniadis recently publised ([5]) and in Amato & Vuza [2]. In smoothing splines, a popular method for nonparametric regression problems such as the ones treated here a νth order smoothing spline $g_\lambda(x)$ is defined to be that function with square integrable νth derivative which minimizes over the Sobolev space $H^\nu[0,1]$ the "discrete" functional:

$$\frac{1}{n}\sum_{i=0}^{n-1}(Y_i - g(t_i))^2 + \lambda\int_0^1 \left(g^{(\nu)}(t)\right)^2 dt,$$

where $g^{(\nu)}$ indicates the νth derivative of g (see for example Wahba [86]). The "curvature" term $\int_0^1 \left(g^{(\nu)}(t)\right)^2 dt$ is a penalty term for lack of smoothness. Noting that the details of the wavelet coefficients of a function g at high resolution levels correspond to rough parts of the function, this problem can be generalized by seeking at the minimizer of an expression similar to

$$\|\widehat{\Pi}_n g - f\|^2_{L^2([0,1])} + \lambda J^p_{spp}(P_{V_{J_0}} f) \tag{11}$$

where J_0 is a coarse resolution level, $\widehat{\Pi}_n g$ is the interpolation estimate based on coiflets and J_{spp} is the equivalent norm of the Besov space $B^2_{pp}([0,1])$. In Antoniadis [5] as well as in Amato & Vuza [2] the particular choice $p = 2$ is made. This choice and the use of wavelet decompositions of f and g, allows one to find an optimal solution to the variational problem given in (11). This is possible because the norms $\|f\|^2_{L^2([0,1])}$ and $J^2_{s22}(f)$ can be determined simultaneously by examining the wavelet coefficients of f and g.

The solution to the variational problem is the function

$$g_\lambda = \sum_{k=0}^{2^{J_0}-1} c_{j_0,k}\varphi_{J_0,k} + \sum_{j=J_0}^{n}\sum_{k=0}^{2^j-1}\hat{\beta}_{j,k}\psi_{j,k}, \tag{12}$$

where $c_{j_0,k},\quad k = 0,\dots,2^{J_0}-1$ denote the empirical scaling coefficients of the discrete wavelet transform of the data vector $\mathbf{Y}$ and

$$\hat{\beta}_{j,k} = \frac{d_{j,k}}{1+\lambda 2^{2sj}},\quad j \geq j_0, k = 0,\dots,2^j-1$$

with $d_{j,k}$ being the empirical wavelet coefficients of $\mathcal{W}_{n,J_0}\mathbf{Y}$. The estimator g_λ appears as 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 $\widehat{\Pi}_n g$ through a low pass filter controlled by the parameters λ and s. It can be shown that, asymptotically as $n \to \infty$, if $s > 1/2$ and if $\lambda = \mathcal{O}\left(n^{-2s/(2s+1)}\right)$ then the mean squared error of the estimate g_λ behaves like $\mathcal{O}\left(n^{-2s/(2s+1)}\right)$.

For the practical application of the method, it is of course necessary to have an objective rule for the choice of the "coarse" resolution J_0 and the penalty parameter λ. As done for previous estimates the resolution level J_0 is chosen as $\log_2 n/(2s+1)$. Noting that the only term involving λ in the upper bound of the risk of the estimator g_λ is the expectation of $\|\Pi_n g - g_\lambda\|^2_{L^2([0,1])}$, the data-driven determination of λ is based on minimization of an appropriate estimate of this expectation and the knowledge of the noise variance σ^2. A possible choice for an estimate of the noise variance is the one suggested by Müller [64]:

$$\hat{\sigma}^2 = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} [Y_i - \frac{1}{2}(Y_{i-1} + Y_{i+1})]^2,$$

obtained by fitting constants to successive triples of the data. Lemma 1 of Müller shows $\hat{\sigma}^2$ is almost surely consistent and

$$|\hat{\sigma}^2 - \sigma^2| = O\Big(\frac{(\log n)^{\frac{1}{2}+\epsilon}}{n^{\frac{1}{2}}}\Big)$$

a.s. as $n \to \infty$ for any $\epsilon > 0$.

In spline smoothing, another method for providing, via a further approximation, an objective estimate for the minimizers of the integrated mean squared error of the estimates is generalized cross-validation. It is easy to see that the wavelet estimator introduced by regularization appears as a particular diagonal linear shrinker (see Donoho and Johnstone [33]). For each resolution $j \geq J_0$, the wavelet coefficients $d_{j,k}$ are shrinked by a factor $1/(1+\lambda 2^{2sj})$ which is level dependent. Assuming that g is a periodic function, Amato & Vuza [2] use $J_0 = 0$ and choose λ as the minimizer of the "GCV" function

$$V_n(\lambda) = \frac{\|(I_n - R_n(\lambda))\mathbf{Y}\|^2}{\left[\frac{1}{n-1}\mathrm{Tr}(I_n - R_n(\lambda))\right]^2}$$

where $R_n(\lambda)$ denotes this diagonal shrinkage operator.

The linear estimate suggested by Devore and Lucier [30] approximately minimizes the penalized functional (11) by a factor of 2 and is obtained by projecting the data vector on V_K where K is chosen such that

$$2^{2K} = \left(\frac{\|g\|^2_{B^s_{22}} 2^{2n}}{\sigma^2}\right)^{1/(s+1)}.$$

While a reasonable estimate of the unknown variance can be obtained, it seems more difficult to estimate the norm $\|g\|^2_{B^s_{22}}$ unless one has a precise upper bound for this norm.

To end this subsection, we briefly mention an interesting result of Donoho [32] where a linear wavelet estimator for an equidistant regression model with independent Gaussian errors is shown to attain the best asymptotic minimax rate $(n^{-1}\log n)^{\beta/(2\beta+1)}$ in the sup-norm for the class of functions

$$\{f \;:\; \sup_{x,y\in[0,1]} \frac{|f^m(x) - f^m(y)|}{|x-y|^\alpha} \leq L\} \cap \{f \;:\; \sup_{x\in[0,1]} |f(x)| \leq B\},$$

for $L > 0$, $B > 0$ and $1/2 < \beta = m + \alpha$ with $0 < \alpha \leq 1$. A more elementary and transparent proof of his result is also given in the paper of Oudshoorn [71]. This result is interesting because it shows that with linear wavelet estimators one can attain minimimax rates for sup-norm loss.

3.2 Nonparametric density estimation

The estimation of probability density functions from data is another example of basic problems in applied statistics. The idea to use a wavelet series expansion for the estimation of probability functions was first considered by Doukhan and Léon [40], Antoniadis and Carmona [6], Kerkyacharian and Picard [56] and Walter [87]. These works are motivated by the multiresolution decomposition associated with wavelet orthonormal bases and the localized character of wavelet expansions. Specialized versions of histograms constructed via Haar basis decompositions are described in Chapter 12 of Walter [88] and some interesting properties of such Haar-based estimators on the interval $[0,1]$ are discussed in Engel [41]. All these papers assume i.i.d. observations. In Antoniadis and Carmona [6], the unknown density belongs to the Sobolev space B^s_{22}, $s>0$, whereas in Kerkyacharian and Picard [56] f belongs to the Besov space B^s_{pp}, $p \geq 1$, $s>0$. The consistency rates obtained by these authors for linear wavelet estimators are respectively

$$\mathbb{E}\|f - \hat{f}_n\|_2^2 = \mathcal{O}(n^{-2s/(2s+1)})$$

and

$$\mathbb{E}\|f - \hat{f}_n\|_p^p = \mathcal{O}(n^{-ps/(2s+1)}).$$

A precise asymptotic expression for $\mathbb{E}\|f - \hat{f}_n\|_2^2$ in the case of $f \in B^s_{22}$, $s>0$ was given later by Masry [60] for more general stationary processes. In the i.i.d. case the asymptotic expression is shown to be exactly of the form

$$n^{-2s/(2s+1)}\mathbb{E}\|f - \hat{f}_n\|_2^2 \to 1$$

as $n \to \infty$.

