Statistica Sinica **12**(2002), 1275-1290

SENSITIVITY AND OTHER PROPERTIES OF WAVELET REGRESSION AND DENSITY ESTIMATORS

Olivier Renaud

University of Geneva

Abstract: We give a unified, non-iterative formulation for wavelet estimators that can be applied in density estimation, regression on a regular grid and regression with a random design. This formulation allows us to better understand the bias due to a given method of coefficients estimation at high resolution. We also introduce functional representations for estimators of interest. The proposed formulation is well suited for the study of estimation bias and sensitivity analysis and, in the second part, we compute the influence function of various wavelet estimators. This tool allows us to see how the influence of observations can differ strongly depending on their locations. The lack of shift-invariance can be investigated and the influence function can be used to compare different approximation schemes for the wavelet estimator. We show that a local linear regression-type approximation for the higher resolution coefficients induces more extreme and variable influence of the observations on the final estimator than does the standard approximation. New approximation schemes are proposed.

Key words and phrases: Approximation kernel, influence function, irregular design, functional, sensitivity to the design, shift invariance.

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 Vidakovic (1999). The majority of papers deal with regression estimation based on an equispaced design and estimation only at these equispaced points. 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 this approximation scheme and an unbiased estimator is often negligible for large samples.

Larger errors arise when we use the same kind of approximation for regression based on a random design or for density estimation. Here, additional sources of variation are induced by the design, by the choice of the origin, and the choice of the initial scale of the wavelet basis.

In Section 2 we define the different estimators for both regression and density estimation. In Section 3, a direct formula for the linear wavelet estimators

is given. This allows us to capture their basic properties. The formula 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 an arbitrary approximation kernel, including the linear wavelet density estimators, of the regression estimators and their thresholded counterparts. This can be used to better understand wavelet estimators, and to compare different methods of estimation. In Section 4.3, we compare the standard regression estimator with a more involved local linear regression type of approximation for the coefficients at the highest frequency and we propose new estimators.

2. Wavelets and Estimation

We provide here the basic definitions and theorems of the wavelet theory in statistics. For a complete coverage, see Vidakovic (1999). A wavelet basis allows us to expand any function $f \in \mathsf{L}^2$ on the orthonormal basis consisting of $\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$ and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$:

$$f(x) = \sum_{k\in\mathbb{Z}} \alpha[j_0,k]\ \varphi_{j_0k}(x) + \sum_{j=j_0}^{\infty}\sum_{k\in\mathbb{Z}} \beta[j,k]\ \psi_{jk}(x), \tag{1}$$

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. The above properties demand a special form to the functions φ and ψ, known as the 2-scale equations: for any j and k,

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

where h and g are (discrete) filters with a finite ℓ^2-norm. Under very mild conditions, this implies the following links between the coefficients:

$$\alpha[j,k] = \sum_{m\in\mathbb{Z}} h[m-2k]\ \alpha[j+1,m], \qquad \beta[j,k] = \sum_{m\in\mathbb{Z}} g[m-2k]\ \alpha[j+1,m], \tag{3}$$

$$\alpha[j+1,k] = \sum_{l\in\mathbb{Z}} h[k-2l]\ \alpha[j,l] + \sum_{l\in\mathbb{Z}} g[k-2l]\ \beta[j,l], \tag{4}$$

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

$$\hat f(x) = \sum_{k\in\mathbb{Z}} \hat\alpha[j_0,k]\ \varphi_{j_0k}(x) + \sum_{j=j_0}^{J-1}\sum_{k\in\mathbb{Z}} \hat\beta[j,k]\ \psi_{jk}(x). \tag{5}$$

Usually, one first computes a *raw* estimator (subscript r) $\hat\alpha_r[J,k]$ of the coefficients at some high resolution level J and then, by means of (3), obtains the coefficients $\hat\alpha_r[j_0,k]$ and $\hat\beta_r[j,k]$. In this process, no real noise reduction has been made,

and the function $\hat{f}$ has to be regularised. The simplest approach is a projection that sets all the β to zero. The linear wavelet estimator (subscript l) is then defined by setting $\hat{\alpha}_l[j_0,k] = \hat{\alpha}_r[j_0,k]$ and $\hat{\beta}_l[j,k] = 0$. A regularisation that shrinks some of the $\hat{\beta}_r[j,k]$ towards zero leads to the thresholded wavelet estimator (subscript t). The coefficients are $\hat{\alpha}_t[j_0,k] = \hat{\alpha}_r[j_0,k]$, and $\hat{\beta}_t[j,k] = \xi_\lambda\left(\hat{\beta}_r[j,k]\right)$. The two basic choices for the shrinkage rule are $\xi_\lambda(u) = u\ I(|u| > \lambda)$ for hard thresholding and $\xi_\lambda(u) = \operatorname{sgn}(u)\,(|u| - \lambda)_+$ for soft thresholding, where $(v)_+$ denotes the positive part of v. The parameter λ has to be selected, and regulates the trade-off between the bias and variance of the estimator. Once the regularisation has been done, one can compute the estimator $\hat{f}_t$ either by (5) or by approximation, obtaining $\hat{\alpha}_t[J,k]$ using (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 require an explicit form for φ or ψ.

