L’estimation non paramétrique de l’entropie en présence de données censurées.

Bouzebda Salim, Elhattab Issam

A Strong Consistency of a Nonparametric Estimate of Entropy under Random Censorship

March 31, 2009

Salim BOUZEBDA and Issam ELHATTAB

Abstract

The purpose of this Note is to provide the rate of strong consistency for a nonparametric estimator of entropy under random censorship. We also establish an uniform-in-bandwidth consistency for this estimator.

Résumé

Loi du logarithme pour un estimateur non paramétrique de l'entropie en présence de données censurées.
Dans cette Note, nous obtenons la consistance forte pour un estimateur non paramétrique de l'entropie en présence de données censurées. Nous montrons que ce résultat demeure valable uniformément en terme de la fenêtre.

Version française abrégée

Soient (X_i, Y_i), $i = 1, 2, \ldots,$ des copies indépendantes et identiquement distribuées du vecteur aléatoire (X, Y). Les $\{X_i, i = 1, \ldots, n\}$ s'interprètent comme étant *des durées de survie* et les $\{Y_i, i = 1, \ldots, n\}$ désignent *des temps de censure*. Nous allons travailler sous la condition d'indépendance entre X et Y. Le modèle que nous adoptons ici est basé sur les observations (Z_i, δ_i) pour $i = 1, \ldots, n$, où $Z_i = \min(X_i, Y_i)$ et $\delta_i = \mathbb{1}_{\{X_i \leq Y_i\}}$ avec $\mathbb{1}_S$ désigne la fonction indicatrice de l'ensemble S. Pour tout $x \in \mathbb{R}$, soit $F(x) = \mathbb{P}(X \leq x)$ et $G(x) = \mathbb{P}(Y \leq x)$. Nous supposons que $F(\cdot)$ admette une densité $f(\cdot)$ par rapport à la mesure de Lebesgue. L'entropie $H(f)$ associée à $f(\cdot)$ est définie par $H(f) = -\mathbb{E}(\log f(X))$.
Soit $L(x) = \mathbb{P}(V \leq x)$ la fonction de répartition d'une variable aléatoire V. On note $T_L = \sup\{x : L(x) < 1\}$ la borne supérieure du support de $L(\cdot)$. Nous

supposons ici que les bornes supérieures T_F et T_G des supports de $F(\cdot)$ et de $G(\cdot)$, respectivement, sont telles que $0 < T_F \leq T_G$ et

$$H(f) = -\int_0^{T_F} f(x) \log \big(f(x)\big)\, dx. \tag{1}$$

Nous supposons que $|H(f)| < \infty$, donc pour tout $\varepsilon > 0$, il existe des constantes a et b, telles que $0 < a < b < T_F$ et

$$\int_a^b f(x) \log \big(f(x)\big) dx \geq \int_0^{T_F} f(x) \log \big(f(x)\big) dx - \varepsilon.$$

Nous supposons que $F(\cdot)$ et $G(\cdot)$ satisfont les conditions (F1) et (F2) ci-dessous. Nous définissons l'estimateur du maximum de vraisemblance généralisée non paramétrique de $F(\cdot)$ par

$$F_n(x) = 1 - \prod_{i:Z_{i,n} \leq x, 1 \leq i \leq n} \left\{ 1 - \frac{\delta_{i,n}}{n-i+1} \right\},$$

où $Z_{1,n}, \leq \cdots \leq Z_{n,n}$ est la statistique d'ordre de $Z_1, \ldots, Z_n$, et pour tout $i = 1, \ldots, n, \delta_{i,n}$ est le δ_j correspondant à $Z_{i,n} = Z_j, 1 \leq j \leq n$ (on utilise la convention $\prod_\emptyset = 1$). Pour tout $x \in \mathbb{R}$, l'estimateur à noyau $f_{n,h_n}(x)$ de $f(x)$ est défini par

$$f_{n,h_n}(x) = \int_{\mathbb{R}} \frac{1}{h_n} K\left(\frac{t-x}{h_n}\right) dF_n(t), \tag{2}$$

