A note on the adaptive estimation of a conditonal continuous-discrete multivariate density by wavelet methods

Christophe Chesneau, Hassan Doosti

A note on the adaptive estimation of a conditonal continuous-discrete multivariate density by wavelet methods

Christophe Chesneau[1], **Hassan Doosti**[2]

[1] *Laboratoire de Mathématiques Nicolas Oresme, Université de Caen BP 5186, F 14032 Caen Cedex, France. e-mail:* christophe.chesneau@unicaen.fr

[2] *Mashhad University of Medical Sciences, Mashhad, Iran. e-mail:* hassandoosti1353@yahoo.com

Abstract: In this note we investigate the estimation of a multivariate continuous-discrete conditional density. We develop an adaptive estimator based on wavelet methods. We prove its good theoretical performance by determining sharp rates of convergence under the $\mathbb{L}_p$ risk with $p \geq 1$ for a wide class of unknown conditional density. A simulation study illustrates the good practical performances of our estimator.

AMS 2000 subject classifications: 62G07, 62G20.
Keywords and phrases: Conditional density estimation, Density estimation, Wavelet methods.

1. Introduction

The estimation of conditional densities is an important statistical challenge with applications in many practical problems, especially those connected with forecasting (economics ...). There is a vast literature in this area. We refer to the papers of Li and Racine (2007), Akakpo and Lacour (2011), Chagny (2013) and the references therein. In this note we focus our attention on a specific problem: the estimation of a multivariate continuous-discrete conditional density. The considered model is described as follows. Let d, d_*, ν and n be positive integers and $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ be n *iid* random vectors defined on the probability space $(\Omega, \mathcal{A}, P)$. We suppose that $\mathbf{X}_1$ is continuous with support $[0,1]^d$ and $\mathbf{Y}_1$ is discrete with support $\{0, 1, \ldots, \nu\}^{d_*}$. Let f be the density of $(\mathbf{X}_1, \mathbf{Y}_1)$. We define the density function of $\mathbf{X}_1$ conditionally to the event $\{\mathbf{Y} = \mathbf{m}\}$ by

$$g(\mathbf{x}, \mathbf{m}) = f(\mathbf{x}|\ \mathbf{Y}_1 = \mathbf{m}) = \frac{f(\mathbf{x}, \mathbf{m})}{P(\mathbf{Y_1} = \mathbf{m})}, \tag{1.1}$$

$(\mathbf{x}, \mathbf{m}) \in [0,1]^d \times \{0, 1, \ldots, \nu\}^{d_*}$. We aim to estimate $g(\mathbf{x}, \mathbf{m})$ from $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$. The most common approach is based on the kernel methods developed by Li and Racine (2003). Applications and recent developments for these methods are described in details in Li and Racine (2007).

In this note we develop a new estimator $\widehat{g}(\mathbf{x},\mathbf{m})$ based on wavelet methods. It is now an established fact: in comparison to kernel methods, wavelet methods have the advantage to achieve a high degree of adaptivity for a large class of unknown functions, with possible complex discontinuities (jumps, spikes...). See, for instance, Antoniadis (1997), Härdle *et al.* (1998) and Vidakovic (1999). This fact motivates our interest to develop wavelet methods for the considered conditional density estimation problem. The main ingredients in the construction of $\widehat{g}(\mathbf{x},\mathbf{m})$ are: an estimation of $f(\mathbf{x},\mathbf{m})$ with a new wavelet estimator $\widehat{f}(\mathbf{x},\mathbf{m})$, an estimation of $P(\mathbf{Y_1}=\mathbf{m})$ by an empirical estimator and a global thresholding technique developed by Vasiliev (2014). In particular, the considered estimator $\widehat{f}(\mathbf{x},\mathbf{m})$ can be viewed as a multivariate (but "non smooth") version of the one introduced in the univariate case, i.e., $d=d_*=1$, in Chesneau *et al.* (2014). We prove that $\widehat{g}(\mathbf{x},\mathbf{m})$ is both adaptive and efficient; it don't dependent on the smoothness of $g(\mathbf{x},\mathbf{m})$ in its construction and, under mild assumptions on the smoothness of $g(\mathbf{x},\mathbf{m})$ (we assume that it belongs to a wide class of functions, the so-called Besov balls), it attains fast rates of convergence under the $\mathbb{L}_p$ risk (with $p\geq 1$). These theoretical guarantees are illustrated by a numerical study showing the good practical performances of our estimator.

The remainder of the note is set out as follows. Next, in Section 2, we briefly describe the considered multidimensional wavelet bases and Besov balls. Our wavelet estimator and some of its theoretical properties are presented in Section 3. A short numerical study can be found in Section 4. Finally, the proofs are postponed to Section 5.

2. Multidimensional wavelet bases and Besov balls

Let d be positive integers and $p\geq 1$. First of all, we define the $\mathbb{L}_p([0,1]^d)$ spaces as $\mathbb{L}_p([0,1]^d)=\left\{f:[0,1]^d\to\mathbb{R};\ \ \int_{[0,1]^d}|f(\mathbf{x})|^p d\mathbf{x}<\infty\right\}$.

In this study, we consider a wavelet bases on $[0,1]^d$ based on the scaling and wavelet functions ϕ and ψ respectively from Daubechies family (see Daubechies (1992)). For any $\mathbf{x}=(x_1,\ldots,x_d)\in[0,1]^d$, we set

$$\Phi(\mathbf{x})=\prod_{v=1}^{d}\phi(x_v),\ \text{and}\ \Psi_u(\mathbf{x})=\begin{cases}\psi(x_u)\displaystyle\prod_{\substack{v=1\\v\neq u}}^{d}\phi(x_v) & \text{for } u\in\{1,\ldots,d\},\\ \displaystyle\prod_{v\in A_u}\psi(x_v)\prod_{v\notin A_u}\phi(x_v) & \text{for } u\in\{d+1,\ldots,2^d-1\},\end{cases}$$

where $(A_u)_{u\in\{d+1,\ldots,2^d-1\}}$ forms the set of all non void subsets of $\{1,\ldots,d\}$ of cardinality greater or equal to 2.

