C. R. Acad. Sci. Paris, Ser. I 335 (2002) 375–380

Statistique/*Statistics*
(Probabilités/*Probability Theory*)

Almost sure convergence of a tail index estimator in the presence of censoring

Emmanuel Delafosse, Armelle Guillou

Université Paris VI, L.S.T.A., boîte 158, 4, place Jussieu, 75252 Paris cedex 05, France

Received 8 April 2002; accepted after revision 28 June 2002

Note presented by Paul Deheuvels.

Abstract In Beirlant and Guillou [1] an exponential regression model was introduced on the basis of scaled log-spacing between subsequent extreme order statistics from a Pareto-type distribution in the presence of censoring. From this representation, they derived an estimator for the Pareto index. In this note, we revisit this adaptation of the popular Hill [5] estimator for heavy-tailed distributions, generalizing the almost sure convergence of this estimator under very general conditions on N_r, the number of non-censored observations. ***To cite this article: E. Delafosse, A. Guillou, C. R. Acad. Sci. Paris, Ser. I 335 (2002) 375–380.***

Convergence presque sûre d'un index de queue en présence de censure

Résumé Dans Beirlant et Guillou [1] un modèle de régression exponentiel basé sur l'écart du logarithme de statistiques d'ordres consécutives d'un échantillon issu d'une loi de type Pareto a été introduit en présence de censure. De cette représentation, ils obtiennent un estimateur de l'index de Pareto. Dans cette note, nous revisitons cette adaptation de l'estimateur de Hill [5] en établissant en particulier sa convergence presque sûre sous des conditions très générales sur le nombre N_r de données non censurées. ***Pour citer cet article : E. Delafosse, A. Guillou, C. R. Acad. Sci. Paris, Ser. I 335 (2002) 375–380.***

Version française abrégée

Nous considérons un échantillon $X_1, \ldots, X_n$ de variables aléatoires positives de loi F. Nous supposons F de type Pareto et nous nous intéressons à l'estimation de l'index de queue γ. Plus précisément, nous étudions, en présence de censure, la convergence presque sûre d'un estimateur de γ défini par Beirlant et Guillou [1] de la façon suivante :

$$\hat{\gamma}_{k_n,n} = \frac{1}{k_n - n + N_r} \left\{ \sum_{j=n-N_r+1}^{k_n} \log \frac{X_{n-j+1,n}}{X_{n-k_n,n}} + (n - N_r) \log \frac{X_{N_r,n}}{X_{n-k_n,n}} \right\},$$

E-mail addresses: delafos@ccr.jussieu.fr (E. Delafosse); guillou@ccr.jussieu.fr (A. Guillou).

S1631-073X(02)02486-X/FLA

où $X_{j,n}$ désigne la $j^{\text{ème}}$ statistique d'ordre de l'échantillon initial et $k_n = n - N_r + 1, \ldots, n-1$. Cet estimateur est fonction des N_r données non censurées et coïncide avec l'estimateur de Hill [5] dans le cas d'absence de censure. Nous établissons les conditions sur N_r et caractérisons les suites k_n telles que notre estimateur soit consistent presque sûrement. La condition principale sur la censure est $(n - N_r)/k_n \to C \in [0, 1)$ p.s. Elle est par conséquent moins restrictive que celle de Beirlant et Guillou [1] correspondant au cas particulier où $C = 0$. De plus les conditions sur k_n, similaires à celles imposées par Deheuvels et al. [4], sont également moins fortes que celles faites dans Beirlant et Guillou [1]. Les preuves s'appuient essentiellement sur des théorèmes généraux de processus empiriques.

