A note on the adaptive estimation of the differential entropy by wavelet methods

Christophe Chesneau, Fabien Navarro, Oana Silvia Serea

Abstract. In this note we consider the estimation of the differential entropy of a probability density function. We propose a new adaptive estimator based on a plug-in approach and wavelet methods. Under the mean $\mathbb{L}_p$ error, $p \geq 1$, this estimator attains fast rates of convergence for a wide class of functions. We present simulation results in order to support our theoretical findings.

Keywords: entropy; wavelet estimation; rate of convergence; mean $\mathbb{L}_p$ error

Classification: 62G07, 62G20

1. Introduction

Entropy is a measure of uncertainty which plays a fundamental role in many applications, such as goodness-of-fit tests, quantization theory, statistical communication theory, source-coding, econometrics, and many other areas (see, e.g., [Beirlant *et al.* (1997)]).

In this paper, we focus our attention on the concept of differential entropy, originally introduced by [Shannon (1948)]. More precisely, we explore the estimation of the differential entropy of a probability density function $f : [0,1]^d \to [0,\infty)$, $d \geq 1$. Recall that the entropy is defined by

$$H = -\int_{[0,1]^d} f(\mathbf{x}) \log(f(\mathbf{x})) d\mathbf{x}. \tag{1.1}$$

The literature on the estimation of H is extensive, see, e.g., [Beirlant *et al.* (1997)] and the references cited therein. Among the existing estimation methods, we consider a plug-in integral estimator of the form:

$$\hat{H} = -\int_{\hat{A}} \hat{f}(\mathbf{x}) \log(\hat{f}(\mathbf{x})) d\mathbf{x},$$

where $\hat{f}$ denotes an estimator for f, and $\hat{A} \subseteq \{\mathbf{x} \in [0,1]^d;\ \hat{f}(\mathbf{x}) > 0\}$. This type of plug-in integral estimators was introduced by [Dmitriev and Tarasenko (1973)], in the context of kernel density estimation. The authors showed strong consistency of the estimator, but other aspects have been studied as well, e.g.,

DOI 10.14712/1213-7243.2015.191

by [Prakasa Rao (1983)], [Joe (1989)], [Mokkadem (1989)], [Györfi and van der Meulen (1990)], [Györfi and van der Meulen (1991)] and [Mason (2003)]. Recent developments can be found, e.g., in [Bouzebda and Elhattab (2011)], [Bouzebda and Elhattab (2009)], [Bouzebda and Elhattab (2010)].

The contributions of this paper are twofold. Firstly, we establish a new general upper bound for the mean $\mathbb{L}_p$ error of $\hat{H}$, i.e., $R(\hat{H}, H) = \mathbb{E}(|\hat{H} - H|^p)$, expressed in terms of mean integrated $\mathbb{L}_{2p}$ error of $\hat{f}$, i.e., $R_*(\hat{f}, f) = \mathbb{E}(\int_{[0,1]^d}(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^{2p} d\mathbf{x})$. The obtained bound illustrates that the more efficient $\hat{f}$ is under the mean integrated $\mathbb{L}_{2p}$ error, the more efficient is $\hat{H}$ under the mean $\mathbb{L}_p$ error. The advantage of this result is its great flexibility with respect to both the model and the estimation method for $\hat{f}$. This result can also be viewed as an extension of the mean $\mathbb{L}_p$ error of $\hat{H}$ obtained by [Mokkadem (1989)] for the standard density model and kernel method.

Secondly, we introduce a new integral estimator $\hat{H}$ based on a multidimensional hard thresholding wavelet estimator for $\hat{f}$. Such a wavelet estimator $\hat{f}$ was introduced by [Donoho *et al.* (1996)] and [Delyon and Juditsky (1996)]. The construction of this estimator does not depend on the smoothness of f, and it is efficient under the mean integrated $\mathbb{L}_q$ error (with $q \geq 1$). Further details on wavelet estimators in various statistical setting can be found, e.g., in [Antoniadis (1997)], [Härdle *et al.* (1998)] and [Vidakovic (1999)]. Applying our general upper bound, we prove that $\hat{H}$ attains fast rates of convergence under mild assumptions on f: we only suppose that f belongs to a wide set of functions, the so-called Besov balls. Consequences of our results are $\mathbb{L}_p$ as well as a.s. convergence of our estimator. To the best of our knowledge, in this statistical context, $\hat{H}$ constitutes the first adaptive estimator for H based on wavelets. We also propose a short simulation study to support our theoretical findings.

The remainder of this paper is organized as follows. In the next section, we present an upper bound for the mean $\mathbb{L}_p$ error of $\hat{H}$. Section 3.1 is devoted to our wavelet estimator and its performances in terms of rate of convergence under the mean $\mathbb{L}_p$ error over Besov balls. Section 4 contains a short simulation study illustrating the performance of our wavelet estimator. For the convenience of the reader, the proofs are postponed to Section 5.