For any integer j and any $\mathbf{k}=(k_1,\ldots,k_d)$, we consider

$$\Phi_{j,\mathbf{k}}(\mathbf{x})=2^{jd/2}\Phi(2^jx_1-k_1,\ldots,2^jx_d-k_d),$$
$$\Psi_{j,\mathbf{k},u}(\mathbf{x})=2^{jd/2}\Psi_u(2^jx_1-k_1,\ldots,2^jx_d-k_d),\ \text{for any } u\in\{1,\ldots,2^d-1\}.$$

Let $\mathbf{D}_j=\{0,\ldots,2^j-1\}^d$. Then, with an appropriate treatment at the boundaries, there exists an integer τ such that the collection

$\{\Phi_{\tau,\mathbf{k}}, \mathbf{k} \in \mathbf{D}_\tau;\ (\Psi_{j,\mathbf{k},u})_{u\in\{1,\ldots,2^d-1\}},\quad j \in \mathbb{N} - \{0,\ldots,\tau-1\},\ \mathbf{k} \in \mathbf{D}_j\}$ forms an orthonormal basis of $\mathbb{L}_2([0,1]^d)$. A function $f \in \mathbb{L}_2([0,1]^d)$ can be expanded into a wavelet series as

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbf{D}_\tau} c_{\tau,\mathbf{k}}\Phi_{\tau,\mathbf{k}}(\mathbf{x}) + \sum_{u=1}^{2^d-1}\sum_{j=\tau}^{\infty}\sum_{\mathbf{k}\in\mathbf{D}_j} d_{j,\mathbf{k},u}\Psi_{j,\mathbf{k},u}(\mathbf{x}), \qquad \mathbf{x} \in [0,1]^d, \quad (2.1)$$

where

$$c_{\tau,\mathbf{k}} = \int_{[0,1]^d} f(\mathbf{x})\Phi_{\tau,\mathbf{k}}(\mathbf{x})d\mathbf{x}, \qquad d_{j,\mathbf{k},u} = \int_{[0,1]^d} f(\mathbf{x})\Psi_{j,\mathbf{k},u}(\mathbf{x})d\mathbf{x}. \qquad (2.2)$$

All the details about these wavelet bases, including the expansion into wavelet series as described above, can be found in, e.g., Meyer (1992), Daubechies (1992), Cohen *et al.* (1993) and Mallat (2009).

Let $M > 0$, $s \in (0, N)$, $p \geq 1$ and $r \geq 1$. We say that a function $f \in \mathbb{L}_2([0,1]^d)$ belongs to the Besov balls $\mathbf{B}^s_{r,q}(M)$ if and only if there exists a constant $M^* > 0$ such that the associated wavelet coefficients (2.2) satisfy

$$\left(\sum_{\mathbf{k}\in\mathbf{D}_\tau} |c_{\tau,\mathbf{k}}|^r\right)^{1/r} + \left(\sum_{j=\tau}^{\infty}\left(2^{j(s+d(1/2-1/r))}\left(\sum_{u=1}^{2^d-1}\sum_{\mathbf{k}\in\mathbf{D}_j}|d_{j,\mathbf{k},u}|^r\right)^{1/r}\right)^q\right)^{1/q} \leq M^*$$

and with the usual modifications for $r = \infty$ or $q = \infty$.

These sets contain function classes of significant spatial inhomogeneity, including Sobolev balls, Hölder balls.... Details about Besov balls can be found in, e.g., Meyer (1992) and Härdle *et al.* (1998).

3. Conditional density estimation

We formulate the following assumptions.

(B1) There exists a known constant $C > 0$ such that

$$\sup_{\mathbf{x}\in[0,1]^d}\sup_{\mathbf{m}\in\{0,1,\ldots,\nu\}^{d_*}} f(\mathbf{x},\mathbf{m}) \leq C.$$

(B2) There exists a known constant $c \in (0,1)$ such that

$$c \leq \inf_{\mathbf{m}\in\{0,1,\ldots,\nu\}^{d_*}} P(\mathbf{Y_1} = \mathbf{m}).$$

We propose the following "ratio-thresholding estimator" $\widehat{g}(\mathbf{x},\mathbf{m})$ for $g(\mathbf{x},\mathbf{m})$:

$$\widehat{g}(\mathbf{x},\mathbf{m}) = \frac{\widehat{f}(\mathbf{x},\mathbf{m})}{\widehat{\rho}_{\mathbf{m}}}\mathbf{1}_{\{\widehat{\rho}_{\mathbf{m}}\geq c/2\}}, \qquad (3.1)$$

$(\mathbf{x},\mathbf{m}) \in [0,1]^d \times \{0,1,\ldots,\nu\}^{d_*}$, where $\mathbf{1}$ denotes the indicator function, c refers to the constant in **(B2)**, $\widehat{f}(\mathbf{x},\mathbf{m})$ is defined by

$$\begin{aligned}\widehat{f}(\mathbf{x},\mathbf{m}) &= \sum_{\mathbf{k}\in\mathbf{D}_\tau} \widehat{c}_{\tau,\mathbf{k}}(\mathbf{m})\Phi_{\tau,\mathbf{k}}(\mathbf{x}) \\ &+ \sum_{u=1}^{2^d-1}\sum_{j=\tau}^{j_1}\sum_{\mathbf{k}\in\mathbf{D}_j} \widehat{d}_{j,\mathbf{k},u}(\mathbf{m})\mathbf{1}_{\left\{|\widehat{d}_{j,\mathbf{k},u}(\mathbf{m})|\geq\kappa\sqrt{\frac{\ln(n)}{n}}\right\}}\Psi_{j,\mathbf{k},u}(\mathbf{x}),\end{aligned} \tag{3.2}$$

where

$$\widehat{c}_{\tau,\mathbf{k}}(\mathbf{m}) = \frac{1}{n}\sum_{i=1}^{n}\Phi_{\tau,\mathbf{k}}(\mathbf{X}_i)\mathbf{1}_{\{\mathbf{Y}_i=\mathbf{m}\}}, \tag{3.3}$$

$$\widehat{d}_{j,\mathbf{k},u}(\mathbf{m}) = \frac{1}{n}\sum_{i=1}^{n}\Psi_{j,\mathbf{k},u}(\mathbf{X}_i)\mathbf{1}_{\{\mathbf{Y}_i=\mathbf{m}\}}, \tag{3.4}$$

