Estimation de quantiles extrêmes pour des observations tronquées

Laurent Gardes [1] & Gilles Stupfler [2]

[1] *Université de Strasbourg & CNRS, IRMA, UMR 7501, 7 rue René Descartes, 67084 Strasbourg Cedex, France*
[2] *Aix Marseille Université, CERGAM, EA 4225, 15-19 allée Claude Forbin, 13628 Aix-en-Provence Cedex 1, France*

Résumé. Dans ce travail, nous proposons une famille d'estimateurs non-paramétriques du quantile extrême ainsi que de l'indice des valeurs extrêmes pour des lois à queues lourdes et tronquées à droite. Nous montrons la consistance faible et établissons la normalité asymptotique des estimateurs. Nous terminons par une illustration sur des données simulées.

Mots-clés. Loi à queue lourde, indice des valeurs extrêmes, quantile extrêmes.

Abstract. The goal of this paper is to provide estimators of the tail index and extreme quantiles of a heavy-tailed random variable when the data is right-truncated. The weak consistency and asymptotic normality of the estimators are established and we illustrate the finite sample performance of our estimators on a simulation study.

Keywords. Heavy-tailed distribution, tail index, extreme quantile.

1 Présentation du modèle et définition des estimateurs

On considère n copies indépendantes $(Y_1, T_1), \ldots, (Y_n, T_n)$ d'un vecteur aléatoire $(Y, T) \in [y_0, \infty) \times [t_0, \infty)$, où $y_0 \geq 0$ et $y_0 \leq t_0$. La variable T est une variable de troncature que l'on supposera indépendante de la variable d'intérêt Y. Nous supposerons de plus que les fonctions de répartition marginales F et G des variables Y et T sont à queue lourde.

(M) Soient $0 < \gamma_F < \gamma_G$. Pour tout $\lambda > 0$, lorsque $y \to \infty$,

$$\frac{1 - F(\lambda y)}{1 - F(y)} \to \lambda^{-1/\gamma_F} \text{ et } \frac{1 - G(\lambda y)}{1 - G(y)} \to \lambda^{-1/\gamma_G}.$$

L'objectif de ce travail est de proposer une famille d'estimateurs du quantile extrême d'ordre $\alpha_n \in]0, 1[$ défini par $q(\alpha_n) := \inf\{y \geq y_0 \,|\, 1 - F(y) \leq \alpha_n\}$ avec $\alpha_n \to 0$ lorsque

$n \to \infty$. L'estimation de $q(\alpha_n)$ passe par l'estimation de la fonction de répartition empirique. Cependant, la variable Y étant tronquée par T, nous disposons uniquement des observations (Y_i, T_i) pour lesquelles $Y_i \le T_i$. Ainsi, l'estimateur non-paramétrique classique de la fonction de répartition ne peut pas être utilisé dans notre situation. Nous allons construire notre estimateur à l'aide du résultat suivant. Posons

$$F^*(y) = \mathbb{P}(Y \le y | Y \le T) \text{ et } G^*(t) = \mathbb{P}(T \le t | Y \le T),$$

les fonctions de répartition conditionnelles de Y et T sachant $Y \le T$. On peut montrer que pour $y > y_0$,

$$-\log F(y) = \int_y^\infty \frac{dF^*(z)}{F^*(z) - G^*(z)}.$$

Ainsi, pour estimer la fonction de répartition F, il suffit d'estimer les fonctions de répartition conditionnelles F^* et G^*. En posant

$$N := \sum_{i=1}^n \mathbb{I}_{\{Y_i \le T_i\}},$$

et en notant (Y_i^*, T_i^*), $1 \le i \le N$ les couples observés, les estimateurs classiques de F^* et G^* sont donnés par

$$\widehat{F}_N^*(y) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{Y_i^* \le y\}} \text{ et } \widehat{G}_N^*(t) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{T_i^* \le t\}}.$$

Remarquons que N suit une loi binomiale de paramètres n et $p := \mathbb{P}(Y \le T)$. On en déduit facilement des estimateurs de la fonction de répartition et du quantile :

$$\widehat{F}_N(y) = \exp\left(-\frac{1}{N} \sum_{i=1}^N \frac{\mathbb{I}_{\{Y_i^* > y\}}}{\widehat{F}_N^*(Y_i^*) - \widehat{G}_N^*(Y_i^*)}\right) \quad \text{et} \quad \widehat{q}_N(\alpha) = \inf\left\{y \ge y_0 \,|\, 1 - \widehat{F}_N(y) \le \alpha\right\}$$