Finally we discuss how to obtain an estimator of the raw coefficients at high resolution $\hat{\alpha}_r[J,k]$ from the data. Different solutions are possible with different properties (biased or not) and different computational complexities. They are defined in the following subsections. The impact of this choice on the quality of the final estimator is investigated in the remaining sections.

2.1. Density estimation

Let $X_1, \ldots, X_N$ be independent and identically distributed observations with distribution F and density $f \in \mathsf{L}^2$. Since $\alpha[J,k] = E\left(\varphi_{Jk}(X)\right)$, the empirical moment (subscript m) gives us the *raw* estimators of the coefficients:

$$\hat{\alpha}_{r,m}[J,k] = \int \varphi_{Jk}(y)\ dF_N(y) = \frac{1}{N}\sum_n \varphi_{Jk}(X_n), \tag{6}$$

where F_N is the empirical distribution function. Note that a good approximation of the scaling function is needed, see Section 3. A faster way to compute the coefficients that avoids using φ gives the (biased) estimator (the subscript b stands for box)

$$\hat{\alpha}_{r,b}[J,k] = \frac{2^{J/2}}{N}\sum_n I(X_n \in B_{Jk} = [k/2^J,\ (k+1)/2^J]). \tag{7}$$

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

Both the linear wavelet density estimator and the convolution kernel can be viewed as operators associated with approximation kernels, defined as follows.

Definition 1. For a distribution F, the operator associated with a kernel $K(x,y)$ is defined by $\mathcal{K}^x(F) = \int K(x,y)\ dF(y)$. The convolution kernel has

$K(x,y) = K(x-y)$. 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 $K_j(x,y) = 2^j K(2^j x, 2^j y)$ and $\mathcal{K}_j^x(F)$ are defined accordingly.

2.2. Regression

Let $(X_1,Y_1),\ldots,(X_N,Y_N)$ be independent and identically distributed observations with X-marginal distribution G and with the conditional distribution of $Y|X=x$ centered at $f(x)$, our estimand. Let $P(x,y)$ denote the induced joint distribution. One usually sets $\hat{\alpha}_{r,d0}[J,k]$ to be $2^{-J/2}$ times the average of the Y's in the interval $B_{Jk} = [k/2^J,\ (k+1)/2^J]$, compare with (7). The subscript $d0$ stands for 0-degree polynomial (i.e., average). Kovac and Silverman (2000) compute a linear regression with the observations in this interval and set $\hat{\alpha}_{r,d1}[J,k]$ to be $2^{-J/2}$ times value of the regression at the middle point. They utilize known covariance structure of these coefficients to define an improved thresholding rule.

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

3. Unified Formulation of the Estimators and Interval Based Approximation

It can be readily shown that for any $j_0 \le J$, the linear density estimator with the empirical moment (6) can be written as

$$\hat{f}(x) = \sum_{k\in\mathbb{Z}} \frac{1}{N} \sum_{n=1}^{N} \varphi_{j_0 k}(X_n)\ \varphi_{j_0 k}(x) = \mathcal{K}_{j_0}^x(F_N). \tag{8}$$

Theorem 1 shows that most wavelet density and regression estimators have a similar representation for all points x of interest.

Let us 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 equivalent to computing the coefficients by the moment method using the Haar basis. Thus it is as if one uses the Haar basis at this first step, and then possibly 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 estimator of (5), but with the Haar basis. Such interpretation of these two approximations will turn out to be useful for the understanding of Theorem 1 and of the influence function.

A closely related topic is a particular method for the construction (approximation) of the scaling function φ, due to Daubechies (1988), it consists of iterating the two-scale equation

$$\varphi^{\{i\}}(x) = \sum_{m\in\mathbb{Z}} h[m]\ \sqrt{2}\varphi^{\{i-1\}}(2x-m) \tag{9}$$

with the indicator function (i.e., the Haar scaling function) $\varphi^{\{0\}}(x) = I(0 \le x < 1)$ as starting point. The next theorem is the central part of this section. It gives a simple form for wavelet based estimators and it is computationally economic for fixed and random designs regardless of the design being dyadic or not. This formulation is first useful to obtain the wavelet-based estimators on all points and not only at points of the form $k/2^J$. It also permits a common form for all interval-based density and regression estimators. In addition, it shows directly the impact and bias of different approximation schemes in the final estimators.

Theorem 1. *Suppose the raw coefficients at level J can be written as*

$$\hat{\alpha}_r[J,k] = \sum_{n=1}^{N} U_n \varphi_{Jk}^{\{0\}}(X_n), \tag{10}$$

where $\varphi_{Jk}^{\{0\}}(x) = 2^{J/2} I(x \in B_{Jk} = [k/2^J, (k+1)/2^J])$ is the Haar scaling function, and where U_n can depend on anything except k. Then for any $j < J$,

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

where $\varphi^{\{J-j\}}$ and $\psi^{\{J-j\}}$ are Daubechies approximation as in (9)*. With the above coefficients, the linear estimator at level j_0 has a hybrid form*

$$\hat{f}(x) = \sum_{k \in \mathbb{Z}} \sum_{n=1}^{N} U_n \varphi_{j_0 k}^{\{J-j_0\}}(X_n)\, \varphi_{j_0 k}(x). \tag{11}$$

If the Haar approximation is applied also at the final step, the linear estimator at level j_0 is instead

$$\hat{f}(x) = \sum_{k \in \mathbb{Z}} \sum_{n=1}^{N} U_n \varphi_{j_0 k}^{\{J-j_0\}}(X_n)\, \varphi_{j_0 k}^{\{J-j_0\}}(x). \tag{12}$$

The proof is given in the appendix. As the next subsections show, most wavelet estimators have the form required by this theorem. It is also useful in practice, since the functions $\varphi^{\{i\}}$ and $\psi^{\{i\}}$ can be computed with the fast cascade algorithm. It is also the key to computing properties like the influence function in Section 4.