κ is a large enough constant and j_1 is an integer such that $n/\ln(n) \leq 2^{j_1 d} \leq 2n/\ln(n)$, and $\widehat{\rho}_{\mathbf{m}}$ is defined by

$$\widehat{\rho}_{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{\mathbf{Y}_i=\mathbf{m}\}}.$$

The estimator (3.2) uses a hard thresholding technique of the wavelet coefficients estimators 3.4. Such a selection rule is at the heart of the adaptive nature of wavelet methods which have the ability to capture the most important wavelet coefficients of a function, i.e., those with the high magnitudes. We refer to Antoniadis (1997), Härdle *et al.* (1998) and Vidakovic (1999) for further details. The definition of the threshold, i.e. $\lambda_n = \kappa\sqrt{\ln(n)/n}$, corresponds to the universal one proposed by Donoho and Johnstone (1994) and Donoho *et al.* (1996). It is based on technical considerations ensuring good convergence properties of the hard thresholding wavelet estimator (see also Theorem 5.1 in Appendix).

Note that (3.2) can be viewed as a non smooth multivariate version of the estimator proposed by Chesneau *et al.* (2014). The main advantage of this estimator is to be more easy to implement for a practical point of view (see Section 4 below for a numerical comparison in the univariate case). Concerning $\widehat{\rho}_n$, let us mention that it is a natural unbiased estimator for $P(\mathbf{Y}_1 = \mathbf{m})$ with nice convergence properties. They will be used in the proof of our main result.

The global construction of (3.1) follows the idea proposed by Vasiliev (2014) for other statistical contexts. Note that a control on the lower bound of $\widehat{\rho}_{\mathbf{m}}$ is necessary; it must be large enough to ensure good statistical properties for (3.1).

The following result investigates the rates of convergence attained by (3.1) under the $\mathbb{L}_p$ risk with $p \geq 1$.

Theorem 3.1. *Let* $\mathbf{m} \in \{0, 1, \ldots, \nu\}^{d_*}$, $p \geq 1$, $g(\mathbf{x}, \mathbf{m})$ *be* (1.1) *and* $\widehat{g}(\mathbf{x}, \mathbf{m})$ *be defined by* (3.1) *with a large enough* κ *(the exact condition is described in* (5.3)*). Suppose that* **(B1)** *and* **(B2)** *hold and* $f(\mathbf{x}, \mathbf{m}) \in \mathbf{B}^s_{r,q}(M)$ *with* $M > 0$, $s > d/r$, $r \geq 1$ *and* $q \geq 1$. *Then there exists a constant* $C > 0$ *such that, for* n *large enough,*

$$E\left(\int_{[0,1]^d} |\widehat{g}(\mathbf{x}, \mathbf{m}) - g(\mathbf{x}, \mathbf{m})|^p d\mathbf{x}\right) \leq C\Theta_n,$$

where

$$\Theta_n = \begin{cases} \left(\dfrac{\ln(n)}{n}\right)^{\frac{sp}{2s+d}}, & \text{for} \quad 2rs > d(p-r), \\ \left(\dfrac{\ln(n)}{n}\right)^{\frac{(s-d(1/r-1/p))p}{2s-2d/r+d}}, & \text{for} \quad 2rs < d(p-r), \\ \left(\dfrac{\ln(n)}{n}\right)^{\frac{(s-d(1/r-1/p))p}{2s-2d/r+d}} (\ln(n))^{\left(p-\frac{2r}{q}\right)_+}, & \text{for} \quad 2rs = d(p-r). \end{cases}$$

The proof of Theorem 3.1 is based on several technical inequalities and the application of a general result derived from (Kerkyacharian and Picard, 2000, Theorem 5.1) and (Delyon and Juditsky, 1996, Theorem 1) (see Theorem 5.1 in Appendix).

Theorem 3.1 provides theoretical guaranties on the convergence of (3.1) under mild assumptions on the smoothness of $f(\mathbf{x}, \mathbf{m})$, and a fortiori $g(\mathbf{x}, \mathbf{m})$, under the $\mathbb{L}_p$ risk. The obtained rates of convergence are sharp. However, since the lower minimax bounds are not established in our setting, we do not claim that they are the optimal ones in the minimax sense. An important benchmark is that they correspond to the optimal ones in the minimax sense for the standard multivariate density estimation problem, corresponding to $d_* = 1$ and $\mathbf{Y}_1$ is constant almost surely, up to a logarithmic term (see Donoho *et al.* (1996)).

4. A short numerical study

In this section we investigate some practical aspects of our wavelet methods. For ths sake of simplicity, we focus our attention on the univariate case, i.e., $d = d_* = 1$ (so $\mathbf{x} = x$, $\mathbf{m} = m$, $\mathbf{Y}_1 = Y \ldots$). The codes are written in Matlab and are adopted from Ramirez and Vidakovic (2010). First we compare the performance of new estimators of density functions $f(x, m)$ with those proposed in our former publication, Chesneau *et al.* (2014) in two styles, accuracy and speed of computation. In order to illustrate the rate of decrease of errors, as Chesneau *et al.* (2014), we employ the indicator defined by

$$L_2Norm = \frac{1}{100nN} \sum_{i=1}^{N} \sum_{j=1}^{n} \left(\widehat{f}_i\left(\frac{j}{n}\right) - f\left(\frac{j}{n}\right)\right)^2,$$

where n and N are sample size and the number of replications respectively, f represents the true density and $\widehat{f}$ an estimator. We consider three estimators based on our statistical methodology: the linear wavelet estimator, i.e.,

$$\widehat{f}_L(x,m) = \sum_{k=0}^{2^{j_0}-1} \widehat{c}_{j_0,k}(m)\phi_{j_0,k}(x), \tag{4.1}$$

