# Detecting abrupt changes by wavelet methods

Anestis Antoniadis
Laboratoire LMC - IMAG
Université Joseph Fourier
BP 53, 38041 Grenoble Cedex 09, France

Irène Gijbels
Institut de Statistique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

**Abstract**

The objective of this paper is to contribute to the methodology available for dealing with the detection and the estimation of the location of discontinuities in one dimensional piecewise smooth regression functions observed in white Gaussian noise over an interval.

Our approach is nonparametric in nature because the unknown function is not assumed to have any specific form. Our method relies upon a wavelet analysis of the observed signal and belongs to the class of "indirect" methods, where one detects and locates the change points prior to fitting the curve, and then uses one's favorite function estimation technique on each segment to recover the curve. We show that, provided discontinuities can be detected and located with sufficient accuracy, detection followed by wavelet smoothing enjoys optimal rates of convergence.

## 1 Introduction

Wavelet based nonparametric regression has become a well known and mathematically founded technique for estimating smooth functions adaptively. Optimal rates of convergence have been thoroughly examined for different observation schemes and $L_2$ losses by many authors. In many applications, however, the function to be estimated presents isolated discontinuities, and is only piecewise smooth. To cite only a few examples, such change point problems arise in quality control, medical monitoring, econometrics, study of growth in biology, signal processing and edge detection in image processing. Since the basic aim of wavelet analysis is to represent a function as a linear superposition of wavelets centred on a sequence of time points, it forms a natural tool for the investigation of "local" properties of time varying functions. The object of this paper is to examine how well wavelet analysis addresses the detection of jump points in time varying functions observed with noise.

There already exists a considerable body of literature on discontinuity detectors with good location properties that may be employed. These include semiparametric smoothing splines methods proposed by Laurent and Utreras (1986), Girard (1990); kernel based methods by Canny (1986), Korostelev (1987), Hall and Titterington (1992), Müller (1992), Wu and Chu (1993), Speckman (1994), Eubank and Speckman (1994), Bunt, Koch and Pope (1995); local polynomial methods by McDonald and Owen (1986), Leclerc and Zucker (1987), Loader (1996), Qui and Yandell (1994); wavelet methods proposed by Mallat and Hwang (1992), Potier and Vercken (1994), Wang (1995). See also Gijbels, Hall and Kneip (1995) who considered a two-step estimation procedure. Most of these papers deal with a generally piecewise smooth function with a finite number of change-points. Moreover they focus on estimation of locations and magnitude of jumps. When one is concerned with the estimation of the function itself, the detection of change-points is just a part of the estimation procedure. Oudshoorn (1995) addresses this situation using a segmented Fourier expansions in eigenfunctions arising from particular differential boundary value problems with Neumann conditions and obtains under reasonable conditions an estimation procedure that is asymptotically optimal for piecewise regression functions with a finite number of change points. The pilot estimators of the locations of change-points that are still required as a technical step in her work are based on the difference between two one-sided kernel smoothers, which makes the computation quite expensive. From a practical point of view, the segmented Fourier expansions that are used, while optimal, are not easy to implement.

Our benchmark in this paper will be Oudshoorn's (1995) work. However, instead of using differences of one-sided kernel smoothers, we will use a refinement of Wang's (1995) procedure for detecting and locating the discontinuities. We will then use a segmented multiresolution analysis to obtain almost optimal mean-square rates in the context of a Sobolev-type setting with Normal errors. The resulting nonparametric regression estimator, regardless of the presence of a finite number of jumps in the unknown regression function, is then shown to enjoy similar asymptotic convergence rates as in the case of smooth functions without discontinuities.

The paper is organized in the following way. Section 2 contains a description of the wavelet transform that we are going to use. In Section 3 we describe our setting, and in particular our segmentation strategy. We also describe known results for linear wavelet estimators adapted to an interval. Our main results, paralleling those of Oudshoorn, are given in Section 4, in which we show that our segmented wavelet estimation procedure reaches almost the same asymptotic rate of convergence of MISE as in the smooth case.

The applicability of our method on multiple change-point data is illustrated in Section 5 by simulations and the analysis of a data set that has been analyzed previously in the statistical literature. Proofs of our results are given in the Appendix.

# 2 Wavelet decompositions of signals with isolated singularities

This section is devoted to the description of the various wavelet transforms and a discussion of the properties they share in addressing the detection of singularities in time-varying signals. Most of the facts stated in this section are known and are just gathered together for the convenience of the reader.

There exist different types of wavelet transforms that provide time-scale representations of square integrable signals. The most natural way to obtain a time-scale representation of a finite energy signal is to define a family of functions

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{t-b}{a}\right), \quad a>0,\ b\in \mathbb{R},$$

where $\psi$ is a fixed function in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ with $\int \psi(x)dx = 0$, called "wavelet". The factor $1/\sqrt{a}$ ensures that the functions $\psi_{a,b}$ have a constant norm in the space $L^2(\mathbb{R})$ of square integrable functions. The *continuous wavelet transform* (CWT for short) of a function $f \in L^2(\mathbb{R})$ is defined as

$$\mathrm{CWT}_f(a,b) = \langle f, \psi_{a,b}\rangle = \frac{1}{\sqrt{a}} \int f(t)\overline{\psi}\left(\frac{t-b}{a}\right)\,dt.$$

The existence of an inverse transform depends on the choice of $\psi$. If $\psi$ is such that

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|}\,d\omega < +\infty,$$

where $\hat{\psi}$ denotes the Fourier transform of $\psi$, then $f$ can be reconstructed (in the weak sense) by

$$f = C_\psi^{-1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{da\,db}{a^2}\mathrm{CWT}_f(a,b)\psi_{a,b}.$$

Note that condition $\int \psi(t)\,dt = 0$ implies that $\psi$ oscillates. More generally, one can impose more cancellations on $\psi$, in the sense that $\int t^k \psi(t)\,dt = 0$, for $k = 0, 1, \ldots, N$. An important consequence of the cancellations of $\psi$ is that the "wavelet coefficients" $\mathrm{CWT}_f(a,b)$ are influenced by the local regularity of the function $f$. To be more specific, for any $\alpha > 0$, we say that $f$ is $\alpha$-Hölder regular at some point $t_0$ *if and only if* there exist a polynomial of degree $n$, $n < \alpha \leq n+1$, $P_n(t)$, and a function $f_{\text{loc}}$ such that we may write

$$f(t_0 + t) = P_n(t) + f_{\text{loc}}(t), \qquad (1)$$

with

$$f_{\text{loc}}(t) = \mathcal{O}(t^\alpha), \qquad \text{as} \quad t \to 0.$$

Note that this property is satisfied in particular when $f$ is $m$-times differentiable in a neighborhood of $t_0$ with $m > \alpha$ by taking $P_n(t) = \sum_{0 \leq k < \alpha} f^{(k)}(t_0)\,\frac{(t-t_0)^k}{k!}$. Note also that for $0 < \alpha < 1$, one can only take $P_0(t) = f(t_0)$. If the local estimation (1) holds uniformly in $t_0$, we say that $f$ is uniformly $\alpha$-Hölder regular. If $\psi$ has cancellations up to some order $N \geq \alpha - 1$ and sufficient decay at infinity so that $\int \psi(t)|t|^\alpha dt < \infty$, one can characterize the global and the local Hölder regularity of a function by the decay of the modulus of its wavelet transform across scales. More precisely, we have :

**Theorem 1 (Mallat (1989))** *If $\psi$ has cancellations up to the order $N \geq [\alpha] - 1$ and is such that $\int \psi(t)|t|^\alpha dt < \infty$, then the wavelet transform of an $\alpha$-Hölder regular function $f$ at $t_0$ satisfies*

$$|CWT_f(a, t_0 + b)| = \mathcal{O}(\sqrt{a}(|a|^{1/2} + |b|^{1/2})), \quad (a,b) \to 0.$$

*Conversely, if furthermore $\psi$ is compactly supported and for $\alpha < N$ a noninteger positive number, a square integrable function $f$ is $\alpha$-Hölder regular at $t_0$, if the following two conditions hold :*

- *for some $\gamma > 0$ and some constant $C$ and for all $t$ in a neighborhood of $t_0$ and any scale $a$,*
$$|CWT_f(a,b)| \leq C|a|^{\gamma+1/2},$$
- *there exists a constant $B$, such that for all $t$ in a neighborhood of $t_0$ and any scale $a$,*
$$|CWT_f(a,t)| \leq B\,|a|^{1/2}\left(|a|^\alpha + \frac{|t-t_0|^\alpha}{|\log|t-t_0||}\right).$$

The decay of the modulus of the scale-space coefficients as the scale becomes small completely characterizes the Hölder regularity of a function $f$ when it is uniform. In such a case, it is known (see Holschneider (1995), Theorem 2.1.1) that

$$|\text{CWT}_f(a,b)| \leq C|a|^{\alpha+1/2} \Rightarrow f \text{ is uniformly } \alpha\text{-Hölder regular.}$$

Among the important properties of the continuous wavelet transform is its time translation invariance, which means that the wavelet transform of a translated function is obtained by translating the wavelet transform of the original function.