3.1. Density case

Theorem 1 trivially applies to the linear density estimator based on the box approximation (7) with $U_n = N^{-1}$. It follows that the difference between the

moment estimator (6) and the box approximation (7) is given by the difference between φ and $\varphi^{\{J-j_0\}}$. As $J-j_0$ tends to infinity, estimators given in (11) and (12) converge to the moment estimator given in (8). However, for the level computed in practice, this difference can be large. In Section 4, we see the impact of this difference through the influence function.

Note that the estimators given in (11) or (12) are still associated with approximation kernels, which can easily be deduced. The kernels are essentially approximation of the kernel associated to a wavelet basis given in Definition 1. However, the kernels from (11) and (12) vary with the level considered.

3.2. Regression case

The standard definition of the estimators does not satisfy the condition in (10). For instance, the estimator based on average has the form $U_n = 2^{-J}Y_n/\sum_m I(X_m \in B_{Jk})$. It depends on k, as it counts the observations in the interval B_{Jk}.

However since in (10) U_n is multiplied by 0 for all X_n not in B_{Jk}, instead of counting the observations in the box B_{Jk}, we can count the observations in the same box as X_n and remove the dependency on k. Thus we can write

$$\begin{aligned}\hat{\alpha}_{r,d0}[J,k] &= \sum_{n=1}^{N} 2^{-J}Y_n \frac{1}{\sum_{t\in\mathbb{Z}}\sum_m I(X_m \in B_{Jt})I(X_n \in B_{Jt})}\varphi_{Jk}^{\{0\}}(X_n)\\ &= \sum_{n=1}^{N} 2^{-J}Y_n \sum_{t\in\mathbb{Z}} \frac{I(X_n \in B_{Jt})}{\sum_m I(X_m \in B_{Jt})}\varphi_{Jk}^{\{0\}}(X_n), \qquad (13)\end{aligned}$$

which is the form required by (10). By Theorem 1, we replace $\varphi_{Jk}^{\{0\}}$ by $\varphi_{j_0k}^{\{J-j_0\}}$ to obtain the value of $\hat{\alpha}_{r,d0}[j_0,k]$. The functional form for the estimator at x becomes

$$\mathcal{R}_{d0}^{x}(P) = 2^{-J}\sum_{t\in\mathbb{Z}} p_t^{-1}\sum_{k\in\mathbb{Z}} \varphi_{j_0k}^{\{J-j_0\}}(x)\int f(v)\varphi_{j_0k}^{\{J-j_0\}}(v)I(v\in B_{Jt})\,dG(v), \quad (14)$$

where $p_t = \int I(v\in B_{Jt})\,dG(v)$, and G is the marginal distribution of X. For the estimator based on a local linear regression, the coefficient $\hat{\alpha}_{r,d1}[j_0,k]$ is

$$\hat{\alpha}_{r,d0}[j_0,k] + \sum_{n=1}^{N} 2^{-J}Y_n \sum_{t\in\mathbb{Z}} \frac{(a_t-C_t)(X_n-C_t)I(X_n\in B_{Jt})}{\sum_m (X_m-C_t)^2 I(X_m\in B_{Jt})}\varphi_{j_0k}^{\{J-j_0\}}(X_n), \quad (15)$$

where $a_t = (t+1/2)/2^J$ is the center of B_{Jt} and $C_t = \sum_m X_m I(X_m\in B_{Jt})/\sum_m I(X_m\in B_{Jt})$ is the center of gravity of B_{Jt} with respect to the empirical distribution of X. The functional form of the estimator follows easily.

This new representation adds two summation signs at the high resolution level. However, it allows a non-iterative and unified formulation for the estimator at any level, that is similar to the high resolution level. The form is very convenient both for computational and theoretical reasons, as will be exemplified in the next section.

4. Influence Functions

We discuss some properties of the regression and the density estimators defined in Section 2. We consider the sensitivity of the estimator to the design, to the choice of the origin of the wavelet basis, and to the method for the raw estimation of the coefficients. A fundamental tool for this purpose is the influence function.