1. Introduction

Let $X_1, \ldots, X_n, \ldots$ be a sequence of positive independent and identically (i.i.d.) distributed random variables from a distribution function F. We suppose that F is of Pareto-type, that is, that there exists a positive constant γ for which

$$1 - F(x) = x^{-1/\gamma} \ell_F(x), \tag{1}$$

where $\ell_F(x)$ is a so-called slowly varying function at infinity satisfying

$$\frac{\ell_F(\lambda x)}{\ell_F(x)} \to 1, \quad \text{as } x \to \infty \text{ for all } \lambda > 0.$$

The present model is well-known to be equivalent to the following model

$$U(x) = x^{\gamma} \ell_U(x),$$

where $U(x) = \inf\{y : F(y) \geqslant 1 - 1/x\},\ x > 1$, and with $\ell_U(x)$ again a slowly varying function.

Unfortunately no estimation method for γ exists which provides a prescribed rate of convergence for members of this model in its full generality. Therefore, in general, a condition on the slowly varying functions ℓ_U must be imposed and the classical one is the following:

ASSUMPTION (R_{ℓ_U}). – There exists a real constant $\rho \leqslant 0$ and a positive rate function b satisfying $b(x) \to 0$ as $x \to \infty$, such that for all $\lambda \geqslant 1$, as $x \to \infty$,

$$\log \frac{\ell_U(\lambda x)}{\ell_U(x)} \sim b(x) k_\rho(\lambda)$$

with $k_\rho(\lambda) = \int_1^\lambda v^{\rho-1}\,\mathrm{d}v$.

Since (1) involves only the upper tail of the distribution F, it is reasonable to construct estimators of γ based on the top extreme values of the sample $X_1, \ldots, X_n$ of size $n \geqslant 1$. The most commonly used estimator of this kind was proposed by Hill [5] and is given as follows:

$$H_{k,n} = \frac{1}{k} \sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n},$$

where $X_{j,n},\ j = 1, \ldots, n$, denotes the order statistics based on the first n observations. In this note, we examine the a.s. behaviour of a tail index estimator in the context of censored data. Let us illustrate this situation with a concrete example from the insurance field. Suppose that contract stipulate an upper limit to the amount to be paid out. If the data are expressed in claim ratio percentage (i.e., claim size over sum insured) then some data will be censored at 100%, while the actuarial statistician will be interested in the real claim ratio. One then observes a random number N_r of claims to their full extent, while the remaining $n - N_r$ observations are equal to the upper limit. In such cases, the estimation of the Pareto index for the real loss process should proceed differently. Since we have observed the N_r claims $X_{1,n} \leqslant \cdots \leqslant X_{N_r,n}$, all smaller than the maximum value M, the Pareto quantile plot will then end with a set of points situated at the same height $\log M$. When estimating γ (of the real loss process) one should of course put less weight on these data values. In such settings, Beirlant and Guillou [1] proposed the following adaptation of the popular Hill [5] estimator

$$\hat{\gamma}_{k,n} = \frac{1}{k - n + N_r}\left\{ \sum_{j=n-N_r+1}^{k} \log \frac{X_{n-j+1,n}}{X_{n-k,n}} + (n - N_r)\log \frac{X_{N_r,n}}{X_{n-k,n}} \right\},$$

for $k = n - N_r + 1, \ldots, n - 1$. In the absence of censoring, this estimator reduces to the Hill estimator. They derived in particular the basic asymptotic properties of $\hat{\gamma}_{k,n}$ under the assumption that $(n - N_r)/k \to 0$ in probability or a.s. and proposed an optimal choice of k based on a MSE criterion. In this note, we revisit this adaptation of the Hill estimator, generalizing the almost sure convergence of $\hat{\gamma}_{k,n}$ under the following general conditions on N_r : $(n - N_r)/k \to C \in [0, 1)$ a.s. Note that the convergence to 1 has no practical interest since it means that asymptotically the number of observations retained (k) is equal to the number of censored data. Moreover the conditions imposed on k are less stringent than those in Beirlant and Guillou [1], but similar to those imposed by Deheuvels et al. [4]. The case of the absence of censoring ($N_r = n$) has been excluded in our result given below since we have imposed the condition $(n - N_r)/(\log\log n) \to \infty$ a.s. as $n \to \infty$, but this situation has already been studied in detail in Deheuvels et al. [4].

