ON SOME EXACT PROPERTIES OF WAVELET REGRESSION AND DENSITY ESTIMATION

by

Olivier Renaud

SEPTEMBER 2000

Department of Statistics
STANFORD UNIVERSITY
Stanford, California
ON SOME EXACT PROPERTIES OF WAVELET REGRESSION AND DENSITY ESTIMATION

by

OLIVIER RENAUD
Visiting Scholar
Department of Statistics
Stanford University

SEPTEMBER 2000

This research was supported by the
Swiss National Science Foundation

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
On Some Exact Properties of Wavelet Regression and
Density Estimation *

Olivier RENAUD,
Visiting Scholar, Stanford University.
E-mail address: renaud@stat.stanford.edu

Abstract

In this paper, we give an exact formulation for the wavelet estimator, that can be applied in density, regression on a regular grid and regression with a random design. This formulation is handy and allows us to better understand the type of bias due to a given method for the estimation of the coefficients at the high resolution. In the second part of the paper, we use the result obtained previously to compute the influence function of various wavelet estimators. This tool allows us to see how the influences of observations can be different depending on their locations. The lack of shift-invariance can also be investigated. The influence function is also useful to compare two different approximation schemes for the wavelet estimator. We show that a local linear regression type approximation for the higher resolution coefficients induces a more extreme and variable influence of the observations on the final estimator than the more usual approximation.

1 Introduction

Wavelet methods have been used in Statistics for a few years now and are quite powerful in estimating objects of unknown smoothness, see for example Härdle, Kerkyacharian, Picard, and Tsybakov (1998) for their properties. The vast majority of papers deals only with regression estimation based on an equispaced design and estimate the curve only at these abscissa. Moreover, the methods presented often rely on a first approximation of the coefficients at the highest frequency. Although important in moderate samples, the difference between the approximate and the unbiased estimate is however often small in large sample.

*Most of the research was supported by the Swiss National Science Foundation.
Larger errors arise when we use the same kind of estimator for regression based on a random design or for density. Here, additional sources of variation are present, like the ones induced by the choice of the origin of the wavelet bases and the initial scale.

In this article, we first define in Section 2 the estimators, unbiased and approximate, for both regression and density. We also provide a few tools to quantify the estimates and the error of approximation when using a scheme as outlined above. The first step, given in Section 3, is to obtain an explicit and exact formula for all linear wavelet estimates. This formula is handy and allows us to better capture the basics of the estimates. It also has an interesting link with an algorithm to obtain a numerical approximation of the wavelet and the scaling function.

In Section 4 we compute the influence function of any approximation kernel, including the linear wavelet density (unbiased and approximation), of the regression estimate, and their thresholded counterparts. The influence function, usually used in the robustness context, is a very demonstrative tool that shows the influence of an observation on the estimate. It can be used to better understand the wavelet estimators, and to compare different estimators. It is useful to see the impact of different approximation scheme. In Section 4.3, we compare the usual regression estimator with a more involved loess-type approximation for the high frequency coefficients. Note that most of the computations of the influence functions are made possible thanks to the formula of the wavelet estimate given in Section 3.

## 2 Wavelets and Estimation

We provide here the minimal definitions and theorems of the wavelet theory in statistics. For a complete coverage, we refer to H"ardle, Kerkyacharian, Picard, and Tsybakov (1998) and Vidakovic (1999).

A wavelet basis is defined by a scaling function \( \varphi \in L^2 \) and its associated wavelet \( \psi \in L^2 \). One uses their translated and dilated versions: \( \varphi_{jk}(x) = 2^{j/2} \varphi(2^j x - k) \) and \( \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k) \), \( j, k \in \mathbb{Z} \). These functions are such that \( \{ \varphi_{j_0 k}, \psi_{j k}, j \geq j_0, k \in \mathbb{Z} \} \) is an orthonormal basis of \( L^2 \). This means that, for any \( f \in L^2 \) and \( j_0 \in \mathbb{Z} \),

\[
f(x) = \sum_{k \in \mathbb{Z}} \alpha[j_0, k] \varphi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}} \beta[j, k] \psi_{j k}(x),
\]

(1)

where \( \alpha[j, k] = \langle f; \varphi_{jk} \rangle \) and \( \beta[j, k] = \langle f; \psi_{jk} \rangle \). Here \( \langle \cdot ; \cdot \rangle \) stands for the \( L^2 \) inner product. This requirement demand a special form to the functions \( \varphi \) and \( \psi \), known
as the 2-scale equations: for any $j$ and $k$,

$$
\varphi_{jk}(x) = \sum_{m \in \mathbb{Z}} h[m - 2k] \varphi_{j+1,m}(x), \quad \psi_{jk}(x) = \sum_{m \in \mathbb{Z}} g[m - 2k] \varphi_{j+1,m}(x),
$$

where $h$ and $g$ are (discrete) filter with a finite $\ell^2$-norm. With very mild conditions, this implies the following link between the coefficients

$$
\alpha[j, k] = \sum_{m \in \mathbb{Z}} h[m - 2k] \alpha[j + 1, m], \quad \beta[j, k] = \sum_{m \in \mathbb{Z}} g[m - 2k] \alpha[j + 1, m],
$$

(3)

$$
\alpha[j + 1, k] = \sum_{l \in \mathbb{Z}} h[k - 2l] \alpha[j, l] + \sum_{l \in \mathbb{Z}} g[k - 2l] \beta[j, l].
$$

(4)

known as the cascade algorithm, see Vetterli and Kovacevic (1995). For density or regression estimation, the wavelet estimator of $f$ will be a truncation of (1) of the form:

$$
\hat{f}(x) = \sum_{k \in \mathbb{Z}} \hat{\alpha}[j_0, k] \varphi_{j_0,k}(x) + \sum_{j = j_0}^{J-1} \sum_{k \in \mathbb{Z}} \hat{\beta}[j, k] \psi_{jk}(x).
$$

(5)