où $K(\cdot)$ est une fonction measurable vérifiant les conditions (K1)-(K4) (ci-dessous) et $\{h_n\}_{n\geq 1}$ est une suite de constantes positives vérifiant (H1)-(H3) données dans le paragraphe suivant. Dans un deuxième temps, soient $\{a_n\}_{n\geq 1}$ et $\{b_n\}_{n\geq 1}$ deux suites positives telles que $a_n \downarrow a$ et $b_n \uparrow b$. Nous définissons l'estimateur de $H(f)$ par

$$H_{n,h_n,\beta,\gamma} = -\int_{A_{n,\beta}^{\gamma}} f_{n,h_n}(x) \log f_{n,h_n}(x) dx, \tag{3}$$

où $A_{n,\beta}^{\gamma} := \{x \in [a_n, b_n] : f_{n,h_n}(x) \geq \gamma (\log_+ n)^{-\beta}\}$ avec $\beta \in [0, 1/4)$ et $\gamma \in (0, 1)$ sont des constantes spécifiques. Soit pour tout $x \in \mathbb{R}$ et $n \geq 1$,

$$\widehat{\mathbb{E}} f_{n,h_n}(x) = \int_{\mathbb{R}} \frac{1}{h_n} K\left(\frac{t-x}{h_n}\right) dF(t). \tag{4}$$

Nous considérons pour tout $n \geq 1$, la notation suivante,

$$\widehat{\mathbb{E}} H_{n,h_n\beta,\gamma} = -\int_{A_{n,\beta}^{\gamma}} \widehat{\mathbb{E}} f_{n,h_n}(x) \log \big(\widehat{\mathbb{E}} f_{n,h_n}(x)\big) dx. \tag{5}$$

Dans cette Note, nous obtenons les résultats suivants :

1. Sous les conditions du théorème 2.1 ci-dessous, on a, presque sûrement, pour une constante $C > 0$,

$$\limsup_{n\to\infty} \left\{ \frac{nh}{2\{\log_+ n\}^{4\beta}\{\log(1/h)\}} \right\}^{1/2} \mid H_{n,h_n,\beta,\gamma} - \widehat{\mathbb{E}} H_{n,h_n,\beta,\gamma} \mid \leq C. \quad (6)$$

2. Sous les conditions du théorème 2.2 ci-dessous, on a, presque sûrement, pour une constante $C > 0$,

$$\limsup_{n\to\infty} \sup_{h'_n \leq h \leq h''_n} \left\{ \frac{nh}{2\{\log_+ n\}^{4\beta}\{\log(1/h)\}} \right\}^{1/2} \mid H_{n,h_n,\beta,\gamma} - \widehat{\mathbb{E}} H_{n,h_n,\beta,\gamma} \mid \leq C. \quad (7)$$

3. Sous les conditions du corollaire 2.3 ci-dessous, pour tout $\beta \in (0, 1/4)$, et toutes suites $0 < h'_n < h''_n \leq 1$ où $h''_n \downarrow 0$ et $nh'_n/(\log n)^{1+4\beta} \to \infty$, on a

$$\sup_{h'_n \leq h \leq h''_n} |H_{n,h_n,\beta,\gamma} - H(f)| \to 0 \qquad p.s. \quad (8)$$

1 Introduction and estimation

The notion of entropy is of interest from the theoretical viewpoint as well as with respect to applications. We refer to Beirlant *et al.* (1997) for an exposition of the subject. In this work, we are concerned with the estimation of entropy based upon censored observations. Let the *lifetimes* $\{X_i, i = 1, \ldots, n\}$ and the *censoring times* $\{Y_i, i = 1, \ldots, n\}$ be defined as independent sequences of independent and identically distributed random replicæ, respectively, of the mutually independent nonnegative random variables X and Y. In the model that we consider, we observe (Z_i, δ_i) for $i = 1, \ldots, n$, where $Z_i = \min(X_i, Y_i)$ and $\delta_i = \mathbb{1}_{\{X_i \leq Y_i\}}$ with $\mathbb{1}_S$ denoting the indicator function of S. Set $F(x) = \mathbb{P}(X \leq x)$ and $G(x) = \mathbb{P}(Y \leq x)$, for $x \in \mathbb{R}^+$. Our goal is to estimate, out of the data set (Z_i, δ_i) for $i = 1, \ldots, n$, the entropy $H(f)$ of the *lifetime density function* $f(x) = (d/dx)F(x)$, defined by

