Un estimateur des moments pour l'indice des valeurs extrêmes conditionnel

Gilles Stupfler

Université d'Aix-Marseille, CERGAM, 15-19 allée Claude Forbin,
13628 Aix-en-Provence Cedex 1, France
E-mail : gilles.stupfler@univ-amu.fr

Résumé. En théorie des valeurs extrêmes, l'indice des valeurs extrêmes est un paramètre contrôlant le comportement asymptotique d'une fonction de répartition. La connaissance de ce paramètre est donc primordiale lorsqu'on souhaite étudier les extrêmes d'une variable aléatoire. On introduit ici un estimateur de l'indice des valeurs extrêmes en présence d'une covariable aléatoire. On examine quelques propriétés asymptotiques de cet estimateur sans condition sur le domaine d'attraction (DA) de la variable réponse et on illustre le comportement de l'estimateur sur simulations.

Mots-clés. Indice des valeurs extrêmes, covariable aléatoire, consistance, normalité asymptotique ponctuelle.

Abstract. In extreme value theory, the so-called extreme-value index is a parameter that controls the behavior of a distribution function in its right tail. Knowing this parameter is thus essential to solve many problems related to extreme events. In this paper, the estimation of the extreme-value index is considered in the presence of a random covariate. The pointwise weak consistency and asymptotic normality of the proposed estimator are established without any condition on the domain of attraction of the variable of interest. We examine the finite sample performance of our estimator in a simulation study.

Keywords. Extreme-value index, random covariate, consistency, pointwise asymptotic normality.

1 Introduction

Le théorème fondamental de la théorie des valeurs extrêmes est le théorème de Fisher-Tippett-Gnedenko (Fisher et Tippett [3], Gnedenko [5]). Ce théorème dit que si (Y_n) est une suite de copies indépendantes de la variable aléatoire Y de fonction de répartition F telles qu'il existe deux suites déterministes (a_n) et (b_n), avec $a_n > 0$ et telles que

$$\frac{1}{a_n}\left(\max_{1\leq i\leq n} Y_i - b_n\right)$$

converge en loi vers une variable aléatoire Z non constante, alors la fonction de répartition de Z est de la forme $y \mapsto G_\gamma(ay+b)$, où $a>0$ et b, $\gamma \in \mathbb{R}$ où

$$G_\gamma(y) = \begin{cases} \exp\left(-(1+\gamma y)^{-1/\gamma}\right) & \text{si } \gamma \neq 0 \text{ et } 1+\gamma y > 0, \\ \exp\left(-\exp(-y)\right) & \text{si } \gamma = 0. \end{cases}$$

On écrit alors $F \in \mathcal{D}(G_\gamma)$. Le paramètre γ, appelé indice des valeurs extrêmes de Y, contrôle le comportement asymptotique de G_γ et donc celui de la fonction de répartition de Y :

- si $\gamma > 0$, autrement dit Y appartient au DA de Fréchet, alors $1-G_\gamma$ est à décroissance polynomiale ;
- si $\gamma < 0$, autrement dit Y appartient au DA de Weibull, alors $1-G_\gamma$ est à support borné à droite ;
- si $\gamma = 0$, autrement dit Y appartient au DA de Gumbel, alors $1-G_\gamma$ est à décroissance exponentielle.

En pratique, la variable Y est souvent reliée à une covariable X. Dans ce cas, γ dépend de la valeur de la covariable et est appelé indice des valeurs extrêmes conditionnel. Dans la plupart des travaux récents sur ce sujet, son estimation a été considérée dans le cas où $\gamma(x) > 0$; il semble que le seul travail qui ne fait pas cette hypothèse soit celui de Daouia *et al.* [1], qui étudie une adaptation de l'estimateur de Pickands [7].

Notre but est de présenter une adaptation de l'estimateur des moments de l'indice des valeurs extrêmes conditionnel, qui soit convergent dans les trois domaines d'attraction. On étudie ses propriétés asymptotiques et on illustre son comportement sur simulations.