Definition 2. Suppose $\mathcal{T}(F_N)$ is the estimator of $\mathcal{T}(F)$, where F_N and F are the empirical and the underlying distribution functions, respectively. Then, if it exists, the influence function of $\mathcal{T}$ at F at a point $\mathbf{z} = (z_1, \ldots, z_q)'$ is

$$\mathbf{IF}(\mathbf{z}; \mathcal{T}, F) = \lim_{\epsilon \searrow 0} \frac{\mathcal{T}(F_{\epsilon,\mathbf{z}}) - \mathcal{T}(F)}{\epsilon},$$

with $F_{\epsilon,\mathbf{z}} = (1-\epsilon)F + \epsilon\Delta_{\mathbf{z}}$, where $\Delta_{\mathbf{z}}$ is the point-mass distribution at $\mathbf{z}$.

The influence function is usually designed for a parametric model with a finite dimensional parameter, but interpret our case as having to compute one **IF** for each x where we estimate $f(x)$. It is used generally to assess the robustness properties of estimators, see Hampel, Ronchetti, Rousseeuw and Stahel (1986). Here we are rather interested in sensitivity in a broad sense. The general shape of the influence function reveals the behavior of the estimator. Two global properties of the **IF** give a first insight into its sensitivity. The first one is boundedness. Indeed, this ensures that a single observation cannot have immoderate consequences on the estimator. A second is a finite local-shift sensitivity, defined as $\sup_{\mathbf{z}\neq\mathbf{y}} |\mathbf{IF}(\mathbf{y}; \mathcal{T}, F) - \mathbf{IF}(\mathbf{z}; \mathcal{T}, F)|/\|\mathbf{y} - \mathbf{z}\|$. It detects the (standardised) maximal change in the estimator due to a wiggling of the sample. Different approximations for the wavelet estimator will be quite different as regards local-shift sensitivity.

In Section 4.1, we treat the case of density estimation, which displays phenomena also present in the regression case. The thresholded wavelet density estimator is treated in Section 4.2, and finally, the regression case is discussed in Section 4.3.

4.1. Approximation kernels

We give the influence function for any estimator associated with an approximation kernel.

Theorem 2. *Let $K_j(x,y)$ be an approximation kernel and $\mathcal{K}_j^x(F)$ its associated operator evaluated at a point x and distribution F. Then the influence function of $\mathcal{K}_j^x$ is* $\mathbf{IF}(z;\mathcal{K}_j^x,F) = K_j(x,z) - \mathcal{K}_j^x(F)$.

The proof is straightforward, since the estimator is linear with respect to the distribution F. Note that the first term does not depend on F and the second term does not depend on z. For convolution kernels, $K(x,z) = K(x-z)$, which depends only on the distance between the perturbation point z and the estimation point x. Up to a translation all influence functions are the same for different values of x, and the influence function is bounded. Most of the convolution kernels have a bounded local-shift sensitivity, the same for all values of x and independent of F. Figure 1(a) provides an example with a Gaussian kernel K. The distribution F is $\mathcal{N}(0.5,(0.15)^2)$ and the **IF** is evaluated for 8 estimation points x ranging from 0.375 to 0.5.

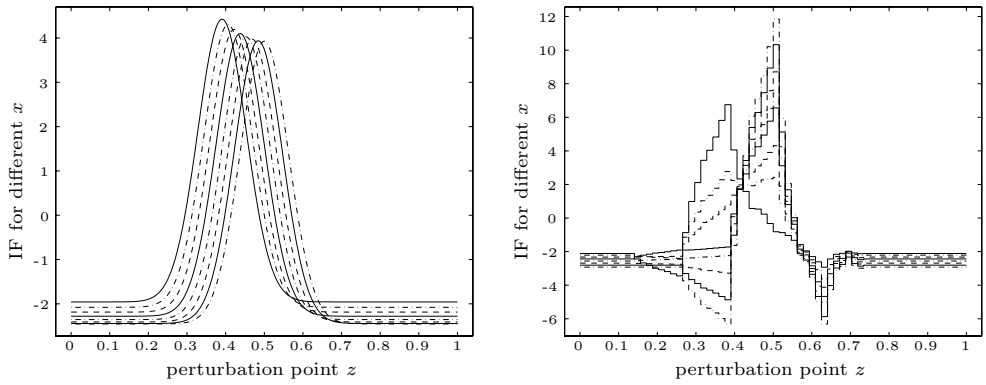

Figure 1. (a) Influence functions for the Gaussian convolution kernel for x values between 0.375 and 0.5 where $\hat{f}$ is evaluated. The distribution F is $\mathcal{N}(0.5,(0.15)^2)$. (b) Influence functions for the same settings as in (a), but for the Daub2 linear wavelet estimator with the box approximation.

For the linear wavelet estimator, the first term of its influence function depends also on the location of x and z with respect to the dyadic grid. Figure 2(a) shows the influence function for the same setting as Figure 1(a) but with the linear density estimator. Here the Daub2 basis has been used and the level j is 3. All the influence functions are bounded. However, influence functions may have high local-shift sensitivity. The differences between the influence functions show that, depending on the location of an observation with respect 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 a second or higher order convolution kernel, whereas others will

have an influence closer to the first order kernel. Note however that the global influence, if measured as $\int \mathbf{IF}(z; \mathcal{K}^x, F)\, dx$, is equal to zero for all approximation kernels.

