Available online at www.sciencedirect.com
SCIENCE DIRECT®

C. R. Acad. Sci. Paris, Ser. I 338 (2004) 493–498

COMPTES RENDUS
MATHEMATIQUE

Statistics/Probability Theory

A new extreme quantile estimator for heavy-tailed distributions

Amélie Fils, Armelle Guillou

Université Paris VI, Laboratoire de statistique théorique et appliquée, boîte 158, 175, rue du Chevaleret, 75013 Paris, France

Received 30 October 2003; accepted after revision 20 January 2004

Presented by Paul Deheuvels

Abstract

The classical estimation method for extreme quantiles of heavy-tailed distributions was presented by Weissman (J. Amer. Statist. Assoc. 73 (1978) 812–815) and makes use of the Hill estimator (Ann. Statist. 3 (1975) 1163–1174) for the positive extreme value index. This index estimator can be interpreted as an estimator of the slope in the Pareto quantile plot in case one considers regression lines passing through a fixed anchor point. In this Note we propose a new extreme quantile estimator based on an unconstrained least squares estimator of the index, introduced by Kratz and Resnick (Comm. Statist. Stochastic Models 12 (1996) 699–724) and Schultze and Steinebach (Statist. Decisions 14 (1996) 353–372) and we study its asymptotic behavior. ***To cite this article: A. Fils, A. Guillou, C. R. Acad. Sci. Paris, Ser. I 338 (2004).***

Résumé

Un nouvel estimateur des quantiles extrêmes pour des distributions à queues lourdes. La méthode classique d'estimation de quantiles extrêmes dans le cas de distributions à queues lourdes a été introduite par Weissman (J. Amer. Statist. Assoc. 73 (1978) 812–815) et fait usage de l'estimateur de Hill (Ann. Statist. 3 (1975) 1163–1174) comme estimateur de l'index positif des valeurs extrêmes. Cet estimateur de l'index peut être interprété comme un estimateur de la pente dans le « Pareto quantile plot » dans le cas où on considère une régression linéaire passant par un point fixe. Dans cette Note nous proposons un nouvel estimateur des quantiles extrêmes basé sur un estimateur des moindres carrés classique de l'index, qui a été introduit par Kratz et Resnick (Comm. Statist. Stochastic Models 12 (1996) 699–724) et Schultze et Steinebach (Statist. Decisions 14 (1996) 353–372) et nous étudions son comportement asymptotique. ***Pour citer cet article : A. Fils, A. Guillou, C. R. Acad. Sci. Paris, Ser. I 338 (2004).***

Version française abrégée

Nous considérons un échantillon $X_1, \ldots, X_n$ de variables aléatoires positives de loi F. Nous supposons F de type Pareto, ce qui revient à supposer l'existence d'une constante positive γ telle que

$$1 - F(x) = x^{-1/\gamma} \ell_F(x),$$

E-mail addresses: fils@ccr.jussieu.fr (A. Fils), guillou@ccr.jussieu.fr (A. Guillou).

1631-073X/$ – see front matter © 2004 Académie des sciences. Published by Elsevier SAS. All rights reserved.
doi:10.1016/j.crma.2004.01.019

où $\ell_F(x)$ est une fonction à variations lentes à l'infini. L'estimateur le plus connu de γ est alors l'estimateur de Hill (voir [4]). Celui-ci, ainsi que beaucoup d'autres proposés dans la littérature, peuvent être interprétés comme des estimateurs de la pente dans le « Pareto quantile plot » correspondant au graphe de $(\log(\frac{n+1}{j}), \log X_{n-j+1,n})$, $j = 1, \ldots, n$. En effet, dans le cas de distributions de type Pareto ce graphe sera approximativement linéaire de pente γ pour les points extrêmes. On peut donc obtenir différents estimateurs de l'index à partir de ce graphe suivant le type de régression utilisée au-delà d'un seuil $X_{n-k,n}$. Une régression linéaire peut être faite en forçant la droite à passer par un point fixe. On obtient alors l'estimateur de Hill et par extrapolation le long de cette droite, on obtient l'estimateur de Weissman (voir [7]) comme estimateur des quantiles extrêmes. Une autre façon d'aborder le problème consiste à faire une régression des moindres carrés classique basée sur les k plus hauts points dans le « Pareto quantile plot ». On aboutit alors à l'estimateur de l'index proposé par Kratz et Resnick [5] et Schultze et Steinebach [6]. Dans cet article, nous proposons un nouvel estimateur d'un quantile extrême obtenu par extrapolation le long de cette nouvelle droite et nous étudions son comportement asymptotique.

1. Introduction and statement of the result

Let $X_1, \ldots, X_n$ be a sample of n independent and identically distributed (i.i.d.) random variables according to some continuous distribution function F. Our concern is how to obtain with such a limited sample a good estimate for a quantile

$$Q(1-p) = \inf\{y\colon\ F(y) \geqslant 1-p\},$$

where p is small, such that the quantile to be estimated is situated on the border of or beyond the range of the data. Estimating such high quantiles is directly linked to the accurate modelling of the tail of the distribution

$$\overline{F}(x) := \mathbb{P}(X > x)$$

for large thresholds x.