$x \in [0,1]$, the hard thresholding wavelet estimator defined by (3.2) and the smooth version of the linear wavelet estimator after local linear regression (see, e.g., Fan (1993)). The practical construction of this smooth version of linear wavelet estimators was proposed by Ramirez and Vidakovic (2010). Several studies confirm that this version of estimators have nice performance in different fields (see, for instance, Abbaszadeh*et al.* (2013) and Chesneau *et al.* (2016)). We adopt similar set up from Chesneau *et al.* (2014) for our example, i.e., we use Daubechies's compactly supported "Daubechies 3" and we take $j_0 = 6$. Also, we generate different sample sizes $n = 20, 50, 100, 200, 500$ and 1000 data points $X_1, \ldots, X_n$, from Beta(2,3) distribution. The discrete random sample is generated from Binomial(1, x_i); the bivariate density function is

$$f(x,m) = 12x^{1+m}(1-x)^{3-m},$$

$(x,m) \in [0,1] \times \{0,1\}$. Table 1 gives the value of L_2Norm computed from 100 simulations for different sample size. This table should be compare with Table 1 in page 70 Chesneau *et al.* (2014). As we see, similar results could be obtained; L_2Norm decreases while the sample size's increasing. The performance of the smooth version of linear wavelet estimator is the best. As we see there is no significant difference between the new version of estimators with former versions in Chesneau *et al.* (2014).

On the other hands, Table 2 depicts the speed of computation for two groups of estimators in seconds. The codes are run with an ordinary laptop with 4.3 RAM. As we see the speed of new version of estimators is much less than the formers. For example when the sample size is 1000, the speed of computation is about 200 times less than the former version of wavelets estimators of densities. This differences will be much bigger when the sample size increases.
In the second part of this section we show the performance of proposed estimators of conditional density functions. Note that the conditional density function in above examples satisfies

$$g(x,m) = f(x|Y=m) = \begin{cases} 20x(1-x)^3, & \text{for} \quad m=0, \\ 30x^2(1-x)^2, & \text{for} \quad m=1, \end{cases}$$

$x \in [0,1]$. Figures 1, 2 depict the $g(x,0)$ and $g(x,1)$, respectively. In each cases the true conditional density function is shown in black line, the linear wavelet estimator is blue (dashed curve), the hard thresholding wavelet estimator is red (dotted curve) and the smooth version of linear one is green.

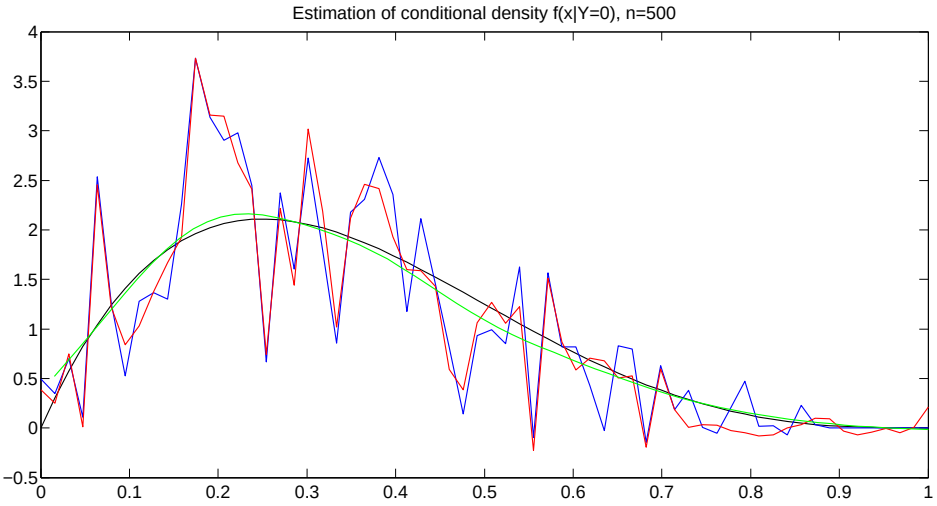

FIG 1. *The true conditional density function $g(x,0)$ is shown in black line, the wavelet linear estimator is blue, the wavelet hard thresholding estimator is red and its smooth version is green with $n = 500$.*

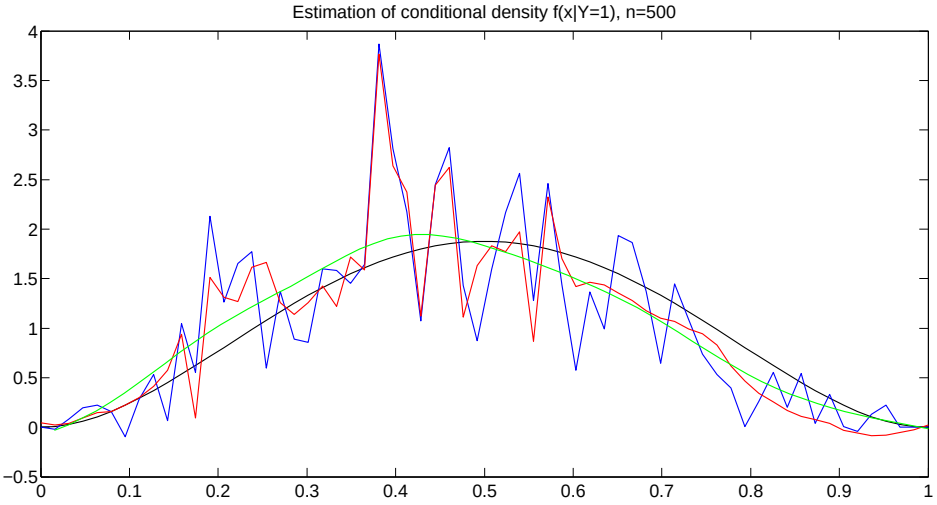

FIG 2. *The true conditional density function $g(x,1)$ is shown in black line, the wavelet linear estimator is blue, the wavelet hard thresholding estimator is red and its smooth version is green with $n = 500$.*

TABLE 1
Computed values for L_2Norm for various sample sizes.