$$H(f) = -\mathbb{E}\big\{ \log\big(f(X)\big)\big\}, \quad (9)$$

whenever (9) is meaningful, namely subject to the condition that

$$|H(f)| < \infty. \quad (10)$$

The following assumptions and notations will be needed. Set $L(x) = \mathbb{P}(V \leq x)$ for the distribution of some given random variable V. We denote by $T_L = \sup\{x : L(x) < 1\}$ the upper endpoint of the distribution of V. Throughout the sequel, we shall assume that the upper endpoints T_F and T_G of the distribution of X and Y are such that $\min(T_F, T_G) = T_F := \Theta > 0$ and that $H(f)$ is properly defined by

$$H(f) = -\int_0^\Theta f(x) \log\big(f(x)\big)\, dx. \quad (11)$$

We refer to Földes *et al.* (1981) for a discussion on the choice of $T_F \leq T_G$.
In view of (10-11), for each $\varepsilon > 0$, there exist constants a and b, such that $0 < a < b < \Theta$ and

$$\int_a^b f(x) \log\big(f(x)\big) dx \geq \int_0^\Theta f(x) \log\big(f(x)\big) dx - \varepsilon. \tag{12}$$

Below, we fix an $\varepsilon > 0$, and select constants a and b, such that (12) holds. We assume that F and G fulfill the following conditions.

(F1) $F(0) = G(0) = 0$;

(F2) (i) F and G are continuous on $[a, b]$;
(ii) $f(x) = (d/dx)F(x)$ is defined, continuous, bounded and strictly positive on $[a, b]$.

The main purpose of the present paper is to provide the strong consistency of an estimator of the entropy $H(f)$, based on a kernel-type method, in the above defined setup of randomly right-censored data. To define our entropy estimator we introduce, in a first step, a kernel density estimator. Towards this aim, we consider the nonparametric maximum likelihood estimator of $F(\cdot)$ based upon the data set. This is given by the product-limit (PL) estimator $F_n(\cdot)$, introduced in Kaplan and Meier (1958), and defined by

$$F_n(x) = 1 - \prod_{i: Z_{i,n} \leq x, 1 \leq i \leq n} \left\{ 1 - \frac{\delta_{i,n}}{n - i + 1} \right\},$$

where $Z_{1,n}, \leq \cdots \leq Z_{n,n}$ denote the ordered statistics of $Z_1, \ldots, Z_n$, and, for each $i = 1, \ldots, n, \delta_{i,n}$ is the δ_j corresponding to $Z_{i,n} = Z_j, 1 \leq j \leq n$ (we use the convention that $\prod_\emptyset = 1$). For all $x \in \mathbb{R}$, the kernel estimator $f_{n,h_n}(x)$ of $f(x)$ (see, e.g., Watson and Leadbetter (1964a), Watson and Leadbetter (1964b), Földes *et al.* (1979); Földes and Rejtő (1981), Tanner and Wong (1983), Winter (1987), Diehl and Stute (1988) and Deheuvels and Einmahl (1996, 2000)) is given by

$$f_{n,h_n}(x) = \int_{\mathbb{R}} \frac{1}{h_n} K\left(\frac{t - x}{h_n}\right) dF_n(t), \tag{13}$$

where $K(\cdot)$ denote a measurable function fulfilling the assumptions:

(K1) $K(\cdot)$ is of bounded variation on $\mathbb{R}$;

(K2) For some $0 < M < \infty$, $K(u) = 0$ for all $|u| \geq \frac{1}{2}M$;

(K3) $\int_{\mathbb{R}} K(s) ds = 1$;

(K4) $K(\cdot)$ is right continuous on $\mathbb{R}$, i.e., $K(t) = \lim_{\varepsilon \downarrow 0} K(t + \varepsilon)$;

and $\{h_n\}_{n \geq 1}$ is a sequence of positive constants satisfies the condition:

(H1) (i) $h_n \downarrow 0$; (ii) $nh_n \uparrow \infty$;

where " $\uparrow$" (resp. " $\downarrow$") stands for non-decreasing (resp. non-increasing).