In practical applications, in particular those involving fast algorithms, the continuous wavelet transform can only be computed on a discrete grid of points $(a_j, b_k)_{j,k \in \mathbb{Z}}$. The efficient choice of this sampling grid, so that it contains all the information on the function to be analyzed, relies upon the theory of *frames*. For a detailed account on this theory the reader is refered to Daubechies (1992). If the sampling is too tight, the decomposition of $f$ into its wavelet coefficients will be redundant. By adding more restrictions on the sampling as well as in the choice of the mother wavelet $\psi$ it is possible to remove the redundancy in the reconstruction formula

$$f = \sum_{j,k} \langle f, \psi_{j,k} \rangle \, \psi_{j,k}$$

so that it may be regarded as the expansion of $f$ in a basis. Such a natural and admissible sampling for the continuous wavelet transform is given by $\mathrm{CWT}_f(2^j, k2^j)$, $k, j \in \mathbb{Z}$. One can then define

$$\psi_{j,k}(t) = 2^{j/2} \, \psi(2^j t - k),$$

and, provided that the mother wavelet $\psi$ is well chosen, the system $\{\psi_{j,k}, \; j, \; k \in \mathbb{Z}\}$ defines an orthonormal basis of $L^2(\mathbb{R})$. The map that associates to an $f \in L^2$, its wavelet coefficients $< f, \psi_{j,k} >= \mathrm{CWT}_f(2^j, k2^j)$ is called the *discrete wavelet transform* (DWT for short). The DWT can be computed efficiently through the famous pyramidal discrete-time algorithm of Mallat (1989).

One of the main weakenesses in using the Discrete Wavelet Transform lies in fact in the nonuniformity of the time sampling across the scales. For example, if a discontinuity occurs at a bad point of the grid at the finest scale then the localisation of the discontinuity through the wavelet coefficients across the coarser levels will not be consistently located at the same time location. Hence, one can not localize the discontinuity across scales. This is mainly due to the lack of translation invariance of the DWT (if the signal is slightly shifted, the wavelet coefficients of the translated signal can be completely different than these of the original signal). This is illustrated in the example displayed in Figure 1 below.

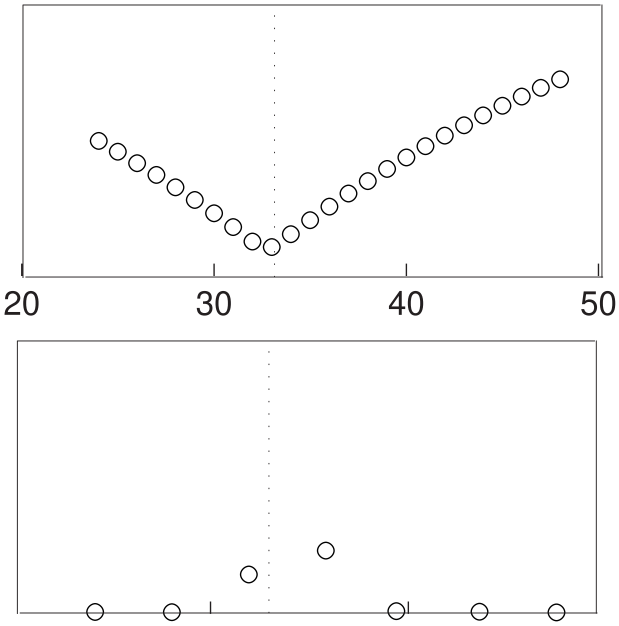



Figure 1: The signal to be decomposed presents a cusp. The wavelet coefficients (details) for a lower resolution are displayed as circles on the bottom panel. As one can see, the DWT, due to the lack of translation invariance, presents on the cusp location a weak coefficient.

Moreover, Hölder regularity characterization does not hold in general for the discrete wavelet transform (see Abry (1994)). This is due to the lack of translation invariance of the discrete wavelet transform. It is therefore imperative, in a discrete and practical setting, to use the continuous wavelet transform for estimating the location of the jump points. However, the use of the continuous wavelet transform (CWT) should be avoided since no fast computational algorithm exists. Mallat (1989) remarks that by discretizing the scale on a dyadic grid one obtains a dyadic wavelet transform (DyWT) and if the analyzing wavelet is appropriately chosen then the characterizing theorems for the local Hölder regularity of the function to analyze, available for the CWT, still hold true for the DyWT. Hence, the Hölder regularity of a function is characterised by the decay across dyadic scales of the wavelet transform modulus, $|\mathrm{CWT}_f(2^{-j}, t)|$. It follows that, for a sampled signal on an equidistant grid, it is sufficient to investigate the wavelet transform modulus on the dyadic scales at the finest grid (this means that we only have to look at $|\mathrm{CWT}_f(2^{-j}, k)|$). This type of transform is refered to as the continuous discrete wavelet transform, i.e. CDWT and can be shown to be time translation invariant.

Abry (1994) proposes a fast algorithm for calculating the CDWT. The proposed algorithm is based on the same ideas as Mallat's pyramidal algorithm and is a variant of the "algorithm à trous". The computation time for the proposed algorithm is of order $O(Mn)$, where $M$ denotes the number of scales and $n$ the sample size. In comparison: for the DWT the algorithm is of the order $O(n)$ whereas for the CWT the algorithm is of order $O(Mn^2)$.

Another essential difference between the CWT and the DWT is the redundance of the transform. More precisely, the information contained in the transformed wavelet "coefficient" $\mathrm{CWT}_f(a, b)$ is also shared by the neighboring coefficients. The above reveals another advantage of the CWT for detecting irregularities in the signal. In the presence of a discontinuity the discrete continuous wavelet transformation will exhibit large coefficients in modulus along many scales, all synchronized at the position of the discontinuity at the time scale. This enables to detect more easily a discontinuity, by looking for large coefficients (in modulus) not only at a particular level where they appear but also at neighboring finer levels. This motivates the strategy that we are going to use for detecting discontinuities, which will be introduced later on.

To end this section, note that when a function is approximated at a particular finite resolution level, it is practically impossible to distinguish jumps and cusps from other regularities of small order (i.e. with small Hölder exponents). So it is the decay of the wavelet transform across scales which will have to play the crucial role here. Consequently, for mathematical convenience, jumps will be modeled as points of local Hölder regularity with very small Hölder regularity index $\alpha$.

# 3 Estimation in a white noise model

One frequently-used formulation of the change point problem is the following. Consider observing a process $Y_\tau(t)$ satisfying the following stochastic differential equation

$$dY_\tau(t) = f(t)dt + \tau\, dW(t), \quad 0 < t \leq 1, \tag{2}$$

with some prescribed initial value $Y(0)$, which is either a random variable independent of the noise $W$ or constant, where $W$ is a standard Wiener process, $\tau$ is a formal noise level parameter which we think of as small, and $f$ is an unknown piecewise smooth function. By piecewise smooth function on $[0, 1]$ we mean that, given a positive integer $m$ and constants $K > 0$, $S > 0$ and $\kappa > 0$:

**M1:** $f \in L_2([0, 1]) \cap L_\infty([0, 1])$,

**M2:** there are $d$ (unknown) points ($d$ bounded by $D$), $0 < t_1 < \cdots < t_d < 1$ points at which $f$ has a discontinuity and $t_{j+1} - t_j > \kappa$, $j = 0, \ldots, d$, where $\kappa$ is an unknown positive constant,

**M3:** at each discontinuity point $t_i$, the left and right limits $f(t_i-)$ and $f(t_i+)$ exist and are such that $|f(t_i+) - f(t_i-)| \geq S$, where $S$ is a positive constant not necessarily known.

**M4:** $f$ belongs to a piecewise Sobolev-ellipsoid, i.e.

$$\sum_{j=0}^{d} \int_{t_j}^{t_{j+1}} [f^{(m)}(t)]^2 \, dt \leq K, \quad t_0 = 0, \; t_{d+1} = 1.$$

The case $d = 0$ is allowed and means of course that $f$ has no discontinuities. For this setting we refer to Oudshoorn (1995), where Fourier expansions in eigenfunctions arising from a particular differential boundary value problem with Neumann conditions were used to obtain results similar to the ones of the present paper. For convenience, we denote by $\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$ the class of functions $f$ that satisfy assumptions **M1–M4**.