THEOREM. – *Under Assumption* (R_{ℓ_U}), *if* (1) *holds, whenever* $\frac{k}{n} \to 0$, $\frac{k}{\log\log n} \to \infty$, $\frac{n-N_r}{k} \to C \in [0, 1)$ *a.s. and* $\frac{n-N_r}{\log\log n} \to \infty$ *a.s. as* $n \to \infty$, *we have*

$$\lim_{n\to\infty} \hat{\gamma}_{k,n} = \gamma \quad \textit{a.s.}$$

The proof of this theorem requires several lemmas which will be postponed in the next section. Note also that the weak convergence and the asymptotic normality of $\hat{\gamma}_{k,n}$ under this more general condition on N_r is only a direct adaptation of the same result in the case $C = 0$ and it can be found in Matthys et al. [6]. Under this general condition, Matthys et al. [6] proposed also two estimators of extreme quantiles for censored observations and they established some asymptotic results. Then, they compared their small sample properties in a simulation study and they discussed the optimal choice of k when using the extreme quantile which is the analogue of Weissman's [8] estimator in the censoring case.

2. Proof of the almost sure convergence of $\hat{\gamma}_{k,n}$

Let $U_1, U_2, \ldots$ be a sequence of independent and uniformly distributed random variables on $(0, 1)$ with $U_{j,n}$ the order statistics based on the first n observations. Also G_n denotes the right continuous empirical distribution function based on $U_1, \ldots, U_n$ and $Q(s) = \log(\inf\{y : F(y) \geqslant s\})$.

It is clear that our estimator can be rewritten as

$$\hat{\gamma}_{k,n} = \frac{n}{k - n + N_r} \int_{U_{n-k,n}}^{U_{N_r,n}} \bigl(1 - G_n(s)\bigr)\,\mathrm{d}Q(s).$$

Set

$$\mu_{n,N_r} = \frac{n}{k-n+N_r} \int_{1-k/n}^{N_r/n} (1-u)\,\mathrm{d}Q(u).$$

We can establish that μ_{n,N_r} converges a.s. to γ under the assumption that $(n-N_r)/k \to C \in [0,1)$ a.s. using Lemma 3.2 in Csörgő and Mason [3]. Therefore it is sufficient to study the difference $\hat{\gamma}_{k,n} - \mu_{n,N_r}$ to conclude. Now, we consider a decomposition of this last quantity as follows.

$$\hat{\gamma}_{k,n} - \mu_{n,N_r} =: T_{1,n} + T_{2,n} + T_{3,n},$$

where

$$\begin{cases} T_{1,n} := \frac{n}{k-n+N_r} \int_{1-k/n}^{N_r/n} \big(u - G_n(u)\big)\,\mathrm{d}Q(u), \\ T_{2,n} := -\frac{n}{k-n+N_r} \int_{1-k/n}^{U_{n-k,n}} \big(1 - G_n(u)\big)\,\mathrm{d}Q(u), \\ T_{3,n} := \frac{n}{k-n+N_r} \int_{N_r/n}^{U_{N_r,n}} \big(1 - G_n(u)\big)\,\mathrm{d}Q(u). \end{cases}$$

We split now $T_{1,n}$ into two parts:

$$\begin{cases} T_{1,n}^{(1)} := \frac{n}{k-n+N_r} \int_{1-k/n}^{1-(\log\log n)/n} \big(u - G_n(u)\big)\,\mathrm{d}Q(u), \\ T_{1,n}^{(2)} := \frac{n}{k-n+N_r} \int_{1-(\log\log n)/n}^{N_r/n} \big(u - G_n(u)\big)\,\mathrm{d}Q(u) \end{cases}$$

and in the two following lemmas, we study each term separately.