With the basic introduction of wavelets in Section 2 we can examine more closely the way these estimators are constructed. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample and let f be the probability density of X_1 which is assumed to exist and satisfy $f \in L_2(\mathbb{R})$. Using an orthonormal wavelet basis, the wavelet representation of f is then given by

$$f = \sum_{k\in\mathbb{Z}} c_{J_0,k}\varphi_{J_0,k} + \sum_{j=J_0}^{\infty}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k},$$

where J_0 represents again a coarse level of approximation. The first issue in estimating f involves in estimating the coefficients in the above decomposition. This can be accomplished by using their empirical counterparts, that is

$$\hat{c}_{J_0,k} = n^{-1}\sum_{i=1}^{n}\varphi_{J_0,k}(X_i), \tag{13}$$

and

$$\hat{d}_{j,k} = n^{-1}\sum_{i=1}^{n}\psi_{j,k}(X_i). \tag{14}$$

Given these estimates, one then estimates f by

$$\hat{f}_n = \sum_{k\in\mathbb{Z}} \hat{c}_{J_0,k}\varphi_{J_0,k} + \sum_{j=J_0}^{J_1}\sum_{k\in\mathbb{Z}} \hat{d}_{j,k}\psi_{j,k}, \tag{15}$$

where $J_1 \geq J_0$ is a resolution suitably chosen. Note that the estimator $\hat{f}_n$ defined in (15) belongs to V_J with $J = J_1 + 1$ and can be written as

$$\hat{f}_n = \sum_{k\in\mathbb{Z}} \hat{c}_{J,k}\varphi_{J,k}. \tag{16}$$

For linear wavelet density estimators the smoothing parameter is the index J of the highest level to be considered. Several strategies for the automatic choice of the tuning parameter have been suggested in the literature. Walter [88] discusses an automatic algorithm to choose the most appropriate level J by using the integrated mean square error criterion

$$\text{IMSE} = \int \mathbb{E}\left(\hat{f}_n(t) - f(t)\right)^2 dx.$$

The algorithm begins by computing the $\hat{c}_{K,k}$ at a high and non optimal level, estimating the IMSE of the resulting estimate, and then recursively computing lower-level coefficients and the associated estimated error. The level J chosen by Walter is the one at which the estimated error increases most rapidly when moving from a level to the next coarser.

Another method that is considered to be optimal with respect to the IMSE criterion is the one discussed by Tribouley [82]. The choice is based on the cross-validation principle and results in the minimization with respect to J of the expression:

$$\text{CV}(J) = \sum_k \left[\frac{2}{n(n-1)}\sum_{i=1}^{n}(\varphi_{J,k}(X_i))^2 - \frac{n+1}{n^2(n-1)}(\sum_{i=1}^{n}\varphi_{J,k}(X_i))^2\right]^2.$$

For densities that are compactly supported within a known interval $]a, b[$ and that are continuously differentiable, a method for choosing J involving the Fisher functional of the density f, defined by

$$F(f) = \int_{\mathbb{R}} \frac{1}{f(t)}\left[\frac{d}{dt}f(t)\right]^2 dt$$

has been introduced recently by Vannuci and Vidakovic [83]. The "optimal" J is chosen such that the wavelet estimator $\hat{f}_n$ has an estimated Fisher information close to the theoretical minimal bound $4\pi^2/(b-a)^2$.

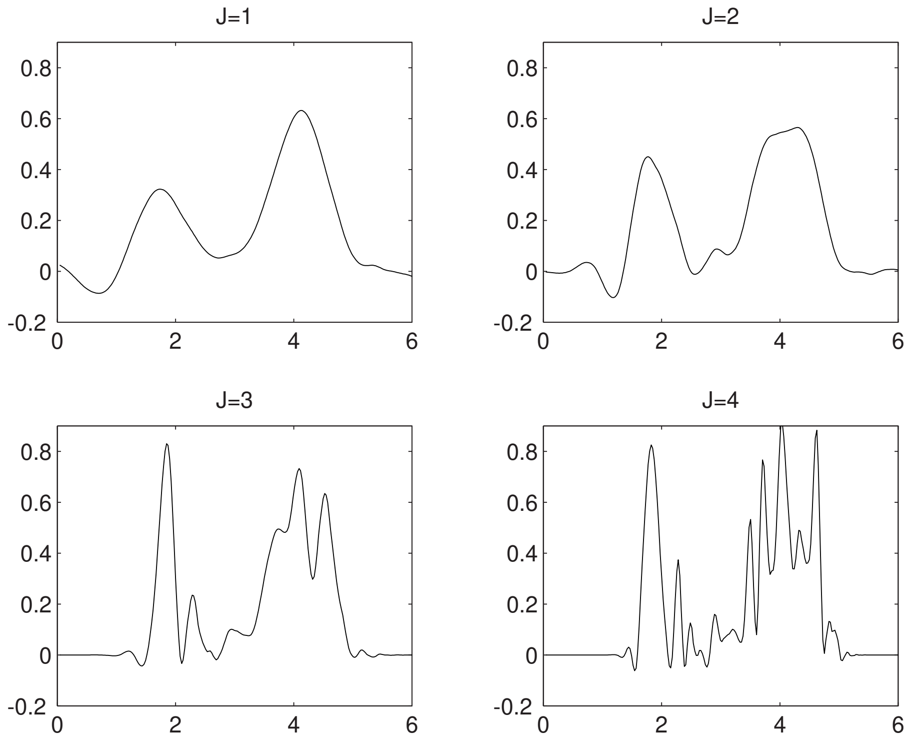

Figure 2: Linear wavelet-based density estimates for the duration times of eruption from the Old Faithfull geyser data set using Daubechies filters of order 5 and four choices of J.

Figure 2 displays smooth wavelet-based density estimates of the duration times of eruption from the Old Faithfull geyser in Yellowstone National Park. The Old Faithfull data set has been often used as a benchmark for density estimators.