Usually, one first compute a raw estimate (subscript $r$) $\hat{\alpha}_r[J, k]$ of the coefficients at the high resolution level $J$, and by means of Equation (3) obtain the coefficients $\hat{\alpha}_r[j_0, k]$ and $\hat{\beta}_r[j, k]$ of Equation (5). No real noise reduction has been made yet, and the function has to be regularised. The simplest is a projection that sets all the $\beta$ to zero. The estimator called linear wavelet estimator (subscript $l$) is then defined by $\hat{\alpha}_l[j_0, k] = \hat{\alpha}_r[j_0, k]$ and $\hat{\beta}_l[j, k] = 0$. A regularisation that shrinks the $\hat{\beta}_r[j, k]$ towards zero leads to the thresholded wavelet estimator (subscript $t$). The transformation writes $\hat{\alpha}_t[j_0, k] = \hat{\alpha}_r[j_0, k]$, and $\hat{\beta}_t[j, k] = \xi_\lambda \left( \hat{\beta}_r[j, k] \right)$, and there are two basic choices for the shrinkage rule

$$
\xi_\lambda(u) = \begin{cases} 
  u & \{u| > \lambda \} \\
  \text{sgn}(u)(|u| - \lambda)_+ & \text{soft threshold,}
\end{cases}
$$

where $(v)_+$ denotes the positive part of $v$. The parameter $\lambda$ has to be selected (either globally, or as a function of the level $j$, or individually for each $\{j; k\}$), and represents the trade-off between bias and variance. Once the regularisation has been done, one can either compute the estimate $\hat{f}_t$ by Equation (5) or compute an approximation by obtaining $\hat{\alpha}_t[J, k]$ by means of Equation (4) and setting $\hat{f}_t(k/2^j) = 2^{j/2} \hat{\alpha}_t[J, k]$. The advantage of the latter is that it does not need an explicit form for $\varphi$ or $\psi$ (if at the first step, the computation of $\hat{\alpha}_r[J, k]$ doesn’t either).
2.1 Density

The last thing not discussed above is how to obtain from the data an estimate of the raw coefficients at the high resolution \( \hat{\alpha}_r[J,k] \). Different solutions are possible, which have different properties (biased or not) and different computational difficulties (need of a good approximation of \( \varphi \) or not). The next two subsection define the possibilities in both the density and the regression cases. The consequences of this choice on the quality of the final estimate will be investigated in the remaining sections.

2.1 Density

For the density setting, suppose that \( X_1, \ldots, X_N \) are independent and identically distributed observations with distribution \( F \) and density \( f \in L^2 \). To estimate \( f \), since \( \alpha[J,k] = E(\varphi_{Jk}(X)) \), the moment condition (subscript \( m \)) gives us the the raw estimators of the coefficients:

\[
\hat{\alpha}_{r,m}[J,k] = \int \varphi_{Jk}(y) \, dF_N(y) = \frac{1}{N} \sum_n \varphi_{Jk}(X_n),
\]

(7)

where \( F_N \) is the empirical distribution function. Note, however, that an explicit form – or at least a good approximation – for the scaling function is needed, see Section 3.

There is a faster way to compute the coefficients to avoid using the explicit form of \( \varphi \). It computes the histogram of the observations \( X_1, \ldots, X_N \) with bandwidth \( h = 1/2^J \) and views the height of the histogram bars as regression points. One then obtains the (biased) estimator (the subscript \( b \) stands for box)

\[
\hat{\alpha}_{r,b}[J,k] = \frac{2^{J/2}}{N} \cdot \{ \# X_n \in I_{jk} = [k/2^J, (k+1)/2^J] \}.
\]

(8)

In a series of papers by Kerkyacharian, Picard and Johnstone, summarised in Härdle et al. (1998), the properties of the linear and the thresholded estimators are explored.

The other density estimator we will use is the Rosenblatt–Parzen kernel or convolution kernel density estimator, that is defined as \( \hat{f}_h(x) = (Nh)^{-1} \sum_n K((x - X_n)/h) \), where \( K \), the kernel, integrates to 1, and is symmetric around 0.

It is very convenient to view both the linear wavelet density estimate and the convolution kernel as approximation kernels, as shown by the following definition.

**Definition 1** A kernel is a function \( K(x,y) \) that takes two real arguments. For a distribution \( F \), the associated operator is defined by \( K^\pi(F) = \int K(x,y) \, dF(y) \). The convolution kernel is a special case where \( K(x,y) = K(x - y) \). The kernel associated to a wavelet basis is given by \( K(x,y) = \sum_{k \in \mathbb{Z}} \varphi(x - k) \varphi(y - k) \), and we define \( K^\pi_j(F) \) accordingly.
2.2 Regression

For the regression setting, suppose that \((X_1, Y_1), \ldots, (X_N, Y_N)\) are independent and identically distributed observations with \(X\) marginal distribution \(G\) and with the conditional distribution of \(Y | X = x\) is centered in \(f(x)\), which is the function we want to estimate. \(P(x, y)\) will denote the induced joint distribution. One usually set \(\hat{\alpha}_{r,d0}[J, k]\) to be \(2^{-J/2}\) times the average of the points in the interval \(I_{jk} = [k/2^J, (k + 1)/2^J]\). The subscript \(d0\) stands for 0-degree polynomial, since the average can be viewed as the fitting of a 0-degree polynomial in each box. Note that this method is very close to the box approximation (8). Kovac and Silverman (2000) proposed instead to do a linear regression with the observations in this interval and to set \(\hat{\alpha}_{r,d1}[J, k]\) to be \(2^{-J/2}\) times value of the regression at the middle point. Then they use the known covariance structure of the starting value to define a better thresholding rule.

Both of these methods simplify to \(\hat{\alpha}_r[J, k] = 2^{-J/2} Y_k\) if the observations are on a (non-random) regular grid at dyadic points.

