Analyse statistique des valeurs extrêmes pour des échantillons aléatoirement censurés

Julien Worms (1) & Rym Worms (2)

(1) *Laboratoire de Mathématiques de Versailles, Université Versailles-St-Quentin, CNRS, Université Paris-Saclay, 78035 Versailles*
(2) *UPEMLV, UPEC, Université Paris-Est, Laboratoire d'Analyse et de Mathématiques Appliquées (CNRS UMR 8050), F-94010 Créteil*

Résumé : L'estimation de l'indice des valeurs extrêmes et des quantiles extrêmes pour des données incomplètes a fait l'objet de plusieurs publications ces dernières années. Cet exposé sera consacré, dans le cas de données censurées aléatoirement à droite, à l'une des approches possibles, qui consiste à estimer l'indice des valeurs extrêmes par une fonctionnelle de Kaplan-Meier, ou par une combinaison de telles fonctionnelles. Une adaptation de l'estimateur de Hill ou de l'estimateur dit des moments de Dekkers-Einmahl-de Haan sera présentée dans le cas de données à queue lourde, et dans le cas de données à loi dans le domaine de Weibull l'estimateur des moments et des variantes seront également adaptés au cadre censuré.

Abstract : The estimation of the extreme value index and extreme quantiles for incomplete datasets has been the topic of several publications these recent years. This talk will be devoted, in the framework of data randomly censored from the right, to one of the possible approaches, which consists in estimating the extreme value index by a Kaplan-Meier functional, or by a combination of such functionals. A version of the Hill estimator or of the so-called moment estimator by Dekkers-Einmahl-de Haan, adapted to this censored context, will be presented in the case of heavy-tailed data, and in the case of data with distribution in the Weibull maximum domain of attraction, the moment estimator and some of its variants will also be adapted.

Mots-clés : Valeurs extrêmes, données censurées, fonctionnelles de Kaplan-Meier.

1 Introduction

L'analyse statistique des valeurs extrêmes pour des données réelles indépendantes mais formant un échantillon incomplet (données censurées ou tronquées, aléatoirement ou de manière fixe) a bénéficié d'une certaine attention dans la littérature récente. En ce qui concerne en particulier le cas de données censurées aléatoirement à droite (qui sera le cadre de cet exposé), le point de départ est le travail novateur contenu dans Beirlant, Dierckx,

Fils-Villetard et Guillou (2007), dont le principe a été généralisé dans Einmahl, Fils-Villetard et Guillou (2008), puis exploité, dans le cas de données à queues lourdes, dans Diop, Dupuy et Ndao (2014) (avec présence de covariables) et dans Brahimi, Meraghni et Necir (2015). On peut citer également Grama, Petiot et Tricot (2014) sur le sujet.

Les résultats des références citées ci-dessus (sauf le dernier) sont essentiellement basés sur une astuce permettant de séparer les aspects valeurs extrêmes et censure du problème, et permettant de définir des estimateurs de l'indice des valeurs extrêmes (de la variable censurée) qui soient adaptés à la situation de censure aléatoire. Dans cet exposé nous allons présenter une autre approche, plus proche des principes usuels de l'analyse de survie, développée dans Worms et Worms (2014a, 2014b), et valable dans le cas d'observations à queue lourde ou bien bornées.

On considère le cadre classique de l'observation d'un échantillon censuré aléatoirement à droite. Soient $(X_i)_{i\leq n}$ et $(C_i)_{i\leq n}$ deux suite i.i.d. indépendantes positives de fonctions de répartition respectives F et G continues, de points terminaux x_F^+ et x_G^+ (où $x_F^+ := \sup\{x,\ F(x) < 1\}$). On considère qu'on n'observe que les couples (Z_i, δ_i) $(1 \leq i \leq n)$ définis par

$$Z_i = \min(X_i, C_i) \ \text{ et } \ \delta_i = \mathbb{I}_{X_i \leq C_i}.$$

On note alors H la fonction de répartition commune des Z_i, et $Z_{1,n} \leq \cdots \leq Z_{n,n}$ les statistiques d'ordres associées, ainsi que $\delta_{1,n}, \ldots, \delta_{n,n}$ les indicateurs de censure correspondants.