2. A general upper bound

2.1 Notations and assumptions. We define the $\mathbb{L}_p([0,1]^d)$-spaces with $p \geq 1$ by

$$\mathbb{L}_p([0,1]^d) = \left\{ h : [0,1]^d \to \mathbb{R}; \quad \left(\int_{[0,1]^d} |h(\mathbf{x})|^p d\mathbf{x} \right)^{1/p} < \infty \right\}$$

with the usual modification if $p = \infty$.

We formulate the following assumptions:

(A1): There exists a constant $c_* > 0$ such that

$$\inf_{\mathbf{x}\in[0,1]^d} f(\mathbf{x}) \geq c_*.$$

(A2(p)): Let $p \geq 1$ and $q = p/(p-1)$. We have

$$f \in \mathbb{L}_{2p}([0,1]^d), \qquad \log(f) \in \mathbb{L}_q([0,1]^d), \qquad f\log(f) \in \mathbb{L}_q([0,1]^d).$$

(A3): There exists a constant $C_* > 0$ such that

$$\sup_{\mathbf{x}\in[0,1]^d} f(\mathbf{x}) \leq C_*.$$

These assumptions are satisfied by a wide family of probability density functions. They have ever been used in the context of estimating the differential entropy see, for instance, [Beirlant *et al.* (1997)]. Note that **(A1)** and **(A3)** imply **(A2(p))** (since $|\log(f(\mathbf{x}))| \leq \max(|\log(c_*)|, |\log(C_*)|)$).

2.2 An auxiliary result. In this section, we adopt a general estimation setting: let $\hat{f} : [0,1]^d \to \mathbb{R}$ be an estimator of f constructed from random vectors defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Various estimation methods can be found in, e.g., [Tsybakov (2004)]. Suppose that **(A1)** is satisfied. We study the following plug-in integral estimator for H (1.1):

$$\hat{H} = -\int_{\hat{A}} \hat{f}(\mathbf{x})\log(\hat{f}(\mathbf{x}))d\mathbf{x}, \tag{2.1}$$

where

$$\hat{A} = \left\{\mathbf{x} \in [0,1]^d;\ \hat{f}(\mathbf{x}) \geq \frac{c_*}{2}\right\}.$$

Such plug-in integral estimator was introduced by [Dmitriev and Tarasenko (1973)] with a kernel density estimator (and a different $\hat{A}$). Related results can be found in [Beirlant *et al.* (1997)] and the references cited therein. In particular, [Mokkadem (1989)] has investigated the mean $\mathbb{L}_p$ error of $\hat{H}$ for the standard density model and a kernel estimator (with a different $\hat{A}$).

Without the specification of the model and for any estimator $\hat{f}$ for f, Proposition 2.1 establishes a general upper bound for the mean $\mathbb{L}_p$ error of $\hat{H}$ in terms of the mean integrated $\mathbb{L}_{2p}$ error of $\hat{f}$.

Proposition 2.1. *Let $p \geq 1$. Suppose that* **(A1)** *and* **(A2(p))** *are satisfied and $\hat{f} \in \mathbb{L}_{2p}([0,1]^d)$. Let $\hat{H}$ be defined by* (2.1) *and H be defined by* (1.1). *Then we* *have the following upper bound for the mean* $\mathbb{L}_p$ *error of* $\hat{H}$*:*

$$\mathbb{E}(|\hat{H}-H|^p) \leq K\left(\sqrt{\mathbb{E}\left(\int_{[0,1]^d}(\hat{f}(\mathbf{x})-f(\mathbf{x}))^{2p}d\mathbf{x}\right)}+\mathbb{E}\left(\int_{[0,1]^d}(\hat{f}(\mathbf{x})-f(\mathbf{x}))^{2p}d\mathbf{x}\right)\right),$$

where

$$K = 2^{p-1}\max\left((C')^p, (c_*)^{-p}\right),$$

$$C' = \frac{2}{c_*}\left(\int_{[0,1]^d}(f(\mathbf{x})|\log(f(\mathbf{x}))|)^q d\mathbf{x}\right)^{1/q} + \left(\int_{[0,1]^d}(|\log(f(\mathbf{x}))|+1)^q d\mathbf{x}\right)^{1/q}$$

with $q = p/(p-1)$.

Proposition 2.1 illustrates the intuitive idea that more $\hat{f}$ is efficient in terms of mean integrated $\mathbb{L}_{2p}$ error, more $\hat{H}$ is efficient in terms of mean $\mathbb{L}_p$ error. The obtained bound has the advantage of enjoying a great flexibility on the model and the choice of $\hat{f}$.

In order to highlight this flexibility, one can consider the standard density model: f is the common probability density function of n *iid* $[0,1]^d$-valued random vectors, $d \geq 1$, $\mathbf{X}_1,\ldots,\mathbf{X}_n$, or with no *iid* assumption, or f can be a probability density function emerging from a more sophisticated density model as the convolution one (see, e.g., [Caroll and Hall (1988)], [Devroye (1989)] and [Fan (1991)]). On the other hand, one can consider several type of estimators as kernel, spline, Fourier series or wavelet series, as soon as they enjoy good mean integrated $\mathbb{L}_{2p}$ error properties.