Estimator		Sample Size					
	Density	n=20	n=50	n=100	n=200	n=500	n=1000
Linear	f(x,0)	0.0188	0.0080	0.0038	0.0019	0.00080	0.00041
Linear	f(x,1)	0.0132	0.0050	0.0027	0.0013	0.00051	0.00026
Hard Thresh.	f(x,0)	0.0178	0.0073	0.0033	0.0016	0.0006	0.00034
Hard Thresh.	f(x,1)	0.0123	0.0044	0.0021	0.00096	0.00036	0.00017
Smooth	f(x,0)	0.0011	0.00047	0.00026	0.00017	.00012	0.00009
Smooth	f(x,1)	0.00063	0.00028	0.00014	0.00007	0.00004	0.00002

TABLE 2
Elapsed time (Seconds)

Estimator	Sample Size					
	n=20	n=50	n=100	n=200	n=500	n=1000
Chesneau *et al.* (2014)	64.3205	157.546	176.094	369.802	771.294	2277.76
New estimators	0.641954	0.89170	1.481340	2.62647	5.78828	11.267064

All the figures illustrate the good performances of our proposed linear and nonlinear estimators of conditional density functions. It should be remind that the hard thresholding one has no tuning parameter, it is entirely adaptive. The smooth version of our wavelet linear estimator has the best performance. Furthermore, Table 3 represents the impact of sample size on performance of our estimators. This table also compares the performance of three estimators. The number of replication is 500. As the sample size increases the value of indicator decrease and the performance of smooth version of linear wavelet estimators are the best.

TABLE 3
Computed values for $100L_2Norm$ *for various sample sizes.*

Estimator		Sample Size					
	Conditional density	n=20	n=50	n=100	n=200	n=500	n=1000
Linear	$g(x,0)$	5.3407	1.1357	1.0883	0.5421	0.2150	0.1131
Linear	$g(x,1)$	8.8589	3.2498	1.6118	0.7908	0.3255	0.1614
Hard Thresh.	$g(x,0)$	5.0716	1.9346	0.9430	0.4576	0.1745	0.0900
Hard Thresh.	$g(x,1)$	8.1867	2.7903	1.2612	0.5851	0.2287	0.1064
Smooth	$g(x,0)$	0.2303	0.1064	0.0682	0.0433	0.0286	0.0261
Smooth	$g(x,1)$	0.3351	0.1297	0.0684	0.0377	0.0196	0.0145

5. Proof of Theorem 3.1

In what follows, C denotes any constant that does not depend on j, $\mathbf{k}$ and n. Its value may change from one term to another. For the sake of simplicity, we set $\rho_{\mathbf{m}} = P(\mathbf{Y}_1 = \mathbf{m})$. Observe that

$$\begin{aligned}\widehat{g}(\mathbf{x},\mathbf{m}) - g(\mathbf{x},\mathbf{m}) &= \frac{\widehat{f}(\mathbf{x},\mathbf{m})}{\widehat{\rho}_{\mathbf{m}}}\mathbf{1}_{\{\widehat{\rho}_{\mathbf{m}}\geq c/2\}} - \frac{f(\mathbf{x},\mathbf{m})}{\rho_{\mathbf{m}}} \\ &= \frac{1}{\widehat{\rho}_{\mathbf{m}}\rho_{\mathbf{m}}}\Big(\rho_{\mathbf{m}}(\widehat{f}(\mathbf{x},\mathbf{m}) - f(\mathbf{x},\mathbf{m})) + f(\mathbf{x},\mathbf{m})(\rho_{\mathbf{m}} - \widehat{\rho}_{\mathbf{m}})\Big)\mathbf{1}_{\{\widehat{\rho}_{\mathbf{m}}\geq c/2\}} \\ &- \frac{f(\mathbf{x},\mathbf{m})}{\rho_{\mathbf{m}}}\mathbf{1}_{\{\widehat{\rho}_{\mathbf{m}}< c/2\}}.\end{aligned}$$

Owing to **(B2)**, we have $\{\widehat{\rho}_{\mathbf{m}} < c/2\} \subseteq \{|\widehat{\rho}_{\mathbf{m}} - \rho_{\mathbf{m}}| > c/2\}$ implying $\mathbf{1}_{\{\widehat{\rho}_{\mathbf{m}}<c/2\}} \leq (2/c)|\widehat{\rho}_{\mathbf{m}} - \rho_{\mathbf{m}}|$ and $(1/(\widehat{\rho}_{\mathbf{m}}\rho_{\mathbf{m}}))\mathbf{1}_{\{\widehat{\rho}_{\mathbf{m}}\geq c/2\}} \leq 2/c^2$. Moreover, note that $\rho_{\mathbf{m}} \leq 1$ and, thanks to **(B1)**, $f(\mathbf{x},\mathbf{m}) \leq C$. It follows from the triangular inequality and the above inequalities that

$$|\widehat{g}(\mathbf{x},\mathbf{m}) - g(\mathbf{x},\mathbf{m})| \leq C(|\widehat{f}(\mathbf{x},\mathbf{m}) - f(\mathbf{x},\mathbf{m})| + |\widehat{\rho}_{\mathbf{m}} - \rho_{\mathbf{m}}|).$$

By the inequality: $|x+y|^p \leq 2^{p-1}(|x|^p + |y|^p)$, $(x,y) \in \mathbb{R}^2$, we obtain

$$E\left(\int_{[0,1]^d} |\widehat{g}(\mathbf{x},\mathbf{m}) - g(\mathbf{x},\mathbf{m})|^p d\mathbf{x}\right) \leq C(S+T), \tag{5.1}$$

where

$$S = E\left(\int_{[0,1]^d} |\widehat{f}(\mathbf{x},\mathbf{m}) - f(\mathbf{x},\mathbf{m})|^p d\mathbf{x}\right), \qquad T = E\left(|\widehat{\rho}_{\mathbf{m}} - \rho_{\mathbf{m}}|^p\right).$$

Let us now bound S and T in tun.