Comme on le verra dans le Théorème 1, l'estimateur $\widehat{q}_N(\alpha_n)$ est consistant si α_n ne converge pas trop rapidement vers zéro. Pour s'affranchir de cette limitation, on remarque que sous le modèle **(M)** et si $\beta_n < \alpha_n$ sont deux suites positives qui convergent vers zéro et telles que $\beta_n/\alpha_n \to 0$ alors $q(\beta_n) \approx q(\alpha_n)\,(\alpha_n/\beta_n)^{\gamma_F}$. Ainsi, pour estimer des quantiles d'ordre arbitrairement proche de zéro, on doit dans un premier temps estimer l'indice des valeurs extrêmes γ_F. On propose la famille d'estimateurs

$$\widehat{\gamma}_{N,F}(k_N, k'_N) = \frac{\widehat{\gamma}_{N,F^*}(k_N)\widehat{\gamma}_{N,G}(k'_N)}{\widehat{\gamma}_{N,G}(k'_N) - \widehat{\gamma}_{N,F^*}(k_N)} \tag{1}$$

où $\widehat{\gamma}_{N,F^*}(k_N)$ et $\widehat{\gamma}_{N,G}(k'_N)$ sont les estimateurs de Hill (voir Hill, 1975)

$$\widehat{\gamma}_{N,F^*}(k_N) = \frac{1}{k_N} \sum_{i=1}^{k_N} \log \frac{Y^*_{N-i+1,N}}{Y^*_{N-k_N,N}} \quad \text{et} \quad \widehat{\gamma}_{N,G}(k'_N) = \frac{1}{k'_N} \sum_{i=1}^{k'_N} \log \frac{T^*_{N-i+1,N}}{T^*_{N-k'_N,N}}.$$

Les suites (k_n) and (k'_n) appartiennent à l'ensemble $\{1, \ldots, n-1\}$ et $Y^*_{1,N} \leq \ldots \leq Y^*_{N,N}$, $T^*_{1,N} \leq \ldots \leq T^*_{N,N}$ sont les statistiques d'ordre. On en déduit l'estimateur de type Weissman (voir Weissman, 1978) suivant

$$\widehat{q}_N^W(\beta_n|\alpha_n, k_N, k'_N) = \widehat{q}_N(\alpha_n)\left(\frac{\alpha_n}{\beta_n}\right)^{\widehat{\gamma}_{N,F}(k_N,k'_N)}. \tag{2}$$

2 Propriétés asymptotiques des estimateurs

Le résultat suivant établit la normalité asymptotique de l'estimateur $\widehat{q}_N(\alpha_n)$ où $\alpha_n \to 0$.

Théorème 1 *On suppose la fonction de répartition F dérivable, telle que $yF'(y)/(1-F(y)) \to 1/\gamma_F$ et $n\alpha_n(1-G(q(\alpha_n)) \to \infty$. Sous le modèle* **(M)**

$$\sqrt{n\alpha_n(1-G(q(\alpha_n))}\left(\frac{\widehat{q}_N(\alpha_n)}{q(\alpha_n)}-1\right) \xrightarrow{d} \mathcal{N}(0,\gamma_F^2).$$

Pour établir la normalité asymptotique de l'estimateur de type Weissman, on introduit la condition du second ordre ci-dessous :

(C) Les fonctions de répartition F et G sont dérivables et telles que

$$F'(y) = \left[\frac{1}{\gamma_F} - \Delta_F(y)\right]\frac{1-F(y)}{y} \quad \text{et} \quad G'(t) = \left[\frac{1}{\gamma_G} - \Delta_G(t)\right]\frac{1-G(t)}{t},$$

où Δ_F, Δ_G sont des fonctions bornées asymptotiquement de signe constant, qui convergent vers zéro et telles que $|\Delta_F|$ et $|\Delta_G|$ sont asymptotiquement monotones et à variations régulières à l'infini d'indices respectifs $\rho_F/\gamma_F \leq 0$ et $\rho_G/\gamma_G \leq 0$.