3 Explicit Formulation of the Estimators

To ease the computation on the above estimators, it is important to have an handy explicit form. For the linear density estimator with the moment estimate, it is not too difficult to show that for any \(j_0 \leq J\), it can be written as

\[
\hat{f}_{l,m}(x) = \sum_{k \in \mathbb{Z}} \frac{1}{N} \sum_{n=1}^{N} \varphi_{j_0 k}(X_n) \varphi_{j_0 k}(x) = K_{j_0}^x (F_N).
\]  

(9)

In this section we will give a similar form for the linear density estimator with the box estimate, and for the regression estimates.

Let’s first note that for density and for regression, using the box approximation in the first step to compute the \(\hat{\alpha}_r[J, k]\), is in fact equivalent to compute the coefficients by the moment method, but using the Haar basis. So it is as if one uses the Haar basis at this first step and then another wavelet basis for the cascade and the inverse cascade. Likewise, using the approximation at the last step to obtain \(\hat{f}(k/2^J)\), is equivalent to using the exact estimate of Equation (5) but with the Haar basis at the last step. This different interpretation of these two approximations will turn out to be useful for the understanding of the next result and for the phenomenon described in the next section.

The above interpretation of the approximations turns out to be closely related to one of the method for the construction (approximation) of the scaling function \(\varphi\).
3. Explicit Formulation of the Estimators

We review here this way to obtain \( \varphi \) to any desired precision from the filters. It is due to Daubechies (1988) and works as follows. It iterates the two-scale equation, which gives

\[
\varphi^{(i)}(x) = \sum_{m \in \mathbb{Z}} h[m] \sqrt{2} \varphi^{(i-1)}(2x - m),
\]

with the indicator function (i.e. the Haar scaling function) \( \varphi^{(0)}(x) = \{0 \leq x < 1\} \) as starting point. This is equivalent to the application of the filter \( h \) to itself – discrete convolution without reversing the time – leading to the filters \( h^{(i)} \). At iteration \( i \) the approximation of \( \varphi \) is given as a piecewise constant function with knots at point \( k/2^i, \ k \in \mathbb{Z} \), where the values at these points are equal to \( h^{(i)}[k] \). It can be shown that the process preserves all the properties, such as orthogonality.

The next theorem is the central part of this section. It gives a simple form for all estimators that use an interval-based approximation of the coefficients. The result is that all the estimators discussed above have an explicit form that is handy and that shows to what extent an approximation scheme for the coefficients at the high resolution implies in the final estimate.

**Theorem 1** Suppose the raw coefficients at level \( J \) can be written as

\[
\hat{\alpha}_r[J, k] = \sum_{n=1}^{N} U_n \varphi_{jk}^{[0]}(X_n),
\]

where \( \varphi_{jk}^{[0]}(x) = 2^{j/2} \{x \in I_{jk} = [k/2^j, (k + 1)/2^j]\} \) is the Haar scaling function, and where \( U_n \) can depend on anything except \( k \). Then, for any \( j < J \), the coefficients are equal to

\[
\hat{\alpha}_r[j, k] = \sum_{n=1}^{N} U_n \varphi_{jk}^{[J-j]}(X_n), \quad \hat{\beta}_r[j, k] = \sum_{n=1}^{N} U_n \psi_{jk}^{[J-j]}(X_n).
\]

In addition, with the above coefficients, the linear estimator at level \( j_0 \) has the hybrid form

\[
\hat{f}(x) = \sum_{k \in \mathbb{Z}} \sum_{n=1}^{N} U_n \varphi_{j_0 k}^{[J-j_0]}(X_n) \varphi_{j_0 k}(x).
\]

If one applies the Haar approximation also at the final step, the linear estimator at level \( j_0 \) is instead

\[
\hat{f}(x) = \sum_{k \in \mathbb{Z}} \sum_{n=1}^{N} U_n \varphi_{j_0 k}^{[J-j_0]}(X_n) \varphi_{j_0 k}^{[J-j_0]}(x).
\]

The proof is given in the appendix.
3.1 Density

The previous theorem applies to the linear density estimate based on the histogram estimate of Equation (8) trivially with $U_n = N^{-1}$. It follows from the theorem that the difference between the estimate based on the moment condition (7) and the box approximation (8) is given by the difference between $\varphi$ and $\varphi^{(J-j_0)}$. This shows that as $J - j_0$ tends to infinity, both estimators given by Equations (12) and (13) converge to the moment estimate given in Equation (9). However, for the number of level computed in practice, the difference can be large. In Section 4.1.1, we see the consequence of this difference through the influence function.

Another conclusion that one can draw from the theorem is that even with the approximation at the first step, and with or without the Haar approximation at the last step, the estimator is still an approximation kernel. The corresponding kernel can easily be deduced from Equation (12) or (13) and is essentially an approximation of the kernel associated to a wavelet basis given in Definition 1. However, this shows also that the kernel changes with the number of level considered.

3.2 Regression

The usual definition of the estimate in the regression case does not satisfy the condition of Equation (11). For instance, the simple estimate that average on all value in the interval $I_{jk}$ has the form $U_n = 2^{-J}Y_n / \sum_m \{X_m \in I_{jk}\}$. It clearly depends on $k$, as it counts the number of observations in the interval $I_{jk}$.

However, as $U_n$ is multiplied by 0 for all $X_n$ not in $I_{jk}$, instead of counting the number of observations in the box $I_{jk}$, we can count the number of observations in the same box as $X_n$. This adds a summation, but does not depend on $k$ anymore.

We thus can write

$$\hat{\alpha}_{r,d0}[J,k] = \sum_{n=1}^{N} 2^{-J} Y_n \frac{1}{\sum_{t \in \mathbb{Z}} \sum_{m} \{X_m \in I_{jt}\} \{X_n \in I_{jt}\}} \varphi^{(0)}_{jk}(X_n)$$

$$= \sum_{n=1}^{N} 2^{-J} Y_n \sum_{t \in \mathbb{Z}} \frac{\{X_n \in I_{jt}\}}{\sum_{m} \{X_m \in I_{jt}\}} \varphi^{(0)}_{jk}(X_n),$$

(14)

that has the form required on Equation (11). By Theorem 1, we can simply replace $\varphi^{(0)}_{jk}$ by $\varphi^{(J-j_0)}_{jk}$ to obtain the value of $\hat{\alpha}_{r,d0}[j_0,k]$. The functional form for the estimate at $x$ become

$$R_{d0}^x(P) = 2^{-J} \sum_{t \in \mathbb{Z}} p_t^{-1} \sum_{k \in \mathbb{Z}} \varphi^{(J-j_0)}_{jk}(x) \int f(v) \varphi^{(J-j_0)}_{jk}(v) \{v \in I_{jt}\} dG(v),$$

(15)
where \( p_t = \int \{ v \in I_{Jt} \} \, dG(v) \).

