Journal of Statistical Planning and Inference 185 (2017) 69–93

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On the Kozachenko–Leonenko entropy estimator

Sylvain Delattre[a,*], Nicolas Fournier[b]

[a] Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599, Université Paris Diderot, Case courrier 7012, avenue de France, 75205 Paris Cedex 13, France

[b] Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599, Université Pierre-et-Marie Curie, Case 188, 4 place Jussieu, F-75252 Paris Cedex 5, France

ARTICLE INFO

Article history:
Received 31 August 2016
Received in revised form 18 January 2017
Accepted 23 January 2017
Available online 30 January 2017

MSC:
62G05

Keywords:
Non-parametric estimation
Entropy
Nearest neighbors

ABSTRACT

We study in detail the bias and variance of the entropy estimator proposed by Kozachenko and Leonenko (1987) for a large class of densities on $\mathbb{R}^d$. We then use the work of Bickel and Breiman (1983) to prove a central limit theorem in dimensions 1 and 2. In higher dimensions, we provide a development of the bias in terms of powers of $N^{-2/d}$. This allows us to use a Richardson extrapolation to build, in any dimension, a root-n consistent entropy estimator satisfying a central limit theorem which allows for explicit (asymptotic) confidence intervals. To our knowledge, all the previous general root-n consistency results were concerning dimension 1.

1. Introduction and main results

1.1. The setting

Consider a probability measure F on $\mathbb{R}^d$ with density f. We are interested in its entropy defined by

$$H(f) = -\int_{\mathbb{R}^d} f(x)\log f(x)\mathrm{d}x.$$

For $N \geq 1$ and for $X_1, \dots, X_{N+1}$ an i.i.d. sample of F, we consider, for each $i = 1, \dots, N+1$,

$$R_i^N = \min\{|X_i - X_j| : j = 1, \dots, N+1,\ j \neq i\} \quad \text{and} \quad Y_i^N = N(R_i^N)^d. \tag{1}$$

Here $|\cdot|$ stands for any norm on $\mathbb{R}^d$. For $x \in \mathbb{R}^d$ and $r \geq 0$, we set $B(x, r) = \{y \in \mathbb{R}^d : |y - x| \leq r\}$ and we introduce $v_d = \int_{B(0,1)} \mathrm{d}x$. We also denote by $\gamma = -\int_0^\infty e^{-x}\log x\mathrm{d}x \simeq 0.577$ the Euler constant. We finally set

$$H_N = \frac{1}{N+1}\sum_{i=1}^{N+1}\log Y_i^N + \gamma + \log v_d. \tag{2}$$

The estimator H_N of $H(f)$ was proposed by Kozachenko and Leonenko (1987). The object of the paper is to study in detail the bias, variance and asymptotic normality of H_N.

* Corresponding author.
E-mail addresses: sylvain.delattre@univ-paris-diderot.fr (S. Delattre), nicolas.fournier@upmc.fr (N. Fournier).

http://dx.doi.org/10.1016/j.jspi.2017.01.004
0378-3758/

1.2. Heuristics

Let us explain briefly why H_N should be consistent.

The conditional law of Y_i^N knowing X_i is approximately $\mathrm{Exp}(v_d f(X_i))$ for N large: for $r > 0$, $\Pr(Y_i^N > r \mid X_i) = [1 - F(B(X_i, (r/N)^{1/d}))]^N \simeq \exp(-NF(B(X_i, (r/N)^{1/d}))) \simeq \exp(-v_d f(X_i) r)$.

Consequently, we expect that $Y_i^N = \xi_i/(v_d f(X_i))$, for a family $(\xi_i)_{i=1,\dots,N+1}$ of approximately Exp(1)-distributed random variables, hopefully not too far from being independent.

We thus expect that $(N+1)^{-1}\sum_{i=1}^{N+1} \log Y_i^N \simeq \mathbb{E}[\log(\xi_1/(v_d f(X_1)))] = \mathbb{E}[\log \xi_1] - \log v_d - \mathbb{E}[\log f(X_1)] \simeq -\gamma - \log v_d + H(f)$ and thus that $H_N \simeq H(f)$.

1.3. Motivation

Estimating the entropy given some observations seems to be useful in various applied sciences and engineering. Searching for works on this topic, one finds a considerable number of applied papers that we do not try to summarize. Let us mention various fields such as independent component analysis, image analysis, genetic analysis, speech recognition, manifold learning, kinetic physics, molecular chemistry, computational neuroscience, etc. Beirlant et al. (1997) also mention applications to quantization, econometrics and spectroscopy.