The statistical inference problem then consists in:

- estimating the number and position(s) of change-point(s).
- recovering the piecewise smooth regression function $f$ without degrading its discontinuities.

**Remark 1** . *The white noise model (2) is a generalization of the diffusion model used by Pollak & Siegmund (1985) to study change-point problems. It is also closely related to the following discrete regularly spaced nonparametric regression model:*

$$Z_i = f(x_i) + \sigma \epsilon_i \quad i = 1, \ldots, n, \tag{3}$$

*with $x_i = i/n$, the $\epsilon$'s independent standard normal errors, $\sigma > 0$ and $f$ an unknown function. By defining as in Wang (1995) or Donoho & Johnstone (1995), the regression process $\{Y_n(t),\ t \in [0,1]\}$ via $t_0 = 0$, $Y_n(0) = 0$ and $Y_n(x_i) = \sum_{j=1}^{i} Z_j$ for $i = 1, \ldots, n$, with interpolation by a Wiener process $W$ for $x_i \leq t < x_{i+1}$, $Y_n$ appears as a white noise process with the function $f_n(t) = f(x_i)$ for $x_i \leq t < x_{i+1}$ and $\tau = \sigma n^{-\frac{1}{2}}$.*

Since the assumptions **M1–M4** mean in effect that we observe $d + 1$ nonparametric regression problems which have been joined end-to-end without marking the endpoints, a natural strategy for estimation of $f$ is to attempt to reverse the process, that is:

**S1:** Determine estimates $\hat{t}_j$, $j = 1, \ldots, \hat{d}$ of jump points of $f$; set $\hat{t}_0 = 0$ and $\hat{t}_{\hat{d}+1} = 1$.

**S2:** Break the process $\{Y_\tau(t),\ t \in [0,1]\}$ into $\hat{d}+1$ segments $\{Y_\tau(t), t \in [\hat{t}_j, \hat{t}_{j+1}[, j = 0, \ldots, \hat{d}\}$.

**S3:** Use a nonparametric linear wavelet estimator adapted on each segment separately to estimate $f$ on that part of its domain.

**S4:** Join the estimates together in the obvious way.

In order to achieve this task two steps in the previous process need to be specified: the discontinuity detector/locator (step **S1**) and the nonparametric regression estimator (for step **S3**). We shall concentrate on how well it is possible to estimate $f$ by this strategy.

The quadratic risk ($L_2$-loss) for an estimator $\hat{f}$ of $f$ is defined as

$$R(\hat{f}, f) = \mathbb{E}\{\|\hat{f} - f\|_2^2\}. \tag{4}$$

The minimax quadratic risk in estimating $f$ with respect to the class $\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$ is then defined by

$$R_\tau\left(\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})\right) = \inf_{\hat{f}} \sup_{f \in \mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})} R(\hat{f}, f), \tag{5}$$

where the infimum is taken over all estimators $\hat{f}$.

The exact asymptotic behavior of the minimax risk for the class $\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$ satisfies

$$\lim_{\tau \downarrow 0} \tau^{-\frac{4m}{2m+1}} R_\tau\left(\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})\right) = \gamma(m, K), \tag{6}$$

where $\gamma(m, K) = (K(2m+1))^{\frac{1}{2m+1}} \left(\frac{m}{\pi(m+1)}\right)^{\frac{2m}{2m+1}}$. We refer to Oudshoorn (1995) for the derivation of this result.

Our purpose is to show that it is possible to estimate $f$ by wavelet methods with a risk that is nearly equal to this optimal asymptotic minimax risk. To do so we first use linear wavelet smoothers to the data, in case the jump points are known. Results from Antoniadis (1994), produce an estimator $\tilde{f}$ with a risk similar to (6), within a constant factor. Of course this estimator depends on the (unknown) jump points. We realize step **S1** by estimating these jump points in an optimal fashion. To finally produce our estimator we replace in $\tilde{f}$ the jump points $t_j$ by their estimates, resulting in an asymptotically almost optimal estimator $\hat{f}$.

## 3.1 Quasi-optimal estimation with known jump-points

We follow the ideas of Antoniadis (1994) in deriving an optimal estimator of $f$ when the location of the jump points are assumed to be known. We will be using an orthogonal multiresolution analysis that "lives" on the interval as the one developed by Cohen, Daubechies and Vial (1993) or the one developed by Andersson, Hall, Jawerth and Peters (1993). The starting point for such analyses, is any one of the compactly supported orthogonal bases in Daubechies, with $N$ vanishing moments and support $\psi=$ support $\phi = [0, 2N-1]$. The construction given by these authors involves the construction of multiscale spaces $V_j[a,b]$ and detail spaces $W_j[a,b]$ with properties analogous to those associated to multiresolution analyses on the real line, as well as collections of basis functions $\{\phi_{j,k}\}_k$ and $\{\psi_{j,k}\}_k$ spanning these spaces. At sufficiently fine scales, the approximation spaces are generated by exactly $2^j$ orthonormal basis functions and the same is true for the detail spaces $W_j[a,b]$. The result of these constructions is an orthonormal family of wavelets on $[a,b]$, with $N$ vanishing moments; together with an orthonormal family of scaling functions on $[a,b]$ at the coarsest scale under consideration, these adapted wavelets constitute an orthonormal basis of $L_2([a,b])$. In addition, their regularity and vanishing moment properties ensure that they are unconditional bases for the Hölder spaces $C^s([a,b])$ for all $s < r$, where $r$ is the regularity of the original wavelet basis, $\psi \in C^r$. The above constructions provide also fast wavelet transform of sequences of arbitrary length (see the above cited papers for details). For simplicity of notation, given a multiresolution decomposition $V_\ell([t_j,t_{j+1}]) \oplus \oplus_{q \geq \ell} W_q([t_j,t_{j+1}])$, we hereafter identify $\psi_k^j$ with the sequence

$$\phi_{\ell,0}, \ldots, \phi_{\ell,2^\ell-1}, \psi_{\ell,0}, \ldots, \psi_{\ell,2^\ell-1}, \ldots, \psi_{q,0}, \ldots, \psi_{q,2^q-1}, \ldots.$$

Rewrite the segmented observation process in step **S2** with known segments $[t_j,t_{j+1}]$ into an equivalent sequence of observations $(j = 0, \ldots, d)$

$$Y_{k,(j)} = \theta_{k,(j)} + \tau\epsilon_{k,(j)}, \quad j = 0, \ldots, d \quad k \in I\!N,$$

where

$$Y_{k,(j)} = \int_{t_j}^{t_{j+1}} \psi_k^j(t)\, dY_\tau(t), \qquad \theta_{k,(j)} = \int_{t_j}^{t_{j+1}} f(t)\psi_k^j(t)\, dt,$$

and

$$\epsilon_{k,(j)} = \int_{t_j}^{t_{j+1}} \psi_k^j(t)\, dW(t).$$

The orthonormality of the above wavelet transform has a fundamental statistical consequence: the random variables $\epsilon_{k,(j)}$ are independent and $N(0,1)$ distributed. Moreover, the wavelet transform transforms estimators in one domain in the other domain with isometry of risks. The problem of recovering a smooth $f$ can then be seen as the one recovering those few coefficients of $f$ in each segment that are significantly different than 0 against a Gaussian white noise background. Linear wavelet projectors rules are an instance of this principle. As an estimate of the unknown function $f$ we now consider the following wavelet estimator:

$$\tilde{f}(t) = \sum_{k=0}^{\infty} \tilde{\theta}_{k,(j)}\psi_k^j(t), \quad t \in [t_j, t_{j+1}[, \quad j = 0, \ldots, d \tag{7}$$

where

$$\tilde{\theta}_{k,(j)} = \begin{cases} Y_{k,(j)} & \text{if } k \leq 2^J - 1 \\ 0, & \text{for } k \geq 2^J, \end{cases}$$

with $J = \log_2(\tau^{-2})/(1+2m)$. The resolution $J$ and the method of estimation are optimal in the sense that the maximal quadratic risk of the pseudo-estimator $\tilde{f}$ over $\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$ is asymptotically nearly equal to the bound given in (6). In order to explain this, observe first that the resolution $J(\tau)$ defining our estimator is asymptotically equivalent to the one used by Antoniadis for finite sample regression models. Now, by the Parseval relation for the wavelet basis, the additivity of coordinate risks and independence of the noise, and the fact that on each segment we are dealing with a Sobolev ellipsoid, Theorem 3.1 of Antoniadis (1994) applies on each segment, provided that the wavelet basis we are using is of regularity $r$ larger than $m$. The last mentioned theorem (and its proof), states that the quadratic risk for the sequence space is attained by the wavelet estimator and that it is equivalent, up to some constants that depend only on the wavelet that is used, to the quadratic risk in the function problem. Using the fact that the number of segments for the segmented problem is bounded by $D+1$ and using again the additivity of coordinate quadratic risks, we conclude easily that the maximal quadratic risk for $\tilde{f}$ over $\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$ is bounded above by $\mathcal{O}(\tau^{\frac{4m}{2m+1}})$. It is therefore, up to some constant, nearly minimal.