In a second step, let $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ be two sequences of constants such that $a_n \downarrow a$ and $b_n \uparrow b$. We then estimate $H(f)$ by setting

$$H_{n,h_n,\beta,\gamma} = -\int_{A^{\gamma}_{n,\beta}} f_{n,h_n}(x) \log f_{n,h_n}(x) dx, \tag{14}$$

where $A^{\gamma}_{n,\beta} := \{x \in [a_n, b_n] : f_{n,h_n}(x) \geq \gamma(\log n)^{-\beta}\}$ with $\beta \in [0, 1/4)$ and $\gamma \in (0, 1)$ are specified constants.

Remark 1 In the literature, Carbonez et al. *(1991) proposed to estimate $H(f)$, in a censored framework, by (14), where $f_n(\cdot)$ is a histogram estimate of the density $f(\cdot)$, $A^{\gamma_n}_{n,1} := \{x \in \mathbb{R}^+ : f_{n,h_n}(x) \geq \gamma_n\}$ and $\{\gamma_n\}_{n\geq 1}$ is a sequence of positive constants with $\gamma_n \downarrow 0$. They proved the strong consistency of $H_{n,1,\gamma_n}(f)$ under the condition of fair censoring, assuming only the condition (10) on $f(\cdot)$, and under additional conditions on the sequences h_n and γ_n.*

The following assumptions on the sequence $\{h_n\}_{n\geq 1}$ will be needed in addition to (H1). Set $\log_2 u = \log_+(\log_+ u)$ and $\log_+ u = \log(u \vee e)$, for $u \in \mathbb{R}$.

(H2) $nh_n/(\log_+ n)^{1+4\beta} \to \infty$;

(H3) $\log(1/h_n)/\log_2 n \to \infty$.

Set, for all $x \in \mathbb{R}$ and $n \geq 1$,

$$\widehat{\mathbb{E}} f_{n,h_n}(x) = \int_{\mathbb{R}} \frac{1}{h_n} K\left(\frac{t-x}{h_n}\right) dF(t). \tag{15}$$

In the uncensored case $\widehat{\mathbb{E}} f_{n,h_n}(x) = \mathbb{E} f_{n,h_n}(x)$, where $\mathbb{E}$ denotes the usual expectation. To prove the strong consistency of $H_{n,h_n,\beta,\gamma}$, we shall consider another, but more appropriate, centering factor than the expectation $\mathbb{E}H_{n,h_n,\beta,\gamma}$, which is delicate to handle. This is given by

$$\widehat{\mathbb{E}} H_{n,h_n,\beta,\gamma} = -\int_{A^{\gamma}_{n,\beta}} \widehat{\mathbb{E}} f_{n,h_n}(x) \log\left(\widehat{\mathbb{E}} f_{n,h_n}(x)\right) dx. \tag{16}$$

The remainder of this Note is organized as follows. In Section 2, we state our main results concerning the limiting behavior of $H_{n,h_n,\beta,\gamma}$. The sketch of the proofs of our main results are postponed until Section 3.

2 Main results

The following theorem establishes the almost sure consistency of $H_{n,h_n,\beta,\gamma}$.

Theorem 2.1 *Let the conditions* (F1)-(F2), (K1)-(K3) *and* (H1)-(H3) *be satisfied. Then we have, with probability one, for some constant $C > 0$,*

$$\limsup_{n\to\infty}\left\{\frac{nh_n}{2\{\log_+ n\}^{4\beta}\{\log(1/h_n)\}}\right\}^{1/2} \mid H_{n,h_n,\beta,\gamma} - \widehat{\mathbb{E}}H_{n,h_n,\beta,\gamma} \mid\leq C. \quad (17)$$

It is well known that the limiting behavior of kernel-type estimators depends crucially upon the choice of the bandwidth h. For the entropy estimate (14), we obtain the uniform in bandwidth consistency of $H_{n,h_n,\beta,\gamma}$ to $H(f)$.