Upper bound for S. We investigate an upper bound for S by using Theorem 5.1 in the Appendix. First of all, thanks to **(B1)** implying $f(\mathbf{x},\mathbf{m}) \in \mathbb{L}_2([0,1]^d)$, let us expand the density $f(\mathbf{x},\mathbf{m})$ on the considered wavelet basis :

$$f(\mathbf{x},\mathbf{m}) = \sum_{\mathbf{k}\in\mathbf{D}_\tau} c_{\tau,\mathbf{k}}(\mathbf{m})\Phi_{\tau,\mathbf{k}}(\mathbf{x}) + \sum_{u=1}^{2^d-1}\sum_{j=\tau}^{j_1}\sum_{\mathbf{k}\in\mathbf{D}_j} d_{j,\mathbf{k},u}(\mathbf{m})\Psi_{j,\mathbf{k},u}(\mathbf{x}),$$

where $c_{\tau,\mathbf{k}}(\mathbf{m}) = \int_{[0,1]^d} f(\mathbf{x},\mathbf{m})\Phi_{\tau,\mathbf{k}}(\mathbf{x})d\mathbf{x}$ and $d_{j,\mathbf{k},u}(\mathbf{m}) = \int_{[0,1]^d} f(\mathbf{x},\mathbf{m})\Psi_{j,\mathbf{k},u}(\mathbf{x})d\mathbf{x}$. Let us now prove that the wavelet coefficients estimators $\widehat{c}_{j,\mathbf{k}}(\mathbf{m})$ and $\widehat{d}_{j,\mathbf{k},u}(\mathbf{m})$ satisfy Assumptions **(C1)** and **(C2)** of Theorem 5.1.

First of all, observe that $\widehat{c}_{\tau,\mathbf{k}}(\mathbf{m})$ and $\widehat{d}_{j,\mathbf{k},u}(\mathbf{m})$ are unbiased estimators for

$c_{\tau,\mathbf{k}}(\mathbf{m})$ and $d_{j,\mathbf{k},u}(\mathbf{m})$ respectively:

$$\begin{aligned} E(\widehat{d}_{j,\mathbf{k},u}(\mathbf{m})) &= E\left(\Psi_{j,\mathbf{k},u}(\mathbf{X}_1)\mathbf{1}_{\{\mathbf{Y}_1=\mathbf{m}\}}\right) \\ &= \sum_{\mathbf{v}\in\{0,1,\ldots,q\}^{d_*}} \int_{[0,1]^d} \Psi_{j,\mathbf{k},u}(\mathbf{x})\mathbf{1}_{\{\mathbf{v}=\mathbf{m}\}} f(\mathbf{x},\mathbf{v})d\mathbf{x} \\ &= \int_{[0,1]^d} \Psi_{j,\mathbf{k},u}(\mathbf{x}) f(\mathbf{x},\mathbf{m})d\mathbf{x} = d_{j,\mathbf{k},u}(\mathbf{m}). \end{aligned}$$

We prove similarly that $E(\widehat{c}_{\tau,\mathbf{k}}(\mathbf{m})) = c_{\tau,\mathbf{k}}(\mathbf{m})$

Investigation of **(C1)**. Let us focus on the second inequality in **(C1)**; the first one can be prove with similar arguments. For any $i \in \{1,\ldots,n\}$, set $V_i = \Psi_{j,\mathbf{k},u}(\mathbf{X}_i)\mathbf{1}_{\{\mathbf{Y}_i=\mathbf{m}\}} - d_{j,\mathbf{k},u}(\mathbf{m})$. Then $V_1,\ldots,V_n$ be n zero mean *iid* random variables with, by **(B1)** and $2^{dj} \leq 2n$,

$$\begin{aligned} E(|V_1|^{2p}) &\leq CE(|\Psi_{j,\mathbf{k},u}(\mathbf{X}_1)\mathbf{1}_{\{\mathbf{Y}_1=\mathbf{m}\}}|^{2p}) \\ &= C\sum_{\mathbf{v}\in\{0,1,\ldots,q\}^{d_*}} \int_{[0,1]^d} |\Psi_{j,\mathbf{k},u}(\mathbf{x})\mathbf{1}_{\{\mathbf{v}=\mathbf{m}\}}|^{2p} f(\mathbf{x},\mathbf{v})d\mathbf{x} \\ &= C\int_{[0,1]^d} |\Psi_{j,\mathbf{k},u}(\mathbf{x})|^{2p} f(\mathbf{x},\mathbf{m})d\mathbf{x} \leq C\int_{[0,1]^d} |\Psi_{j,\mathbf{k},u}(\mathbf{x})|^{2p} d\mathbf{x} \\ &\leq C2^{jd(p-1)} \int_{[0,1]^d} |\Psi_{j,\mathbf{k},u}(\mathbf{x})|^2 d\mathbf{x} \leq Cn^{p-1}. \end{aligned} \tag{5.2}$$

It follows from the Rosenthal inequality (see Appendix) that

$$\begin{aligned} E\left(|\widehat{d}_{j,\mathbf{k},u} - d_{j,\mathbf{k},u}|^{2p}\right) &= E\left(\left|\frac{1}{n}\sum_{i=1}^n V_i\right|^{2p}\right) \\ &= \frac{1}{n^{2p}} E\left(\left|\sum_{i=1}^n V_i\right|^{2p}\right) \leq C\frac{1}{n^{2p}} \max\left(nE(|V_1|^{2p}), n^p\left(E(V_1^2)\right)^p\right) \\ &\leq C\frac{1}{n^{2p}} \times n^p \leq C\frac{1}{n^p} \leq C\left(\frac{\ln(n)}{n}\right)^p. \end{aligned}$$

