A de-randomization argument for estimating extreme value parameters of heavy tails

Abdelaati Daouia[1] & Joseph Hachem[2] & Gilles Stupfler[3]

[1] *Université Toulouse Capitole, TSE, 1 Esplanade de l'Université, 31000 Toulouse, France,* abdelaati.daouia@tse-fr.eu

[2] *Université Toulouse Capitole, TSE, 1 Esplanade de l'Université, 31000 Toulouse, France,* joseph.hachem@tse-fr.eu

[3] *Université d'Angers, CNRS, LAREMA, SFR MATHSTIC, F-49000 Angers, France,* gilles.stupfler@univ-angers.fr

Résumé. En analyse des valeurs extrêmes, il a été récemment montré qu'on peut utiliser une technique de dé-randomisation, consistant à remplacer un seuil aléatoire dans l'estimateur d'intérêt par son homologue déterministe, afin d'estimer simultanément plusieurs risques extrêmes, mais seulement pour des données i.i.d.. Dans cet exposé, nous montrerons comment cette méthode peut être utilisée pour estimer plusieurs quantités extrêmes (indice de queue, expected shortfalls...) dans des contextes généraux de données dépendantes/hétéroscédastiques/hétérogènes, sous une hypothèse L^1 pondérée sur l'écart entre la loi moyenne des données et la loi dominante. Cette technique peut également être utilisée pour traiter des données hétérogènes multivariées, ce que la littérature actuelle ne permet pas de faire.

Mots-clés. Valeurs extrêmes, dé-randomisation, modèles à queue lourde, estimateur de Hill, données hétérogènes, statistiques d'ordre.

Abstract. In extreme value analysis, it has recently been shown that one can use a de-randomization trick, replacing a random threshold in the estimator of interest with its deterministic counterpart, in order to estimate several extreme risks simultaneously, but only in an i.i.d. context. In this talk, I will show how this method can be used to handle the estimation of several tail quantities (tail index, expected shortfall...) in general dependence/heteroskedasticity/heterogeneity settings, under a weighted L^1 assumption on the discrepancy between the average distribution of the data and the prevailing distribution. This technique can also be used to deal with multivariate heterogeneous data, which cannot be handled with current methods.

Keywords. Extreme values, de-randomization, heavy tails, Hill estimator, heterogeneity, order statistics.

Summary of the presentation

Let $n \geq 1$, $X_1^{(n)}, \ldots, X_n^{(n)}$ be (almost surely finite) random variables, and denote by $X_{1:n}^{(n)} \leq X_{2:n}^{(n)} \leq \cdots \leq X_{n:n}^{(n)}$ their order statistics. The original motivation for this work is the analysis

of the asymptotic behavior of the quantity

$$\widehat{e}_{f,n}(k) = \frac{1}{k}\sum_{i=1}^{k} f(X^{(n)}_{n-i+1:n}) - f(X^{(n)}_{n-k:n})$$

with $k(n) = k \to \infty$ such that $k/n \to 0$. In the special situation when the $X_i^{(n)}$ have the same distribution as a random variable X having quantile function q, the quantity $\widehat{e}_{f,n}(k)$ is a natural estimator of $\mathbb{E}(f(X) - f(q(1-k/n)) \,|\, X > q(1-k/n))$. This quantity is the mean excess value of $f(X)$ when $X > q(1-k/n)$, which motivates the name *mean f-excess*, and we will call $\widehat{e}_{f,n}(k)$ the *empirical mean f-excess* above the order statistic $X^{(n)}_{n-k:n}$ throughout. Prominent among these quantities is the one obtained with $f = \log$,

$$\widehat{e}_{\log,n}(k) = \frac{1}{k}\sum_{i=1}^{k} \log X^{(n)}_{n-i+1:n} - \log X^{(n)}_{n-k:n},$$