The estimation of the relative entropy (or Kullback–Leibler divergence) of f with respect to some known g is deduced from the entropy estimation, since $H(f|g) = -H(f) - \int_{\mathbb{R}^d} f(x)\log g(x)\mathrm{d}x$ and since $\int_{\mathbb{R}^d} f(x)\log g(x)\mathrm{d}x$ is naturally estimated by the root-N asymptotically normal estimator $N^{-1}\sum_{i=1}^N \log g(X_i)$, at least if $\int_{\mathbb{R}^d} f(x)\log^2 g(x)\mathrm{d}x < \infty$.

Concerning applications to statistics, let us mention a few goodness-of-fit tests based on the entropy estimation: see Vasicek (1976) for normality (Gaussian laws maximize the entropy among all distributions with given variance), Dudewicz and van der Meulen (1981) for uniformity (uniform laws maximize the entropy among all distributions with given support), Mudholkar and Lin (1987) for exponentiality (exponential laws maximize the entropy among all $\mathbb{R}_+$-supported distributions with given mean). Also, Robinson (1991) proposed an independence test, based on the fact that $f \otimes g$ maximizes $H(h)$ among all densities h with marginals f and g.

1.4. Available results

We now list a few mathematical results. Let us first mention the review paper (Beirlant et al., 1997) by Beirlant, Dudewicz, Györfi and van der Meulen.

Levit (1978) has shown that $\mathbb{V}\mathrm{ar}\,(\log f(X_1)) = \int_{\mathbb{R}^d} f(x)\log^2 f(x)\mathrm{d}x - (H(f))^2$ is the smallest possible normalized (by N) asymptotic quadratic risk for entropy estimators in the local minimax sense.

Essentially, there are two types of methods for the entropy estimation. The *plug in* method consists in using an estimator of the form $H_N = -\int_{\mathbb{R}^d} f_N(x)\log f_N(x)\mathrm{d}x$, where f_N is an estimator of f. One then needs to use something like a kernel density estimator and this requires to have an idea of the tail behavior of f. Joe (1989) considers the case where f is bounded below on its (compact) support, while Hall and Morton (1993) propose some root N and asymptotically normal estimators assuming that $f(x) \sim a|x|^{-\alpha}$ (with α known) or $f(x) \sim a\exp(-b|x|^{-\alpha})$ (with α known).

The second class of methods consists in using *spacings* if $d = 1$, see Vasicek (1976), or *neighbors* as proposed by Kozachenko and Leonenko (1987). In Kozachenko and Leonenko (1987), a consistency result is proved (for H_N defined by (2)), in any dimension, under rather weak conditions on f and this is generalized to other notions of entropies by Leonenko et al. (2008) and Leonenko and Pronzato (2010). Instead of using *nearest* neighbor, we can use kth nearest neighbors with either k fixed or $1 \ll k \ll N$ (similarly, in dimension 1, we can use k-spacings).

In dimension 1 and assuming that f is bounded below on its (compact) support, Hall (1984, 1986) and van Es (1992) show some root N consistency and asymptotic normality for the entropy estimator based on k spacings (in both cases where k is fixed or tends to infinity at some suitable rate). Tsybakov and van der Meulen (1996) are the first to prove some root N consistency for some entropy estimator for general densities with unbounded support, in dimension 1. They consider a modified version of (2) and assume that f is sufficiently regular, positive and has some sub-exponential tails. Still in dimension 1, El Haje and Golubev (2009) prove some root N consistency and asymptotic normality for the entropy estimator based on 1-spacings. Furthermore, their assumptions on f are weaker than those of Tsybakov and van der Meulen (1996). In particular, they allow f to have some zeros and some fat tails.

Bickel and Breiman (1983) prove a very general central limit theorem for nonlinear bounded functionals of nearest neighbors in any dimension. Unfortunately, this does not apply to H_N and anyway, they do not study the bias. The work of Bickel and Breiman has been generalized in many directions, see Chatterjee (2008), Penrose and Yukich (2013) and Biau and Devroye (2015), with new ideas of proofs. However, these generalizations do not help us much, either because the moment conditions are too strong or because f is supposed to be bounded from below on its compact support. To prove our central limit theorem, the simplest is thus to start from the results of Bickel and Breiman (1983), because they are the closest to our framework.

Let us finally mention the paper of Pál et al. (2010): they study other notions of entropy, work in dimension $d \geq 1$, use estimators based on nearest neighbors and give some bounds on their rates of convergence.