Remark 2.1. If n is such that $c_* \geq 1/\log(n)$, one can define $\hat{H}$ (2.1) by replacing c_* in $\hat{A}$ by $1/\log(n)$ and Proposition 2.1 is still valid with $1/\log(n)$ instead of c_*, implying that $K \leq C(\log(n))^p$.

In the rest of the study we focus our attention on a nonlinear wavelet estimator having the features to be adaptive and efficient under the mean integrated $\mathbb{L}_{2p}$ error for a wide class of functions f.

3. An adaptive wavelet estimator

Before introducing our main estimator, let us present some basics on wavelets and the considered function spaces characterizing the unknown smoothness of f; the Besov balls.

3.1 Wavelet bases on $[0,1]$**.** We consider an orthonormal wavelet basis generated by dilations and translations of the scaling and wavelet functions ϕ and

ψ from the Daubechies family db_{2R}, with $R \geq 1$ (see [Daubechies (1992)]). We define the scaled and translated version of ϕ and ψ by

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), \qquad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k).$$

Then, with an appropriate treatment at the boundaries, there exists an integer τ satisfying $2^\tau \geq 2R$ such that, for any integer $j_* \geq \tau$, the collection

$$\{\phi_{j_*,k},\ k \in \{0,\dots,2^{j_*}-1\};\ \psi_{j,k};\ j \in \mathbb{N} - \{0,\dots,j_*-1\},\ k \in \{0,\dots,2^j-1\}\},$$

forms an orthonormal basis of $\mathbb{L}_2([0,1])$. See [Meyer (1992)], [Daubechies (1992)], [Cohen *et al.* (1993)] and [Mallat (2009)].

3.2 Wavelet tensor product bases on $[0,1]^d$. We use compactly supported tensor product wavelet bases on $[0,1]^d$ based on the Daubechies family. Their construction is recalled below. For any $\mathbf{x} = (x_1,\dots,x_d) \in [0,1]^d$, we set

$$\Phi(\mathbf{x}) = \prod_{v=1}^{d} \phi(x_v),$$

and

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

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

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

$$\Phi_{j,\mathbf{k}}(\mathbf{x}) = 2^{jd/2}\Phi(2^j x_1 - k_1,\dots,2^j x_d - k_d),$$
$$\Psi_{j,\mathbf{k},u}(\mathbf{x}) = 2^{jd/2}\Psi_u(2^j x_1 - k_1,\dots,2^j x_d - k_d), \quad \text{for any } \ u \in \{1,\dots,2^d-1\}.$$

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

$$\{\Phi_{\tau,\mathbf{k}}, \mathbf{k} \in D_\tau;\ (\Psi_{j,\mathbf{k},u})_{u\in\{1,\dots,2^d-1\}}, \quad j \in \mathbb{N} - \{0,\dots,\tau-1\},\ \mathbf{k} \in D_j\}$$

forms an orthonormal basis of $\mathbb{L}_2([0,1]^d)$.

For any integer j_* such that $j_* \geq \tau$, a function $h \in \mathbb{L}_2([0,1]^d)$ can be expanded into a wavelet series as

$$h(\mathbf{x}) = \sum_{\mathbf{k}\in D_{j_*}} \alpha_{j_*,\mathbf{k}}\Phi_{j_*,\mathbf{k}}(\mathbf{x}) + \sum_{u=1}^{2^d-1}\sum_{j=j_*}^{\infty}\sum_{\mathbf{k}\in D_j} \beta_{j,\mathbf{k},u}\Psi_{j,\mathbf{k},u}(\mathbf{x}), \qquad \mathbf{x} \in [0,1]^d,$$

where

$$(3.1)\quad \alpha_{j,\mathbf{k}} = \int_{[0,1]^d} h(\mathbf{x})\Phi_{j,\mathbf{k}}(\mathbf{x})d\mathbf{x}, \qquad \beta_{j,\mathbf{k},u} = \int_{[0,1]^d} h(\mathbf{x})\Psi_{j,\mathbf{k},u}(\mathbf{x})d\mathbf{x}.$$

3.3 Besov balls. Let $M > 0$, $s \in (0, R)$, $p \geq 1$ and $q \geq 1$. We say that a function h in $\mathbb{L}_r([0,1]^d)$ belongs to $\mathbf{B}^s_{r,q}(M)$ if, and only if, there exists a constant $M^* > 0$ (depending on M) such that the associated wavelet coefficients (3.1) satisfy

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

In this expression, s is a smoothness parameter and p and q are norm parameters. Besov spaces include many traditional smoothness spaces as the standard Hölder and Sobolev balls. See [Meyer (1992)], [Härdle *et al.* (1998)] and [Mallat (2009)].