The discussion that follows next focuses on the way in which appropriate estimates of the discontinuity locations may be fitted into the definition of $\tilde{f}$, in order to obtain asymptotically the same quality of estimation as that of $\tilde{f}$.

## 3.2 A discontinuity locator

As already mentioned in the introduction, many examples of estimators which detect and locate discontinuities exist in the literature. However we need here discontinuity locators that converge fast enough (in the mean-square sense) to the true discontinuity location. Condition **M2** prevents the discontinuity detection and location step from finding discontinuities less than the prescribed distance $\kappa$ apart or too close to the boundary of the data. Assuming that such a condition is adopted, one can segment the interval $[0,1]$ into intervals of size $h = \tau^2(\log(\tau^{-2}))^{1+\eta}$ for some fixed $\eta > 0$ and such that $2h < \kappa$ for $\tau$ small enough. Note here that such a segmentation is suggested by the fact that for intervals of length $h$ in which the regression function is differentiable the orders of $\mathrm{CWT}_f(h,x) = Tf(h,x)$ and of $\tau TW(h,x)$ are balanced (see Wang (1995)). Oudshoorn (1995) has used a similar grid to compute local left and right kernel smoothers for estimating the left and right limits of $f$ at these grid points, in order to locate the jumps. The method adopted here relies upon the continuous wavelet transform evaluated at these grid points and is much easier to compute and analyze. We denote by $G = \{g_i = i\tau^2, i = \lceil h\tau^{-2}\rceil, \ldots, \lfloor(1-h)\tau^2\rfloor\}$ the corresponding grid.

Let $\psi$ be a compactly supported Daubechies wavelet regular enough to be differentiable and define $\psi_h(x) = h^{-1/2}\psi(x/h)$. The continuous wavelet transform of $f$ at $(h,x)$ is then given by $Tf(h,x) = \int \psi_h(x-u)f(u)du$. According to Mallat's theorem stated in the introduction, at small scales $h$, $Tf(h,x)$ characterizes the local regularity of $f$. We therefore expect that $|Tf(h,g_i)|$ will be "large" if there is a discontinuity location $t_j$ at the vicinity of $g_i$. Indeed this is the case and we have the following proposition :

**Proposition 1** *If $\psi$ is differentiable, then for every $\delta > 0$ and $0 < \alpha < (1+\eta)/2$, uniformly in $f \in \mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$, one has*

*(i) $I\!\!P\{|TY(h,g_i)| > (\log\tau^{-2})^{-\alpha}\} = o(\tau^\delta)$, as $\tau \downarrow 0$, for every point $g_i \in G$ such that $[g_i - h, g_i + h]$ does not contain any discontinuity location $t_j$,*

*(ii) $I\!\!P\{|TY(h,g_i)| < (\log\tau^{-2})^{-\alpha}\} = o(\tau^\delta)$, as $\tau \downarrow 0$, for every point $g_i \in G$ such that for some $j = 1,\ldots,d$, $|g_i - t_j| < h/2$.*

**Proof of Proposition 1:** Consider first a grid point $g_i$ such that no jump location $t_j$ is covered by the interval $[g_i - h, g_i + h]$. Since $m \geq 1$, the function $f$ is at least of Lipschitz regularity 1 at $g_i$. Now, by Mallat's theorem, we have

$$|Tf(h,g_i)| \leq \text{const} \cdot h \leq \frac{1}{2}(\log\tau^{-2})^{-\alpha}, \tag{8}$$

for $\tau$ sufficiently small. At the scale $h$, the wavelet transform of the white noise, is a stationary Gaussian process with zero mean and variance $1/h$. Therefore, $TY(h,g_i)$ is a Normal random variable with mean $Tf(h,g_i)$ and variance $\tau^2/h$. Using the fact that for the standard normal cumulative distribution function $\Phi$ we have $1 - \Phi(c) \leq \phi(c)/c$ as soon as $c > 1$, where $\phi$ denotes the standard normal density function, we get

$$\begin{aligned}
I\!\!P\{|TY(h,g_i)| > (\log\tau^{-2})^{-\alpha}\} &\leq I\!\!P\{|TY(h,g_i) - Tf(h,g_i)| > (\log\tau^{-2})^{-\alpha} - |Tf(h,g_i)|\} \\
&\leq I\!\!P\{|TY(h,g_i) - Tf(h,g_i)| > 1/2(\log\tau^{-2})^{-\alpha}\} \\
&\leq \sqrt{\frac{2}{\pi}}(\log\tau^{-2})^{\alpha - \frac{\eta}{2} - \frac{1}{2}} \exp\{-\frac{1}{8}(\log\tau^{-2})^{1+\eta-2\alpha}\} = o(\tau^\delta),
\end{aligned}$$

provided that $\alpha < (1+\eta)/2$.

For a grid point $g_i$ such that $|g_i - t_j| < h/2$, expression (A4) of Wang (1995) applies and we have

$$|Tf(h,g_i)| = 2^{-1}|f(t_j+) - f(t_j-)|\{|\int \psi(u-t)sign(u)\,du| + o(1)\}$$

which by **M3**, we bound below by $S \cdot C/4$ where $C = |\int \psi(u-t)\mathrm{sign}(u)du|$. It then follows that

$$\begin{aligned} I\!P\{|TY(h,g_i)| < (\log\tau^{-2})^{-\alpha}\} &\leq I\!P\{|TY(h,g_i) - Tf(h,g_i)| > S/4\} \\ &\leq \frac{2\sqrt{2}}{S\sqrt{\pi}}(\log\tau^{-2})^{-\frac{1+\eta}{2}} \exp\{-\frac{S^2}{32}(\log\tau^{-2})^{1+\eta}\} = o(\tau^{\delta}). \end{aligned}$$

Since more than one grid point $g_i$ may be close to a true discontinuity $t_j$, we find it necessary to match up as much as possible the true discontinuity points of $f$ with the points of the grid that detect them. We describe now such a matching algorithm. Let $\alpha < (1+\eta)/2$. Pair of adjacent points $(g_i, g_{i+1})$ in the grid $G$ will be called *linked* if both $|TY(h,g_i)|$ and $|TY(h,g_{i+1})|$ exceed $(\log\tau^{-2})^{-\alpha}$. According to our Proposition 1, linked pairs consist of potential representatives of a true discontinuity point and their associated estimated values. If there exist more than one link for a point, which is typically the case when using the nondecimated wavelet transform in our introduction, we choose the best link as follows. Let $U$ be the subset of the grid $G$ consisting of all elements linked to some other element. Define a chain of length $\ell$ in $U$ to be a sequence $g_r < g_{r+1} < \ldots < g_{r+\ell}$ of elements in $U$ such that $g_i$ is linked to $g_{i+1}$ and no further elements can be added to the chain without destroying this property. Consider now all chains of length greater than $\lceil(\log\tau^{-2})^{1+\eta}\rceil$ and denote them by $C_j$, $j = 1, \ldots, \hat{d}$. If such a chain does not exist we set $\hat{d} = 0$. Finally we can now define the following estimates for possible jump points:

$$\hat{t}_j = \mathrm{argmax}_{g_i \in C_j}\{|TY(h,g_i)|\}, \qquad j = 1, \ldots, \hat{d}, \tag{9}$$

with $\hat{t}_0 = 0$ and $\hat{t}_{\hat{d}+1} = 1$. Having defined these estimators we now wish to control the probability of introducing spurious discontinuities (false positives). Such points are grid-points $g_i$ for which there is no $t_j$ in $[g_i - h, g_i + h]$ and nevertheless $|TY(h,g_i)| > (\log\tau^{-2})^{-\alpha}$ and therefore introducing false positives is equivalent to the event $\hat{d} > d$. Using assertion (i) of Proposition 1 with $\delta = 2 + \frac{4m}{2m+1}$, we have for this probability:

$$I\!P\{\hat{d} > d\} \leq \sum_{g_i \in G;\ \text{no}\ t_j \in [g_i - h, g_i + h]} I\!P\{|TY(h,g_i)| > (\log\tau^{-2})^{-\alpha}\} = o\left((\tau^2)^{\frac{2m}{2m+1}}\right). \tag{10}$$