LEMMA 1. – *Under Assumption* (R_{ℓ_U}), *if* (1) *holds, whenever* $\frac{k}{n} \to 0$, $\frac{k}{\log\log n} \to \infty$ *and* $\frac{n-N_r}{k} \to C \in [0,1)$ *a.s. as* $n \to \infty$, *we have*

$$T_{1,n}^{(1)} \to 0 \quad \text{a.s. as } n \to \infty.$$

Proof. – It is clear that

$$\big|T_{1,n}^{(1)}\big| \leqslant \left(\frac{\log\log n}{k}\right)^{1/2} \left(\frac{k}{n}\right)^{-1/2} \int_{1-k/n}^{1} (1-u)^{1/2}\,\mathrm{d}Q(u)$$
$$\times \sup_{0 \leqslant s \leqslant 1-(\log\log n)/n} \frac{\sqrt{n}|G_n(s)-s|}{\sqrt{\log\log n}\sqrt{1-s}}\, \frac{k}{k-n+N_r}.$$

Lemma 1 follows now from Lemma 3.2 in Csörgő and Mason [3] and Theorem 3.2 in Csáki [2]. □

LEMMA 2. – *Under Assumption* (R_{ℓ_U}), *if* (1) *holds, whenever* $\frac{k}{n} \to 0$, $\frac{k}{\log\log n} \to \infty$ *and* $\frac{n-N_r}{k} \to C \in [0,1)$ *a.s. as* $n \to \infty$, *we have*

$$T_{1,n}^{(2)} \to 0 \quad \text{a.s. as } n \to \infty.$$

Proof. – Note that

$$\big|T_{1,n}^{(2)}\big| \leqslant \frac{n}{k-n+N_r} \left(\int_{1-(\log\log n)/n}^{1} (1-u)\,\mathrm{d}Q(u) + \int_{1-(\log\log n)/n}^{1} \big(1 - G_n(u)\big)\,\mathrm{d}Q(u) \right).$$

Using again Lemma 3.2 in Csörgő and Mason [3], combining with the a.s. convergence to zero of the quantity $nk^{-1}\int_{1-(\log\log n)/n}^{1} (1-G_n(u))\,\mathrm{d}Q(u)$ established in Deheuvels et al. [4], Lemma 2 follows. □

Now, we study the second term in the decomposition.

LEMMA 3. – *Under Assumption* (R_{ℓ_U}), *if* (1) *holds, whenever* $\frac{k}{n} \to 0$, $\frac{k}{\log\log n} \to \infty$ *and* $\frac{n-N_r}{k} \to C \in [0,1)$ *a.s. as* $n \to \infty$, *we have*

$$T_{2,n} \to 0 \quad \textit{a.s. as } n \to \infty.$$

Proof. – We can rewrite $T_{2,n}$ as follows:

$$T_{2,n} = -\frac{k}{k-n+N_r} n k^{-1} \int_{1-k/n}^{U_{n-k,n}} \big(1 - G_n(u)\big)\,\mathrm{d}Q(u).$$

Using the convergence of $(n - N_r)/k$ to $C \in [0,1)$ a.s. and Lemma 7 in Deheuvels et al. [4], the proof of Lemma 3 is achieved. □

Under an additional assumption, which is $(n - N_r)/(\log\log n) \to \infty$ a.s. as $n \to \infty$, we establish in the following lemma the a.s. convergence to zero of the last term in the decomposition.