The above "linear" method of viewing wavelet-based density estimators might not be seen so much as an alternative to the kernel approach but as a way of enhancing that technique. Indeed, the wavelet estimators described above are nothing else than generalized kernel estimators based on kernels of the form E_J. The resolution J permits a global level of smoothing in terms of the frequency of the scaling function and 2^{-J} is analogous to bandwidth for a kernel estimator. However, contrary to the case of classical kernel estimators, the terms representing bias and variance of wavelet-based density estimators oscillate erratically with a wavelength of the same order as 2^{-J}. Indeed, as proved by Hall & Patil [48], the classical pointwise bias and variance formulae,

$$\text{bias}(t) = \mathbb{E}\hat{f}_n(t) - f(t) \simeq a_1(t)2^{-Js}$$

and

$$\text{variance}(t) = \text{var}\{f_n(t)\} \simeq a_2(t)2^J/n$$

for smooth functions a_1 and a_2 are no longer valid. They are replaced by

$$\text{bias}(t) \simeq a_1(t)a_3(2^J t)2^{-Js}$$

and

$$\text{variance}(t) \simeq a_2(t)a_4(2^J t)2^J/n$$

for new nondegenerate functions a_3 and a_4. The erratic oscillations represented by $a_3(2^J t)$ and $a_4(2^J t)$ can be clearly observed in the estimates displayed in figure 2. One way to reduce these oscillations, thus resulting in a smaller mean-squared error, is to not insist on choosing a smoothing frequency that is a power of 2 for the wavelet estimator. Hall & Patil [48], suggest using the family of orthonormal scaling functions $\phi_k(x) = p^{1/2}\varphi(px+k)$ where $p > 0$ denotes an arbitrary positive number. It is easy to see that when $p = 2^J$ one has $\phi_k = \varphi_{J,k}$. This generalization permits a wider range of choices for the smoothing parameter in applications of curve estimation. A quantification of the advantages of non-integer resolution levels as well as some techniques for choosing the smoothing parameter by cross-validation as is done for kernel estimation is given in Hall & Nason [47]. When $p = p2^j$ the resulting estimator may be seen as a classical wavelet-based estimator applied to a preliminary binned data with binwidth proportional to p (see Antoniadis & Pham [8]). This is also the approach taken by by Antoniadis, Grégoire and Vial [10], to generalize the fast linear wavelet estimators to general design nonparametric regression and density estimation.

There is a potential problem in using wavelets for density estimation. When using general scaling functions there is no guarantee that the estimates are positive or integrate to 1. Indeed, it does happens that they are often negative in the tails of the distribution. Moreover there is no easy way to norm the wavelet estimator, except to numerically integrate the estimate in order to work out the norming constant. Walter [88] considers estimating the density function indirectly, by using wavelets to estimate the Fourier transform of the density, and then transforming back but he points out that the rate of convergence of such an estimate may be relatively slow.

Another approach used in the literature, that will be discussed further when nonlinear wavelet estimation methods will be presented, is to estimate the square root of the density and square back the estimate after. The idea of the above transformation can be found in Good & Gaskins [46] in the context of penalized likelihood methods. The condition $\int f(x)dx = 1$ becomes $\int(\sqrt{f}(x))^{1/2}dx = 1$, so that $\sqrt{f} \in L^2$. Pinheiro and Vidakovic [73] do exploit this idea of estimating the square root of the density in a wavelet setting, but, in order to get estimators of the needed wavelet coefficients they use a rough but consistent pre-estimator of the unknown density. There is no theoretical or convincing numerical evidence in their paper that optimal asymptotic rates can be obtained in this way.

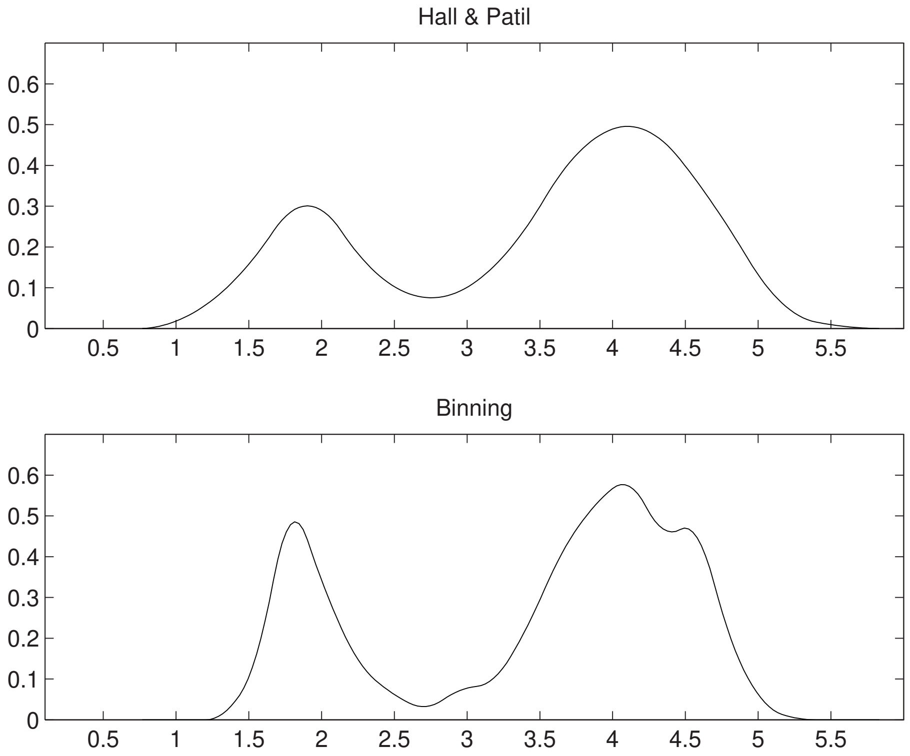

Figure 3: Linear wavelet-based density estimates for the duration times of eruption from the Old Faithfull geyser data set using Hall and Patil's scaling functions (left) and Antoniadis *et al.*'s pre-binning (64 bins), both based on Daubechies wavelet filters of order 4 adapted to the interval.

Figure 3 displays the linear estimates corresponding to the Hall & Patil approach as well as on the binning + smoothing approach of Antoniadis *et al.* To avoid negative values of the estimates Pinheiro and Vidakovic's [73] idea was used.

Along the same lines, but using a different approach and different estimators, is the research completed by Penev & Dechevsky [72]. Since their method deals principally with nonlinear thresholding methods, it will be discussed in the next Section.

4 Nonlinear wavelet methods for curve estimation

In the previous section the nonparametric estimation of regression functions and probability density functions has been restricted to the context of linear wavelet-based estimators. The application of these methods has provided only asymptotic upper bounds to the integrated squared error for functions that are traditionally smooth. For functions that might not be smooth in the classical sense, nonlinear wavelet-based estimation methods provide levels of smoothing which automatically adapt to local variations of roughness of the curve. Nonlinear wavelets methods in statistics were introduced by Donoho and Johnstone [33], [34], [36] and Donoho, Johnstone, Kerkyacharian and Picard [37], to cite only few of their papers. They permit two non-overlapping levels of smoothing, one global, via the frequency of the scaling function, and the other one local, via the scale of the wavelet function. This section deals with the ability of the nonlinear component of wavelet methods to adapt to local features of an unknown curve, and thus to correct for more erratic features of the curve no taken into account by the linear component.

4.1 Nonparametric regression

In this subsection, contrary to the assumptions used in the linear case, we restrict our attention to nonparametric regression models on the unit interval with an equidistant deterministic design and a Gaussian noise. Possible extensions that might be possible for a random design and other types of noise will be discussed later.