which is reminiscent of the Hill [1975] estimator of a positive extreme value index for heavy-tailed distributions. The behavior of the empirical mean f-excess is hard to assess in a non-i.i.d. scenario due to its formulation in terms of order statistics. It is however well understood in the i.i.d. case, as seen in Stupfler [2019], using a de-randomization technique to write $\widehat{e}_{f,n}(k)$ as the sum of a quotient of random sums of i.i.d. terms and a negligible remainder; the proof involves a Gaussian approximation of the underlying empirical process $\widehat{F}_n$. A recent advance has been made in Einmahl and He [2023] for the Hill estimator for independent heterogeneous data, using the empirical process theory. The main difficulty of this approach is that, on one hand, it does not allow for deriving the asymptotic bias of the Hill estimator and, on the other hand, it cannot be easily adapted to multivariate or dependent data, and that the assumptions made require some sort of uniform boundedness of the mean survival function of the sample, restricting the scope of such a technique.
To address these issues in a general framework, we shall adapt the de-randomization technique of Stupfler [2019]. First, we will show, under the hypothesis that there exist a positive sequence x_n (which will typically be a quantile) tending to infinity and constants $c_1, c_2 > 0$ and $c_3 \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \frac{\frac{1}{k}\sum_{i=1}^{n} \mathbb{E}((f(X_i^{(n)}) - f(x_n))\mathbb{1}\{X_i^{(n)} > x_n\})}{x_n f'(x_n)} = c_1 \tag{1}$$

$$\text{and } \forall t \in \mathbb{R},\ \lim_{n\to\infty} \sqrt{k}\left(\frac{k/n}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X_i^{(n)} > (1+t/\sqrt{k})x_n)} - 1\right) = c_2 t + c_3, \tag{2}$$

that the asymptotic behavior of the empirical mean f-excess is very closely linked to the joint asymptotic behavior of its de-randomized counterpart

$$\overline{e}_{f,n}(k) = \frac{1}{k}\sum_{i=1}^{n} (f(X_i^{(n)}) - f(x_n))\mathbb{1}\{X_i^{(n)} > x_n\},$$

and, for $t \in \mathbb{R}$ fixed,

$$\widehat{\overline{F}}_n((1+t/\sqrt{k})x_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i^{(n)} > (1+t/\sqrt{k})x_n\},$$

which are way easier to handle in a general scenario, including the multivariate or dependent cases, which cannot be handled with current techniques; more specifically, (1) means that $\mathbb{E}(\overline{e}_{f,n}(k))$ has an appropriate asymptotic behavior, and (2) means that $X^{(n)}_{n-k:n,n}$ can basically be replaced with $\widehat{F}_n((1+t/\sqrt{k})x_n)$ for any $t \in \mathbb{R}$, using an argument inspired by Koenker [2005]. Then, we will show that these two estimators are jointly asymptotically normal in a broad heterogeneous scenario for independent data with simple hypotheses, assuming that there exist a distribution $\mathbb{P}_X$ of a heavy-tailed random variable X and $\gamma > 0$ such that, for any sequence (x_n) tending to infinity with $n\mathbb{P}(X > x_n) \to \infty$ (for example, $x_n = q(1-k/n)$ with q the quantile function of X), and for all $u > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{P}(X_i^{(n)} > x_n u)}{\mathbb{P}(X > x_n)} = u^{-1/\gamma}.$$

The proof of the joint asymptotic normality of the above two estimators relies on elementary tools of probability theory such as the Lyapounov CLT and the Cramér-Wold device. It also uses elementary techniques in extreme value analysis such as the Potter bounds and Drees inequality, as seen in de Haan and Ferreira [2006], without resorting to Gaussian approximations. As a result, this approach can be used for multivariate data, and could be adapted for dependent data as well.

Acknowledgments

This research was supported by the French National Research Agency under the grants ANR-19-CE40-0013 (ExtremReg project), ANR-18-EURE-0023 (EUR MINT), ANR-17-EURE-0010 (EUR CHESS) and ANR-11-LABX-0020-01 (Centre Henri Lebesgue). A. Daouia and G. Stupfler acknowledge financial support from the TSE-HEC ACPR Chair.

References

L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, 2006.

J. H. J. Einmahl and Y. He. Extreme value inference for heterogeneous power law data. *Annals of Statistics*, to appear, 2023.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174, 1975.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

G. Stupfler. On a relationship between randomly and non-randomly thresholded empirical average excesses for heavy tails. *Extremes*, 22(4):749–769, 2019.