LEMMA 4. – *Under Assumption* (R_{ℓ_U}), *if* (1) *holds, whenever* $\frac{k}{n} \to 0$, $\frac{k}{\log\log n} \to \infty$, $\frac{n-N_r}{k} \to C \in [0,1)$ *a.s. and* $\frac{n-N_r}{\log\log n} \to \infty$ *a.s. as* $n \to \infty$, *we have*

$$T_{3,n} \to 0 \quad \textit{a.s. as } n \to \infty.$$

Proof. – Using the fact that $G_n(\cdot)$ is an increasing function, we have

$$T_{3,n} \leqslant \frac{n-N_r}{k-n+N_r} \left| Q(U_{N_r,n}) - Q\left(\frac{N_r}{n}\right)\right| \left(\frac{n}{n-N_r} \left| \frac{N_r}{n} - G_n\left(\frac{N_r}{n}\right)\right| + 1\right).$$

The assumptions of Lemma 4 and Theorem 3.2 in Csáki [2] give the a.s. convergence to zero of the quantity

$$\frac{n}{n-N_r} \left| \frac{N_r}{n} - G_n\left(\frac{N_r}{n}\right)\right|.$$

Therefore, it is sufficient to study the a.s. behaviour of $|Q(U_{N_r,n}) - Q(N_r/n)|$. Using the Karamata representation

$$Q(1-s) = -\gamma \log s + \log a(s) + \int_s^1 \frac{f(u)}{u}\,\mathrm{d}u,$$

where $a(s) \to a_0 \in (0,\infty)$ and $f(s) \to 0$ as $s \to 0$, combining with the fact that $1 - U_{N_r,n} \to 0$ a.s., we only have to prove that $|\log \frac{n-N_r}{n} - \log(1 - U_{N_r,n})|$ tends to zero a.s. First, we note that $\{-\log(1-U_i),\ i \geqslant 1\} =^d \{Y_i,\ i \geqslant 1\}$ where Y_i are i.i.d. random variables from an exponential distribution with parameter 1. Therefore, setting H the distribution function of $Y_1, \ldots, Y_n$ and $\mathbb{H}_n$ the empirical distribution function associated, the quantity of interest can be rewritten as

$$\begin{aligned}\left|\mathbb{H}_n^{-1}\left(\frac{N_r}{n}\right) - H^{-1}\left(\frac{N_r}{n}\right)\right| &= \left| H^{-1}\left(G_n^{-1}\left(\frac{N_r}{n}\right)\right) - H^{-1}\left(\frac{N_r}{n}\right)\right| \\ &= \left| \frac{G_n^{-1}(N_r/n) - N_r/n}{1 - N_r/n}\big(1 + \mathrm{o}(1)\big)\right| \\ &\leqslant \frac{|V_n(N_r/n)|}{\sqrt{(N_r/n)(1-N_r/n)}} \sqrt{\frac{N_r/n}{n-N_r}}\big(1 + \mathrm{o}(1)\big),\end{aligned}$$

where $V_n(\cdot)$ denotes the empirical uniform quantile process.

Then, using Theorem 1 in Shorack and Wellner [7] (p. 616) combining with the assumptions on N_r, Lemma 4 follows. □

Combining the four lemmas, the proof of Theorem 2 is achieved. □

References

[1] J. Beirlant, A. Guillou, Pareto index estimation under moderate right censoring, Scand. Actuar. J. 2 (2001) 111–125.
[2] E. Csáki, The law of the iterated logarithm for normalized empirical distribution function, Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete 38 (1977) 147–167.
[3] S. Csörgő, D.M. Mason, Central limit theorems for sums of extreme values, Math. Proc. Cambridge Philos. Soc. 98 (1985) 547–558.
[4] P. Deheuvels, E. Haeusler, D.M. Mason, Almost sure convergence of the Hill estimator, Math. Proc. Cambridge Philos. Soc. 104 (1988) 371–381.
[5] B.M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Statist. 3 (1975) 1163–1174.
[6] G. Matthys, E. Delafosse, A. Guillou, J. Beirlant, Estimating high quantiles, Technical Report, 2002.
[7] G.R. Shorack, J.A. Wellner, Empirical Processes with Applications to Statistics, Wiley, New York, 1986.
[8] I. Weissman, Estimation of parameters and large quantiles based on the k largest observations, J. Amer. Statist. Assoc. 73 (1978) 812–815.