For the second estimator, based on a local linear regression, the same trick can be applied and one obtains

\[
\hat{\alpha}_{r,d1}[j_0,k] = \hat{\alpha}_{r,d0}[j_0,k] + \sum_{n=1}^{N} 2^{-J} Y_n \sum_{i \in \mathbb{Z}} \frac{(a_i - C_i)(X_n - C_i)\{X_n \in I_{Jt}\}}{\sum_{m}(X_m - a_i)^2\{X_m \in I_{Jt}\}} \varphi_{j_0,k}^{(j - j_0)}(X_n),
\]

(16)

where \( a_i = (t + 1/2)/2^J \) is the center of the box \( I_{Jt} \) and \( C_i = \sum_m X_m \{X_m \in I_{Jt}\} / \sum_m \{X_m \in I_{Jt}\} \) its center of gravity with respect to the empirical distribution. Its functional form follows easily.

This new writing adds two summation signs at the high resolution level. However, it allows to have an explicit formulation for the estimator at any level, that is almost the same as for the high resolution level. It is a very convenient form, both for computational and theoretical reasons. It can have various application and next section is an example where the explicit formulation is the key to obtain the stated results.

## 4 Influence Functions

In this section, we will characterize some properties of the regression and the density estimators previously defined. Amongst them the sensitivity of the estimator to the design, to the choice of the origin of the wavelet basis, to the choice of the method for the raw estimation of the coefficients, and so on. A very useful tool for this purpose is the influence function, which is formally defined as the Gâteaux derivative of the functional of interest in the direction of a mass distribution. Note that usually the influence function is used to assess the robustness properties of estimators. Although it is possible to deduce these properties for our estimators, we will rather be interested in their sensitivity in a broad sense.

**Definition 2** Suppose that the estimator of interest can be set in a functional form, such that \( T(F_N) \) is the estimator of \( T(F) \), where \( F_N \) and \( F \) are the (p-dimensional) empirical and the underlying distribution function, respectively. Then, if it exists, the (p-dimensional) influence function of \( T \) at \( F \) in a point \( z = (z_1, \ldots, z_p) \) is given by

\[
IF(z; T, F) = \lim_{\epsilon \searrow 0} \frac{T(F_{\epsilon,z}) - T(F)}{\epsilon},
\]

with \( F_{\epsilon,z} = (1 - \epsilon)F + \epsilon \Delta_z \), where \( \Delta_z \) is the mass distribution at the point \( z \).
4.1 Approximation Kernels

This function defines the influence on the estimator of an infinitesimal contamination at \( z \). It allows us to deduce various characteristics of the estimator, like its maximal bias and its asymptotic variance, see Hampel, Ronchetti, Rousseeuw, and Stahel (1986) for a complete coverage of the matter. The general shape of the influence function is very instructive on the behavior of the estimator. Two global properties of the IF can give a first insight on the sensitivity of the estimate. The first desirable property for any estimator is given by a bounded influence function. Indeed, this ensures that a single observation cannot have immoderate consequences on the estimator. Another important property is the local-shift sensitivity defined as \( \sup_{x \neq y} \| \text{IF}(y; T, F) - \text{IF}(x; T, F) \| \| y - x \| \). It shows the (standardised) maximal change in the estimator due to a wiggling of the sample. The different approximations for the wavelet estimator will reveal to be quite different on this matter.

Note that the influence function is usually designed for parametric model with a finite dimensional parameter. Here, we have an infinite dimensional estimator as \( f \) is estimated for every point \( x \in \mathbb{R} \). We can look at this as having to compute an infinite number of influence functions – one for each \( x \). In some cases they will all look alike, whereas in most cases they will depend on \( x \).

We will first treat the case of density estimation. A lot of phenomenon that appears in the regression case are already present in density, but in a more tractable and interpretable fashion. We will first compute the influence function for any approximation kernel in Section 4.1. The thresholded wavelet density estimator is treated in Section 4.2 and finally, the regression case can be found in Section 4.3.

4.1 Approximation Kernels

The next result gives the influence function for any approximation kernel – including both linear wavelet density estimators and the convolution kernel.

**Theorem 2** Let \( K(x, y) \) be an approximation kernel and \( K^z(F) \) its associated operator (as in Definition 1) evaluated at a point \( x \) and for the distribution \( F \). Then, the influence function of the kernel at a point \( x \) is given by

\[
\text{IF}(z; K^z, F) = K(z, z) - K^z(F).
\]

The same result holds if one replaces \( K \) by \( K_j \) and \( K \) by \( K_j \).

The proof is obvious, since the estimator is linear with respect to the distribution. The important part in the influence function given in Theorem 2 is \( K(x, z) \) since the second term is the estimator itself and does not depend on \( z \). For the convolution
4.1 Approximation Kernels

\textbf{Figure 1} (a) Influence functions for the convolution kernel for 8 different places \( x \) where \( \tilde{f} \) is evaluated. (b) Same as (a), but for the Daub2 linear wavelet estimator.

kernels, \( K(x, z) = K(x - z) \) which means that the influence of an observation at \( z \) on the estimation at \( x \) depends only on the distance between \( x \) and \( z \). Up to a translation, all the influence functions are the same for different \( x \). We can thus treat \( x \) as fixed. Provided that the kernel is bounded, the influence is bounded, and dies away rapidly as \( z \) moves away from \( x \). For compact convolution kernels, the influence function is even zero outside an interval. Most of the convolution kernels have a bounded local-shift sensitivity. From the point of view of stability of the estimator towards small changes in the sample, the convolution kernels represent an ideal case. Figure 1(a) give an example with a Gaussian kernel \( K \).