For the case of false negatives (undetected discontinuities) there will be at least one location $t_j$ such that there exists a grid point $g_i$ within a distance $h/2$ of $t_j$ having a $|TY(h,g_i)|$-value smaller than $(\log\tau^{-2})^{-\alpha}$. Using assertion (ii) of Proposition 1 with $\delta = \frac{4m}{2m+1}$ gives

$$I\!P\{\hat{d} < d\} \leq \sum_{t_j} I\!P\{|TY(h,g_i)| < (\log\tau^{-2})^{-\alpha}\} = o\left((\tau^2)^{\frac{2m}{2m+1}}\right), \tag{11}$$

since the number of terms in the sum is bounded by $D$. It follows from inequalities (10) and (11), that

$$I\!P\{\hat{d} \neq d\} = o\left((\tau^2)^{\frac{2m}{2m+1}}\right) \qquad \text{as } \tau \downarrow 0. \tag{12}$$

Moreover, using the fact that $I\!E(\hat{t}_j - t_j)^2 \leq (2h)^2 + I\!P\{|\hat{t}_j - t_j| > 2h\}$ and Proposition 1, it is easy to see that

$$I\!E(\hat{t}_j - t_j)^2 \leq \mathcal{O}\left((\tau^2(\log \tau^{-2})^{1+\eta})^2\right) = o\left((\tau^2)^{\frac{2m}{2m+1}}\right). \tag{13}$$

# 4 Estimating the regression function

To define the estimator in the case of unknown jump points it suffices now to substitute in subsection 3.1, the unknown values $t_j$ and $t_{j+1}$ by their estimates $\hat{t}_j$ and $\hat{t}_{j+1}$. Such a substitution leads to the segmented observation process in step **S2** with segments $[\hat{t}_j, \hat{t}_{j+1}]$ into the equivalent sequence of observations ($j = 0, \ldots, \hat{d}$)

$$\hat{Y}_{k,(j)} = \hat{\theta}_{k,(j)} + \tau\hat{\epsilon}_{k,(j)}, \quad j = 0, \ldots, \hat{d} \quad k \in I\!N,$$

where

$$\hat{Y}_{k,(j)} = \int_{\hat{t}_j}^{\hat{t}_{j+1}} \hat{\psi}_k^j(t)\, dY_\tau(t), \qquad \hat{\theta}_{k,(j)} = \int_{\hat{t}_j}^{\hat{t}_{j+1}} f(t)\hat{\psi}_k^j(t)\, dt,$$

and

$$\hat{\epsilon}_{k,(j)} = \int_{\hat{t}_j}^{\hat{t}_{j+1}} \hat{\psi}_k^j(t)\, dW(t).$$

As suggested in subsection 3.1, we propose then the following estimator for $f$,

$$\hat{f}(t) = \sum_{k=0}^{\infty} \hat{\tilde{\theta}}_{k,(j)}\hat{\psi}_k^j(t), \quad t \in [\hat{t}_j, \hat{t}_{j+1}[, \quad j = 0, \ldots, \hat{d} \tag{14}$$

where

$$\hat{\tilde{\theta}}_{k,(j)} = \begin{cases} \hat{Y}_{k,(j)} & \text{if } k \leq 2^J - 1 \\ 0, & \text{for } k \geq 2^J, \end{cases}$$

with $J = \log_2(\tau^{-2})/(1 + 2m)$. The resulting estimator $\hat{f}$ is nearly minimax, because it shares almost the same quadratic risk as $\tilde{f}$. More precisely we have the following proposition whose proof is given in the appendix.

**Proposition 2** *With the same assumptions as in Proposition 1, let $\hat{t}_j$, $j = 0, \ldots, \hat{d}$ be the estimates of the change-points obtained through Proposition 1 and let $\hat{f}$ be the estimator of $f \in \mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$ as defined in expression (14), with $J = \log_2(\tau^{-2})/(1 + 2m)$. Then, as $\tau \to 0$,*

$$MISE(\hat{f}) \leq \mathcal{O}\left(\tau^{\frac{4m}{2m+1}}\right).$$

**Remarks**

1. In applied problems, the process $Y(t)$ is usually not observed at every point $t$ but is characterized by a uniform sampling which approximates $Y$ at a resolution that depends upon the sampling rate (see Remark 1 in Section 3). Now, we have already seen that the discrete regularly spaced nonparametric regression model given in (3) is equivalent to the white noise model $Y_n(t)$ given in (2) with $\tau = \sigma/\sqrt{n}$, provided that we take $f_n(t) = \sum_{i=0}^{n-1} f(t_i)\mathbf{I}_{\{t_i \leq t < t_{i+1}\}}(t)$. Hence, on a common probability space, the process $(Y_n, \mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S}))$ differs from $(Y, \mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S}))$ by the difference between $f_n$ and $f$. Now, since $\mathcal{H}(\Updownarrow, \mathcal{K}, \kappa, \mathcal{S})$ is a piecewise Sobolev ellipsoid, it follows by Theorems 15 and 16 in Donoho and Johnstone (1995, Section 9) that the risks of the discrete nonparametric regression problem and of the white noise model are equivalent. Our estimation strategy therefore also holds for the discrete nonparametric regression model.

2. For locating the discontinuities in Proposition 1, we have used the continuous wavelet transform. For efficient numerical implementation, no fast computational algorithm exists, and we need to discretize the scaling parameter $h$ along a sparse sequence. When the scale is discretized along a dyadic sequence, the wavelet transform can be computed with the fast dyadic wavelet transform of Mallat (see Abry (1994)). We already mentioned in Section 1 that the characterization theorems for local Lipschitz regularities for a sampled signal via the discrete dyadic wavelet transform remain valid and that such Lipschitz regularity characterization does not hold in general for the discrete wavelet transform. It is therefore imperative, in a discrete and practical setting, to use the discrete dyadic wavelet transform for estimating the location of the jump points.

3. Usually $\tau$ is only known up to a scale factor. For example, using the relation between the discrete model and the white noise model we have $\tau = \sigma/\sqrt{n}$, and $\sigma$ is rarely known in practice. One way to estimate the unknown scale is to follow Donoho and Johnstone (1994), that is to estimate $\sigma$ by the median absolute deviation of the $n-D$ smallest wavelet coefficients at the finest level available, divided by 0.6745. We are using here the $n-D$ smallest coefficients because in the presence of discontinuities it are exactly the largest coefficients at the finest level that carry the singularities, thus inducing a possibly biased estimation of $\sigma^2$. An alternative approach for the noise regression problem is to use a device suggested by Lombard (1988), that is to transform the original white noise model into a white noise model driven by rank statistics. Important to note is that the structure of the regression function

transformed in this way remains the same as in the original problem, but the scale parameter $\sigma^2$ is now equal to 1.

# 5 Simulations and an illustrative example

## 5.1 Simulations

To illustrate the behaviour of the estimation procedures developed in the previous sections, an empirical study was carried out. Data were generated from a regression model with two jumps. We used the following mean function

$$f(t) = \begin{cases} \exp\{-2(x-0.35)\} - 1 & \text{if} \quad x \in [0, 0.35) \\ \exp\{-2(x-0.35)\} & \text{if} \quad x \in [0.35, 0.65) \\ \exp\{2(x-0.65)\} + \exp(-0.6) - 2 & \text{if} \quad x \in [0.65, 1], \end{cases}$$

with an equidistant design and pseudo random Gaussian errors with variance $\sigma^2 = 0.1$ or $\sigma^2 = 0.25$ and sample sizes $n = 256$. This regression model was also considered in Gijbels, Hall and Kneip (1995). The basic experiment was repeated 100 times for each value of the noise variance. Figure 2 depicts a typical simulated data set with $\sigma^2 = 0.1$, together with the mean function as the solid curve.