3.4 Wavelet estimations. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n *iid* $[0,1]^d$-valued random vectors, $d \geq 1$, with common probability density function f. We aim to estimate the differential entropy of f defined by

$$H = -\int_{[0,1]^d} f(\mathbf{x})\log(f(\mathbf{x}))d\mathbf{x},$$

from $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Under **(A1)**, we consider the following estimator for H:

$$(3.2)\qquad \hat{H} = -\int_{\hat{A}} \hat{f}(\mathbf{x})\log(\hat{f}(\mathbf{x}))d\mathbf{x},$$

where

$$\hat{A} = \left\{\mathbf{x} \in [0,1]^d;\ \hat{f}(\mathbf{x}) \geq \frac{c_*}{2}\right\}$$

and $\hat{f}$ is the following hard thresholding wavelet estimator for f:

$$(3.3)\qquad \begin{aligned}\hat{f}(\mathbf{x}) = &\sum_{\mathbf{k}\in D_\tau}\hat{\alpha}_{\tau,\mathbf{k}}\Phi_{\tau,\mathbf{k}}(\mathbf{x}) \\ &+ \sum_{u=1}^{2^d-1}\sum_{j=j_*}^{j_1}\sum_{\mathbf{k}\in D_j}\hat{\beta}_{j,\mathbf{k},u}\mathbf{1}_{\left\{|\hat{\beta}_{j,\mathbf{k},u}|\geq\kappa\lambda_n\right\}}\Psi_{j,\mathbf{k},u}(\mathbf{x}),\end{aligned}$$

where

$$\hat{\alpha}_{j,\mathbf{k}} = \frac{1}{n}\sum_{i=1}^{n}\Phi_{j,\mathbf{k}}(\mathbf{X}_i), \qquad \hat{\beta}_{j,\mathbf{k},u} = \frac{1}{n}\sum_{i=1}^{n}\Psi_{j,\mathbf{k},u}(\mathbf{X}_i),$$

j_1 is the resolution level satisfying $2^{dj_1} = [n/\log(n)]$ (the integer part of $n/\log(n)$), $\mathbf{1}$ is the indicator function, κ is a large enough constant and λ_n is the threshold

$$\lambda_n = \sqrt{\frac{\log(n)}{n}}.$$

This estimator was introduced by [Donoho *et al.* (1996)] for $d = 1$ and generalized to the multidimensional case by [Delyon and Juditsky (1996)]. The central idea is to estimate only the wavelet coefficients with a high magnitude because they contain all the necessary information inherent to f. The others, less important, are suppressed instead of being estimated in order to avoid the accumulation of superfluous errors in their estimation.

This estimator is adaptive; its construction does not depend on the unknown smoothness of f.

Let us mention that $\alpha_{j,\mathbf{k}}$ and $\beta_{j,\mathbf{k},u}$ are unbiased estimators for the wavelet coefficients $\alpha_{j,\mathbf{k}}$ and $\beta_{j,\mathbf{k},u}$ respectively. They also satisfied powerful moment inequalities and concentration inequalities. Further details on wavelet estimation can be found in, e.g., [Antoniadis (1997)], [Härdle *et al.* (1998)] and [Vidakovic (1999)].

Theorem 3.1 below investigates the rates of convergence attained by $\hat{H}$ under the mean $\mathbb{L}_p$ error over Besov balls for f.

Theorem 3.1. *Let $p \geq 1$. Suppose that* **(A1)** *and* **(A3)** *are satisfied. Let $\hat{H}$ be* (3.2). *Suppose that $f \in \mathbf{B}^s_{r,q}(M)$ with $s > d/r$, $r \geq 1$ and $q \geq 1$. Then there exists a constant $C > 0$ such that, for n large enough,*

$$\mathbb{E}(|\hat{H} - H|^p) \leq C\varphi_n(p),$$

where

$$\varphi_n(p) = \begin{cases} \left(\dfrac{\log(n)}{n}\right)^{sp/(2s+d)}, \\ \text{for } \ 2rs > d(2p-r), \\ \left(\dfrac{\log(n)}{n}\right)^{(s-d/r+d/(2p))p/(2s-2d/r+d)}, \\ \text{for } \ 2rs < d(2p-r), \\ \left(\dfrac{\log(n)}{n}\right)^{(s-d/r+d/(2p))p/(2s-2d/r+d)} (\log(n))^{\max(2p-r/q,0)}, \\ \text{for } \ 2rs = d(2p-r). \end{cases}$$

The proof of Theorem 3.1 is based on Proposition 2.1 and a result on the rates of convergence of $\hat{f}$ under the mean integrated $\mathbb{L}_{2p}$ error.

The rate of convergence $\varphi_n(p)$ is closed to the one attains by $\hat{f}$ (3.3) under the mean integrated $\mathbb{L}_p$ error. We do not claim that $\varphi_n(p)$ is the optimal one for the estimation of H in the minimax sense. However, Theorem 3.1 is enough to prove that:

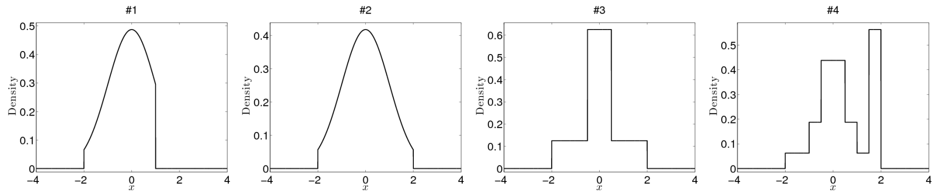

FIGURE 1. Test densities.