On suppose que F et G appartiennent au domaine d'attraction du maximum $D(H_{\gamma_F})$ et $D(H_{\gamma_G})$ respectivement (où H_γ désigne la loi des valeurs extrêmes de paramètre γ). Dans cet exposé, nous ne considérerons que les cadres suivants :

— Cadre 1 : $\gamma_F > 0$ et $\gamma_G > 0$, $x_F^+ = x_G^+ = +\infty$

— Cadre 2 : $\gamma_F < 0$ et $\gamma_G < 0$, $x_F^+ = x_G^+ = x^+ < +\infty$

Le cas où F et G appartiennent au domaine d'attraction de Gumbel n'est pas traité, tandis que tous les autres cadres (par exemple $\gamma_F < 0$ et $\gamma_G > 0$, ou bien l'inverse, ou bien encore $\gamma_F < 0$, $\gamma_G < 0$, mais $x_F^+ < x_G^+$) ne sont pas intéressants à étudier : en pratique, les données dans la queue seraient alors soit toutes censurées soit toutes non-censurées.

2 Méthodologie

Le principe utilisé dans ce travail est que, si ϕ est une fonction mesurable de $\mathbb{R}_+$ dans $\mathbb{R}$ telle que $\mathbb{E}(\phi(X_1)) = \int \phi\, dF < \infty$, alors cette espérance là peut être estimée par

$$\int \phi\, d\hat{F}_n \ = \ \sum_{i=1}^{n} W_{i,n}\phi(Z_i) \quad \text{où} \quad W_{i,n} = \frac{\delta_i}{n(1-\hat{G}_n(Z_i^-))},$$

les notations $\hat{F}_n$ et $\hat{G}_n$ désignant les estimateurs de Kaplan-Meier de F et de G, définis pour tout $t < Z_{n,n}$ par

$$1 - \hat{F}_n(t) = \prod_{Z_{i,n} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{i,n}} \quad \text{et} \quad 1 - \hat{G}_n(t) = \prod_{Z_{i,n} \leq t} \left(\frac{n-i}{n-i+1} \right)^{1-\delta_{i,n}}.$$

Le principe en question est également lié à la propriété $\mathbb{E}(\phi(X_1)) = \mathbb{E}\left(\frac{\delta_1}{1-G(Z_1)} \phi(Z_1) \right)$, qu'il est aisé de démontrer. Dans notre cas d'étude des valeurs extrêmes de X, ce principe est exploité en considérant, pour $\alpha = 1$ ou bien pour $\alpha > 1$ à choisir, les fonctions

$$\phi_{n,\alpha}(x) = \frac{1}{1 - \hat{F}_n(Z_{n-k_n,n})} \log^\alpha \left(\frac{x}{Z_{n-k_n,n}} \right) \mathbb{I}_{x > Z_{n-k_n,n}}$$

où $\log^\alpha(x) = (\log(x))^\alpha$, et k_n est le nombre d'observations dans la queue que l'on choisira de conserver pour calculer nos estimations. En effet, dans le cadre $\gamma_F > 0$, il est connu que si $t_n \to +\infty$, alors

$$\frac{1}{1 - F(t_n)} \int_{t_n}^{+\infty} \log \left(\frac{x}{t_n} \right) \, dF(x) \longrightarrow \gamma_F.$$