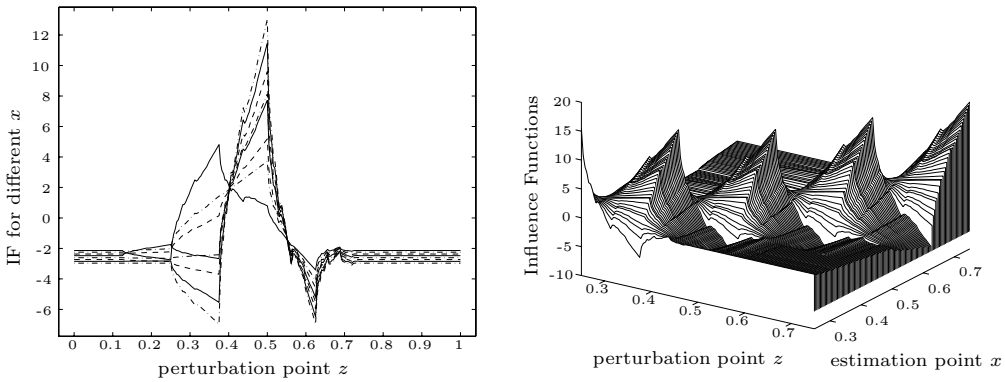

Figure 2. (a) Same as Figure 1(a), but for the Daub2 linear wavelet estimator. (b) Influence functions as a function of x and z for the same settings as in (a).

In Figure 2(b), influence functions are plotted as a function of two arguments z and x. Each individual curve corresponds to the **IF** evaluated at a given x. Although the influence functions are different from one another, the plot in Figure 2(b) shows that they depend continuously on x, provided that the scaling function and the density are continuous. Note that the spikes in Figure 2 are due to the Daub2 wavelet basis. For a smoother basis, the peaks will be replaced by smooth bumps.

It is worthwhile to look at the Haar basis as well. In this case, the influence function is $\mathbf{IF}(z; \mathcal{K}^x, F) = I(\lfloor x \rfloor \le z < \lfloor x \rfloor + 1) - (F(\lfloor x \rfloor + 1) - F(\lfloor x \rfloor))$, constant over z in intervals. 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 arbitrarily close to a point x has as little influence on the estimator as observations far away. Moreover, the **IF** of two neighboring points can be very different. Finally, the influence function is not continuously varying with x, since abrupt changes happen at integers. This is a well-known problem, since this estimator is actually equivalent to a histogram.

As the estimator in (6) is often replaced in practice by the box approximation in (7), it is instructive to check this case as well. As seen in Section 3.1, the estimator in (7) is associated with a kernel that is an approximation of the kernel of the wavelet basis. By Theorem 2, its influence function is an approximation of the discussed **IF**. As an example, in Figure 1(b), the **IF** of the box approximation is shown for the same setting as Figure 2(a). The approximate estimator inherits

both the shortcomings of the Haar basis and the shortcomings of the wavelet basis defined by φ. Indeed, the influence functions are piecewise constant, hence the local-shift sensitivity is infinite, and not continuous in x or in z. Moreover, the influence functions are constant approximations of the influence functions in Figure 2 and are quite different depending on the location of z with respect to the dyadic grid.

One may believe that the approximation of the coefficients by (7) acts as a stabilisation of the moment, since $\varphi_{Jk}(X_n)$ varies rapidly, inducing a significant variance. The above study shows that this is not the case, and that the estimator is in fact less stable: its **IF** is not continuous, indicating that jittering or rounding points can change the estimator in a significant way. It has been found in Renaud (1999) that the box approximation is much more sensitive to the choice of the origin of the wavelet basis than the moment estimator. The **IF** explains this phenomenon.

4.2. Thresholded density estimator

The functional form of the thresholded estimator at x is

$$\begin{aligned}\mathcal{T}_\lambda^x(F) = f_t(x) &= \sum_k \alpha[j_0,k]\ \varphi_{j_0k}(x) + \sum_{j=j_0}^{J-1}\sum_k \xi_\lambda(\beta[j,k])\ \psi_{jk}(x)\\ &= \mathcal{K}_{j_0}^x(F) + \sum_{j=j_0}^{J-1}\sum_k \xi_\lambda\left(\int \psi_{jk}(y)\ dF(y)\right)\ \psi_{jk}(x).\end{aligned}$$

For hard thresholding, it corresponds to the linear estimator plus the projection of the true density on a subspace adaptively selected for F.

Theorem 3. *For a given distribution F and a fixed value of the threshold λ, the influence function for the hard thresholded density estimator at a point x is given by*

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

For soft thresholding, $\psi_{jk}(z)I(|\beta[j,k]|>\lambda)$ is replaced by $(\psi_{jk}(z)-\lambda)I(\beta[j,k]>\lambda)+(\psi_{jk}(z)+\lambda)I(\beta[j,k]<-\lambda)$. The hard thresholding is required to satisfy the additional condition that no coefficient $\beta[j,k]$ is $+\lambda$ or $-\lambda$, since the discontinuity of this thresholding implies that the functional is not Gâteaux differentiable for those F.