The influence function of the linear wavelet estimator on the other hand is not as simple. It does not depend merely on the difference \( (x - z) \) but also on the location of \( x \) and \( z \) compared to the dyadic grid. We must thus look at several influence functions, for different values of \( x \). Each curve on Figure 1(b) shows the kernel \( K(x, z) \) for the linear density estimator as a function of \( z \) for a given values of \( x \). Here the Daub2 basis has been used. All the influence functions are bounded and die away rapidly. For compact scaling functions, the influence function is zero outside an interval, which means that observations far away from \( x \) have absolutely no influence on the estimation at \( x \). However, the influence functions may have high local-shift sensitivity and are quite different depending on the location of \( x \) compared to the dyadic grid. The influence functions are similar to the functions \( \varphi \) and \( \psi \). One of the influence functions would be equal to \( \varphi \) if the basis was shift-interpolating (i.e. such that there exists \( \tau \) with \( \varphi(\tau + n) = \delta[n] \)). The Daubechies’ bases are only \textit{almost shift-interpolating}, see (Vidakovic, 1999). The flattest curve in Figure 1(b) is quite similar to \( \varphi \) itself. The differences between the influence functions show that depending on the location of an observation compared to the dyadic grid, its influence on the general bearing of the curve is quite different. Some observations will have important positive and negative influence on parts of the curve, much like
4.1 Approximation Kernels

Figure 2 Influence functions for the Daub2 linear wavelet estimator as a function of \( x \) and \( z \).

an high order convolution kernel, whereas other will have an influence closer to a first order kernel. Note however that the global influence, if measured as \( \int K(x, z) \, dx \), is constant, since this integral is actually equal to 1 for all approximation kernels.

Since the influence functions are different for different points of evaluation \( x \), it is natural to plot \( K(x, y) \) as a function of its two arguments: the location of the contamination \( z \) and the point \( x \) where the estimator is evaluated. This is done in Figure 2. Each individual curve corresponds to the influence function evaluated at a given \( x \). This plot shows that although the influence functions are different from one another, they are continuously depending on \( x \). This is of course not true if the scaling function is discontinuous. Note that the spikes are due to the fact that we represent here the functions for the Daub2 basis. For a smoother basis, which should be the choice in practice, the peaks are replaced by smooth bumps.

It is worthwhile looking at the Haar basis as well. In this case, the influence function is \( \text{IF}(z; K^x, F) = \{ |x| \leq z < |x| + 1 \} - (F(|x| + 1) - F(|x|)) \). The first part of the influence functions is therefore the same for all \( z \) with the same integer value. However, this does not mean that the estimation is more stable. On the contrary,
the local-shift sensitivity is infinite and it may happen that an observation that is arbitrary close to a point \( x \) has no influence on the estimation at this point. Moreover the \( \text{IF} \) an the influence range of two neighboring points can be very different. The kernel \( K(x, y) \), as a function of both \( x \) and \( z \), is a set of cubes along the diagonal with support \([k, k+1] \times [k, k+1]\). This shows that the influence function is not even continuously varying with \( z \), since abrupt changes happen at integers. This situation is more critical than for smoother wavelet bases. It is however a well-known problem, since the linear wavelet estimator with the Haar basis is actually equivalent to the histogram.

### 4.1.1 Box approximation

As the estimate of Equation (7) is often replaced in practice by the approximation of Equation (8), it is worth looking at the influence function of the estimator based on the latter and compare it to the former. Its characteristics will also be useful in the regression context, where the estimators are also based on a box approximation.

As seen in Section 3.1 the corresponding kernel for the approximation of Equation (8) is an approximation of the kernel of the wavelet basis. The influence function being merely the kernel, the \( \text{IF} \) is also an approximation of the regular \( \text{IF} \). As an example, on Figure 3 are shown some \( \text{IF} \) of the box approximation in the case of the Daub2 wavelet. From the influence function point of view, the approximate estimators inherit both the shortcomings of the Haar basis and the shortcomings of the wavelet basis defined by \( \varphi \), without capturing their strengths. Indeed, the influence function is stepwise constant, hence the local-shift sensitivity is infinite and the influence function is not continuous as a function of \( x \) or of \( z \). Moreover, the influence function is a sort of constant approximation of the influence function given in Theorem 2 and shown on Figure 2 for the Daub2 case. The influence functions are quite different depending on the location of \( z \) compared to the dyadic grid and, therefore, the same phenomenon as for continuous wavelet bases holds.

The approximation of the coefficients by Equation (8) may be viewed at first sight as a stabilisation of the moment condition, since \( \varphi_{jk}(X_n) \) seems to vary too rapidly, inducing an important variance. The above study shows that it is not the case, and that the estimator is in fact less stable, since the \( \text{IF} \) is not even continuous. It has been found in Renaud (1999) that the box approximation is much more sensible to the choice of the origin of the wavelet basis than the moment estimate. The form of the \( \text{IF} \) explain this phenomenon.
4.2 Thresholded density estimator

We turn now to the influence function for the thresholded wavelet estimator. We first have to find the functional notation of the thresholded estimator at \( x \), noted \( \mathcal{T}_x^\lambda \), which is slightly more complicated than the linear case. It can be written as

\[
\mathcal{T}_x^\lambda(F) = f(x) = \sum_k \alpha_{[j_0, k]} \varphi_{j_0 k}(x) + \sum_{j=j_0}^{J-1} \sum_k \xi_{\lambda}(\beta[j, k]) \psi_{jk}(x) \\
= \mathcal{K}_{j_0}^\lambda(F) + \sum_{j=j_0}^{J-1} \sum_k \xi_{\lambda} \left( \int \psi_{jk}(y) \, dF(y) \right) \psi_{jk}(x),
\]

where \( \xi_{\lambda} \), the threshold rule, is defined as in Equation (6). For the hard threshold, it correspond to the linear estimator plus the projection of the true density on a subspace adaptively selected for \( F \). Of course, when the threshold \( \lambda \) is set to 0, \( \mathcal{T}_0^\lambda \) is the functional form of the linear wavelet estimator. The following theorem gives the influence function.