De plus, posons

$$R_{F^*}(y) = |\Delta_F(y)| \vee |\Delta_G(y)| \quad \text{et} \quad R_{G^*}(t) = |\overline{F}(t)| \vee |\Delta_G(t)|$$

où U_{F^*}, U_{G^*} sont les inverses généralisées des fonctions of $1/(1-F^*)$ and $1/(1-G^*)$. On a le résultat suivant :

Théorème 2 *Soient $\alpha_n \to 0$, $\beta_n \to 0$, $k_n \wedge k'_n \to \infty$ et $(k_n/n \vee k'_n/n) \to 0$. Sous le modèle* **(M)**, *si $\rho_F/\gamma_F \neq \rho_G/\gamma_G$, $(\rho_F/\gamma_F) \vee (\rho_G/\gamma_G) \neq -1/\gamma_F$, $\beta_n/\alpha_n \to 0$, $n\alpha_n(1-G(q(\alpha_n)) \to \infty$, $n\alpha_n(1-G(q(\alpha_n))\Delta_F^2(q(\alpha_n)) \to 0$, $k_n R^2_{F^*}(U_{F^*}(n/k_n)) \vee k'_n R^2_{G^*}(U_{G^*}(n/k'_n)) \to 0$, alors*

$$\frac{k_{\lfloor np \rfloor} \wedge k'_{\lfloor np \rfloor}}{n\alpha_n(1-G(q(\alpha_n))} \to 1 \ \textit{et} \ \sup_{q,r\in I_n}\left|\frac{k_q \wedge k'_q}{k_r \wedge k'_r} - 1\right| \to 0.$$

Alors, si $k_n/k'_n \to 0$ ou $k'_n/k_n \to 0$,

$$\frac{\sqrt{n\alpha_n(1-G(q(\alpha_n))}}{\log(\alpha_n/\beta_n)}\left(\frac{\widehat{q}_N^W(\beta_n|\alpha_n,k_N,k'_N)}{q(\beta_n)}-1\right) \xrightarrow{d} \mathcal{N}\left(0,\sigma_F^2\right),$$

avec $\sigma_F^2 = p\gamma_F^2[1+\gamma_F/\gamma_G]^3$ si $k_n/k'_n \to 0$ et $\sigma_F^2 = p\gamma_F^4/\gamma_G^2$ si $k'_n/k_n \to 0$.

3 Simulations

On s'intéresse ici au comportement à taille d'échantillon finie des estimateurs $\widehat{q}_N$ et $\widehat{q}_N^W$. Pour ce faire, on considère le modèle :

$$\forall y, t > 0,\ \overline{F}(y) = (1 + y^{1/\delta})^{-\delta/\gamma_F} \quad \text{et} \quad \overline{G}(t) = (1 + t^{1/\delta})^{-\delta/\gamma_G},$$

avec $\delta > 0$ et $0 < \gamma_F < \gamma_G$. Dans ce cas, la probabilité de troncature est $1 - p$, avec $p = \gamma_G/(\gamma_F + \gamma_G)$. En pratique, pour calculer l'estimateur $\widehat{q}_N$, on utilise l'expression

$$\widehat{q}_N(\alpha) = \sum_{i=1}^{N-1} Y_{i,N}^* \mathbb{I}_{\{\alpha \in [\Theta_{i+1}, \Theta_i)\}} + Y_{N,N}^* \mathbb{I}_{\{\alpha < \Theta_N\}},$$

où pour tout $i = 1, \ldots, N$,

$$\Theta_i = 1 - \exp\left(-\frac{1}{N}\sum_{j=i}^{N} \frac{1}{\widehat{F}_N^*(Y_{j,n}^*) - \widehat{G}_N^*(Y_{j,n}^*)}\right).$$

L' estimateur $\widehat{q}_N^W$est défini pour $\beta_n \to 0$ par

$$\widehat{q}_N^W(\beta_n|\alpha_n) = \widehat{q}_N(\alpha_n)\left(\frac{\alpha_n}{\beta_n}\right)^{\widehat{\gamma}_{n,F}(\alpha_n)},$$