The proof is provided in the appendix. For hard thresholding, if all the $\beta[j,k]$ are greater than λ in absolute value, the influence function would equal

the **IF** of the linear estimator at the higher resolution level J. The **IF** depends on coefficients $\beta[j,k]$ that are kept in the estimator. This dependence relies on F, not on z. The influence function is thus equal to the linear kernel $K_{j_0}(x,z)$ at the low resolution j_0, plus terms that are part of a kernel based on ψ, but that are selected depending on the distribution F. The influence function will be bounded, but if the underlying density is irregular, it may have sharp edges. Note that this is what we expect from thresholding. This can be illustrated in the following special case.

For the Haar basis, apart from $\mathcal{T}_\lambda^x$, the influence function is restricted to an interval of the form $B = [m/2^{j_0}; (m+1)/2^{j_0}]$ 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 is suitably chosen, the coefficients $\beta[j,k]$ will be smaller than the threshold and consequently the **IF** will be proportional to the indicator function of the interval B for any z in this interval. This shows that when the underlying function is very smooth, the thresholding procedure takes advantage of all the observations in the interval to estimate the average value of f on this interval with greater precision. On the other hand, if an abrupt change in the density happens somewhere in the interval, some $\beta[j,k]$ are large. As a result, the influence function becomes important only locally, in a neighboring region, and is smaller 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 $x \geq (m+1/2)/2^{j_0}$, $a, b \geq 0$. If $\lambda > |a-b|2^{-j_0}$, the only non-zero term in the sum of Theorem 3 is $\psi_{j_0 m}(x)\ \psi_{j_0 m}(z)$. As a result, if x and z are in the same subinterval, apart from $\mathcal{T}_\lambda^x$, the **IF** is equal to 2^{j_0+1}, and the **IF** is reduced to 0 when x and z are not in the same subinterval. Informally, the procedure detects two different regimes and therefore completely separates the groups.

In summary, from the influence function point of view, the convolution kernel is better than the linear wavelet estimator, the latter being more variable and lacking shift-invariance. However, the weakness of the latter is the necessary price to pay for adaptivity of the thresholded procedure. A way to circumvent the lack of shift-invariance and to lower the variance of the estimator, while preserving adaptivity, is shown in Renaud (1999).

For the computation of the influence function, the threshold has been supposed fixed. The influence function with a data-based threshold cannot be computed since the threshold cannot be represented in a functional form. One of the reasons is that λ has to go to 0 for increasing values of N. Nevertheless the influence function with a fixed threshold still makes sense as, for standard procedures, one observation can only have a bounded influence on the data-driven choice of the threshold.

4.3. Regression

The regression case is more complicated but often of more interest.

Theorem 4. *With the same notation as in Sections* 2.2 *and* 3.2, *let* $\mathcal{R}^x_{d0}(P)$ *and* $\mathcal{R}^x_{d1}(P)$ *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* $\mathbf{z} = (z_1, z_2)'$ *are given, respectively, by*

$$\mathbf{IF}(\mathbf{z}; \mathcal{R}^x_{d0}, P) = 2^{-J} \sum_{t\in\mathbb{Z}} p_t^{-2} I(z_1 \in B_{Jt}) \sum_{k\in\mathbb{Z}} \varphi_{j_0k}^{\{J-j_0\}}(x) \int \left(z_2 \varphi_{j_0k}^{\{J-j_0\}}(z_1) - f(v)\varphi_{j_0k}^{\{J-j_0\}}(v) \right) I(v \in B_{Jt})\, dG(v), \quad (16)$$

$$\begin{aligned}
&\mathbf{IF}(\mathbf{z}; \mathcal{R}^x_{d1}, P) \\
&= \mathbf{IF}(\mathbf{z}; \mathcal{R}^x_{d0}, P) + 2^{-J} \sum_{t\in\mathbb{Z}} q_t^{-1}(z_1 - c_t) I(z_1 \in B_{Jt}) \sum_{k\in\mathbb{Z}} \varphi_{j_0k}^{\{J-j_0\}}(x) \\
&\quad \Big(z_2(a_t - c_t)\varphi_{j_0k}^{\{J-j_0\}}(z_1) - p_t^{-1} \int f(v)((v-c_t)+(a_t-c_t))\varphi_{j_0k}^{\{J-j_0\}}(v) I(v \in B_{Jt})\, dG(v) \\
&\quad - q_t^{-1}(z_1 - c_t)(a_t - c_t) \int f(v)(v - c_t)\varphi_{j_0k}^{\{J-j_0\}}(v) I(v \in B_{Jt})\, dG(v) \Big), \qquad (17)
\end{aligned}$$

where $c_t = p_t^{-1} \int v I(v \in B_{Jt})\, dG(v)$ *and* $q_t = \int (v - c_t)^2 I(v \in B_{Jt})\, dG(v)$.

The proofs of both results are not displayed here as they are long and technical, but they are similar to those of previous theorems. We 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 unbounded with increasing values of z_2. The thresholded versions also have unbounded **IF** with respect to z_2. This comes as no surprise, since the linear and thresholded wavelet estimators can be represented as the solution of a least-squares, respectively a penalized least-squares, problem. If the sample is at risk concerning outliers, one should avoid them. To circumvent this problem, Sardy (1998) proposes a robust thresholding. Kovac and Silverman (2000) propose an ad-hoc outlier detection method.