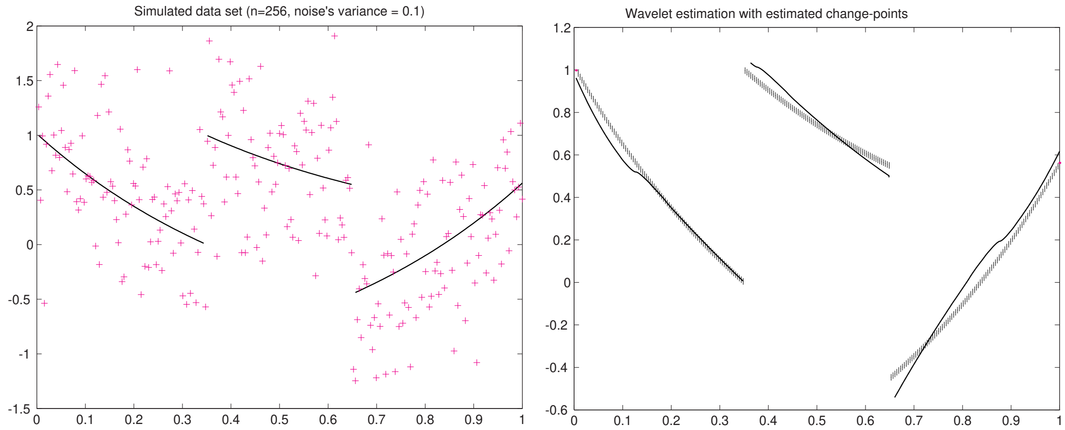



Figure 2 (left): A typical simulated data set with $\sigma^2 = 0.1$, together with the mean function as a solid curve. Figure 2 (right): The wavelet estimator $\hat{f}$, adapted to the estimated change-points (solid curve) plotted together with the original mean function (dotted curve).

The detector/locator of Section 3 with a dyadic wavelet transform based on a cubic B-spline wavelet (Mallat's Wave1 C-code) was applied to determine how many change-points (if any) are present and to estimate the exact location of the change-points. This wavelet satisfies the conditions imposed in Section 3. The grid that we used was a 64-point grid of the interval $[0,1]$ and $\alpha$ in Proposition 1 was taken equal to 1/8. For each simulated sample, the noise variance was estimated by the median absolute deviation of the $256-6$ smallest wavelet coefficients at the finest level ($J=8$), divided by 0.6745. The number of identified change-points over the 100 simulations and for each setting is given in Table 1. For correctly identified change points, the bias of the estimates for the location of change points and their empirical standard deviation are also reported. Moreover, for such samples, the wavelet estimator $\hat{f}$ with estimated jump points was computed (an example of such an estimation is reported in Figure 2) using a a Daubechies D6 adjusted to the interval wavelet. The average mean squared error AMSE was estimated by computing the average squared errors

$$ASE(\hat{f}) = \frac{1}{n}\sum_k (\hat{f}_n(t_k) - f(t_k))^2$$

for each sample and by averaging over the samples presenting a correct identification of change points.

| Estimation results | | |
|---|---|---|
| | $\sigma^2 = 0.1$ | $\sigma^2 = 0.5$ |
| $\hat{d}$ | | |
| 1 | 0.00 | 0.12 |
| 2 | 0.94 | 0.83 |
| 3 | 0.06 | 0.05 |
| MSE($\hat{t}_1$) | 7.16E-03 | 4.39E-02 |
| MSE($\hat{t}_2$) | 9.84E-03 | 1.26E-02 |
| AMSE($\hat{f}$) | 7.71E-02 | 1.14E-01 |

Table 1. The results on the estimated quantities from a total of 100 repetitions for each variance $\sigma^2 = 0.1$ and $\sigma^2 = 0.5$.

As one can see the detection/location procedure works quite well even for finite samples with a considerable variance. As expected both detection of jumps and estimation of the regression curve deteriorates as the noise variance increases.

## 5.2 An illustrative example

The example concerns the Nile data given in Cobb (1978) and displayed in Figure 3. The data set consists of measurements of the annual volume of discharge from the Nile river at Aswan for the years 1870 to 1930. The measurements are historically significant as evidence of a possible abrupt change in annual volume near the turn of the last century.

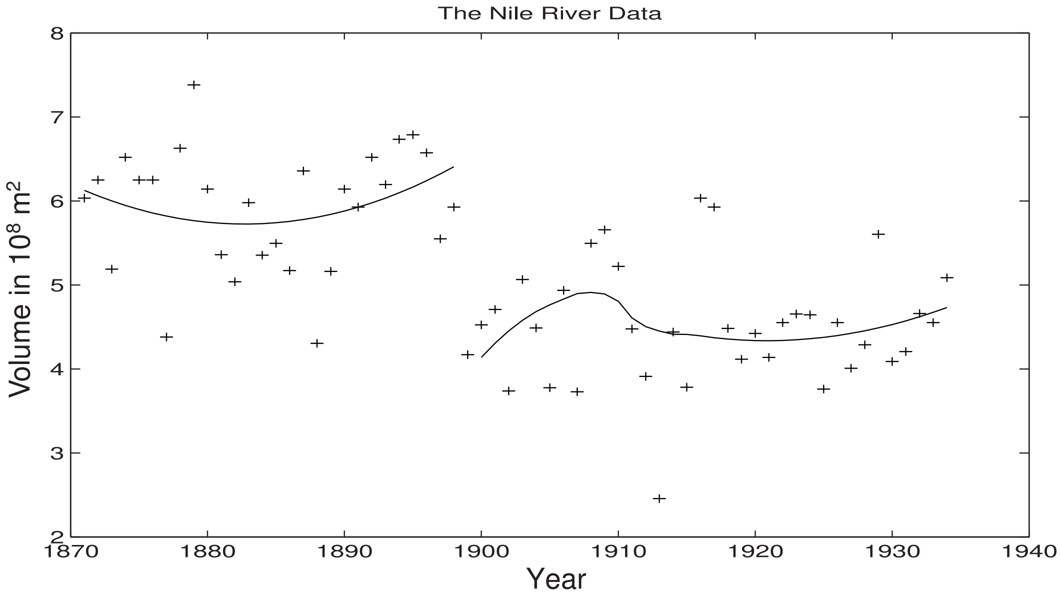



Figure 3: The Nile river flow data (dots), with the wavelet estimator $\hat{f}$, adapted to the estimated change-point (solid curve).

Figure 3 shows the fit from our method with a change-point for the year 1898. The strategy implemented for treating this example had also two steps. First, the detector/locator of Section 3 with a dyadic wavelet transform based on a cubic B-spline wavelet was applied to determine how many change-points (if any) are present and to estimate the exact location of the change-points. The grid that we used was a 64-point grid of the interval $[1870, 1930]$ and $\alpha$ in Proposition 1 was taken equal to $1/2$. One point only was identified as a change point, occurring between 1898 and 1899. Independent evidence that a change point at 1898 is real can be found in Cobb (1978). Finally the entire function was fitted with an estimated change point at 1898, using a Daubechies D8 adjusted to the interval wavelet and a hard thresholding rule with an estimated $\hat{\sigma} = 111$.

Despite the fact that this situation is no longer an asymptotic situation, we can see that the method performs quite well.

# 6 Appendix: Proof of Proposition 2.

Denote by

$$A = \{\hat{d} = d\} \cap \{\cap_{j=1}^{d}\{|\hat{t}_j - t_j| < 2h\}\}$$

and write

$$R(\hat{f}, f) = \mathbb{E}\{\|\hat{f} - f\|_2^2\} = \mathbb{E}\{\|\hat{f} - f\|_2^2 \mathbf{I}_A\} + \mathbb{E}\{\|\hat{f} - f\|_2^2 \mathbf{I}_{A^c}\}.$$

Note that

$$\mathbb{E}\{\|\hat{f}-f\|_2^2\mathbf{I}_{A^c}\} = \mathbb{E}\{\mathbb{E}[\|\hat{f}-f\|_2^2\mathbf{I}_{A^c}|\hat{t}_1,\ldots,\hat{t}_{\hat{d}},\hat{d}]\} = \mathbb{E}\{\mathbf{I}_{A^c}\mathbb{E}[\|\hat{f}-f\|_2^2|\hat{t}_1,\ldots,\hat{t}_{\hat{d}},\hat{d}]\}.$$

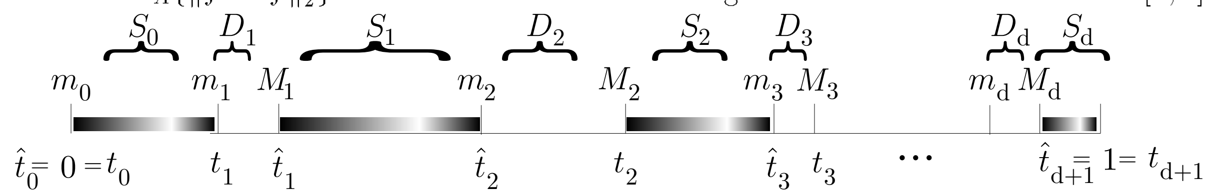
By the results of Section 3, we know that $\mathbb{E}[\|\hat{f}-f\|_2^2|\hat{t}_1,\ldots,\hat{t}_{\hat{d}},\hat{d}] = \mathcal{O}(1)$ and therefore,

$$\mathbb{E}\{\|\hat{f}-f\|_2^2\mathbf{I}_{A^c}\} = \mathcal{O}(P(A^c)).$$

Combining inequalities (12) and (13) together, it is easy to see that the complement of the event $A$ has a probability $o\left((\tau^2)^{\frac{2m}{2m+1}}\right)$ as $\tau$ goes to 0 and therefore

$$R(\hat{f},f) \leq \mathbb{E}\{\|\hat{f}-f\|_2^2\mathbf{I}_A\} + \mathcal{O}\left((\tau^2)^{\frac{2m}{2m+1}}\right).$$