2 Construction de l'estimateur

Soient $(X_1,Y_1),\ldots,(X_n,Y_n)$ des copies aléatoires d'un couple aléatoire (X,Y) à valeurs dans $E\times(0,\infty)$ où (E,d) est un espace métrique. Pour tout $x\in E$, on suppose que la fonction de survie conditionnelle $\overline{F}(\cdot|x) = 1-F(\cdot|x)$ de Y sachant $X=x$ appartient à $\mathcal{D}(G_{\gamma(x)})$. On peut montrer que cette hypothèse se met sous la forme suivante (voir de Haan et Ferreira [6]) :

(M_1) Y est strictement positive et pour tout $x\in E$, il existe $\gamma(x)\in\mathbb{R}$ et une fonction strictement positive $a(\cdot|x)$ telle que l'inverse généralisé $U(\cdot|x)$ de $1/\overline{F}(\cdot|x)$, défini par $U(z|x) = \inf\{y\in\mathbb{R}\,|\,1/\overline{F}(y|x)\geq z\}$ pour tout $z\geq 1$, vérifie

$$\forall z>0,\ \lim_{t\to\infty}\frac{U(tz|x)-U(t|x)}{a(t|x)} = \begin{cases} \dfrac{z^{\gamma(x)}-1}{\gamma(x)} & \text{si } \gamma(x)\neq 0 \\ \log z & \text{si } \gamma(x)=0. \end{cases}$$

Notre estimateur est une adaptation de celui de Dekkers *et al.* [2]. On note, pour $x \in E$ et $h = h(n) \to 0$, $N_n(x,h)$ le nombre d'observations dans la boule fermée $B(x,h)$ de centre x et rayon h:

$$N_n(x,h) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B(x,h)\}} \quad \text{avec} \quad B(x,h) = \{x' \in E \,|\, d(x,x') \leq h\},$$

où $\mathbb{1}_{\{.\}}$ est l'indicatrice. Sachant $N_n(x,h) = p \geq 1$, on note pour $i = 1, \ldots, p$, $Z_i = Z_i(x,h)$ les Y_i dont les covariables associées $W_i = W_i(x,h)$ appartiennent à $B(x,h)$. Soient $Z_{1,p} \leq \cdots \leq Z_{p,p}$ les statistiques d'ordre associées et pour $j = 1, 2$

$$M_n^{(j)}(x,k_x,h) = \frac{1}{k_x} \sum_{i=1}^{k_x} \left[\log(Z_{p-i+1,p}) - \log(Z_{p-k_x,p})\right]^j$$

si $k_x \in \{1, \ldots, p-1\}$ et 0 sinon. Notre estimateur est alors

$$\begin{aligned} \widehat{\gamma}_n(x,k_x,h) &= \widehat{\gamma}_{n,+}(x,k_x,h) + \widehat{\gamma}_{n,-}(x,k_x,h) \\ \text{où} \quad \widehat{\gamma}_{n,+}(x,k_x,h) &= M_n^{(1)}(x,k_x,h) \\ \text{et} \quad \widehat{\gamma}_{n,-}(x,k_x,h) &= 1 - \frac{1}{2}\left(1 - \frac{\left[M_n^{(1)}(x,k_x,h)\right]^2}{M_n^{(2)}(x,k_x,h)}\right)^{-1} \end{aligned}$$

si $\left[M_n^{(1)}(x,k_x,h)\right]^2 \neq M_n^{(2)}(x,k_x,h)$, et $\widehat{\gamma}_{n,-}(x,k_x,h) = 0$ sinon.

3 Résultats principaux

Pour $x \in E$, on note $n_x = n_x(n,h) = n\mathbb{P}(X \in B(x,h))$ le nombre moyen de points dans la boule $B(x,h)$ et on suppose que $n_x(n,h) > 0$ pour tout n. On se donne également $k_x = k_x(n)$ une suite d'entiers strictement positifs. Soit $F_h(\cdot|x)$ la fonction de répartition conditionnelle de Y sachant $X \in B(x,h)$:

$$F_h(y|x) = \mathbb{P}(Y \leq y \,|\, X \in B(x,h))$$

et $U_h(\cdot|x)$ l'inverse généralisé de $1/\overline{F}_h(\cdot|x)$. Pour u, $v \in (1,\infty)$ tels que $u < v$, on note

$$\omega(u,v,x,h) = \sup_{z \in [u,v]} \left|\log \frac{U_h(z|x)}{U(z|x)}\right|.$$

On rappelle les notations $a \wedge b = \min(a,b)$ et $a \vee b = \max(a,b)$ pour a, $b \in \mathbb{R}$.