Theorem 2.2 *Let the conditions* (F1)-(F2) *and* (K1)-(K4) *be satisfied. Let $\{h_n'\}_{n\geq 1}$ and $\{h_n''\}_{n\geq 1}$ fulfilling conditions* (H1)-(H3)*, with $0 < h_n' < h_n'' < 1$. Then for each $0 \leq \beta < 1$, we have with probability one,*

$$\limsup_{n\to\infty}\sup_{h_n'\leq h\leq h_n''}\left\{\frac{nh}{2\{\log_+ n\}^{4\beta}\{\log(1/h)\}}\right\}^{1/2} \mid H_{n,h_n,\beta,\gamma} - \widehat{\mathbb{E}}H_{n,h_n,\beta,\gamma} \mid\leq C. \quad (18)$$

where C is as in Theorem 2.1.

An application of Theorem 2.2 shows that, with probability 1,

$$\sup_{h_n'\leq h\leq h_n''}|H_{n,h_n,\beta,\gamma} - \widehat{\mathbb{E}}H_{n,h_n,\beta,\gamma}| = O\left(\sqrt{\frac{\{\log n\}^{4\beta}(\log(1/h_n')\vee\log\log n)}{nh_n'}}\right).$$

This, in turn, implies that

$$\lim_{n\to\infty}\sup_{h_n'\leq h\leq h_n''}|H_{n,h_n,\beta,\gamma} - \widehat{\mathbb{E}}H_{n,h_n,\beta,\gamma}| = 0 \quad a.s. \quad (19)$$

Thus we have the following corollary of Theorem 2.2.

Corollary 2.3 *Let $K(\cdot)$ satisfy* (K1)-(K4)*, and let $f(\cdot)$ be a uniformly Lipschitz continuous, and strictly positive density, on $[a,b]$, fulfilling* (F1)*. Let $\{h_n'\}_{n\geq 1}$ and $\{h_n''\}_{n\geq 1}$ fulfill* (H1)-(H3)*, with $0 < h_n' < h_n'' < 1$. Then for any $\beta \in (0, 1/4)$, we have*

$$\sup_{h_n'\leq h\leq h_n''}|H_{n,h_n,\beta,\gamma} - H(f)| \to 0 \qquad a.s. \quad (20)$$

Remark 2 Giné and Mason (2008) establish uniform in bandwidth consistency for an entropy estimator, in the uncensored case. Their estimate is somewhat different from $H_{n,h_n,\beta,\gamma}$. It is defined by

$$\widehat{H}_{n,h_n} = -\frac{1}{n}\sum_{i=1}^{n}\log\{f_{n,h_n,-i}(X_i)\},$$

where

$$f_{n,h_n,-i}(X_i) = \frac{1}{(n-1)h_n}\sum_{1\leq j\neq i\leq n}K\left(\frac{X_i - X_j}{h_n}\right).$$

Their results hold subject to the condition that the density $f(\cdot)$ is bounded away from zero on its support.

3 Sketch proof

The proofs of Theorems 2.1 and 2.2 are based essentially on the following decomposition

$$\begin{aligned} H_{n,\beta,\gamma}(f) - \widehat{\mathbb{E}}H_{n,\beta,\gamma}(f) &= -\int_{A^{\gamma}_{n,\beta}} \Big\{ f_{n,h_n}(x) - \widehat{\mathbb{E}}f_{n,h_n}(x) \Big\} \log f_{n,h_n}(x)dx \\ & \quad - \int_{A^{\gamma}_{n,\beta}} \widehat{\mathbb{E}}f_{n,h_n}(x) \Big\{ \log f_{n,h_n}(x) - \log \widehat{\mathbb{E}}f_{n,h_n}(x) \Big\} dx =: \Delta_{1,n} + \Delta_{2,n}. \end{aligned}$$