The paper by Donoho, Johnstone, Kerkyacharian and Picard [37] is perhaps the most significant paper from both a mathematical and practical point of view for the existence of nonparametric function estimators that behave in a (near) asymptotic optimal way simultaneously for a broad range of function spaces (Besov or Triebel spaces) not considered before in statistics and a variety of loss measures (L_p-losses) and whose definitions are independent of the set of function spaces considered. Mathematically it gives a unified treatment of optimal rates of convergence for nonparametric function estimation in a very general setup. This is achieved by using the approximating properties of wavelet bases and the close relation between the problem of minimax estimation and the theory of optimal recovery, a survey of which can be found in a paper by Michelli and Rivlin [63]. The connection with deterministic optimal recovery problems is obtained by means of a simple but powerful thresholding device on the empirical wavelet coefficients, which works reasonably well in practice. We refer the reader to the above papers for the exact assumptions and consistency rates, which are nearly optimal in the sense that they are equal to the optimal asymptotic rates up to a $\log n$ multiplicative factor.

Let us now further describe the regression model and the methods of estimation. The data are discrete and follow the fixed equidistant design regression model on $[0, 1]$:

$$Y_i = f(t_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n = 2^N,$$

where $t_i = i/n$, and the ϵ_i's, the noise in the observations, are i.i.d. $N(0, 1)$ random errors. To this data set, we apply the discrete wavelet transform $\mathcal{W}_{n,J_0} : \mathbb{R}^n \to \mathbb{R}^n$ for some $J_0 < N$ and for simplicity of exposition we will use a periodic version of the transform. Heuristically, the assumption made in the various papers cited above is that

$$2^{N/2} \int_{\mathbb{R}} \varphi((t - i/2^N)2^N) f(t)\, dt \simeq f(i/2^N)/2^{N/2},$$

if N is large. Such an assumption is reasonable when the function f is sufficiently smooth (see Section 3) but the approximation seems questionable when f belongs to classes of functions that may be not even continuous. To overcome this, one may think of the $2^{-N/2}Y_i$'s as noisy versions of the left hand side of the above formula. Let $\mathbf{c} = 2^{-N/2}\mathcal{W}_{n,J_0}\mathbf{Y}$ be the empirical scaling coefficients, let $\boldsymbol{\beta} = 2^{-N/2}\mathcal{W}_{n,J_0}\mathbf{f}$ and let $\mathbf{z} = 2^{-N/2}\mathcal{W}_{n,J_0}\boldsymbol{\epsilon}$. Since the transformation is linear one has, for $j = 0, \ldots, N-1$ and $k = 0, \ldots, 2^j - 1$:

$$c_{j,k} = \beta_{j,k} + 2^{-N/2}\sigma z_{j,k}, \tag{17}$$

and since it is orthonormal, the $z_{j,k}$ are i.i.d. $N(0,1)$. The respective mean squared errors in estimating the wavelet coefficients $\boldsymbol{\beta}$ of f or f are therefore the same. Now, for the large classes of functions considered, and with the use of sufficiently regular wavelets the vector $\boldsymbol{\beta}$ is generally sparse, i.e., relatively few components are large. The noise in the original sequence Y_i is spread out uniformly among all empirical wavelet coefficients. The heuristic idea underlying the Donoho-Johnstone procedure is to choose the set of coefficients that contain significant signal and to remove the noise component from the noisy coefficients. This is achieved by thresholding.

The thresholding estimator of the true coefficient $\beta_{j,k}$, $j \geq J_0$ is defined by

$$\hat{\beta}_{j,k} = \frac{\sigma}{\sqrt{n}}\eta_\lambda\left(\frac{\sqrt{n}c_{j,k}}{\sigma}\right) \tag{18}$$

where the function η_λ in (18) is either the *hard thresholding* function

$$\eta_\lambda^H(x) = \begin{cases} x, & \text{if } |x| > \lambda, \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

or the *soft thresholding* function

$$\eta_\lambda^S(x) = \begin{cases} x - \lambda, & \text{if } |x| > \lambda, \\ 0, & \text{if } |x| \leq \lambda, \\ x + \lambda, & \text{if } x < -\lambda. \end{cases} \tag{20}$$

Once the thresholding is performed, one applies the inverse empirical transform $\mathcal{W}_{n,J_0}^T$ to the estimated thresholded vector, obtaining the estimated regression curve $\hat{f}_n(t)$. The method is therefore simple and practical, with an algorithm that functions in order $\mathcal{O}(n)$ operations. The above arguments produce the regression estimator

$$\hat{f}_n = \sum_k c_{J_0,k}\varphi_{J_0,k} + \sum_{j \geq J_0}\sum_k \hat{\beta}_{j,k}\psi_{j,k}.$$

The first part of the right hand side is identical to the linear wavelet-based estimator studied in the previous Section. The second part enhances the linear estimator by incorporating through thresholding some wavelet terms.

Clearly, when using either type of wavelet thresholding, the choice of the cut-off resolution J_0 and of a threshold λ is a fundamental issue (see figure 4). The typical sparsity of the $\beta_{j,k}$ sequence ensures that most of the appropriately scaled coefficients $\sqrt{n}c_{j,k}/\sigma$ are essentially white noise. Motivated from the "large deviation" nature of the problem, Donoho and Johnstone suggest taking $\lambda = \sqrt{2\log n}$, named *universal threshold*. The procedure is proven to be asymptotically optimal for many classes of functions and makes no *a priori* assumptions on the particular class that f may belong, producing therefore an asymptotically adaptive estimator.

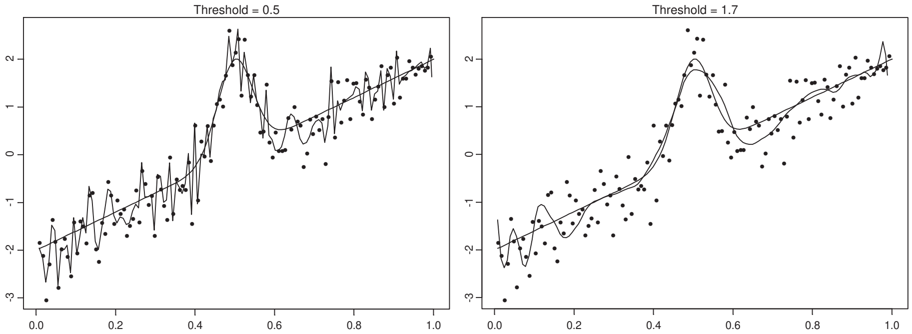

Figure 4: The effect of varying the threshold value on the resulting wavelet estimator from a simulated data set.

Another method for global thresholding proposed by Donoho and Johnstone [33] is labeled *minimax thresholding*. Briefly, the performances of shrinkage estimators are compared to the benchmark

$$B_n(\boldsymbol{\theta}, \sigma^2) = \sigma^2 + \sum_{i=1}^{m} \min(\theta_i, \sigma^2)$$

where, to simplify the notation, $\boldsymbol{\theta}$ denotes the m-dimensional vector of wavelet coefficients, which are observed with a white noise of variance σ^2. The above benchmark is derived by using the fact that if one has knowledge of the true coefficients, an ideal minimal mean-squared error estimator is obtained by setting a noisy coefficient to zero if the variance σ^2 of the noise is larger than the square of the true wavelet coefficient. Note that the benchmark $B_n(\boldsymbol{\theta}, \sigma^2)$ is small in comparison to $m\sigma^2$ (the total variance in the signal observed with noise) if $\boldsymbol{\theta}$ is sparse, and usually serves as a measure for the sparsity of $\boldsymbol{\theta}$. One of the most significant results of Donoho and Johnstone is that, in the case one observes a realization from a Gaussian vector m-dimensional vector $\mathbf{U} \sim \mathrm{N}_m(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_m)$, the following upper bound holds

$$\sup_{\boldsymbol{\theta} \in \mathbb{R}^m} \frac{\mathbb{E}\|\eta^S_{\lambda_n}(\mathbf{U}) - \boldsymbol{\theta}\|^2}{B_n(\boldsymbol{\theta}, \sigma^2)} \leq (1 + 2\log n),$$

when $\lambda_n = \sqrt{2\log n}$ is taken to be the universal threshold. It is also proven that the $2\log n$ factor cannot be improved, that is

$$\lim\,\inf_{n\to\infty}\frac{1}{2\log n}\inf_{\lambda}\sup_{\boldsymbol{\theta}}\frac{\mathbb{E}\|\eta^S_\lambda(\mathbf{U})-\boldsymbol{\theta}\|^2}{B_n(\boldsymbol{\theta},\sigma^2)}\geq 1.$$