De plus, quand $\gamma_F < 0$ et $g(t_n) = \log^\alpha(x^+/t_n)$, ou bien quand $\gamma_F > 0$, $g(t_n) = 1$ et $x^+ = +\infty$, on peut montrer que

$$\frac{1}{g(t_n)(1 - F(t_n))} \int_{t_n}^{x^+} \log^\alpha \left(\frac{x}{t_n} \right) \, dF(x)$$

converge vers une constante explicite dépendant de α et de γ_F. En considérant les fonctionnelles de Kaplan-Meier $\int \phi_{n,\alpha} \, d\hat{F}_n$, qui s'écrivent plus explicitement

$$M_{n,k_n}^{(\alpha)} := \frac{1}{n(1 - \hat{F}_n(Z_{n-k_n,n}))} \sum_{i=1}^{k_n} \frac{\delta_{n-i+1,n}}{1 - \hat{G}_n(Z_{n-i+1,n}^-)} \log^\alpha \left(\frac{Z_{n-i+1,n}}{Z_{n-k_n,n}} \right)$$

il est alors possible de définir des versions adaptées à la censure des estimateurs :

— de Hill dans le cadre 1, en posant $\hat{\gamma}_F = M_{n,k_n}^{(1)}$

— des moments dans le cadre 1 ou 2, en posant $\hat{\gamma}_F = M_{n,k_n}^{(1)} + 1 - \frac{1}{2} \left(1 - \frac{(M_{n,k_n}^{(1)})^2}{M_{n,k_n}^{(2)}} \right)^{-1}$.

Dans l'exposé, nous présenterons des résultats asymptotiques pour ces estimateurs, et pour d'autres combinaisons des moments $M_{n,k_n}^{(\alpha)}$ qui aboutissent à des variantes de l'estimateur des moments. Nous évoquerons également une variante de tous ces estimateurs, basée sur l'utilisation de moments $\widetilde{M}_{n,k_n}^{(\alpha)}$ différents des $M_{n,k_n}^{(\alpha)}$, et inspirée de mécaniques

connues dans des problèmes de régression avec censure. Cette variante produit des estimateurs légèrement différents de ceux définis plus haut, mais montre souvent de meilleures performances en termes d'erreur quadratique moyenne dans les simulations.

Dans cette étude, les difficultés rencontrées sont surtout liées au fait que les intégrales de Kaplan-Meier considérées le sont avec des fonctions non-déterministes (en raison de la présence des estimateurs de Kaplan-Meier de F et de G, et du seuil aléatoire $Z_{n-k_n,n}$), non bornées, et de support "glissant vers l'infini" donc non compact ; ceci se rajoute au fait que le comportement des estimateurs de Kaplan-Meier de F et de G est délicat à gérer aux abords de la queue de F et de G.

Bibliographie

[1] Beirlant, J. ; Dierckx, G. ; Fils-Villetard, A. et Guillou, A. (2007). Estimation of the extreme value index and extreme quantiles under random censoring, *Extremes*, **10**, 151–174.

[2] Brahimi, B. ; Meraghni, D. et Necir, A. (2015). Gaussian approximation to the extreme value index estimator of a heavy-tailed distribution under random censoring, *Mathematical Methods of Statistics*, 24(4), 266–279.

[3] Diop, A. ; Dupuy, J-F. et Ndao, P. (2014). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring, *Computational Statistics & Data Analysis*, **79**, 63–79.

[4] Einmahl, J. ; Fils-Villetard, A. et Guillou, A. (2008). Statistics of extremes under random censoring, *Bernoulli*, **14**, 207–227.

[5] Grama, I. ;Petiot, J-F. et Tricot, J-M. (2014). Estimation of extreme survival probabilities from censored data, *Bulletin of the Academy of Sciences of Moldova*, **1(74)**, 33–62.

[6] Worms, J. et Worms, R. (2014a). New estimators of the extreme value index under random right censoring, for heavy-tailed distributions, *Extremes*, **17**, 2, 337–358.

[7] Worms, J. et Worms, R. (2014b). Moment estimators of the extreme value index for randomly censored data in the Weibull domain of attraction, *accessible sur HAL (hal-01162967)*.