**Theorem 3** For a given distribution \( F \), a fixed value of the threshold \( \lambda \) and provided that no coefficient \( \beta[j, k] \) is exactly equal to \( +\lambda \) or \( -\lambda \), the influence function for the
4.2 Thresholded density estimator

A hard thresholded density estimator at a point \( x \) is given by

\[
IF(z; T^x, F) = K_{j_0}(x, z) + \sum_{j=j_0}^{J-1} \sum_{k \in I_j} \psi_{jk}(x) \psi_{jk}(z) \mathbf{1}\{|\beta[j, k]| > \lambda\} - T^x(F).
\]

For the soft threshold, \( \psi_{jk}(z) \mathbf{1}\{|\beta[j, k]| > \lambda\} \) is replaced by

\[
(\psi_{jk}(z) - \lambda) \mathbf{1}\{|\beta[j, k]| > \lambda\} + (\psi_{jk}(z) + \lambda) \mathbf{1}\{|\beta[j, k]| < -\lambda\}.
\]

The proof is in the appendix. It is not difficult to see that for the hard threshold, if all the \( \beta[j, k] \) are greater than \( \lambda \), the influence function will be equal to the one of the linear estimator \( (K_J(x, z) - K_J^x(F)) \) at the higher resolution level \( J \). Apart from the term \( T^x(F) \), which represents the “true estimator”, the \( IF \) still depends on which coefficients \( \beta[j, k] \) have been kept for the estimator. This dependence relies on \( F \), not on \( x \). The varying part of the influence function is thus equal to the linear kernel \( K_{j_0}(x, z) \) at the low resolution \( j_0 \), added by terms that are part of a wavelet kernel (based on \( \psi \)), but that are selected depending on the distribution \( F \). For a compact wavelet, the influence function will be bounded and compact, but if the underlying density is irregular, it may have sharp edges. Note that this is what we expect from thresholding. It can easily be illustrated in the following special case.

For the Haar basis, and a given \( x \), the influence function is restricted to an interval of the form \( I = [m/2^{j_0}; (m + 1)/2^{j_0}] \) and is piecewise constant with knots at points of the form \( k/2^J \). If the underlying density \( f \) is almost flat in the given interval and the threshold suitably chosen, all the \( \beta[j, k] \) will be smaller than the threshold and consequently the \( IF \) will simply be proportional to the indicator function of the whole interval \( I \) for any \( z \) in this interval. This shows that when the underlying function is very smooth, the thresholding procedure takes advantage to use all the observations in the interval to estimate with greater precision the average value of \( f \) on this interval. On the other hand, if an abrupt change in the density happens somewhere in the interval, some \( \beta[j, k] \) are much greater. As a result, the influence function differentiates between the different values of \( x \). The influence function for a given \( x \) becomes important only locally, in a neighboring region, and is zero or almost zero in other regions of the interval. As an example, suppose that in the given interval, \( f(x) = a \) for \( x < (m + 1/2)/2^{j_0} \) and \( f(x) = b \) for \( x \geq (m + 1/2)/2^{j_0} \), for two values \( a, b \geq 0 \). \( K_{j_0}(x, z) \) is equal to \( 2^{j_0} \) if both \( x \) and \( z \) are in the interval \( I \). Suppose that \( \lambda > |a - b|2^{-j_0} \), so that the only non-zero beta coefficient, \( \beta[j_0, m] \), is not thresholded. This implies that

\[
\sum_{j, k} \psi_{jk}(x) \psi_{jk}(z) \mathbf{1}\{|\beta[j, k]| > \lambda\} = \psi_{jom}(x) \psi_{jom}(z).
\]
Figure 4 (a) Influence functions for the Daub2 linear wavelet estimator for 4 different places \( x \) where \( \hat{f} \) is evaluated. (b) Same as (a), but for the thresholded estimator. Note how some influence functions become less important.

But, this expression is equal to \( 2^{2\lambda} \) if both \( x \) and \( z \) are in the same half interval and \( -2^{2\lambda} \) otherwise. As a result, if \( x \) and \( z \) are in the same subinterval, the IF is equal to \( 2^{2\lambda+1} \), and the IF is reduced to 0 when they are not. Informally, the procedure detects that we are in presence of two different regimes and therefore completely separates the groups, so that observations from one part do not influence the estimation of the second.

In Figure 4 are shown 4 influence functions in the linear case (\( \lambda = 0 \)) and in the thresholded case, with the Daub2 basis. As an example, the highest broken-line IF has dramatically changed between the linear and the threshold case. Its scope (its support) has shrunk a lot, while its maximal value has almost doubled. Only observations very close to 0.5 will have an influence on the value of the estimation at this point.

To summarise this section, it seems that the special form of the influence function for the linear wavelet estimator is a drawback and that other estimators like the convolution kernel are better. However, it is the necessary price to pay for the adaptivity of the thresholded procedure. No method can be as adaptative to the
4.3 Regression

signal as the wavelet basis is, but we must admit as a counterpart the lack of shift-invariance even in the linear case. There is however a way to circumvent this and to lower the variance of the estimate, while keeping the adaptivity, as shown in Renaud (1999).

The IF is a useful tool to compute and compare different wavelet procedures and let us understand exactly the way thresholding works with different signals. We can e.g. compare the IF of the translation invariant de-noising of (Coifman and Donoho, 1995) to the traditional thresholding policies.