In his thesis, Gao [43], proves similar results for i.i.d. variables with exponential tails. Recently, Averkamp and Houdré [14], obtained a stronger result of this type for a wider class of distributions. Using the above, when $\boldsymbol{\theta}$ is the vector of wavelet coefficients of the regression function, the the minimax thresholded wavelet estimator is obtained by computing the threshold λ^*_n that attains the bound

$$\inf_{\lambda}\sup_{\boldsymbol{\theta}}\frac{\mathbb{E}\|\eta^S_\lambda(\mathbf{U})-\boldsymbol{\theta}\|^2}{B_n(\boldsymbol{\theta},\sigma^2)}.$$

Note however that the above results cannot be applied to the empirical coefficients of a regression with non-normal errors since then the noise in the wavelet coefficients is no longer independent nor identically distributed. For some particular non-Gaussian regression models a possible approach, using some large deviation results, is one proposed by Neumann and Spokoiny [69], where a risk equivalence between some non-Gaussian regression models and Gaussian white noise models is established.

The threshold λ^*_n does not exist in analytical form but a numerical approximation for a range of sample sizes are given in Donoho and Johnstone [33]. For a given sample size, the optimal minimax threshold is typically smaller than the universal one, and thus results in less smoothing (see figure 5).

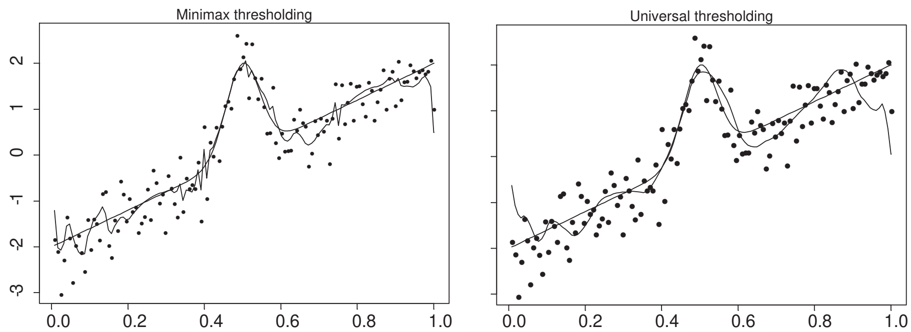

Figure 5: Minimax and universal thresholding applied to the simulated data set in figure 4.

Both thresholding rules require an estimate of the usually unknown variance σ^2. When it is known that the underlying regression function is Hölder continuous an estimator as the one described in Section 3 can be used. Donoho and Johnstone [33] propose a robust estimate of σ by taking the median absolute deviation of the coefficients at the finest level of the empirical decomposition

$$\hat{\sigma}=\frac{\text{median}(|c_{N-1,k}-\text{median}(|c_{N-1,k})|)}{0.6745}$$

since typically there is also some signal present even at the finest level.

The estimators described above, while applicable to a wide range of variable frequency curves, usually provide an excessive amount of smoothing when applied to curves that are piecewise smooth. Their mean-squared errors are asymptotically dominated by bias. To address this problem, Donoho and Johnstone [35] look at a variant with level-dependent thresholds. The method, called *Sureshrink* employs an unbiased risk estimation that is due to Stein [78] and is shown in Ogden [70] to be in relation with Akaike's information criterion (AIC), introduced by Akaike for times series modeling.

Hall and Patil ([48], [49], [50]) studied asymptotic wavelet shrinkage methods in nonparametric curve estimation from the different viewpoint of a fixed target function, as opposed to the minimax approach of Donoho *et al.*. In the case of functions that are smooth or piecewise smooth in the classical sense, using wavelet decompositions which allow non-integer resolution levels, already described in Section 3, they derive necessary and sufficient conditions on the asymptotic form of the threshold and smoothing parameters for their resulting curve estimator to achieve optimal mean square convergence rates.

Most of the methods and results described above are asymptotic in character. As with any asymptotic result, there remain doubts as to how well the asymptotics describe small sample behavior. These issues are addressed by Marron *et al.* [59] using the tools of exact risk analysis, which was developed in Gasser and Müller [45], and first applied to wavelet estimators by Antoniadis *et al.* [7]. Finite sample performance of thresholded wavelet estimators has also been studied by Bruce and Gao [19], where computationally efficient formulas for computing the exact pointwise bias, variance and L^2 risk of thresholded wavelet estimators in finite sample situations are derived, thus complementing the tools of simulation and asymptotic analysis. Comparing hard and soft shrinkage, hard shrink tends to have bigger variance (because of the discontinuity of the shrinkage function) and soft shrink tends to have bigger bias (because of shrinking all big coefficients towards 0 by λ). To remedy these drawbacks, and paralleling the choice of shrinkage functions with that of influence functions in robust statistics, Bruce and Gao [20] introduce a general semisoft shrinkage function

$$\eta_{\lambda_1,\lambda_2}(x) = \begin{cases} 0 & \text{if } |x| \leq \lambda_1 \\ \operatorname{sgn}(x)\frac{\lambda_2(|x|-\lambda_1)}{\lambda_2-\lambda_1} & \text{if } \lambda_1 < |x| \leq \lambda_2 \\ x & \text{if } |x| > \lambda_2, \end{cases}$$

that offers some advantages over both hard shrinkage (uniformly smaller risk and less sensitivity to small perturbations in the data) and soft shrinkage (smaller bias and smaller overall L^2 risk). A drawback of this semisoft rule is that it requires two thresholds, thus making threshold selection problems much harder and computationally more expensive for adaptive threshold selection.

One way to choose the thresholds is by generalized cross-validation proposed first for nonlinear wavelet series estimators by Weyrich and Warhola [91]. Recently, Jansen *et al.* [53] have shown that, under appropriate conditions, this generalized cross-validation choice is asymptotically optimal, in the sense of yielding asymptotically the threshold that minimizes the expected mean squared error. Other data-driven methods for the choice of the smoothing parameter(s) in thresholding wavelet estimators have also been proposed in the literature. For a detailed account and description of these methods the reader is referred to the papers by Nason ([66], [65]) or the book by Ogden [70].

4.2 Density estimation

Nonlinear wavelet-based density estimators in the i.i.d. setting were introduced by Johnstone *et al.* [54] and Donoho *et al.* [38] and parallel exactly the results obtained for the regression case, although the proofs are entirely different. For the appropriate compactly supported wavelet basis, they take the form

$$\hat{f}_n = \sum_{k \in \mathbb{Z}} \hat{c}_{J_0,k} \varphi_{J_0,k} + \sum_{j=J_0}^{J_1} \sum_{k \in \mathbb{Z}} \eta_{\lambda_j}(\hat{d}_{j,k}) \psi_{j,k}, \tag{21}$$

with properly chosen resolutions J_0, J_1 and level dependent thresholds λ_j. As pointed earlier, the estimator (21) may be seen as a coarse approximation of f at level J_0 plus some details that are added to improve the approximation. Using $J_1 = \log_2 n - \log_2(\log n)$, $\lambda_j = A\sqrt{(j - J_0)/n}$, where A is some constant and J_0 is chosen according to the regularity of ϕ and the sample size, the thresholded estimator $\hat{f}_n$ in the papers cited above (see also Delyon and Judistsky [29]) is shown to be asymptotical optimal in the sense that, for $s > 1/p$ and $p^{'} \geq (1 + 2s)p$, it attains the minimax $L_{p'}$ rate

$$\left(\frac{\log n}{n}\right)^{(s-1/p+1/p^{'})/(1+2(s-1/p))}$$

in the class of densities in the Besov space B^s_{pq} with $J_{spq}(f) \leq M$, where M is a given constant. This rate cannot be attained with linear methods. Note, however, that when $p^{'} \to \infty$, Masry [61] has shown that this rate is attained by linear estimators and thus nonlinear estimators do not improve the rate of convergence in this case.

We have already mentioned the approach taken by Penev and Dechevsky [72], to estimate first by wavelet methods the square root of the density before taking its square as the final estimate, in order to preserve the non-negativity while still retaining the asymptotic minimax properties. The nonlinearity of the estimates is justified by the fact that they assume that it is the square root of f that belongs to a Besov ball. However, they prove that there are some reasonable connections between Besov regularity of f and that of $\sqrt{f}$. The advantage of the estimate they propose is that it can be normed to integrate to 1 very easily without numerical integration.

Some data dependent methods for choosing J_0 (Tribouley [82]) and λ_j have been proposed by Pinheiro and Vidakovic [73] and more recently by Vannucci and Vidakovic [83].

5 Related topics

The regression models discussed in the previous sections involve additive white noise of constant level, no weighting and most of the time normality. Antoniadis and Lavergne [9] extend the linear wavelet-based methods to data with heteroscedastic noise. More recently, an extension of Donoho and Johnstone's wavelet shrinkage smoothing technique to handle data with heteroscedastic noise has been given by Gao [44]. Johnstone and Silverman [55] have considered the extension to more general noise models than the white noise model. When the noise is stationary, using appropriately chosen level dependent thresholds, they obtain asymptotic minimax results similar to the ones obtained for the white noise regression model. A simpler proof of the optimality of their thresholding procedure is given by Amato and Vuza [3]. Brillinger ([16], [17]) also presents some inferential aspects of the wavelet technique far a deterministic signal in the presence of additive stationary non necessarily Gaussian noise. Function estimation for nonparametric regression with long-range dependence errors is studied in Wang [90].

Wavelet versions of estimators of a hazard rate function in the context of inference for a counting process multiplicative intensity model have been studied by Antoniadis *et al.* [7]. See also Antoniadis, Grégoire and Nason [12] for a contribution to the methodology available for estimating the density and the hazard rate from randomly censored data.

The problem of estimating the log spectrum of a stationary Gaussian time series by wavelet thresholding techniques has been addressed by Gao [43] in his thesis. More generally Neumann [68] applied the thresholding procedure in the framework of spectral density estimation for a stationary, possibly non Gaussian time series. It has also been applied by von Sachs and Schneider [85] to the periodogram of a locally stationary process for the estimation of its evolutionary spectrum.

A generalization to the problem of recovering f from indirect data $\mathbf{Y} = K\mathbf{f} + \boldsymbol{\epsilon}$, where K is a known operator has been addressed by Kolaczyk [57] in the context of integration, fractional integration and tomography.

Since the basic aim of wavelet analysis is to represent a function as a linear superposition of wavelets centered on a sequence of time points, it forms a natural tool for the investigation of jump points in time varying functions observed with noise. Wavelet methods for detecting and locating the jump points can be found in Vercken and Potier [74], Wang [89] and more recently in Antoniadis and Gijbels [13] and the thesis of Raimondo [75].

Applications of wavelet decompositions in statistical hypothesis testing and model selection appear in particular Fan [42] and Antoniadis *et al.* [11]. Fan shows that traditional nonparametric tests have low power in detecting fine features such as sharp and short aberrant as well as global features such as high frequency components.These drawbacks are repaired via wavelet thresholding and the Neyman truncation test. Antoniadis *et al.* [11] discuss how to use wavelet decomposition to select a regression model. Their methodology relies on a minimum description length criterion which is used to determine the number of non-zero coefficients in the vector of wavelet coefficients. The developed model selection rule is then applied to testing for no effect in nonparametric regression and for martingale structure in time series.

To end this section, let us mention some bayesian methods that have been proposed recently for nonparametric curve estimation, since they offer an interesting and useful alternative to the methods discussed earlier. In Vidakovic [84], the wavelet coefficients $\beta_{j,k}$ in the decomposition (17) as well as the unknown standard deviation σ of the noise are assumed to be independent random variables with an imposed prior distribution. The posterior means of the wavelet coefficients have the shape of standard soft wavelet thresholding rules and are used to estimate the unknown curve. Other papers considering wavelet shrinkage or thresholding within a Bayesian framework are those by Clyde *et al.* [22], Chipman *et al.* [23]. Again, a prior distribution is imposed on wavelet coefficients of the unknown response function, and the function is estimated by computing the mean of the resulting posterior distribution of wavelet coefficients. Recent work on this direction has been done by Abramovich *et al.* [1], with a prior designed to capture the sparseness of the wavelet expansion and a Bayes rule corresponding to the posterior median. Moreover, in the last mentioned paper, the prior model for the underlying regression function is adjusted to give functions falling in any specific Besov space. In order to achieve this, a relation between the hyperparameters of the prior model and the parameters of the Besov spaces is established.

6 Conclusion

So far, we have presented various ways in which univariate orthogonal wavelet series decompositions have been used successfully and realistically in solving theoretical and practical univariate problems of nonparametric statistics. The application of wavelet methods to nonparametric regression has been mostly confined to the context of the normal distribution, with regularly spaced design points and for problems where both sample size and resolution levels are dyadic. Despite some papers addressing ways to remove these restrictions, some progress on alternative approaches to deal with such problems is very desirable in order to apply wavelet methods "naturally" to the general nonparametric regression setting.

A possibility to deal with non-uniform stochastic design would be to apply a discrete wavelet transform for unequally spaced data based on a basis particularly adapted to the irregular grid and constructed via the *lifting scheme* recently proposed by Sweldens ([80], [79]). Here one entirely abandons the idea of translation and dilation. This gives extra flexibility which can be used to construct wavelets adapted to irregular samples. However, to use such an approach some progress is needed on the deeper mathematical properties of the resulting scaling functions and these "second generation" wavelets.

Research is also needed in developing wavelet based methods to carry over likelihood-based models such as generalized linear models occurring often in practice.