Note that for a given point of interest x, both **IF** are zero, except in an interval, if $\varphi_{j_0k}^{\{J-j_0\}}(x)$ is bounded. Thanks to the similarity with the **IF** of the density estimator, all remarks on density estimation apply to regression as well. In particular, the **IF** depends on the piecewise constant function $\varphi^{\{J-j_0\}}$. Figure 3 demonstrates the similarities. Thresholding also produces similar result on the **IF** for density and regression.

As an example, Figure 3(a) gives the influence function $\mathbf{IF}(\mathbf{z}; \mathcal{R}^x_{d0}, P)$ for box estimation and Figure 3(b) displays $\mathbf{IF}(\mathbf{z}; \mathcal{R}^x_{d1}, P)$ for the local linear regression approximation. Here the Daub2 wavelet basis is used, the density of X is triangular between 0 and 1 ($g(x) = 2x$), and the expectation of Y given $X = x$ is

$1 - x - I(x < 0.55)$. The first **IF** is piecewise constant. One could expect that the second estimator would correct this toward an **IF** that has smaller gaps. In fact, the contrary happens: the influence of an observation at a point z is more extreme than in the simple case. First, the maximal value of the **IF** is larger and second, some gaps are even more important than in the simple case. The general shape of the **IF** is surprising. The local linear regression lowers the stability of the estimator and increases its variability by letting some points inside an interval have much greater influence than others.

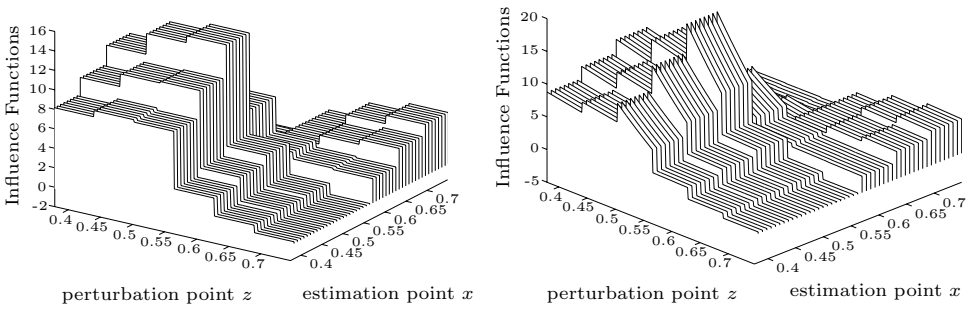

Figure 3. (a) Influence functions for the standard linear wavelet regression estimator. (b) Influence function for the linear wavelet regression estimator with a local linear approximaiton. Note how the gaps are sometimes accentuated.

To increase stability, instead of or in addition to taking a local linear fit, it seems more important to replace the box kernel of both (13) and (15) by a smoother kernel, which gives $\hat{\alpha}[J,k] = 2^{-J}\sum_n Y_n K(2^J X_n - (k + 1/2))/\sum_n K(2^J X_n - (k + 1/2))$, with the constraint that $\sum_k K(x - k)$ is constant for all x. For instance, the trapezoidal kernel is given by $K(x) = 1$ for $x \in [-0.5; 0.5]$, $x + 1.5$ for $x \in [-1.5; -0.5]$, $-x + 1.5$ for $x \in [0.5; 1.5]$ and zero otherwise. Figure 4 shows how the influence function becomes smooth when using this kernel for the same setting as Figure 3. Another possibility is to use the empirical moment $\hat{\alpha}[J,k] = 2^{-J/2}\sum_n Y_n \varphi_{Jk}(X_n)/\sum_n \varphi_{Jk}(X_n)$. In both cases, we can use the idea of Kovac and Silverman (2000) to assess the covariance structure of the coefficients at other levels, which leads to an improved thresholding policy.

We did not treat non-interval-based wavelet regression estimators for random design, as they do not fit in the present framework. They have interesting properties, as shown in Cai and Brown (1998) and Sardy, Percival, Bruce, Gao

and Stuetzle (1999). However, from their construction, they are much more sensitive to small changes in the design and their influence functions, if computable, would be very large.

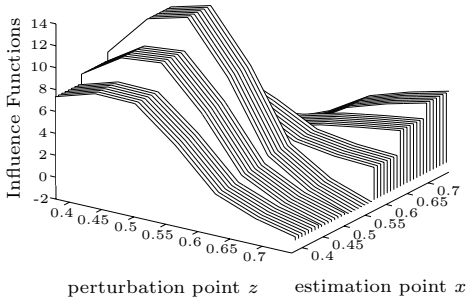

Figure 4. Influence functions for the linear wavelet regression estimator with a kernel-type of approximation with the trapezoidal kernel.

Acknowledgements

Most of the research was supported by the Swiss National Science Foundation. This research was partially carried out during the author's Ph.D. thesis at the Ecole Polytechnique Fédérale de Lausanne under the supervision of Prof. Morgenthaler. I am very thankful for his guidance. The work has been completed while the author was visiting Stanford University. Thanks are also due to the Editor, an associate editor and two referees whose comments improved the presentation of this work.

