C. R. Acad. Sci. Paris, Ser. I 352 (2014) 75–80

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistics

New Kernel-type estimator of Shanonn's entropy

Nouvel estimateur à noyau de l'entropie de Shanonn

Salim Bouzebda [a], Issam Elhattab [b]

[a] *Laboratoire de mathématiques appliquées de Compiègne, Université de technologie de Compiègne, BP 529, 60205 Compiègne cedex, France*
[b] *ENCG–Casablanca, Université Hassan-II Mohammedia, Morocco*

ARTICLE INFO

Article history:
Received 16 February 2012
Accepted after revision 19 November 2013
Available online 20 December 2013

Presented by Paul Deheuvels

ABSTRACT

In the present Note, we propose an estimator of Shanonn's entropy based on smooth estimators of quantile density. The consistency and asymptotic normality of the proposed estimates are obtained.

RÉSUMÉ

Dans cette Note, nous proposons un nouvel estimateur de l'entropie de Shanonn basé sur l'estimateur à noyau de la densité de quantile. Nous obtenons la consistance et la normalité de l'estimateur proposé.

Version française abrégée

Soient $X_1, \ldots, X_n$ des variables aléatoires réelles, indépendantes et de même loi que X. Pour tout $x \in \mathbb{R}$, soit $F(x) = \mathbb{P}(X \leqslant x)$ la fonction de répartition de X. Soit $\mathcal{Q}(\cdot) := F^{-1}(\cdot)$ la fonction *inverse*, ou *de quantiles*, de $F(\cdot)$ définie par :

$$\mathcal{Q}(t) := \inf\{x\colon\ F(x) \geqslant t\}, \quad \text{pour } 0 < t < 1.$$

Posons $x_F = \sup\{x\colon\ F(x) = 0\}$, $x^F = \inf\{x\colon\ F(x) = 1\}$, $\infty \leqslant x_F < x^F \leqslant \infty$. Nous supposons, dans la suite, que $F(\cdot)$ admet une densité $f(\cdot)$ par rapport à la mesure de Lebesgue. Soit $q(x) = \mathrm{d}\mathcal{Q}(x)/\mathrm{d}x = 1/f(\mathcal{Q}(x))$, pour $0 < x < 1$, la fonction de densité du quantile. Nous définissons l'entropie associée à la densité $f(\cdot)$, si elle existe, par :

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

Supposons que $|\mathcal{H}(X)| < \infty$. On peut réécrire $\mathcal{H}(X)$ (voir par exemple [23]) sous la forme suivante :

$$\mathcal{H}(X) = \int_{[0,1]} \log\left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{Q}(x)\right)\mathrm{d}x = \int_{[0,1]} \log\big(q(x)\big)\,\mathrm{d}x.$$

Rappelons que la fonction de quantile empirique $\mathcal{Q}_n(\cdot)$ est définie par :

$$\mathcal{Q}_n(t) := X_{k;n}, \quad (k-1)/n < t \leqslant k/n, \ \text{pour } k = 1, \ldots, n,$$

E-mail addresses: salim.bouzebda@utc.fr (S. Bouzebda), issam.elhattab@gmail.com (I. Elhattab).

1631-073X/$ – see front matter
http://dx.doi.org/10.1016/j.crma.2013.11.011

où $X_{1,n},\dots,X_{n,n}$ désignent les statistiques d'ordre associées à $X_1,\dots,X_n$. Une version lissée de $\mathcal{Q}_n(\cdot)$ peut être définie comme suit, voir [9] :

$$\widehat{\mathcal{Q}}_n(t)=\int_0^1 \mathcal{Q}_n(x)K_n(t,x)\,\mathrm{d}\mu_n(x),\quad \text{pour } t\in(0,1),$$

ce qui permet ensuite d'estimer la densité du quantile $q(\cdot)$ par :

$$\widehat{q}_n(t)=\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\mathcal{Q}}_n(t)=\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \mathcal{Q}_n(x)K_n(t,x)\,\mathrm{d}\mu_n(x),\quad \text{pour } t\in(0,1), \tag{1}$$