Novel bootstrap methods for wavelet-based nonparametric curve estimation, taking advantage of the $\mathcal{O}(n)$ computational efficiency of wavelet decompositions are also highly desirable, since it is known that when the dimension of the unknown parameter exceeds that of the data, most classical (naïve) bootstrap methods for assessing the variability of the estimates and constructing confidence sets fail (see Beran [15]).

The usual wavelet-based approach can be further enhanced by using wavelet packets, a generalization of wavelet bases (see e.g. Wickerhauser [92]). In wavelet packet analysis, a function g is represented as a sum of orthogonal wavelet packet functions $W_{j,b,k}$ at different scales j, oscillations b and location k. By contrast with ordinary wavelet decompositions, in wavelet packet methods, a signal may be represented by many different combinations of wavelet packets. Thus, wavelet packets offer an enormous amount of flexibility in possible sets of basis functions. Adaptative ways to select the most appropriate set of basis functions with which to represent and estimate a density or a regression are particularly important and pose a number of interesting statistical issues. Some results on adaptative model selection using wavelet packets for white noise models already exist (see for example the papers by Donoho and Johnstone [34] and Saito [77]) but their extension to other types of noise are desirable.

Many results in higher dimensions are still incomplete. Theoretical advances in higher dimensional signal approximation bounds, regularity, design techniques, would be very useful in answering some questions that arise in the analysis of additive models in non-parametric regression, slice regression and multivariate density estimation.

To conclude let us say that there is room for substantial improvement of the current state of the art.

References

[1] Abramovich, F., Sapatinas, T. and Silverman, B. W. (1997). Wavelet thresholding via a Bayesian approach. Technical report , University of Bristol, England.

[2] Amato, U. and Vuza, D. T. (1994). Wavelet Regularization for Smoothing Data. Technical report N. 108/94, Instituto per Applicazioni della Matematica, Napoli.

[3] Amato, U. and Vuza, D. T. (1996). An Alternate Proof of a Result of Johnstone and Silverman concerning Wavelet Thresholding Estimators for Data with Correlated Noise. *Revue Roumaine Math. Pures Appl.* **41**, 431–438.

[4] Antoniadis, A. (1994). Smoothing noisy data with coiflets. *Statistica Sinica* **4**(2), 651–678.

[5] Antoniadis, A. (1994). Smoothing noisy data with tapered coiflets series. *Scand. Journal. of Statistics* **23**, 313–330.

[6] Antoniadis, A. and Carmona, R. (1991). Multiresolution analyses and wavelets for density estimation. Technical report, University of California, Irvine.

[7] Antoniadis, A., Grégoire, G. and McKeague, I. (1994). Wavelet methods for curve estimation. *J. Amer. Statist. Assoc.* **89**(428), 1340–1353.

[8] Antoniadis, A. and Pham, D. T. (1995). Wavelet regression for random or irregular design. Technical report. University of Grenoble.

[9] Antoniadis, A. and Lavergne, C. (1995). Variance function estimation in regression with wavelet methods. In A. Antoniadis and G. Oppenheim (eds.), *Wavelets and Statistics*, Lecture Notes in Statistics, **103**, Springer-Verlag.

[10] Antoniadis, A., Grégoire, G. and Vial, P. (1997a). Random design wavelet curve smoothing. *Statistics and Prob. Letters.*. In press.

[11] Antoniadis, A., Gijbels, I. and Grégoire, G. (1997b). Model selection using wavelet decomposition and applications. *Biometrika.* To appear.

[12] Antoniadis, A., Grégoire, G. and Nason, G. (1997c). Density and hazard rate estimation for right censored data using wavelet methods. Technical report. University of Grenoble.

[13] Antoniadis, A. and Gijbels, I. (1997). Detecting abrupt changes by wavelet methods. Technical report. University of Grenoble.

[14] Averkamp, R. and Houdré, C. (1996). Wavelet thresholding for non (necessarily) Gaussian noise: a preliminary report. Technical report, Georgia Institute of Technology, Atlanta.

[15] Beran, R. (1994). Bootstrap variable selection and confidence sets. Technical report, University of California, Berkeley.

[16] Brillinger, D. R. (1994). Some River Wavelets. *Environmetrics* **5**, 211–220.

[17] Brillinger, D. R. (1995). Some uses of Cumulants in Wavelet Analysis. *J. Nonparam. Statistics* **4**.

[18] Bruce, A. G. and Gao, H.-Y. (1994). *S+Wavelets, Users manual.* StatSci, Seatle.

[19] Bruce, A. G. and Gao, H.-Y. (1996). Understanding WaveShrink: Variance and bias estimation. *Biometrika* **83** (4), 727–746.

[20] Bruce, A. G. and Gao, H.-Y. (1997). WaveShrink with firm shrinkage. *Statistica Sinica*, to appear.

[21] Buckheit, J. B. and Donoho, D. (1995). Wavelab and Reproducible research. In A. Antoniadis and G. Oppenheim (eds.), *Wavelets and Statistics*, Lecture Notes in Statistics, **103**, Springer-Verlag.

[22] Clyde, M., Parmigiani, G. and Vidakovic, B. (1995). Multiple shrinkage and subset selection in wavelets. Technical report *DP 95-37*, Duke University.

[23] Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1995). Adaptative Bayesian Wavelet Shrinkage. Technical report , University of Chicago.

[24] Chui, K. (1992). *Wavelets: A Tutorial in Theory and Applications.* Academic Press, Boston.

[25] Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Comp. Harmonic Analysis* **1**(1), 54–81.

[26] Cohen, A. and Ryan, R. D (1995). *Wavelets and Multiscale Signal Processing.* Chapman & Hall, London.

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

[28] Daubechies, I. D. and Lagarias, J. C. (1991). Two-scale difference equations: Existence and global regularity of solutions. *SIAM Journal on Math. Analysis.* **22**, 1388–1410.

[29] Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators. *Applied and Comp. Harmonic Analysis* **3**, 215–228.

[30] DeVore, R. and Lucier, B. J. (1992). Fast wavelet techniques for near-optimal signal processing. In *IEEE Military Communication Conference*, pp. 1129–1135.

[31] DeVore, R. , Jawerth, B. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.* **305**, 397–414.

[32] Donoho, D. (1994). Asymptotic risk for sup-norm loss: solution via optimal recovery. *Prob. Theory and Related Fields.* **99**, 145–170.

[33] Donoho, D. L. and Johnstone, I. M. (1992). Minimax estimation via wavelet shrinkage. Technical report, Stanford University.

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

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

[36] Donoho, D. L. and Johnstone, I. M. (1994b). Ideal denoising in an orthonormal basis chosen from a library of bases. *Compt. Rend. Acad. Sci. Paris A* **319**, 1317–1322.

[37] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia (with discussion)? *J. Roy. Statist. Soc., Ser. B* **57**(2), 301–370.

[38] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**(2), 508–539.

[39] Doukhan, P. (1990). Consistency of delta-sequence estimates of a density or of a regression function for a weakly dependent stationary sequence. Séminaire de Statistique d'Orsay, Université Paris Sud, 1991.

[40] Doukhan, P. and Léon, J. (1990). Déviation quadratique d'estimateurs de densité par projection orthogonale. *Compt. Rend. Acad. Sci. Paris A* **310**, 424–430.

[41] Engel, J. (1990). Density estimation with Haar series. *Statistics and Probability Letters* **9**, 111–117.

[42] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Am. Statist. Assoc.* **91**, 674–688.