où $\widehat{\gamma}_{n,F}(\alpha_n)$ est l'estimateur de l'indice des valeurs extrêmes γ_F défini par (1) avec $k_n = k_n' = \lfloor n\alpha_n \rfloor$. Ainsi, l'estimateur $\widehat{q}_N^W$ dépend uniquement du choix d'une suite α_n.
Nous souhaitons ici comparer les estimateurs $\widehat{q}_N$ et $\widehat{q}_N^W$ et pour ce faire, on simule $R = 1000$ échantillons de taille $n = 200$ dans différentes situations : $\gamma_F \in \{1/4, 1/2, 1\}$ et $p \in \{0.7, 0.8, 0.9, 0.95\}$. Dans chaque cas, pour une suite α_n donnée, on obtient R observations des estimateurs $\widehat{q}_N$ et $\widehat{q}_N^W$ notées $\widehat{q}_N^{(r)}$ et $\widehat{q}_N^{W,(r)}(.|\alpha_n)$, $r = 1, \ldots, R$. La suite α_n est choisie de la façon suivante :

$$\alpha_{opt}^{(r)} := \underset{\alpha \in (0, 0.15]}{\arg\min} \int_{0.07}^{0.15} \log^2\left(\frac{\widehat{q}_N^{(r)}(\beta)}{\widehat{q}_N^{W,(r)}(\beta|\alpha)}\right) d\beta$$

Ce choix est basé sur l'idée simple que, pour des ordres de quantiles β pas trop petits, les estimateurs $\widehat{q}_N^{(r)}$ et $\widehat{q}_N^{W,(r)}(.|\alpha_n)$ doivent être proches si la suite α_n est bien choisie. On calcule ensuite les erreurs

$$E(\tilde{q}^{(r)}) := \int_0^{0.15} \log^2\left(\frac{\tilde{q}^{(r)}(\beta)}{q(\beta)}\right) d\beta,$$

où $\tilde{q}^{(r)}$ correspond soit à l'estimateur $\widehat{q}_N^{(r)}$ soit à l'estimateur $\widehat{q}_N^{W,(r)}(.|\alpha_{opt}^{(r)})$. Pour $\theta = \{0.1, 0.5, 0.9\}$, on note $r(\theta)$ (resp. $s(\theta)$) le numéro de la réplication ayant fourni le quantile

Estimateur $\hat{q}_N^{(r(\theta))}$												
p	0.7			0.8			0.9			0.95		
θ	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
$\gamma_F = 1/4$	0.08	0.10	0.16	0.07	0.08	0.12	0.06	0.07	0.10	0.05	0.06	0.08
$\gamma_F = 1/2$	0.31	0.38	0.60	0.26	0.31	0.45	0.23	0.27	0.36	0.21	0.25	0.32
$\gamma_F = 1$	1.22	1.53	2.27	1.05	1.28	1.82	0.91	1.08	1.49	0.85	0.99	1.29

Estimateur $\hat{q}_N^{W,(s(\theta))}(.\vert\alpha_{opt}^{(s(\theta))})$												
p	0.7			0.8			0.9			0.95		
θ	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
$\gamma_F = 1/4$	0.004	0.03	0.22	0.003	0.02	0.10	0.002	0.01	0.06	0.002	0.01	0.04
$\gamma_F = 1/2$	0.01	0.10	0.50	0.007	0.05	0.27	0.004	0.03	0.16	0.004	0.03	0.12
$\gamma_F = 1$	0.04	0.39	1.71	0.03	0.25	1.15	0.02	0.13	0.61	0.01	0.09	0.39

Table 1: Erreurs associées aux estimateurs $\hat{q}_N^{(r(\theta))}$ et $\hat{q}_N^{W,(s(\theta))}(.|\alpha_{opt}^{(s(\theta))})$.

d'ordre θ de l'ensemble $\{E(\widehat{q}_N^{(r)}),\ r = 1,\ldots,N\}$ (resp. $\{E(\widehat{q}_N^{W,(r)}),\ r = 1,\ldots,N\}$). Le tableau 1 regroupe les valeurs $E(\widehat{q}^{(r(\theta))})$ et $E(\widehat{q}^{W,(s(\theta))}(.|\alpha_{opt}^{(s(\theta))})$ dans chacune des situations. On remarque que l'estimateur $\widehat{q}_N^W$ fournit de meilleurs résultats que l'estimateur empirique $\widehat{q}_N$. De plus, il apparaît que plus la probabilité p est faible, moins bonne est la qualité d'estimation.

Bibliographie

[1] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics* **3**: 1163–1174.
[2] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *Journal of the American Statistical Association* **73**: 812–815.