Investigation of **(C2)**. With the same random variables $V_1,\ldots,V_n$ defined as above, using $2^{jd} \leq 2n/\ln(n)$, note that $|V_1| \leq 2\sup_{\mathbf{x}\in[0,1]^d}|\Psi_{j,\mathbf{k},u}(\mathbf{x})| \leq C2^{jd/2} \leq C\sqrt{n/\ln(n)}$. It follows from the Bernstein inequality (see Appendix) with $\upsilon = (\kappa/2)\sqrt{n\ln(n)}$, $M = C\sqrt{n/\ln(n)}$ and, by (5.2) with $p=1$, $E(V_1^2) \leq$

C, that

$$\begin{aligned} P\left(|\widehat{d}_{j,\mathbf{k},u}-d_{j,\mathbf{k},u}|\geq\frac{\kappa}{2}\sqrt{\frac{\ln(n)}{n}}\right) &= P\left(\frac{1}{n}\left|\sum_{i=1}^{n}V_i\right|\geq\frac{\kappa}{2}\sqrt{\frac{\ln(n)}{n}}\right) \\ &= P\left(\left|\sum_{i=1}^{n}V_i\right|\geq\upsilon\right)\leq 2\exp\left(-\frac{\upsilon^2}{2\left(nE(V_1^2)+\upsilon M/3\right)}\right) \\ &\leq 2\exp\left(-\frac{((\kappa/2)\sqrt{n\ln(n)})^2}{2C\left(n+(\kappa/2)\sqrt{n\ln(n)}\sqrt{n/\ln(n)}/3\right)}\right)\leq 2n^{-\theta(\kappa)}, \end{aligned}$$

where $\theta(\kappa)=\kappa^2/\left(8C\left(1+\kappa/6\right)\right)$. Taking κ such that $\theta(\kappa)=p$, we obtain

$$P\left(|\widehat{d}_{j,\mathbf{k},u}-d_{j,\mathbf{k},u}|\geq\frac{\kappa}{2}\sqrt{\frac{\ln(n)}{n}}\right)\leq Cn^{-p}\leq C\left(\frac{\ln(n)}{n}\right)^p. \tag{5.3}$$

It follows from Theorem 5.1 that

$$S=E\left(\int_{[0,1]^d}|\widehat{f}(\mathbf{x},\mathbf{m})-f(\mathbf{x},\mathbf{m})|^p d\mathbf{x}\right)\leq C\Theta_n. \tag{5.4}$$

Upper bound for T. For any $i\in\{1,\ldots,n\}$, set $V_i=\mathbf{1}_{\{\mathbf{Y}_i=\mathbf{m}\}}-\rho_{\mathbf{m}}$. Then $V_1,\ldots,V_n$ be n zero mean *iid* random variables with $|V_1|\leq 2$. It follows from the Rosenthal inequality (see Appendix) that

$$\begin{aligned} T=E\left(|\widehat{\rho}_{\mathbf{m}}-\rho_{\mathbf{m}}|^p\right) &= E\left(\left|\frac{1}{n}\sum_{i=1}^{n}V_i\right|^p\right) \\ &= \frac{1}{n^p}E\left(\left|\sum_{i=1}^{n}V_i\right|^p\right)\leq C\frac{1}{n^p}\max\left(nE(|V_1|^p),n^{p/2}\left(E(V_1^2)\right)^{p/2}\right) \\ &\leq C\frac{1}{n^p}\times n^{p/2}\leq C\frac{1}{n^{p/2}}. \end{aligned} \tag{5.5}$$

Combining (5.1), (5.4) and (5.5), we obtain

$$E\left(\int_{[0,1]^d}|\widehat{g}(\mathbf{x},\mathbf{m})-g(\mathbf{x},\mathbf{m})|^p d\mathbf{x}\right)\leq C(S+T)\leq\max(\Theta_n,n^{-p/2})\leq C\Theta_n.$$

This complete the proof of Theorem 3.1.

□

Appendix

Here we state the two results that have been used for proving our theorem.

Lemma 5.1 (Rosenthal (1970))**.** *Let n be a positive integer, $p \geq 2$ and $V_1, \ldots, V_n$ be n zero mean iid random variables such that $E(|V_1|^p) < \infty$. Then there exists a constant $C > 0$ such that*

$$E\left(\left|\sum_{i=1}^{n} V_i\right|^p\right) \leq C \max\left(nE(|V_1|^p), n^{p/2}\left(E(V_1^2)\right)^{p/2}\right).$$

Lemma 5.2 (Petrov (1995))**.** *Let n be a positive integer and $V_1, \ldots, V_n$ be n iid zero mean independent random variables such that there exists a constant $M > 0$ satisfying $|V_1| \leq M$. Then, for any $\upsilon > 0$,*

$$P\left(\left|\sum_{i=1}^{n} V_i\right| \geq \upsilon\right) \leq 2\exp\left(-\frac{\upsilon^2}{2\left(nE(V_1^2) + \upsilon M/3\right)}\right).$$

Theorem 5.1. *We consider a general statistical nonparametric framework. Let $p \geq 1$ and $f(.) \in \mathbb{L}_{\max(p,2)}([0,1]^d)$ be an unknown function to be estimated from n observations and (2.1) its wavelet decomposition. Let $\widehat{c}_{j,\mathbf{k}}$ and $\widehat{d}_{j,\mathbf{k}}$ be estimators of $c_{j,\mathbf{k}}$ and $d_{j,\mathbf{k}}$ respectively such that there exist two constants $C > 0$ and $\kappa > 0$ satisfying Assumptions* **(C1)** *and* **(C2)** *below.*