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 it is true for $i - 1$. We have

$$\begin{aligned}
\hat{\alpha}_r[J-i,k] &= \sum_{m\in\mathbb{Z}} h[m-2k]\ \hat{\alpha}_r[J-i+1,m] \\
&= \sum_{m\in\mathbb{Z}} h[m-2k]\ \sum_{n=1}^{N} U_n \varphi^{\{i-1\}}_{J-i+1\,m}(X_n) \\
&= \sum_{n=1}^{N} U_n 2^{(J-i)/2} \sum_{m\in\mathbb{Z}} h[m]\ \sqrt{2}\varphi^{\{i-1\}}\left(2(2^{J-i}X_n) - (m+2k)\right) \\
&= \sum_{n=1}^{N} U_n \varphi^{\{i\}}_{J-i\,k}(X_n),
\end{aligned}$$

which shows the result. The same applies for $\hat{\beta}_r[J-i,k]$ by replacing h by g. The proof of (12) involves an iterative use of (4) and (9). Since the $\hat{\beta}[j,k]$ have been set to 0, we have

$$\begin{aligned}\hat{f}(x) &= \sum_k \hat{\alpha}[J,k]\ \varphi_{Jk}^{\{0\}}(x) = \sum_k \sum_l h[k-2l]\ \hat{\alpha}[J-1,l]\ \varphi_{Jk}^{\{0\}}(x) \\ &= \sum_l \hat{\alpha}[J-1,l]\ \varphi_{J-1\,l}^{\{1\}}(x) = \cdots = \sum_k \hat{\alpha}[j_0,k]\ \varphi_{j_0 k}^{\{J-j_0\}}(x),\end{aligned}$$

which, given the first part of the theorem, shows the result. Equation (11) can be proved in the same way, using (2) instead of (9).

A.2. Proof of Theorem 3

We first note that $\beta_{\epsilon,z}[j,k] = \int \psi_{jk}(y)\ dF_{\epsilon,z}(y) = (1-\epsilon)\beta[j,k] + \epsilon\psi_{jk}(z)$, which is therefore continuous in ϵ. This implies that for each couple $\{j;k\}$ there exists an ϵ_{jk} such that, for any $0 \le \epsilon \le \epsilon_{jk}$, $\beta[j,k]$ and $\beta_{\epsilon,z}[j,k]$ are either both greater than λ or both smaller than $-\lambda$ or both between. As $f \in \mathsf{L}^2$, there are only finitely many $\beta[j,k]$ such that $|\beta[j,k]| > \lambda/2$. We can therefore find an ϵ^* for which the above properties hold for all $\{j;k\}$ and for every $0 \le \epsilon \le \epsilon^*$. For such an ϵ, hard thresholding gives

$$\begin{aligned}\xi_\lambda(\beta_{\epsilon,z}[j,k]) &= \beta_{\epsilon,z}[j,k]\ I(|\beta_{\epsilon,z}[j,k]| > \lambda) = \beta_{\epsilon,z}[j,k]\ I(|\beta[j,k]| > \lambda) \\ &= (1-\epsilon)\xi_\lambda(\beta[j,k]) + \epsilon\psi_{jk}(z)\ I(|\beta[j,k]| > \lambda).\end{aligned}$$

For the functional $\mathcal{T}_\lambda^x(F)$ and for an ϵ such that $0 \le \epsilon \le \epsilon^*$, the influence function numerator is

$$\begin{aligned}\mathrm{num}_\epsilon &= \mathcal{K}_{j_0}^x(F_{\epsilon,z}) + \sum_{j,\,k} \xi_\lambda(\beta_{\epsilon,z}[j,k])\ \psi_{jk}(x) - \mathcal{K}_{j_0}^x(F) - \sum_{j,\,k} \xi_\lambda(\beta[j,k])\ \psi_{jk}(x) \\ &= \epsilon\left\{K_{j_0}(x,z) - \mathcal{K}_{j_0}^x(F)\right\} + \epsilon \sum_{j,\,k}\left\{\psi_{jk}(z)\ I(|\beta[j,k]| > \lambda) - \xi_\lambda(\beta[j,k])\right\}\psi_{jk}(x) \\ &= \epsilon\left\{K_{j_0}(x,z) + \sum_{j,\,k} \psi_{jk}(x)\ \psi_{jk}(z) I(|\beta[j,k]| > \lambda) - \mathcal{T}_\lambda^x(F)\right\}.\end{aligned}$$

The result follows. Similar computations can be done for soft thresholding.

References

Cai, T. T. and Brown, L. D. (1998). Wavelet shrinkage for nonequispaced samples. *Ann. Statist.* **26**, 1783-1799.

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

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Kovac, A. and Silverman, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statist. Assoc.* **95**, 172-183.

Renaud, O. (1999). Shift-invariant wavelet density estimator with smaller variability. Submitted.

Sardy, S. (1998). Regularization techniques for linear regression with a large set of carries. Ph. D. thesis, University of Washington.

Sardy, S., Percival, D., Bruce, A., Gao, H.-Y. and Stuetzle, W. (1999). Wavelet de-noising for unequally spaced data. *Statist. Comput.* **9**, 65-75.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York.

Faculty of Psychology and Education, University of Geneva, 40, bd du Pont d'Arve, CH-1211 Geneva 4.

E-mail: olivier.renaud@pse.unige.ch

(Received October 2000; accepted June 2002)