For the computation of the influence function, the threshold has been supposed fixed. The influence function with a data-based threshold can not be computed for the following reason. Any thresholding rule is typically depending on the finiteness of the sample and writes \( \lambda = \lambda(F_N) \). However, the threshold cannot be put in a functional form \( \lambda = \lambda(F) \). For example, \( \lambda \) has to increase with \( N \) and should therefore be equal to \( \infty \) for any distribution that cannot be written as an empirical distribution. The influence function is a tool that works only with estimators that can be written as functionals. Nevertheless, the influence function with a fixed threshold still makes sense if one can show that one observation can only have a bounded influence on the data-driven choice of the threshold. We cannot conceive any example in density estimation where this is not the case. For instance, the influence is clearly bounded if one uses the universal threshold rule and the MAD to estimate the noise standard error. Note that this is not necessarily true in the regression setting, as we will see in the next section.

4.3 Regression

The regression case is of course more complicated but often of more interest. As shown in Section 3.2, since the estimate at any resolution can be written in the form required to apply Theorem 1, the influence function is computable, as well. The next theorem gives the IF for the two regression estimates discussed in Sections 2.2 and 3.2.

**Theorem 4** With the same notation as in Sections 2.2 and 3.2, let \( R_{d0}^r(P) \) and \( R_{d1}^r(P) \) be the functional forms of the regression estimator that do an average, respectively a linear regression to obtain the starting points. Then, the influence functions of the estimators at a point \( x \), with a perturbation \( z = (z_1, z_2)' \) are given respectively
by

\[ IF(z; R_{\delta_0}, P) = 2^{-J} \sum_{t \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \varphi_{j_0 k}^{(J-j_0)}(x) \sum_{z_1 \in I_{Jt}} (z_2 \varphi_{j_0 k}^{(J-j_0)}(z_1) - f(v) \varphi_{j_0 k}^{(J-j_0)}(v)) \{ v \in I_{Jt} \} dG(v) \]

and

\[ IF(z; R_{\delta_1}, P) = IF(z; R_{\delta_0}, P) + 2^{-J} \sum_{t \in \mathbb{Z}} q_t^{-1} (z_1 - c_t) \{ z_1 \in I_{Jt} \} \sum_{k \in \mathbb{Z}} \varphi_{j_0 k}^{(J-j_0)}(x) \]

\[ - \int f(v)(v - c_t) \varphi_{j_0 k}^{(J-j_0)}(v) \{ v \in I_{Jt} \} dG(v) \]

\[ + q_t^{-1} (a_t - c_t)(a_t - c_t) \int f(v)(v - c_t) \varphi_{j_0 k}^{(J-j_0)}(v) \{ v \in I_{Jt} \} dG(v), \]

where \( c_t = p_t^{-1} \int v \{ v \in I_{Jt} \} dG(v) \) and \( q_t = \int (v - c_t)^2 \{ v \in I_{Jt} \} dG(v) \).

The proof of both results are not displayed here, as they are very long and technical, but based on the same idea as previous theorems. The results have been checked numerically and on the algebraic package Maple.

We can note from the theorem that both estimators are not robust with respect to the response variable, since both \( IF \) are linear in \( z_2 \) and therefore the estimate can be pulled as high as wanted by increasing the value of \( z_2 \). This comes with no surprise, since the linear wavelet estimator can be viewed as the solution of a least-squares problem. This fact will not be improved by hard or soft thresholding, since both can be viewed as penalized least-squares solutions. If the sample is at risk concerning outliers, one shouldn’t use the above estimator. To alleviate this problem, Sardy (1998) proposes a way to robustify the procedure of estimation of the wavelet coefficients that treat at the same time the thresholding and the robust approach. Kovac and Silverman (2000) propose an ad-hoc outlier detection method.

Note also that for a given point of interest \( x \), both \( IF \) are zero except in an interval, provided that \( \varphi_{j_0 k}^{(J-j_0)}(x) \) is bounded. Thanks to the similarity with the \( IF \) of the density estimator, all remarks done on density apply to regression as well. In particular, the dependence of the \( IF \) on the number of level through \( \varphi^{(J-j_0)} \). Figure 5 will demonstrate the similarities. The thresholding will also produce a similar result on the \( IF \) for density and regression. We therefore refer to Section 4.2 for the effect of the threshold.

An example of the two influence functions with the Daub2 wavelet basis is given on Figure 5. Here the density of \( X \) is uniform between 0 and 1, and the expectation of \( Y \) given \( X \) is one for \( X < 0.5 \) and zero otherwise. Figure 5(a) gives the influence
function $\mathbf{IF}(\mathbf{z}; \mathcal{R}_{20}^z, P)$ for the box estimation and Figure 5(b) displays $\mathbf{IF}(\mathbf{z}; \mathcal{R}_{20}^z, P)$ for the local linear regression approximation. Only the middle part of the $\mathbf{IF}$ is shown. As expected the first $\mathbf{IF}$ is piecewise constant. One could expect that the second estimator will correct this towards an $\mathbf{IF}$ that is more regular, in the sense that the gaps are shorter. In fact, it is the contrary that happens: the influence of an observation at a point $z$ is more extreme than in the simple case. First, the maximal value of the $\mathbf{IF}$ is larger, and second, some gaps are even more important than in the simple case. The general form of the $\mathbf{IF}$ seems also strange. It seems from the figure that the local linear regression lowers the stability of the estimator and increases its variability by letting some points inside an interval to be of much greater influence than other.

To increase the stability, it seems to be better to use one of the following algorithm for the estimation of the coefficients at the highest resolution $J$.

- Use the moment estimate $\hat{\alpha}[J,k] = N^{-1} \sum_n Y_n \varphi_{J,k}(X_n)$ or

- Use a kernel-type of estimate of the form $\hat{\alpha}[J,k] = 2^{-J} \sum_n Y_n K(2^J X_n - (k + 1/2))/\sum_n K(2^J X_n - (k+1/2))$. Actually, this simply replaces the box kernel by a smoother one. A possible kernel would be constant over an interval of length 1 and decrease to zero on both side during an other interval of length 1. If the deceeding curve is suitably chosen (central symmetry), it has the additional advantage that the sum of the weight on each observation is constant.
5. Conclusion

In both case, we can use the idea of Kovac and Silverman (2000) to follow the covariance structure of the coefficients at other levels. This improves the thresholding procedure, since we can include the information on the individual variance of the coefficients.