[43] Gao, H.-Y. (1993). *Wavelet estimation of Spectral denstities in Time series analysis.* Ph. D. Thesis, University of California, Berkeley.

[44] Gao, H.-Y. (1997). Wavelet Shrinkage Smoothing For Heteroscedastic Data. Technical report, StatSci, Seatle.

[45] Gasser, T. and Müller, H. (1979). Kernel estimation of regression functions. In Gasser, T. and Müller, H. (eds), *Curve Estimation*, Springer-Verlag, Heidelberg.

[46] Good, I. J. and Gaskins, R. A. (1971). Density estimation and bump haunting by the penalized maximum likelihood method. *J. Am. Statist. Assoc.* **75**, 42–69.

[47] Hall, P. and Nason, G. P. (1996). On choosing a non-integer resolution level when using wavelet methods. Technical report, University of Bristol.

[48] Hall, P. and Patil, P. (1995). On wavelet methods for estimating smooth functions. *Bernoulli* **1**, 41–58.

[49] Hall, P. and Patil, P. (1996a). Effect of threshold rules on performance of wavelet-based curve estimators. *Statistica Sinica* **6**, 331–345.

[50] Hall, P. and Patil, P. (1996b). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods. *J. Roy. Statist. Soc., Ser. B* **58**, 361–377.

[51] Härdle, W. (1990). *Applied nonparametric regression.* Cambridge University Press, Cambridge.

[52] Holschneider, M. (1995). *Wavelets: An analysis tool.* Clarendon Press, Oxford.

[53] Jansen, M., Malfait, M. and Bultheel, A. (1997). Generalized cross-validation for wavelet thresholding. *Signal Processing*, **56** (1). To appear.

[54] Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1992). Estimation d'une densité de probabilité par méthode d'ondelettes. *Compt. Rend. Acad. Sci. Paris A* **315**, 211–216.

[55] Johnstone, I. and Silverman, B. W (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc., Ser. B* **59**. In press.

[56] Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov Spaces. *Statistics and Probability Letters* **13**, 15–24.

[57] Kolaczyk, E. (1994). *Wavelet methods for the inversion of some homogeneous linear operators in the presence of noisy data.* Ph. D. Thesis, Stanford University.

[58] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **11**, 674–693.

[59] Marron, S. J., Adak, S., Johnstone, I., Neumann, M., and Patil, P. (1997). Exact risk analysis of wavelet regression. *Journal of Computational and Graphical Statistics*. In press.

[60] Masry, E. (1994). Probability density estimation from dependent observations using wavelet orthonormal bases. *Statistics and Probability Letters* **21**, 181–194.

[61] Masry, E. (1996). Multivariate probability density estimation by wavelet methods: strong consistency and rates for stationary time series. Technical report, University of California, San Diego.

[62] Meyer, Y. (1990). *Ondelettes et Opérateurs I: Ondelettes.* Hermann, Paris.

[63] Michelli, C. A. and Rivlin, T. J. (1975). A survey of optimal recovery. In Michelli, C. A. and Rivlin, T. J. (eds.), *Optimal estimation in Approximation theory*, pp. 1–54, Plenum, New York.

[64] Müller, H. G. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statist. Decisions* **2**, 193–206.

[65] Nason, G. P. (1996). Wavelet regression using cross-validation. *J. Roy. Statist. Soc., Ser. B* **58**, 463–479.

[66] Nason, G. J. (1995). Choice of the threshold parameter in wavelet function estimation. In A. Antoniadis and G. Oppenheim (eds.), *Wavelets and Statistics*, pp. 261–280, Lecture Notes in Statistics, Springer-Verlag, New York.

[67] Nason, G. J. and Silverman, B. W. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics* **3**, 163–191.

[68] Neumann, M. H. (1994). Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian series. Technical report 99, Institute for applied and stochastic analysis, Berlin.

[69] Neumann, M.H. and Spokoiny, V.G. (1993). On the efficiency of wavelet estimators under arbitrary error distributions. *Discussion Paper No. 4*, Hümboldt Universität zu Berlin.

[70] Ogden, T. R. (1996). *Essential wavelets for statistical applications and data analysis.* Birkhäuser, Basel.

[71] Oudshoorn, C. (1994). Wavelet-based nonparametric regression: optimal rate in the sup-norm. Technical report *848*, University Utrecht.

[72] Penev, S. and Dechevsky, L. (1997). On non-negative wavelet-based estimators. Technical report, University of New South Wales. To appear in *J. of Nonparam. Statistics.*

[73] Pinheiro, A. and Vidakovic, B. (1995). Estimating the square root of a density via compactly supported wavelets. Technical report *DP 95-14*, Duke University.

[74] Potier, C. and Vercken, C. (1994). Spline fitting Numerous Noisy Data with discontinuities. In Laurent et al. (eds.), *Curves and Surfaces*, pp. 477–480, Academic Press, New York.

[75] Raimondo, M. (1996). *Situations non ergodiques et utilisations de méthodes d'ondelettes.* Ph. D. Thesis, University Paris 7.

[76] Ramlau-Hansen, H. (1983). Smoothing counting processes by means of kernel functions. *Ann. Statist.* **11**, 453–466.

[77] Saito, N. (1994). Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. In Foufoula-Georgiou, E. and Kumar, P. (eds), *Wavelets in Geophysics*, Academic Press, New York.

[78] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **10**, 1135–1151.

[79] Sweldens, W. (1996). The lifting scheme: A custom-design construction of biorthogonal wavelets. *Applied and Comp. Harmonic Analysis* **3**, 186-200.

[80] Sweldens, W. (1996). The lifting scheme: A construction of second generation wavelets. Technical report 1995:6, Industrial Mathematics Initiative, Department of Mathematics, University of South Carolina.

[81] Triebel, H. (1992). *Theory of function spaces II.*. Birkhäuser, Basel.

[82] Tribouley, K. (1995). Practical estimation of multivariate densities using wavelet methods. *Statistica Neerlandica* **49**, 41–62.

[83] Vannucci, M. and Vidakovic, B. (1995). Preventing the Dirac disaster: wavelet based density estimation. Technical report *DP 95-24*, Duke University.

[84] Vidakovic, B. (1994). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. Technical report *DP 94-24*, Duke University.

[85] von Sachs, R. and Schneider, K. (1996). Wavelet smoothing of evolutionary spectra by nonlinear thresholding. *Applied and Comp. Harmonic Analysis* **3**(3), 268–282.

[86] Wahba, G. (1990). *Spline models for observational data.* CBMS-NSF regional conferences series in applied mathematics. SIAM, Philadelphia.

[87] Walter, G. G. (1992). Approximation of the Delta Function by Wavelets. *J. Approx. Theory.* **71**, 329–343.

[88] Walter, G. G. (1994). *Wavelets and Other Orthogonal Systems with Applications.* CRC Press, Boca Raton, Florida.

[89] Wang, Y. (1995). Jump and Sharp Cusp Detection by wavelets. *Biometrika* **82**, 385–397.

[90] Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.* **24** (2), 466–484.

[91] Weyrich, N and Warhola, G. T. (1995). Denoising using wavelets and cross-validation. In Singh, S. P. (ed), *Approximation Theory, wavelets and applications*, NATO ASI series C, pp. 523–532.

[92] Wickerhauser, M. V. (1994). *Adapted Wavelet Analysis: From Theory to Software*. AK Peters, Boston.