(C1) *For any $k \in \mathbf{D}_\tau$,*

$$E\left(|\widehat{c}_{\tau,\mathbf{k}} - c_{\tau,\mathbf{k}}|^{2p}\right) \leq C\left(\frac{\ln(n)}{n}\right)^p$$

and for any $j \geq \tau$ such that $2^{jd} \leq n$, $u \in \{1, \ldots, 2^d - 1\}$ and $k \in \mathbf{D}_j$,

$$E\left(|\widehat{d}_{j,\mathbf{k},u} - d_{j,\mathbf{k},u}|^{2p}\right) \leq C\left(\frac{\ln(n)}{n}\right)^p.$$

(C2) *For any $j \geq \tau$ such that $2^{jd} \leq n/\ln(n)$, $u \in \{1, \ldots, 2^d - 1\}$ and $k \in \mathbf{D}_j$,*

$$P\left(|\widehat{d}_{j,\mathbf{k},u} - d_{j,\mathbf{k},u}| \geq \frac{\kappa}{2}\sqrt{\frac{\ln(n)}{n}}\right) \leq C\left(\frac{\ln(n)}{n}\right)^p.$$

Let us define the estimator $\widehat{f}$ by

$$\widehat{f}(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbf{D}_\tau} \widehat{c}_{\tau,\mathbf{k}}\Phi_{\tau,\mathbf{k}}(\mathbf{x}) + \sum_{u=1}^{2^d-1}\sum_{j=\tau}^{\infty}\sum_{\mathbf{k}\in\mathbf{D}_j} \widehat{d}_{j,\mathbf{k},u}\mathbf{1}_{\left\{|\widehat{d}_{j,\mathbf{k},u}| \geq \kappa\sqrt{\frac{\ln(n)}{n}}\right\}}\Psi_{j,\mathbf{k},u}(\mathbf{x}),$$

$\mathbf{x} \in [0,1]^d$, *where j_1 is the integer satisfying $n/\ln(n) < 2^{j_1 d} \leq 2n/\ln(n)$.*

Suppose that $f \in \mathbf{B}^s_{r,q}(M)$ with $M > 0$, $s > d/r$, $r \geq 1$ and $q \geq 1$. Then there exists a constant $C > 0$ such that

$$E\left(\int_{[0,1]^d} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^p d\mathbf{x}\right) \leq C\Theta_n,$$

where

$$\Theta_n = \begin{cases} \left(\dfrac{\ln(n)}{n}\right)^{\frac{sp}{2s+d}}, & \text{for} \quad 2rs > d(p-r), \\ \left(\dfrac{\ln(n)}{n}\right)^{\frac{(s-d(1/r-1/p))p}{2s-2d/r+d}}, & \text{for} \quad 2rs < d(p-r), \\ \left(\dfrac{\ln(n)}{n}\right)^{\frac{(s-d(1/r-1/p))p}{2s-2d/r+d}} (\ln(n))^{\left(p-\frac{2r}{q}\right)_+}, & \text{for} \quad 2rs = d(p-r). \end{cases}$$

Theorem 5.1 can be proved using similar arguments to (Kerkyacharian and Picard, 2000, Theorem 5.1) for a bound of the $\mathbb{L}_p$-risk and the multidimensional framework of (Delyon and Juditsky, 1996, Theorem 1) for the determination of the rates of convergence.

References

Abbaszadeh, M., Chesneau, C. and Doosti, H. (2013). Multiplicative censoring : estimation of a density and its derivatives under the L_p-risk, *Revstat*, 11, 255-276.

Akakpo, N. and Lacour, C. (2011). Inhomogeneous and anisotropic conditional density estimation from dependent data, *Electron. Journal of Statistics*, 5, 1618-1653.

Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion), *Journal of the Italian Statistical Society Series B*, 6, 97-144.

Chagny, G. (2013). Warped bases for conditional density estimation, *Mathematical Methods of Statistics*, 22, 4, 253-282.

Chesneau, C., Dewan, I. and Doosti, H. (2014). Nonparametric estimation of a two dimensional continuous-discrete density function by wavelets, *Statistical Methodology*, 18, 64-78.

Chesneau, C., Dewan, I. and Doosti, H. (2016). Nonparametric estimation of a quantile density function by wavelet methods, *Computational Statistics and Data Analysis*, 94, 161-174.

Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms, *Applied and Computational Harmonic Analysis*, 24, 1, 54-81.

Daubechies, I. (1992). *Ten lectures on wavelets*, SIAM.

Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators, *Applied Computational Harmonic Analysis*, 3, 215-228.

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

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding, *Annals of Statistics*, 24, 508-539.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies, *Annals of Statistics*, 21, 196-216.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelet, Approximation and Statistical Applications*, Lectures Notes in Statistics New York, 129, *Springer Verlag*.

Kerkyacharian, G. and Picard, D. (2000). Thresholding algorithms, maxisets and well concentrated bases (with discussion and a rejoinder by the authors), *Test*, 9, 2, 283-345.

Li, Q. and Racine, J. (2003), Nonparametric estimation of distributions with categorical and continuous data, *Journal of Multivariate Analysis*, 86, 266-292.

Li, Q. and Racine, J. (2007). Nonparametric Econometrics: Theory and Practice, *Princeton University Press*.

Mallat, S. (2009). *A wavelet tour of signal processing*, Elsevier/ Academic Press, Amsterdam, third edition. The sparse way, With contributions from Gabriel Peyré.

Meyer, Y. (1992). *Wavelets and Operators*, Cambridge University Press, Cambridge.

Petrov, V.V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford: Clarendon Press.

Ramirez, P. and Vidakovic, B. (2010). Wavelet density estimation for stratified size-biased sample, *Journal of Statistical planning and Inference*, 140, 419-432.

Rosenthal, H.P. (1970). On the subspaces of $\mathbb{L}^p$ $(p \geq 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8, 273-303.

Vasiliev, V. (2014). A truncated estimation method with guaranteed accuracy, *Ann. Inst. Stat. Math*, 66, 1, 141-163.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*, John Wiley & Sons, Inc., New York, 384 pp.