où les suite de noyaux $\{K_n(t,x),\ (t,x)\in[0,1]\times(0,1)\}_{n\geqslant 1}$ et de mesures $\{\mu_n(\cdot)\}_{n\geqslant 1}$ satisfont certaines conditions de régularité. Dans ce travail, nous caractérisons les propriétés asymptotiques de l'estimateur suivant, pour tout $\varepsilon\in\,]0,1/2[$ petit,

$$\mathcal{H}_{n,\varepsilon}(X)=\varepsilon\log\bigl(\widehat{q}_n(\varepsilon)\bigr)+\varepsilon\log\bigl(\widehat{q}_n(1-\varepsilon)\bigr)+\frac{1}{n}\sum_{i=\lceil\varepsilon n\rceil}^{\lfloor(1-\varepsilon)n\rfloor}\log\widehat{q}_n\bigl(F_n(X_i)\bigr),$$

où $F_n(\cdot)$ désigne la fonction de répartition empirique, i.e.,

$$F_n(t):=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{]-\infty,t]}(X_i),\quad \text{pour } t\in\mathbb{R}.$$

1. Introduction and estimation

Let X be a random variable [r.v.] with cumulative distribution function $F(x)=\mathbb{P}(X\leqslant x)$ for $x\in\mathbb{R}$ and a density function $f(\cdot)$ with respect to Lebesgue measure on $\mathbb{R}$. Then its differential (or Shannon) entropy is defined by:

$$\mathcal{H}(X)=-\int_{\mathbb{R}} f(x)\log f(x)\,\mathrm{d}x. \tag{2}$$

We assume that $\mathcal{H}(X)$ is properly defined by the integral (2), in the sense that:

$$\bigl|\mathcal{H}(X)\bigr|<\infty. \tag{3}$$

We recall from (cf. [2, p. 237], [4, p. 108]) that the finiteness of $\mathcal{H}(X)$ is guaranteed if both $\mathbb{E}\|\mathbf{X}\|^2<\infty$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$ (in which case $\mathcal{H}(X)<\infty$) and $f(\cdot)$ is bounded (in which case $\mathcal{H}(X)>-\infty$). In [2], Ash gives an example of a density function on $\mathbb{R}$ for which $\mathcal{H}(X)=\infty$ and also one for which $\mathcal{H}(X)=-\infty$. We refer to [15, Section 4] for conditions characterizing (3) in terms of $f(\cdot)$. The concept of differential entropy was introduced in Shannon's original paper [18]. Since this early epoch, the notion of entropy has been the subject of great theoretical and applied interest. We refer to [10] (see their Chapter 8) for a comprehensive overview of differential entropy and their mathematical properties. In the literature, several proposals have been made to estimate entropy. Dmitriev and Tarasenko [12] and Ahmad and Lin [1] proposed estimators of the entropy using kernel-type estimators of the density $f(\cdot)$; we refer to [5,6] for more details. Vasicek [23] proposed an entropy estimator based on spacings. Inspired by the work of [23], some authors [22,24,13,7,16] proposed modified entropy estimators, improving in some respects the properties of Vasicek's estimator. The reader will find in [3] detailed accounts of the theory as well as surveys for entropy estimators.

This Note aims to introduce a new entropy estimator and obtains its asymptotic properties. Let us first set out the basic definitions and notation which will be used throughout the sequel. For each distribution function $F(\cdot)$, we define the quantile function by:

$$\mathcal{Q}(t):=\inf\bigl\{x:\ F(x)\geqslant t\bigr\},\quad \text{for } 0<t<1.$$

Let:

$$x_F=\sup\bigl\{x:\ F(x)=0\bigr\}\quad\text{and}\quad x^F=\inf\bigl\{x:\ F(x)=1\bigr\},\quad -\infty\leqslant x_F<x^F\leqslant\infty.$$

We assume that the distribution function $F(\cdot)$ has a density $f(\cdot)$ (with respect to Lebesgue measure in $\mathbb{R}$), and that $f(x)>0$ for all $x\in(x_F,x^F)$. Let:

$$q(x)=\mathrm{d}\mathcal{Q}(x)/\mathrm{d}x=1/f\bigl(\mathcal{Q}(x)\bigr),\quad \text{for } 0<x<1,$$

be the quantile density function (qdf). The entropy $\mathcal{H}(X)$, defined by (2), can be expressed in the form of a quantile-density functional as: