Étude de l'erreur relative d'extrapolation associée à l'estimateur de Weissman pour les quantiles extrêmes.

Clément ALBERT [1], Anne DUTFOY [2] & Stéphane GIRARD [3]

[1] *Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France.*
[2] *EDF R&D dept. Périclès, 91120 Palaiseau, France*
[3] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.*
clement.albert@polytechnique.edu, anne.dutfoy@edf.fr, stephane.girard@inria.fr

Résumé. Nous étudions le comportement asymptotique de l'erreur d'extrapolation (relative) associée à l'estimateur de Weissman, un estimateur semi-paramétrique des quantiles extrêmes dédié au domaine d'attraction de Fréchet. Des conditions sont alors fournies de telle sorte que l'erreur tende vers zéro quand la taille de l'échantillon augmente. Nous montrons que, dans le cas où la loi appartient au domaine d'attraction de Fréchet, sans surprise, l'erreur d'extrapolation relative tend vers zéro sous des conditions très faibles sur l'ordre du quantile. De manière originale, nous montrons également que l'erreur d'extrapolation tend vers zéro pour deux types de lois du domaine d'attraction de Gumbel sous des conditions raisonnables sur l'ordre du quantile. Mieux encore, des équivalents de l'erreur sont établis montrant que l'estimateur de Weissman mène à des erreurs d'extrapolation plus faibles que l'estimateur Exponential Tail pour certains types de lois du domaine d'attraction de Gumbel. Ces résultats sont illustrés numériquement.

Mots-clés. Erreur d'extrapolation, Propriétés asymptotiques, Quantiles extrêmes, Théorie des valeurs extrêmes.

Abstract. We study the asymptotic behaviour of the relative extrapolation error associated with the Weissman estimator, a semi-parametric extreme quantile estimator dedicated to the Fréchet maximum domain of attraction. Conditions are provided such that the error tends to zero as the sample size increases. Not surprisingly, we show that, in the case where the distribution belongs to the Fréchet maximum domain of attraction, the relative extrapolation error tends to zero under mild conditions. Surprisingly, we also show in an original fashion that the extrapolation error tends to zero for two types of distribution of the Gumbel maximum domain of attraction under reasonable conditions on the quantile order. Even better, first order approximations are derived showing that Weissman estimator leads to smaller extrapolation errors than the Exponential Tail estimator for some kind of distribution of the Gumbel maximum domain of attraction. These results are numerically illustrated.

Keywords. Extrapolation error, Asymptotic properties, Extreme quantiles, Extreme-value theory.

1 Introduction

Le point de départ de ce travail est l'étude asymptotique du comportement de l'estimateur de Weissman [8], un estimateur semi-paramétrique des quantiles extrêmes à partir d'une loi inconnue du domaine d'attraction de Fréchet, noté MDA(Fréchet). Étant donné un n-échantillon $X_1, \ldots, X_n$ de fonction de répartition commune F et de fonction de survie associée $\bar{F}$, un quantile extrême est un $(1-p_n)$ quantile $q(p_n)$ de F plus grand que l'observation maximale, *i.e.* de telle sorte que $\bar{F}(q(p_n)) = p_n$ avec $np_n \to 0$ as $n \to \infty$. L'estimation des quantiles extrêmes requiert des méthodes spécifiques. Parmi elles, la méthode Peaks Over Threshold (POT) se base sur une approximation de la loi des excès au dessus d'un seuil donné. Plus précisément, soit u_n un seuil déterministe tel que $\bar{F}(u_n) = \alpha_n$ ou de manière équivalente $u_n = q(\alpha_n)$ avec $\alpha_n \to 0$ et $n\alpha_n \to \infty$ quand $n \to \infty$. Les excès au dessus de u_n sont définis comme $Y_i = X_i - u_n$ pour tout $X_i > u_n$. La fonction de survie d'un excès est donnée par $\bar{F}_{u_n}(x) = \bar{F}(u_n + x)/\bar{F}(u_n)$. Le théorème de Pickands [10] stipule que, sous de faibles conditions, $\bar{F}_{u_n}$ peut être approchée par une loi de Pareto. En conséquence, le quantile extrême $q(p_n)$ peut à son tour être approché par le terme déterministe

$$\tilde{q}_{\mathrm{W}}(p_n; \alpha_n) = q(\alpha_n) \left(\frac{\alpha_n}{p_n}\right)^{\gamma_n}, \tag{1}$$

où $\gamma_n > 0$ est le paramètre de forme de la loi de Pareto. Ensuite, la méthode POT consiste à estimer ce paramètre inconnu. Dès lors, l'estimateur associé [8] est donné par

$$\hat{q}_{\mathrm{W}}(p_n; \alpha_n) = \hat{q}(\alpha_n) \left(\frac{\alpha_n}{p_n}\right)^{\hat{\gamma}_n}$$

où $\hat{q}(\alpha_n) = X_{n-k_n+1,n}$ avec $k_n = \lfloor n\alpha_n \rfloor$ et

$$\hat{\gamma}_n = \frac{1}{k_n} \sum_{i=1}^{k_n} \log X_{n-i+1,n} - \log X_{n-k_n+1,n}$$

est l'estimateur de Hill [9] qui dépend de la suite k_n. Rappelons que $X_{1,n} \leq \cdots \leq X_{n,n}$ désignent les statistiques d'ordre associées à $X_1, \ldots, X_n$. D'autres estimateurs de γ sont envisageables, citons par exemple les estimateurs adaptatifs [3] qui incluent une sélection automatique de k_n. L'erreur $(q(p_n) - \hat{q}_{\mathrm{W}}(p_n; \alpha_n))$ peut-être développée comme la somme de deux termes :

$$q(p_n) - \hat{q}_{\mathrm{W}}(p_n; \alpha_n) = (\tilde{q}_{\mathrm{W}}(p_n; \alpha_n) - \hat{q}_{\mathrm{W}}(p_n; \alpha_n)) + (q(p_n) - \tilde{q}_{\mathrm{W}}(p_n; \alpha_n)),$$

le premier étant une erreur d'estimation aléatoire et le deuxième une erreur d'extrapolation déterministe. Le comportement asymptotique de l'erreur d'estimation dépend en particulier de la loi asymptotique de l'estimateur $\hat{\gamma}_n$ choisi. Ce n'est pas l'objet d'intérêt de cette étude.

Dans cette communication, nous nous concentrons sur le comportement asymptotique de l'erreur d'extrapolation. En effet, au vu de (1), la méthode de Weissman extrapole dans les queues de distribution grâce à une correction multiplicative fonction de (α_n/p_n). Notre but est donc de quantifier la portée jusqu'à laquelle l'extrapolation est faite de manière consistante. Plus précisément, nous donnons des conditions nécessaires et suffisantes sur le couple (p_n, α_n) de telle sorte que l'erreur relative d'extrapolation

$$\varepsilon_{\mathrm{W}}(p_n; \alpha_n) := (q(p_n) - \tilde{q}_{\mathrm{W}}(p_n; \alpha_n))/q(p_n)$$

tende vers zéro quand $n \to \infty$.

2 Résultats asymptotiques

Définissons $H(\cdot) = -\log \overline{F}(\cdot)$ le taux de hasard cumulé, $K(x) = x(\log\log H^{-1})'(x)$, $x > 0$ et ℓ sa limite à l'infini. Dans la suite, on suppose que F est continue et doublement dérivable et que K' est monotone à l'infini. Rappelons qu'une fonction f est dite à variation régulière d'indice θ (noté $f \in RV_\theta$) si et seulement si f est positive à l'infini de telle sorte que $f(tx)/f(x) \to t^\theta$ quand $x \to \infty$ pour tout $t > 0$.

Le résultat qui suit donne une caractérisation du comportement de la queue de distribution de F en fonction de la limite ℓ.

Proposition 1

(i) Si $\exp H \in RV_{1/\gamma}$, $\gamma > 0$, *alors* $K \in RV_0$ *et* $\ell = 1$.

(ii) Si $H \in RV_\beta$, $\beta > 0$, *alors* $K \in RV_0$ *et* $\ell = 0$.

(iii) Si $H(\exp(\cdot)) \in RV_\beta$, $\beta > 0$ *alors* $K \in RV_0$ *et* $\ell = 1/\beta$.

Dans le cas (i) où $\exp H$ est à variation régulière d'indice positif, F est une loi à queue de type Pareto. Les lois de Burr, Cauchy, Fréchet, Pareto, Student sont les plus connues. Les cas (ii) et (iii) correspondent respectivement à des lois à queue de type Weibull et log-Weibull, voir [2, 5].

Proposition 2

(i) Si $F \in$ *MDA(Fréchet) alors* $K \in RV_0$ *et* $\ell = 1$.

(ii) Si $K \in RV_\theta$, $\theta > 0$, *alors* F *n'appartient à aucun domaine d'attraction.*

(iii) Si $K \in RV_\theta$, $\theta < 0$, *alors* F *ne définit pas une fonction de répartition.*

Notons tout d'abord que, dans le cas (i), il n'y a pas équivalence entre $K \in RV_0$ et $F \in$ MDA(Fréchet). Deuxièmement, au vu des Propositions 1 et 2, le seul cas d'intérêt est $\theta = 0$. Le comportement asymptotique de $\varepsilon_{\mathrm{W}}(p_n; \alpha_n)$ est par conséquent étudié dans les trois cas décrits par la Proposition 1 : Lois à queue de type Pareto, Weibull et log-Weibull.

Théorème 1 *Soit* $0 < p_n \leq \alpha_n < 1$ *de telle sorte que* $\limsup \log(1/p_n)/\log(1/\alpha_n) < \infty$. *Soit* $\delta(n) = 1 - \log(1/\alpha_n)/\log(1/p_n)$.

(i) Supposons que $F \in$ *MDA(Fréchet) avec un indice des valeurs extrêmes* $\gamma > 0$. *Soient* $L(t) := t^{-\gamma}H^{-1}(\log t)$, $\eta(t) := tL'(t)/L(t)$, $t > 0$ *et supposons que* $|\eta| \in RV_\rho$ *avec* $\rho < 0$. *Si* $\delta(n) \to \delta_\infty \in [0, 1)$ *alors* $\varepsilon_W(p_n; \alpha_n) \to 0$ *quand* $n \to \infty$.

(ii) Lois à queue de type Weibull. Supposons $H \in RV_\beta$, $\beta > 0$. *Alors,* $\varepsilon_W(p_n; \alpha_n) \to 0$ *si et seulement si* $\delta(n) \to 0$ *quand* $n \to \infty$.

(iii) Lois à queue de type Log-Weibull. Supposons $H(\exp \cdot) \in RV_\beta$, $\beta > 0$ *et* $\beta \neq 1$. *Alors,* $\varepsilon_W(p_n; \alpha_n) \to 0$ *si et seulement si* $\delta^2(n) \log q(p_n) \to 0$ *quand* $n \to \infty$.

Dans le cas (i) avec $F \in$ MDA(Fréchet), la fonction L est à variation lente [7] et η est appelé fonction auxiliaire associée à L. L'hypothèse $|\eta| \in RV_\rho$, $\rho < 0$, est récurrente en statistique des valeurs extrêmes, pour contrôler le biais des estimateurs, ρ étant connu comme le paramètre du second-ordre, voir [4]. Cette hypothèse est vérifiée pour la plupart des lois à queue lourde, telles que les lois Burr, Cauchy, Fréchet, Pareto et Student. Remarquons qu'il est possible dans ce cas de choisir un ordre extrême de telle sorte que $p_n = n^{-\tau}$, $\tau > 1$ tout en continuant à vérifier $\varepsilon_{\mathrm{W}}(p_n; \alpha_n) \to 0$ quand $n \to \infty$. Le Théorème 1(ii,iii) indique par ailleurs que l'erreur d'extrapolation relative induite par l'estimateur de Weissman tend vers zéro pourvu que $\delta(n) \to 0$ dans le cas des lois à queue de type Weibull ou bien pourvu que $\delta^2(n) \log q(p_n) \to 0$ dans le cas des lois à queue de type log-Weibull, même si ces lois n'appartiennent pas au domaine d'attraction de Fréchet.

Théorème 2

(i) Supposons que les hypothèses du Théorème 1(i) soient vérifiées.
Si $\delta(n) \to \delta_\infty \in [0, 1)$ *alors* $\varepsilon_W(p_n; \alpha_n) \sim -\frac{1}{1-\delta_\infty}\delta(n) \log(1/\alpha_n)\eta(1/\alpha_n)$.

(ii) Supposons que les hypothèses du Théorème 1(ii) soient vérifiées.
- *(a) Si* $\delta(n) \to 0$ *alors* $\varepsilon_W(p_n; \alpha_n) \sim -\frac{1}{2\beta}\delta^2(n)$.
- *(b) Si* $\delta(n) \to \delta_\infty \in (0, 1)$ *alors* $\varepsilon_W(p_n; \alpha_n) \to 1 - \exp\left(\frac{1}{\beta}\frac{\delta_\infty^2}{1-\delta_\infty}\right)$.

(iii) Supposons que les hypothèses du Théorème 1(iii) soient vérifiées.
- *(a) Si* $\delta^2(n) \log q(p_n) \to 0$ *alors* $\varepsilon_W(p_n; \alpha_n) \sim \frac{1-\beta}{2\beta^2}\delta^2(n) \log q(p_n)$.
- *(b) Si* $\delta^2(n) \log q(p_n) \to a \in (0, \infty)$ *alors* $\varepsilon_W(p_n; \alpha_n) \to 1 - \exp\left(-\frac{1-\beta}{2\beta^2}a\right)$.

Si $F \in$ MDA(Fréchet), cas (i), il est possible de choisir des ordres extrêmes $p_n = n^{-\tau}$, avec $\tau > 1$ menant à des erreurs d'extrapolation relatives d'ordre polynomial, ce qui est en accord avec les vitesses de convergence apparaissant dans la littérature sur le sujet, voir [6], Paragraphe 3.2. A notre connaissance, les cas (ii) et (iii) n'ont pas été considérés jusqu'à présent. Ils peuvent être illustrés en considérant $p_n = 1/(n \log n)$ et

$\alpha_n = (\log n)/n$. Moyennant ces choix, les équivalents de $\varepsilon_{\mathrm{W}}(p_n;\alpha_n)$ et de son homologue $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$, utilisant l'approximation Exponential Tail (ET) [1], sont donnés Table 1, pour cinq lois différentes. Notons que ces choix impliquent $\delta(n) \to 0$ et $\delta^2(n)\log q(p_n) \to 0$, de telle sorte que le Théorème 2(ii,iii)-(a) peut être utilisé, impliquant $\varepsilon_{\mathrm{W}}(p_n;\alpha_n) \to 0$ quand $n \to \infty$ pour les cinq lois. De manière originale, dans le cas des lois Weibull, Normale, Log-Weibull et Lognormale, la convergence de $\varepsilon_{\mathrm{W}}(p_n;\alpha_n)$ vers zero est du même ordre, voire plus rapide que la convergence de $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$. Dans de telles situations, l'approximation de Weissman est meilleure que celle associée à ET alors que l'estimateur de Weissman n'a pas été initialement conçu pour ce type de lois.

Ces résultats seront illustrés numériquement lors de l'exposé. Dans un premier temps, la qualité des équivalents sera évaluée sur des simulations. Dans un second temps, nous comparerons les erreurs d'extrapolation relatives associées aux méthodes ET et Weissman.

Lois	$\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$	$\varepsilon_{\mathrm{W}}(p_n;\alpha_n)$
Lois à queue de type Weibull		
Gamma($a > 0$)	$2(1-a)\dfrac{(\log\log n)^2}{(\log n)^3}$	$-2\dfrac{(\log\log n)^2}{(\log n)^2}$
Weibull($\beta \neq 1$)	$\dfrac{2(1-\beta)}{\beta^2}\dfrac{(\log\log n)^2}{(\log n)^2}$	$-\dfrac{2}{\beta}\dfrac{(\log\log n)^2}{(\log n)^2}$
Normale	$-\dfrac{1}{2}\dfrac{(\log\log n)^2}{(\log n)^2}$	$-\dfrac{(\log\log n)^2}{(\log n)^2}$
Lois à queue de type Log-Weibull		
Log-Weibull($\beta > 1$)	$\dfrac{2}{\beta^2}\dfrac{(\log\log n)^2}{(\log n)^{2-2/\beta}}$	$\dfrac{2(1-\beta)}{\beta^2}\dfrac{(\log\log n)^2}{(\log n)^{2-1/\beta}}$
Lognormale	$\sigma^2\dfrac{(\log\log n)^2}{\log n}$	$-\dfrac{\sqrt{2\sigma^2}}{2}\dfrac{(\log\log n)^2}{(\log n)^{3/2}}$

Table 1 – Équivalents de $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$ et $\varepsilon_{\mathrm{W}}(p_n;\alpha_n)$ avec $p_n = 1/(n\log n)$ et $\alpha_n = (\log n)/n$ pour cinq lois différentes.

Références

[1] Albert, C., Dutfoy, A., & Girard, S. (2019). Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles, `https://hal.inria.fr/hal-01692544`.

[2] Albert, C., Dutfoy, A., Gardes, L., & Girard, S. (2019). An extreme quantile estimator for the log-generalized Weibull-tail model, *Econometrics and Statistics*, https://doi.org/10.1016/j.ecosta.2019.01.004.

[3] Boucheron, S., & Thomas, M. (2015). Tail index estimation, concentration and adaptivity. *Electronic Journal of Statistics*, 9(2), 2751-2792.

[4] Gardes, L., & Girard, S. (2010). Conditional extremes from heavy-tailed distributions : An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2), 177-204.

[5] Diebolt, J., Gardes, L., Girard, S., & Guillou, A. (2008). Bias-reduced extreme quantile estimators of Weibull tail-distributions. *Journal of Statistical Planning and Inference*, 138(5), 1389-1401.

[6] De Haan, L., & Ferreira, A. (2007). *Extreme value theory : an introduction.* Springer Science & Business Media.

[7] Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1989). *Regular variation* (Vol. 27). Cambridge University Press.

[8] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364), 812-815.

[9] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 1163-1174.

[10] Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1), 119-131.