Théorème 1. On suppose que (M_1) est vérifiée. Soit $x \in E$. On suppose que $n_x \to \infty$, $k_x \to \infty$, $k_x/n_x \to 0$ et pour un certain $\delta > 0$

$$\frac{U(n_x/k_x|x)}{a(n_x/k_x|x)}\omega\left(\frac{n_x}{(1+\delta)k_x}, n_x^{1+\delta}, x, h\right) \to 0 \quad \text{quand} \quad n \to \infty.$$

Alors, si $\gamma_+(x) = 0 \vee \gamma(x)$ et $\gamma_-(x) = 0 \wedge \gamma(x)$, on a

$$\widehat{\gamma}_{n,+}(x, k_x, h) \xrightarrow{\mathbb{P}} \gamma_+(x) \quad \text{et} \quad \widehat{\gamma}_{n,-}(x, k_x, h) \xrightarrow{\mathbb{P}} \gamma_-(x) \quad \text{quand} \quad n \to \infty$$

et en particulier $\widehat{\gamma}_n(x, k_x, h) \xrightarrow{\mathbb{P}} \gamma(x)$ quand $n \to \infty$.

Pour obtenir la normalité asymptotique de notre estimateur, on doit introduire une condition de second-ordre :

(M_2) L'hypothèse (M_1) est vérifiée et pour tout $x \in E$, il existe un réel $\rho(x) \leq 0$ et une fonction $A(\cdot|x)$ de signe constant, convergeant vers 0 à l'infini, telle que

$$\forall z > 0, \ \lim_{t\to\infty} \frac{\frac{U(tz|x)-U(t|x)}{a(t|x)} - \frac{z^{\gamma(x)}-1}{\gamma(x)}}{A(t|x)} = H_{\gamma(x),\rho(x)}(z)$$

où

$$H_{\gamma(x),\rho(x)}(z) = \int_1^z r^{\gamma(x)-1}\left[\int_1^r s^{\rho(x)-1}ds\right]dr.$$

On pose également

$$\ell(x) = \lim_{t\to\infty}\left(U(t|x) - \frac{a(t|x)}{\gamma(x)}\right)$$

et

$$Q(t|x) = \begin{cases} A(t|x) & \text{si} \quad \gamma(x) < \rho(x) \leq 0 \\ & \text{ou} \quad \gamma(x) > 0 \text{ et } \rho(x) = 0 \\ \gamma_+(x) - \dfrac{a(t|x)}{U(t|x)} & \text{si} \quad \rho(x) < \gamma(x) \leq 0 \\ & \text{ou} \quad 0 < \gamma(x) < -\rho(x), \ \ell(x) \neq 0 \\ & \text{ou} \quad 0 < \gamma(x) = -\rho(x) \\ \dfrac{\rho(x)}{\gamma(x)+\rho(x)}A(t|x) & \text{si} \quad 0 < \gamma(x) < -\rho(x), \ \ell(x) = 0 \\ & \text{ou} \quad 0 < -\rho(x) < \gamma(x). \end{cases}$$

Théorème 2. On suppose que (M_1) est vérifiée. Soit $x \in E$. On suppose que $n_x \to \infty$, $k_x \to \infty$, $k_x/n_x \to 0$, $\sqrt{k_x}\,Q(n_x/k_x|x) \to \lambda(x) \in \mathbb{R}$ et pour un certain $\delta > 0$

$$\sqrt{k_x}\frac{U(n_x/k_x|x)}{a(n_x/k_x|x)}\omega\left(\frac{n_x}{(1+\delta)k_x}, n_x^{1+\delta}, x, h\right) \to 0 \quad \text{quand} \quad n \to \infty.$$

Alors, si $\gamma(x) \neq \rho(x)$, la variable aléatoire $\sqrt{k_x}(\widehat{\gamma}_n(x, k_x, h) - \gamma(x))$ est asymptotiquement normale, de biais $\lambda(x)B(\gamma(x), \rho(x))$ et variance $V(\gamma(x))$ qu'on peut expliciter.

4 Etude sur simulations

Pour juger des performances de notre estimateur sur simulations, on considère le cas $E = [0,1] \subset \mathbb{R}$ muni de la distance euclidienne, et une covariable X de loi uniforme sur E. On note $\gamma : E \to \mathbb{R}$ la fonction définie par

$$\forall x \in [0,1],\ \gamma(x) = \frac{2}{3} + \frac{1}{3}\sin(2\pi x).$$

Le premier modèle de fonction de survie conditionnelle pour Y sachant $X = x$ est

$$\forall y > 0,\ \overline{F}_1(y|x) = \left(1 + y^{-\tau}\right)^{1/\tau\gamma(x)},$$

où $\tau \in \{-1.2, -1, -0.8\}$. C'est un exemple de variable appartenant au DA de Fréchet. Le second modèle est

$$\forall y \in [0, g(x)],\ \overline{F}_2(y|x) = \frac{\Gamma(2/\gamma(x))}{\Gamma^2(1/\gamma(x))} \int_{y/g(x)}^{1} t^{1/\gamma(x)-1}(1-t)^{1/\gamma(x)-1} dt$$

où $\Gamma : (0,\infty) \to \mathbb{R}$ est la fonction Gamma d'Euler :

$$\forall \alpha > 0,\ \Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt$$

et la fonction de point terminal est

$$\forall x \in [0,1],\ g(x) = 1 - c + 8cx(1-x)$$

où $c \in \{0.1, 0.2, 0.3\}$. Ce modèle est un cas contenu dans le DA de Weibull. Le dernier modèle est

$$\forall y > 0,\ \overline{F}_3(y|x) = \int_{\log y}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2(x)}} \exp\left(-\frac{(t-\mu(x))^2}{2\sigma^2(x)}\right) dt$$

où μ et σ sont définies par

$$\forall x \in [0,1],\ \mu(x) = \frac{2}{3} + \frac{1}{3}\sin(2\pi x) \ \text{ et } \ \sigma(x) = 0.7 + 2.4x(1-x).$$

Dans ce cas, Y sachant $X = x$ a une loi log-normale de paramètres $\mu(x)$ et $\sigma^2(x)$, ce qui constitue un exemple compris dans le DA de Gumbel.

On estime la fonction γ sur une grille de points $\{x_1, \ldots, x_M\}$ de $[0,1]$. Il y a donc deux paramètres à choisir : la fenêtre h et le nombre k_x. Pour ce faire, on utilise une procédure semblable à celle détaillée dans Gardes et Stupfler [4]. On applique cette procédure à $N = 100$ échantillons indépendants de taille $n = 500$. La grille de points $\{x_1, \ldots, x_M\}$ est constituée de $M = 50$ points de $[0,1]$ régulièrement espacés.

Pour avoir un aperçu du comportement de notre estimateur, on le compare à l'estimateur $\widetilde{\gamma}_D = \widehat{\gamma}_n^{RP,1}$ de Daouia *et al.* [1]. Le calcul de cet estimateur à noyau nécessite également de choisir deux paramètres h et k_x ; pour ce faire, on utilise la procédure décrite dans Daouia *et al.* [1]. Les résultats sont donnés en termes d'erreur quadratique moyenne dans la Table 1 ci-dessous. Ce tableau montre qu'en termes d'erreur quadratique moyenne, notre estimateur a de meilleures performances que l'estimateur de Daouia *et al.* [1] dans les deux premiers exemples, et des performances comparables dans le dernier cas.

Cas	Estimateur des moments $\widehat{\gamma}$	Estimateurr $\widetilde{\gamma}_D$ de Daouia *et al.*
Modèle 1		
$\tau = -0.8$	0.1496	0.1962
$\tau = -1$	0.0781	0.1616
$\tau = -1.2$	0.0553	0.1586
Modèle 2		
$c = 0.1$	0.0686	0.1329
$c = 0.2$	0.0689	0.1257
$c = 0.3$	0.0825	0.1313
Modèle 3	0.3384	0.2801

Table 1: Erreurs quadratiques moyennes commises par les deux estimateurs.

Bibliographie

[1] Daouia, A., Gardes, L., Girard, S. (2013). On kernel smoothing for extremal quantile regression, *Bernoulli*, to appear.

[2] Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Annals of Statistics* **17**(4): 1833–1855.

[3] Fisher, R.A., Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society* **24**: 180–190.

[4] Gardes, L., Stupfler, G. (2013). Estimation of the conditional tail-index using a smoothed local Hill estimator, *Extremes*, to appear.

[5] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics* **44**: 423–453.

[6] de Haan, L., Ferreira, A. (2006). *Extreme value theory: An introduction*. Springer, New York.

[7] Pickands, J. (1975). Statistical inference using extreme order statistics, *Annals of Statistics* **3**: 119–131.