We now deal with the term $\mathbb{E}\{\|\hat{f}-f\|_2^2\mathbf{I}_A\}$ which we will also occasionally denote as $\mathbb{E}_A\{\|\hat{f}-f\|_2^2\}$. We consider the following subdivision of the interval $[0,1]$:




where we use the notation:

$$\begin{aligned} m_i &= \min(\hat{t}_i, t_i), \quad i=1,\ldots,d \\ M_i &= \max(\hat{t}_i, t_i), \quad i=1,\ldots,d, \end{aligned}$$

with the convention $m_0 = M_0 = 0$ and $m_{d+1} = M_{d+1} = 1$.
With the above notation the interval $[0,1]$ is partitionned as

$$[0,1] = \{\cup_{i=0}^d S_i\} \cup \{\cup_{i=1}^d D_i\},$$

where

$$\begin{aligned} S_i &= ]M_i, m_{i+1}[, \quad i=1,\ldots,d-1 \\ S_0 &= [0, m_1[, \quad S_d = ]M_d, 1] \\ D_i &= [m_i, M_i], \quad i=1,\ldots,d. \end{aligned}$$

Note that not any of the sets $S_k$, $k=0,\ldots,d$ contains a change-point. We make use of the following formulae:

$$\mathbb{E}_A\{\|\hat{f}-f\|_2^2\} = \mathbb{E}\{\mathbb{E}[\|\hat{f}-f\|_2^2\mathbf{I}_A|\hat{\mathcal{F}}]\}$$

where $\hat{\mathcal{F}}$ is the $\sigma$-algebra generated by $\hat{t}_j$, $j = 1, \ldots, \hat{d}$. By our partition of the interval $[0, 1]$, we have

$$\mathbb{E}[\|\hat{f}-f\|_2^2 \mathbf{I}_A|\hat{\mathcal{F}}] = \mathbb{E}[\mathbf{I}_A \sum_{i=0}^{d} \int_{S_i} (\hat{f}(t)-f(t))^2 dt|\hat{\mathcal{F}}] + \mathbb{E}[\mathbf{I}_A \sum_{i=1}^{d} \int_{D_i} (\hat{f}(t)-f(t))^2 dt|\hat{\mathcal{F}}] \equiv I + II,$$

and therefore

$$\mathbb{E}_A\{\|\hat{f} - f\|_2^2\} = \mathbb{E}(I) + \mathbb{E}(II).$$

From our definition of the $S_i$'s and of the functions $\hat{\psi}_k^i$, and since $f$ has no discontinuities on $S_i$, using again the results of Section 3, we easily obtain

$$\mathbb{E}(I) = \mathcal{O}(\tau^{\frac{4m}{2m+1}}).$$

The analysis of the second term $II$ requires more attention since the intervals $D_i$ always contain jump-points. We will consider the case where $D_i = [t_i, \hat{t}_i]$. The case where $D_i = [\hat{t}_i, t_i]$ can be treated similarly and is therefore omitted. Write

$$\mathbb{E}[\mathbf{I}_A \int_{D_i} (\hat{f}(t) - f(t))^2 dt|\hat{\mathcal{F}}] = \mathbb{E}[\mathbf{I}_A \int_{D_i} (\hat{f}(t) - \tilde{f}(t) + \tilde{f}(t) - f(t))^2 dt|\hat{\mathcal{F}}]$$

where $\tilde{f}$ is the development in the interval $[t_i, t_{i+1}]$ using the basis $\psi_{k/[t_i,t_{i+1}]}$. We have,

$$\mathbb{E}[\mathbf{I}_A \int_{D_i} (\hat{f}(t)-f(t))^2 dt|\hat{\mathcal{F}}] \leq 2\mathbb{E}[\mathbf{I}_A \int_{D_i} (\hat{f}(t)-\tilde{f}(t))^2 dt|\hat{\mathcal{F}}] + 2\mathbb{E}[\mathbf{I}_A \int_{D_i} (f(t)-\tilde{f}(t))^2 dt|\hat{\mathcal{F}}].$$

It is easy to see that the second term of the right hand side of the above expression $\mathbb{E}[\mathbf{I}_A \int_{D_i} (f(t) - \tilde{f}(t))^2 dt|\hat{\mathcal{F}}]$ is of order $\mathcal{O}(\tau^{4m/(2m+1)})$. For the first term we proceed as follows. Using the development of $\hat{f}$ on the interval $[\hat{t}_{i-1}, \hat{t}_i]$ in terms of $\hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}$ and the development of $\tilde{f}$ on $[t_i, t_{i+1}]$ in terms of $\psi_{k/[t_i,t_{i+1}]}$ we can write

$$\begin{aligned}
\mathbb{E}[\mathbf{I}_A \quad & \textstyle\int_{D_i} \quad (\hat{f}(t) - \tilde{f}(t))^2 dt|\hat{\mathcal{F}}] \\
&= \mathbb{E}[\mathbf{I}_A \int_{t_i}^{\hat{t}_i} \left\{ \sum_{k=0}^{2^J-1} \hat{Y}_{k,(i-1)} \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) - \sum_{k=0}^{2^J-1} Y_{k,(i)} \psi_{k/[t_i,t_{i+1}]}(t) \right\}^2 dt|\hat{\mathcal{F}}] \\
&\leq 2\mathbb{E}[\mathbf{I}_A \int_{t_i}^{\hat{t}_i} \left\{ \sum_{k=0}^{2^J-1} \hat{\theta}_{k,(i-1)} \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) - \sum_{k=0}^{2^J-1} \theta_{k,(i)} \psi_{k/[t_i,t_{i+1}]}(t) \right\}^2 dt|\hat{\mathcal{F}}] \\
&\quad + 2\tau^2 \mathbb{E}[\mathbf{I}_A \int_{t_i}^{\hat{t}_i} \left\{ \sum_{k=0}^{2^J-1} \hat{\epsilon}_{k,(i-1)} \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) - \sum_{k=0}^{2^J-1} \epsilon_{k,(i)} \psi_{k/[t_i,t_{i+1}]}(t) \right\}^2 dt|\hat{\mathcal{F}}] \\
&= \Theta_{i,1} + \Theta_{i,2}.
\end{aligned}$$

Using now the definition of the coefficients $\hat{\theta}_{k,(i-1)}$ and $\theta_{k,(i)}$, after some algebraic manupilation, we get

$$\begin{aligned}\Theta_{i,1} &= 2\mathbb{E}\left[\mathbf{I}_A \int_{t_i}^{\hat{t}_i} \left\{ \sum_{k=0}^{2^J-1} \left( \int_{\hat{t}_{i-1}}^{t_i} f(u)\hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(u)du \right) \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) \right.\right. \\ &\quad - \sum_{k=0}^{2^J-1} \left( \int_{\hat{t}_i}^{t_{i+1}} f(u)\psi_{k/[t_i,t_{i+1}]}(u)du \right) \psi_{k/[t_i,t_{i+1}]}(t) \\ &\quad + \sum_{k=0}^{2^J-1} \left( \int_{t_i}^{\hat{t}_i} f(u)\hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(u)du \right) \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) \\ &\quad \left.\left. - \sum_{k=0}^{2^J-1} \left( \int_{t_i}^{\hat{t}_i} f(u)\psi_{k/[t_i,t_{i+1}]}(u)du \right) \psi_{k/[t_i,t_{i+1}]}(t) \right\}^2 |\hat{\mathcal{F}} \right].\end{aligned}$$

We will now determine the general form and orders of the above terms without writing down all technical details. For the first term in the above expression we note that for $t \in [t_i, \hat{t}_i]$ (the domain of integration) there are no jumps in this interval and hence the wavelet coefficients, denoted say by $\alpha_k^i$ are of order $\mathcal{O}(2^{-J(m+1/2)})$, for a given working resolution level $J$ (see e.g. Antoniadis (1994)). Similarly, the second term in the expression of $\Theta_{i,1}$ is of the form

$$\sum_{k=0}^{2^J-1} \beta_k^i \psi_{k/[t_i,t_{i+1}]}(t),$$

with $\beta_k^i$ of order $\mathcal{O}(2^{-J(m+1/2)})$.