- $\hat{H}$ converges to H under the mean $\mathbb{L}_p$ error, i.e., $\lim_{n\to\infty}\mathbb{E}(|\hat{H}-H|^p)=0$,
- under some restriction on s, r and q, one can find p such that, for any $\epsilon>0$, by the Markov inequality, $\sum_{n=1}^{\infty}\mathbb{P}(|\hat{H}-H|\geq\epsilon)\leq\epsilon^{-p}\sum_{n=1}^{\infty}\varphi_n(p)<\infty$ (convergent Bertrand series). Therefore $\hat{H}$ converges to H a.s. by the Borel-Cantelli lemma.

Remark 3.1. As in Remark 2.1, if n is such that $c_*\geq 1/\log(n)$, one can define $\hat{H}$ (3.2) by replacing c_* in $\hat{A}$ by $1/\log(n)$ and Theorem 3.1 is still valid with the rate of convergence $(\log(n))^p\varphi_n(p)$.

4. Numerical results

We now illustrate these theoretical results by a short simulation study. We have compared the numerical performances of the adaptive wavelet estimator $\hat{H}$ (2.1) to those of the traditional kernel estimator denoted by $\tilde{H}$ and based on the same plug-in approach. All experiments were conducted using a Gaussian kernel and we have been focused on a global bandwidth selector: the *rule of thumb* (rot) bandwidth selector (see, e.g., [Silverman (1986)]). Thus, the optimal bandwidth is given by $h_{\mathrm{rot}}=1.06\min(\hat{\sigma},Q/1.34)n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation and Q is the interquartile range.

In order to satisfy the assumptions (**A1**) and (**A2**), we have considered mixtures of uniform distributions and the two-sided truncated normal distribution on $[a,b]$ denoted by $\mathcal{N}(\mu,\sigma^2,a,b)$, with density

$$f(x;\mu,\sigma,a,b)=\begin{cases}\frac{\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)} & \text{if } a\leq x\leq b\\ 0 & \text{otherwise}\end{cases}$$

where $\varphi(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$ is the probability density function of the standard normal distribution, $\Phi(\cdot)$ is its cumulative distribution function and the parameters μ and σ are respectively the mean and the standard deviation of the distribution.

More precisely we have considered the following examples, see Figure 1

#1 f is the two-sided truncated normal distribution $\mathcal{N}(0,1,-2,1)$.
#2 f is the two-sided truncated normal distribution $\mathcal{N}(0,1,-2,2)$.
#3 f is a mixture of two uniform densities $\frac{1}{2}\mathcal{U}(-\frac{1}{2},\frac{1}{2})+\frac{1}{2}\mathcal{U}(-2,2)$.

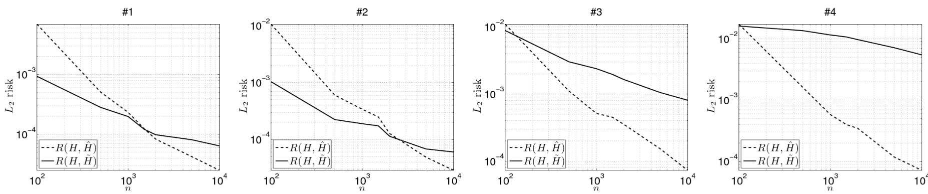

FIGURE 2. $\mathbb{L}_2$ risk as a function of the sample sizes n (in a log-log scale) of $\hat{H}$ (dashed) and $\tilde{H}$ (solid).

#4 f is a mixture of four uniform densities $\frac{1}{4}\mathcal{U}(-\frac{1}{2},\frac{1}{2}) + \frac{1}{4}\mathcal{U}(-2,2) + \frac{1}{4}\mathcal{U}(-1,1) + \frac{1}{4}\mathcal{U}(\frac{3}{2},2)$.

Since our estimation method is adaptive, we have chosen a predetermined threshold $\kappa = \sqrt{2}$ and the density was evaluated at $T = 2^J$ equispaced points $t_i = 2ib_1/T$, $i = -T/2,\dots,T/2-1$ between $-b_1$ and b_1, where J is the index of the highest resolution level and T is the number of discretization points, with $J = 8$, $T = 256$ and $b_1 = 4$. The primary level $j_* = 3$ and the Haar wavelet was used throughout all experiments. For both estimation methods we used the trapezoidal rule to approximate the integral estimate of entropy with $\hat{A} = \{\mathbf{x} \in [0,1]^d;\ \hat{f}(\mathbf{x}) \geq \frac{c_*}{2}\}$. Note that this amounts to evaluating the integral over the grid points located within densities supports (i.e., $[-2,2]$ for #2–#4 and $[-2,1]$ for #1). All simulations have been implemented under Matlab.