5 Conclusion

Thanks to an algorithm close to the one used for the numerical approximation of a wavelet or a scaling function, we give in this paper a direct formulation for wavelet estimators. The formula clearly shows the consequences of the approximation of the coefficients at the high resolution on the final estimate. This result applies for both the regression and the density case, and for a lot of different approximation schemes. We also show how the influence of observations differs according to their locations compared to the dyadic grid. A comparison of how different approximation schemes lead to different weights for the observations is treated as well. More stable algorithms for the estimation of the coefficients are proposed.

A Proofs

A.1 Proof of Theorem 1

The proof of the first assertion is by induction on \( i = J - j \). It is true for \( i = 0 \). Suppose the assertion true for \( i - 1 \). We prove that the relation is then true for \( \hat{\alpha}_r[j, k] = \hat{\alpha}_r[J - i, k] \). We have

\[
\hat{\alpha}_r[J - i, k] = \sum_{m \in \mathbb{Z}} h[m - 2k] \hat{\alpha}_r[J - i + 1, m]
\]

\[
= \sum_{m \in \mathbb{Z}} h[m - 2k] \sum_{n=1}^{N} U_n \varphi_{J-i+1,m}(x_n)
\]

\[
= \sum_{n=1}^{N} U_n 2^{(J-i)/2} \sum_{m \in \mathbb{Z}} h[m] \sqrt{2} \varphi^{(i-1)}(2^{J-i} x_n - (m + 2k))
\]

\[
= \sum_{n=1}^{N} U_n 2^{(J-i)/2} \varphi^{(i)}(2^{J-i} x_n - k)
\]

\[
= \sum_{n=1}^{N} U_n \varphi_{J-i,k}(x_n),
\]
A.2 Proof of Theorem 3

which shows the result. The same applies for \( \hat{\beta}_r[J - i, k] \) by simply replacing \( h \) by \( g \). The proof of Equation (13) is an iterative use of Equations (4) and (10). Since the \( \hat{\beta}[j, k] \) have been set to 0, we have

\[
\hat{f}(x) = \sum_k \hat{\alpha}[J, k] \varphi^{(0)}_{j, k}(x) \\
= \sum_k \sum_l h[k - 2l] \hat{\alpha}[J - 1, l] \varphi^{(3)}_{j, k}(x) \\
= \sum_l \hat{\alpha}[J - 1, l] \varphi^{(1)}_{j-1, l}(x) \\
= \ldots \sum_k \hat{\alpha}[j_0, k] \varphi^{(J-j_0)}_{j_0, k}(x),
\]

which, given the first part of the theorem, shows the result. Equation 12 is proved the same way, using (2) instead of (10).

A.2 Proof of Theorem 3

We first note that

\[
\beta_{\epsilon, z}[j, k] = \int \psi_{j, k}(y) \, dF_{\epsilon, z}(y) \\
= (1 - \epsilon) \int \psi_{j, k}(y) \, dF(y) + \epsilon \int \psi_{j, k}(y) \, d\Delta_z(y) \\
= (1 - \epsilon)\beta[j, k] + \epsilon \psi_{j, k}(z),
\]

which is therefore continuous in \( \epsilon \). This implies that for each couple \( \{j; k\} \) there exists an \( \epsilon_{jk} \) such that, for any \( 0 \leq \epsilon \leq \epsilon_{jk}, \beta[j, k] \) and \( \beta_{\epsilon, z}[j, k] \) are either both greater than \( \lambda \) or both smaller than \( -\lambda \) or both in-between. As \( f \in L^2 \), there are only finitely many \( \beta[j, k] \) such that \( |\beta[j, k]| > \lambda/2 \). We can therefore find an \( \epsilon^* \) for which the above properties hold for all \( \{j; k\} \) and for every \( 0 \leq \epsilon \leq \epsilon^* \). For such an \( \epsilon \), the hard threshold writes

\[
\xi_\lambda(\beta_{\epsilon, z}[j, k]) = \beta_{\epsilon, z}[j, k] \, 1\{|\beta_{\epsilon, z}[j, k]| > \lambda\} \\
= \beta_{\epsilon, z}[j, k] \, 1\{|\beta[j, k]| > \lambda\} \\
= (1 - \epsilon)\beta[j, k] \, 1\{|\beta[j, k]| > \lambda\} + \epsilon \psi_{j, k}(z) \, 1\{|\beta[j, k]| > \lambda\} \\
= (1 - \epsilon)\xi_\lambda(\beta[j, k]) + \epsilon \psi_{j, k}(z) \, 1\{|\beta[j, k]| > \lambda\}.
\]

We now can compute the numerator, which we note \( \text{num}_\epsilon \), that is in the definition of the influence function, for the functional \( T^x_\lambda(F) \) and for an \( \epsilon \) such that \( 0 \leq \epsilon \leq \epsilon^* \).
We have

\[
\text{num}_\varepsilon = \mathcal{K}_{\varphi}(F_{\varepsilon,x}) + \sum_{j,k} \xi(\beta_{[j,k]}) \psi_{jk}(x) - \mathcal{K}_{\varphi}(F) - \sum_{j,k} \xi(\beta_{[j,k]}) \psi_{jk}(x) \\
= \varepsilon \left\{ \mathcal{K}_{\varphi}(x) - \mathcal{K}_{\varphi}(F) \right\} + \varepsilon \sum_{j,k} \left\{ \psi_{jk}(x) \mathbf{1}_{\{\|\beta_{[j,k]}\| > \lambda\}} - \xi(\beta_{[j,k]}) \right\} \psi_{jk}(x) \\
= \varepsilon \left\{ \mathcal{K}_{\varphi}(x) + \sum_{j,k} \psi_{jk}(x) \psi_{jk}(x) \mathbf{1}_{\{|\beta_{[j,k]}| > \lambda\}} - T_{\lambda}(F) \right\}.
\]

The result follows immediately. Similar computations can be done for the soft threshold to give the stated result.

References