and, respectively, the following Facts.
Fact 1. [due to Deheuvels and Einmahl (2000)] Let the conditions (F1)-(F2), (K1)-(K2)-(K3) and (H1)-(H2) be satisfied. Then we have

$$\limsup_{n\longrightarrow\infty} \sup_{a_n \le x \le b_n} \frac{\sqrt{nh} \pm (f_{n,h}(x) - \widehat{\mathbb{E}}f_{n,h}(x))}{\sqrt{2\log(1/h)}} = \sup_{a<x<b} \left\{ \frac{f(x)}{1-G(x)} \right\}^{1/2} \left\{ \int_{-\infty}^{\infty} K^2(u)du \right\}^{1/2} \quad a.s.$$

Fact 2. [due to Viallon (2006)] Under the conditions of Theorem 2.2, we have

$$\limsup_{n\longrightarrow\infty} \sup_{h'_n \le h \le h''_n} \sup_{a_n \le x \le b_n} \frac{\sqrt{nh} \pm (f_{n,h}(x) - \widehat{\mathbb{E}}f_{n,h}(x))}{\sqrt{2\log(1/h)}} = \sup_{a<x<b} \left\{ \frac{f(x)}{1-G(x)} \right\}^{1/2} \left\{ \int_{-\infty}^{\infty} K^2(u)du \right\}^{1/2} \quad a.s.$$

Proof of corollary 2.3. First, we study the bias term. Under the conditions of corollary 2.3, we have

$$\lim_{n\rightarrow\infty} \sup_{h'_n \le h \le h''_n} |\widehat{\mathbb{E}}H_{n,\beta,\gamma}(f) - H(f)| = O\left(h''_n\right). \tag{21}$$

This in combination with Theorem 2.2, completes our proof. □

Acknowledgements

The authors are grateful to the two referees and Professor Paul Deheuvels for their useful suggestions and constructive criticisms on earlier drafts of this work.

References

Beirlant, J., Dudewicz, E. J., Györfi, L., and van der Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, **6**(1), 17–39.

Carbonez, A., Györfi, L., and van der Meulen, E. C. (1991). Nonparametric entropy estimation based on randomly censored data. *Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform.*, **20**(6), 441–451.

Deheuvels, P. and Einmahl, J. H. J. (1996). On the strong limiting behavior of local functionals of empirical processes based upon censored data. *Ann. Probab.*, **24**(1), 504–525.

Deheuvels, P. and Einmahl, J. H. J. (2000). Functional limit laws for the increments of Kaplan-Meier product-limit processes and applications. *Ann. Probab.*, **28**(3), 1301–1335.

Diehl, S. and Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Multivariate Anal.*, **25**(2), 299–310.

Földes, A. and Rejtő, L. (1981). A LIL type result for the product limit estimator. *Z. Wahrsch. Verw. Gebiete*, **56**(1), 75–86.

Földes, A., Rejtő, L., and Winter, B. B. (1979). Strong consistency properties of nonparametric estimators for randomly censored data. The product-limit estimator, estimation of density and failure rate. In *Transactions of the Eighth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Prague, 1978), Vol. C*, pages 105–121. Reidel, Dordrecht.

Földes, A., Rejtő, L., and Winter, B. B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data. II. Estimation of density and failure rate. *Period. Math. Hungar.*, **12**(1), 15–29.

Giné, E. and Mason, D. M. (2008). Uniform in bandwidth estimation of integral functionals of the density function. *To appear in Scand. J. Statist.*

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457–481.

Tanner, M. A. and Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.*, **11**(3), 989–993.

Viallon, V. (2006). *Processus empiriques, estimation non paramétrique et données censurées.* Ph.D. thesis, Université Paris 6.

Watson, G. S. and Leadbetter, M. R. (1964a). Hazard analysis. I. *Biometrika*, **51**, 175–184.

Watson, G. S. and Leadbetter, M. R. (1964b). Hazard analysis. II. *Sankhyā Ser. A*, **26**, 101–116.

Winter, B. B. (1987). Nonparametric estimation with censored data from a distribution with nonincreasing density. *Comm. Statist. Theory Methods*, **16**(1), 93–120.