Each method was applied for sample sizes ranging from 100 to 10,000. The $\mathbb{L}_2-$risk from 100 repetitions are depicted as a function of the sample size in Figure 2. It shows that none of the methods clearly outperforms the others in all cases. However, our estimator outperforms the kernel estimator in many cases especially for the moderate or large sample sizes. In comparison to the kernel method, our method provided much better results on the non-smooth uniform mixture densities. Without any prior smoothness knowledge on the unknown density, $\hat{H}$ provides competitive results in comparison to $\tilde{H}$. Furthermore, as expected, for both methods, and in all cases, the $\mathbb{L}_2-$risk is decreasing as the sample size increases.

5. Proofs

PROOF OF PROPOSITION 2.1: Let $p \geq 1$ and $q = p/(p-1)$. We have

$$\hat{H} - H = -\int_{\hat{A}} (\hat{f}(\mathbf{x})\log(\hat{f}(\mathbf{x})) - f(\mathbf{x})\log(f(\mathbf{x})))d\mathbf{x} + \int_{\hat{A}^c} f(\mathbf{x})\log(f(\mathbf{x}))d\mathbf{x}.$$

The triangular inequality yields

$$|\hat{H} - H| \leq F + G, \tag{5.1}$$

where

$$F = \int_{\hat{A}} |\hat{f}(\mathbf{x}) \log(\hat{f}(\mathbf{x})) - f(\mathbf{x}) \log(f(\mathbf{x}))| d\mathbf{x}, \qquad G = \int_{\hat{A}^c} f(\mathbf{x}) |\log(f(\mathbf{x}))| d\mathbf{x}.$$

Upper bound for G. By the Hölder inequality and **(A2(p))**, we have

$$G \leq \left(\int_{\hat{A}^c} d\mathbf{x} \right)^{1/p} \left(\int_{[0,1]^d} |f(\mathbf{x}) \log(f(\mathbf{x}))|^q d\mathbf{x} \right)^{1/q}. \tag{5.2}$$

Observe that, thanks to **(A1)**, we have

$$\begin{aligned} \hat{A}^c &= \left\{ \mathbf{x} \in [0,1]^d;\ \hat{f}(\mathbf{x}) < \frac{c_*}{2} \right\} \subseteq \left\{ \mathbf{x} \in [0,1]^d;\ f(\mathbf{x}) - \hat{f}(\mathbf{x}) > \frac{c_*}{2} \right\} \\ &\subseteq \left\{ \mathbf{x} \in [0,1]^d;\ |\hat{f}(\mathbf{x}) - f(\mathbf{x})| > \frac{c_*}{2} \right\}. \end{aligned} \tag{5.3}$$

It follows from (5.2), (5.3) and the Markov inequality that

$$\begin{aligned} G &\leq \left(\int_{\hat{A}^c} \left(\frac{2}{c_*} \right)^p |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^p d\mathbf{x} \right)^{1/p} \left(\int_{[0,1]^d} |f(\mathbf{x}) \log(f(\mathbf{x}))|^q d\mathbf{x} \right)^{1/q} \\ &\leq C_o \left(\int_{[0,1]^d} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^p d\mathbf{x} \right)^{1/p}, \end{aligned} \tag{5.4}$$

where $C_o = (2/c_*) \left(\int_{[0,1]^d} (f(\mathbf{x}) |\log(f(\mathbf{x}))|)^q d\mathbf{x} \right)^{1/q}$.

Upper bound for F. Since $f \in \mathbb{L}_{2p}([0,1]^d)$ and $\hat{f} \in \mathbb{L}_{2p}([0,1]^d)$, we have $\max(\hat{f}(\mathbf{x}), f(\mathbf{x})) < \infty$ almost surely. The Taylor theorem with Lagrange remainder applied to $\varphi(y) = y \log(y)$ between $f(\mathbf{x})$ and $\hat{f}(\mathbf{x})$ ensures the existence of a function $\theta(x) \in [\min(\hat{f}(\mathbf{x}), f(\mathbf{x})), \max(\hat{f}(\mathbf{x}), f(\mathbf{x}))] \subseteq [c_*/2, \infty)$ satisfying

$$\begin{aligned} \varphi(\hat{f}(\mathbf{x})) - \varphi(f(\mathbf{x})) &= \varphi'(f(\mathbf{x}))(\hat{f}(\mathbf{x}) - f(\mathbf{x})) + \frac{1}{2} \varphi''(\theta(\mathbf{x}))(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 \\ &= (\log(f(\mathbf{x})) + 1)(\hat{f}(\mathbf{x}) - f(\mathbf{x})) + \frac{1}{2\theta(\mathbf{x})}(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2. \end{aligned}$$

Hence, by the triangular inequality, we have

$$\begin{aligned} &|\hat{f}(\mathbf{x}) \log(\hat{f}(\mathbf{x})) - f(\mathbf{x}) \log(f(\mathbf{x}))| = |\varphi(\hat{f}(\mathbf{x})) - \varphi(f(\mathbf{x}))| \\ &\leq (|\log(f(\mathbf{x}))| + 1)|\hat{f}(\mathbf{x}) - f(\mathbf{x})| + \frac{1}{2\theta(\mathbf{x})}(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 \\ &\leq (|\log(f(\mathbf{x}))| + 1)|\hat{f}(\mathbf{x}) - f(\mathbf{x})| + \frac{1}{c_*}(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2. \end{aligned}$$

By the Hölder inequality and **(A2(p))**, we have

$$
\begin{aligned}
F &\leq \int_{[0,1]^d} (|\log(f(\mathbf{x}))| + 1)\, |\hat{f}(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} + \frac{1}{c_*} \int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} \\
&\leq C_{oo} \left(\int_{[0,1]^d} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^p d\mathbf{x} \right)^{1/p} + \frac{1}{c_*} \int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x},
\end{aligned} \tag{5.5}
$$

where $C_{oo} = (\int_{[0,1]^d} (|\log(f(\mathbf{x}))| + 1)^q d\mathbf{x})^{1/q}$.

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

$$
|\hat{H} - H| \leq C' \left(\int_{[0,1]^d} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^p d\mathbf{x} \right)^{1/p} + \frac{1}{c_*} \int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x},
$$

where $C' = C_o + C_{oo}$.

The inequality: $|x+y|^p \leq 2^{p-1}(|x|^p + |y|^p)$, $(x,y) \in \mathbb{R}^2$, implies that

$$
\begin{aligned}
&|\hat{H} - H|^p \\
&\leq 2^{p-1} \left((C')^p \int_{[0,1]^d} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^p d\mathbf{x} + (c_*)^{-p} \left(\int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} \right)^p \right) \\
&\leq K \left(\int_{[0,1]^d} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^p d\mathbf{x} + \left(\int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} \right)^p \right),
\end{aligned}
$$

where $K = 2^{p-1} \max((C')^p, (c_*)^{-p})$.

The Hölder inequality applied two times gives

$$
|\hat{H} - H|^p \leq K \left(\sqrt{\int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^{2p} d\mathbf{x}} + \int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^{2p} d\mathbf{x} \right).
$$

It follows from the Cauchy-Schwarz inequality that

$$
\begin{aligned}
&\mathbb{E}(|\hat{H} - H|^p) \\
&\leq \quad K \left(\sqrt{\mathbb{E}\left(\int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^{2p} d\mathbf{x} \right)} + \mathbb{E}\left(\int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^{2p} d\mathbf{x} \right) \right).
\end{aligned}
$$

The proof of Proposition 2.1 is complete. □

Proof of Theorem 3.1: First of all, let us present a result on the rates of convergence of $\hat{f}$ (3.3) under the mean $\mathbb{L}_\theta$ error over Besov balls.

Theorem 5.1 (Delyon and Juditsky (1996) & Kerkyacharian and Picard (2000)). *Suppose that* **(A3)** *holds. Let $\theta \geq 1$ and $\hat{f}$ be* (3.3). *Suppose that $f \in \mathbf{B}^s_{r,q}(M)$*

with $s > d/r$, $r \geq 1$ and $q \geq 1$. Then there exists a constant $C > 0$ such that

$$\mathbb{E}\left(\int_{[0,1]^d} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\theta d\mathbf{x}\right) \leq C\Psi_n(\theta),$$

where

$$\Psi_n(\theta) = \begin{cases} \left(\frac{\log(n)}{n}\right)^{s\theta/(2s+d)}, \\ \text{for } 2rs > d(\theta - r), \\ \left(\frac{\log(n)}{n}\right)^{(s-d/r+d/\theta)\theta/(2s-2d/r+d)}, \\ \text{for } 2rs < d(\theta - r), \\ \left(\frac{\log(n)}{n}\right)^{(s-d/r+d/\theta)\theta/(2s-2d/r+d)} (\log(n))^{\max(\theta - r/q, 0)}, \\ \text{for } 2rs = d(\theta - r). \end{cases}$$

Theorem 5.1 can be proved using similar arguments to [Kerkyacharian and Picard (2000), Theorem 5.1] for a bound of the mean integrated $\mathbb{L}_\theta$ error of $\hat{f}$ and [Delyon and Juditsky (1996), Theorem 1] for the determination of the rates of convergence.

It follows from Proposition 2.1 and Theorem 3.1 with $\theta = 2p$ that, for any $f \in \mathbf{B}^s_{r,q}(M)$ with $s > d/r$, $r \geq 1$ and $q \geq 1$, and for n large enough,

$$\begin{aligned} &\mathbb{E}(|\hat{H} - H|^p) \\ &\leq K\left(\sqrt{\mathbb{E}\left(\int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^{2p} d\mathbf{x}\right)} + \mathbb{E}\left(\int_{[0,1]^d} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^{2p} d\mathbf{x}\right)\right) \\ &\leq K \max(\sqrt{C}, C)\left(\sqrt{\Psi_n(2p)} + \Psi_n(2p)\right) \leq 2K\max(\sqrt{C}, C)\sqrt{\Psi_n(2p)} \\ &= C_\star \varphi_n(p), \end{aligned}$$

with $C_\star = 2K\max(\sqrt{C}, C)$. This ends the proof of Theorem 3.1. □

References

[Antoniadis (1997)] Antoniadis A., *Wavelets in statistics: a review (with discussion)*, J. Ital. Statist. Soc. Series B **6** (1997), 97–144.

[Beirlant *et al.* (1997)] Beirlant J., Dudewicz E.J., Gyorfi L., van der Meulen E.C., *Nonparametric entropy estimation: an overview*, Int. J. Math. Stat. Sci. **6** (1997), 17–39.

[Bouzebda and Elhattab (2009)] Bouzebda S., Elhattab I., *A strong consistency of a nonparametric estimate of entropy under random censorship*, C.R. Math. Acad. Sci. Paris **347** (2009), no. 13–14, 821–826.

[Bouzebda and Elhattab (2010)] Bouzebda S., Elhattab I., *Uniform in bandwidth consistency of the kernel-type estimator of the Shannon's entropy*, C.R. Math. Acad. Sci. Paris **348** (2010), no. 5–6, 317–321.

[Bouzebda and Elhattab (2011)] Bouzebda S., Elhattab I., *Uniform-in-bandwidth consistency for kernel-type estimators of Shannon's entropy*, Electron. J. Stat. **5** (2011), 440–459.

[Caroll and Hall (1988)] Caroll R.J., Hall P., *Optimal rates of convergence for deconvolving a density*, J. Amer. Statist. Assoc. **83** (1988), 1184–1186.

[Cohen *et al.* (1993)] Cohen A., Daubechies I., Vial P., *Wavelets on the interval and fast wavelet transforms*, Appl. Comput. Harmon. Anal. **24** (1993), no. 1, 54–81.

[Daubechies (1992)] Daubechies I., *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[Delyon and Juditsky (1996)] Delyon B., Juditsky A., *On minimax wavelet estimators*, Appl. Comput. Harmon. Anal. **3** (1996), 215–228.

[Devroye (1989)] Devroye L., *Consistent deconvolution in density estimation*, Canad. J. Statist. **17** (1989), 235–239.

[Dmitriev and Tarasenko (1973)] Dmitriev Yu.G., Tarasenko F.P., *On the estimation functions of the probability density and its derivatives*, Theory Probab. Appl. **18** (1973), 628–633.

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

[Fan (1991)] Fan J., *On the optimal rates of convergence for nonparametric deconvolution problem*, Ann. Statist. **19** (1991), 1257–1272.

[Györfi and van der Meulen (1990)] Györfi L., van der Meulen E.C., *An entropy estimate based on a kernel density estimation*, in Limit theorems in probability and kernel-type estimators of Shannon's entropy statistics (Pécs, 1989), Colloq. Math. Soc. János Bolyai, 57, North-Holland, Amsterdam, 1990, pp. 229-240.

[Györfi and van der Meulen (1991)] Györfi L., van der Meulen E.C., *On the nonparametric estimation of the entropy functional*, in Nonparametric functional estimation and related topics (Spetses, 1990), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., 335, Kluwer Acad. Publ., Dordrecht, 1991, pp. 81–95.

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

[Joe (1989)] Joe H., *Estimation of entropy and other functionals of a multivariate density*, Ann. Inst. Statist. Math. **41** (1989), no. 4, 683–697.

[Kerkyacharian and Picard (2000)] Kerkyacharian G., Picard D., *Thresholding algorithms, maxisets and well concentrated bases. With comments and a rejoinder by the authors*, Test **9** (2000), no. 2, 283–345.

[Mallat (2009)] Mallat S., *A Wavelet Tour of Signal Processing. The sparse way*, third edition, with contributions from Gabriel Peyré, Elsevier/Academic Press, Amsterdam, 2009.

[Mason (2003)] Mason D.M., *Representations for integral functionals of kernel density estimators*, Austr. J. Stat. **32** (2003), no. 1–2, 131–142.

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

[Mokkadem (1989)] Mokkadem A., *Estimation of the entropy and information for absolutely continuous random variables*, IEEE Trans. Inform. Theory **35** (1989), 193–196.

[Prakasa Rao (1983)] Prakasa Rao B.L.S., *Nonparametric Functional Estimation*, Academic Press, Orlando, 1983.

[Shannon (1948)] Shannon C.E., *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423, 623–656.

[Silverman (1986)] Silverman B.W., *Density estimation: for statistics and data analysis*, Chapman & Hall, London, 1986.

[Tsybakov (2004)] Tsybakov A., *Introduction à l'estimation nonparamétrique*, Springer, Berlin, 2004.

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

Laboratoire de Mathématiques Nicolas Oresme, Université de Caen BP 5186, F 14032 Caen Cedex, France

E-mail: christophe.chesneau@unicaen.fr

CREST-ENSAI, Campus de Ker-Lann, Rue Blaise Pascal - BP 37203, 35172 Bruz Cedex, France

E-mail: fabien.navarro@ensai.fr

Univ. Perpignan Via Domitia, Laboratoire de Mathéematiques et Physique, EA 4217, F-66860 Perpignan, France

E-mail: oana-silvia.serea@univ-perp.fr

(*Received* April 13, 2015, *revised* December 3, 2015)