For the third and the fourth term within the square, we may write these as

$$\begin{aligned}\sum_{k=0}^{2^J-1} & \left( \int_{t_i}^{\hat{t}_i} f(u)\hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(u)du \right) \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) - \sum_{k=0}^{2^J-1} \left( \int_{t_i}^{\hat{t}_i} f(u)\psi_{k/[t_i,t_{i+1}]}(u)du \right) \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) \\ & - \sum_{k=0}^{2^J-1} \left( \int_{t_i}^{\hat{t}_i} f(u)\psi_{k/[t_i,t_{i+1}]}(u)du \right) \left[ \psi_{k/[t_i,t_{i+1}]}(t) - \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t) \right].\end{aligned}$$

The first two terms can be treated with the same remarks as above. For the third term, recalling that the sequences $\psi_{k/[t_i,t_{i+1}]}(t)$ and $\hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t)$ are based on the same resolution level $J$, and the fact that by our assumptions on $\psi$, they are Lipschitz continuous, it is easy to see that the difference $\left[\psi_{k/[t_i,t_{i+1}]}(t) - \hat{\psi}_{k/[\hat{t}_{i-1},\hat{t}_i]}(t)\right]$ is of the order $2^{J/2}(\hat{t}_i - t_i)$. Putting all these considerations together we obtain that the term $\Theta_{i,1}$ behaves as

$$\Theta_{i,1} \leq C_1\mathbb{E}[\mathbf{I}_A \sum_{k=0}^{2^J-1} (\alpha_k^i)^2 |\hat{\mathcal{F}}] + C_2\mathbb{E}[\mathbf{I}_A \sum_{k=0}^{2^J-1} (\beta_k^i)^2 |\hat{\mathcal{F}}] + C_3\mathbb{E}[\mathbf{I}_A 2^{2J}(\hat{t}_i - t_i)^4 |\hat{\mathcal{F}}]$$

where $C_1$, $C_2$ and $C_3$ denote some generic constants. We can now determine the order of $\mathbb{E}(\Theta_{i,1})$. The first two terms lead to the order $2^{-2Jm} = \tau^{4m/(2m+1)}$. For the third term we find the order $2^{2J}h^4 = o(\tau^{4m/(2m+1)})$. It follows that, uniformly in $i$,

$$\mathbb{E}(\Theta_{i,1}) \leq \mathcal{O}(\tau^{4m/(2m+1)}).$$

An upper bound for the term $\mathbb{E}(\Theta_{i,2})$, follows from the following remarks. Define for each indices $\ell, q$ varying from $\lceil h\tau^{-2} \rceil$ to $\lfloor (1-h)\tau^{-2} \rfloor$, the standard Gaussian distributed, but not necessarily independent random variables

$$\epsilon_{k,\ell,q} = \int_{g_\ell}^{g_q} \psi_k(t|[g_\ell, g_q])dW(t)$$

where $g_\ell$ and $g_q$ are grid points from the set $G$ defined in Section 3 and $\psi_k(t|[g_\ell, g_q])$ denotes the $k$th wavelet basis function of $L^2[g_\ell, g_q]$. Using the fact that the estimated locations $\hat{t}_i$ belong to the grid $G$, it follows that for every $i$ there are indices $\ell$ and $q$ such that $\hat{\epsilon}_{k,(i)}$ is equal to $\epsilon_{k,\ell,q}$. Since the $\epsilon_{k,\ell,q}$ are independent of the jump-point estimators and the indicator function of $A$ is bounded by 1, and using Jensen's inequality we obtain

$$\mathbb{E}(\Theta_{i,2}) \leq 2\tau^2 \log(\sqrt{2}N^2)$$

where $N$ is the size of the grid $G$. By construction $N \leq \tau^{-2}$. It follows therefore that

$$\mathbb{E}(\Theta_{i,2}) \leq o(\tau^2 \log^3(\tau^{-2})) \leq \mathcal{O}(\tau^{4m/(2m+1)}).$$

Grouping all the inequalities finishes the proof of Proposition 2. $\square$


### Acknowledgment

This research was supported by the CNRS-INRIA project IDOPT 6 at IMAG-LMC of the University Joseph Fourier of Grenoble. Financial support by the contract "Projet d'Actions de Recherches Concertées", No. 93/98-164 of the Belgian Government and by FNRS-grant (No 1.50.001.95F) from the National Science Foundation (Belgium) is gratefully acknowledged. We wish to thank K. Oudshoorn for providing us with a preprint of her 1995 paper prior to publication. S. Mallat and D. Donoho and I. Johnstone are gratefully acknowledged for providing us respectively the Wavel and Wavelab statistical packages.


## REFERENCES


P. Abry (1994). *Transformées en ondelettes – Analyses multirésolutions et signaux de pression en turbulance.* Thèse de Doctorat. Université Claude Bernard, Lyon, France.

L. Andersson, N. Hall, B. Jawerth, and G. Peters (1993). Wavelets on closed subsets of the real line, tech. report to appear in *Topics in the theory and applications of wavelets*, Eds L. L. Schumacher and G. Webb, Academic Press.

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

M. Bunt, I. Koch and A. Pope (1995). Kernel-based nonparametric regression for discontinuous functions. *Statistics Research Report SRR 012-95*, The Australian National University.

J. Canny (1986). A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **8**, 679–698.

A. Cohen, I. Daubechies and P. Vial (1993). Wavelets on the interval and fast wavelet transforms. *App. Comp. Harm. Anal.*, **1**, 54–82.

G. W. Cobb (1978). The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika*, **65**, 243–251.

I. Daubechies (1992). *Ten Lectures on Wavelets*, CBMS-NSF Series in Applied Mathematics, SIAM, Philadelphia.

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

D. L. Donoho and I. M. Johnstone (1995). Minimax estimation via wavelet shrinkage. *Technical report*, Stanford University.

R. L. Eubank and P. L. Speckman (1994). Nonparametric estimation of Functions with Jump Discontinuities, in *Change-point Problems*, IMS - Lecture Notes, Eds. E. Carlstein, H.-G. Müller and D. Siegmund, **23**, 130 – 143.

I. Gijbels, P. Hall and A. Kneip (1995). On the estimation of jump points in smooth curves. Discussion Paper #9515, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

D. Girard (1990). Détection de discontinuités dans un signal par inf-convolution spline et validation croisée, Technical Report RR 702-I-M, University of Grenoble.

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

M. Holschneider (1995). *Wavelets: An analysis Tool.* Clarendon Press, Oxford.

A. P. Korostelev (1987). On minimax estimation of a discontinuous signal. *Theory of Probab. and Its Appl.*, **32**, 727–730.

P. J. Laurent and F. Utreras (1986). Optimal smoothing of noisy broken data with spline functions. *J. Approximation Theory and its Applications*, **2**, 71 – 94.

Y. C. Leclerc and S. W. Zucker (1987). The local structure of image discontinuities in one dimension. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **9**, 341–355.

C. Loader (1994). Change-point estimation using nonparametric regression. *Technical report*, AT & T Bell Laboratories.

F. Lombard (1988). Detecting change points by Fourier Analysis. *Technometrics*, **30**, 305–310.

S. Mallat (1989). Thoery for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on PAMI*, **11**, 674 – 692.

S. Mallat and W. L. Hwang (1992). Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory*, **2**, 617 – 643.

J. A. McDonald and A. B. Owen (1986). Smoothing with split linear fits. *Technometrics*, **28**, 195–208.

H. -G. Müller (1992). Change-points in nonparametric regression analysis. *Ann. Statist.*, **20**, 737–761.

C. G. M. Oudshoorn (1995). Minimax estimation of a regression function with jumps: attaining the optimal constant. *Technical report 934*, Utrecht University.

M. Pollak and D. Siegmund (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika*, **72**, 267–280.

C. Potier and C. Vercken (1994). Spline fitting Numerous Noisy Data with discontinuities, in *Curves and Surfaces*, Eds P. J. Laurent et al., 477-480, Academic Press.

P. Qui and B. Yandell (1994). A local polynomial jump detection algorithm in nonparametric regression. *Technical report*, Department of Statistics, University of Winsconsin-Madison.

P. L. Speckman (1994). Detection of Change-points in Nonparametric Regression. *Tech. Report*, University of Missouri-Columbia.

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

J. S. Wu and Chu, C. K. (1993). Nonparametric function estimation and bandwidth selection for discontinuous functions. *Statistica Sinica*, **3**, 557–576.