In extreme value theory, the behaviour of such extreme quantiles is known to be governed by one crucial parameter of the underlying distribution, named the extreme value index (EVI). The class with EVI strictly larger than 0 encompasses the so-called "heavy-tailed" distributions, whose tails decay essentially as a negative power function only. For risk analyses, these Pareto-type distributions are unreliable, as they bring on considerably higher quantiles than one would expect when using classical approximations under the assumption of normality, for instance. Therefore, we have decided to focus on this class and consequently we suppose that there exists a positive constant γ for which

$$1 - F(x) = x^{-1/\gamma}\ell_F(x), \tag{1}$$

where $\ell_F(x)$ is a slowly varying function at infinity satisfying

$$\frac{\ell_F(\lambda x)}{\ell_F(x)} \longrightarrow 1, \quad \text{as } x \to \infty \text{ for all } \lambda > 0.$$

The present model is well known to be equivalent to the following model

$$U(x) = x^{\gamma}\ell_U(x), \tag{2}$$

where $U(x) = Q(1 - 1/x)$, $x > 1$, and $\ell_U(x)$ is a slowly varying function.

A prominent estimator for the EVI of distributions with heavy-tails is the Hill estimator (see [4]) defined as:

$$H_{k,n} = \frac{1}{k}\sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n},$$

where $X_{j,n}$, $j=1,\dots,n$, denotes the order statistics based on the observations, and $k=k(n)$ are positive integers which can satisfy some conditions in theoretical asymptotic considerations.

It has been mentioned in the literature (see [1]) that this and many other estimators can be viewed as estimators of the slope in a Pareto quantile plot defined as the graph with coordinates

$$\left(\log\left(\frac{n+1}{j}\right), \log X_{n-j+1,n}\right), \quad j=1,\dots,n, \tag{3}$$

where $\log\frac{n+1}{j}$ is an approximation of $-\log(1-F_n(X_{n-j+1,n}))$, with F_n the classical empirical distribution function. Indeed, for Pareto-type distributions it can be expected that the Pareto quantile plot is approximately linear with slope γ only for the extreme values. Hence in the case of $\gamma>0$, a large group of estimators of γ evolves from different possible regression fits on Pareto quantile plots above specific thresholds levels $X_{n-k,n}$. The linear regression can be done either in a constrained way (considering the threshold point fixed, as an anchor point) or in an unconstrained way. For instance, the Hill estimator can be interpreted as an estimator of the slope when considering regression lines passing through a fixed anchor point. Then, turning to high quantiles, it is natural to estimate these with the Hill estimator by extrapolation along this fitted line. Defining

$$x_p = Q(1-p), \tag{4}$$

this yields the following well-known estimator (see Weissman [7]):

$$\hat{x}^H_{p,k} := X_{n-k,n}\left(\frac{k+1}{(n+1)p}\right)^{H_{k,n}}, \quad k=1,\dots,n-1. \tag{5}$$

Note that others estimators have been proposed in the literature (see [2] and [3]).

Nevertheless, the slope in the Pareto quantile plot can also be calculated based on an unconstrained linear regression for the k highest points, instead of on a regression with a fixed threshold point. For instance, in this spirit, Kratz and Resnick [5] and Schultze and Steinebach [6] proposed the following estimator for the index:

$$\hat{\gamma}_{k,n} := \frac{\frac{1}{k}\sum_{j=1}^{k}\log\frac{k+1}{j}\log X_{n-j+1,n} - \left(\frac{1}{k}\sum_{j=1}^{k}\log\frac{k+1}{j}\right)\left(\frac{1}{k}\sum_{j=1}^{k}\log X_{n-j+1,n}\right)}{\frac{1}{k}\sum_{j=1}^{k}\left(\log\frac{k+1}{j}\right)^2 - \left(\frac{1}{k}\sum_{j=1}^{k}\log\frac{k+1}{j}\right)^2},$$

which follows from the minimization problem

$$\sum_{j=1}^{k}\left(\log X_{n-j+1,n} - a\log\frac{n+1}{j} - b\right)^2 \overset{!}{=} \min_{a,b}.$$

Then, by extrapolation along the fitted line $y=\hat{a}x+\hat{b}$, we deduce a new (as far as we know) quantile estimator for x_p, namely

$$\hat{x}_{p,k} := p^{-\hat{\gamma}_{k,n}}\exp\left(\frac{1}{k}\sum_{j=1}^{k}\log X_{n-j+1,n} - \frac{\hat{\gamma}_{k,n}}{k}\sum_{j=1}^{k}\log\frac{n+1}{j}\right), \tag{6}$$

for which we study its asymptotic behavior.

For this aim, we need some assumptions, commonly used in the literature. The first concerns the slowly varying function ℓ_U and is given by:

Assumption (R_{ℓ_U}). There exists a real constant $\rho\leqslant 0$ and a positive rate function b satisfying $b(x)\to 0$ as $x\to\infty$, such that for all $\lambda\geqslant 1$, as $x\to\infty$,

$$\log\frac{\ell_U(\lambda x)}{\ell_U(x)} \sim b(x)\int_1^{\lambda} v^{\rho-1}\,\mathrm{d}v.$$

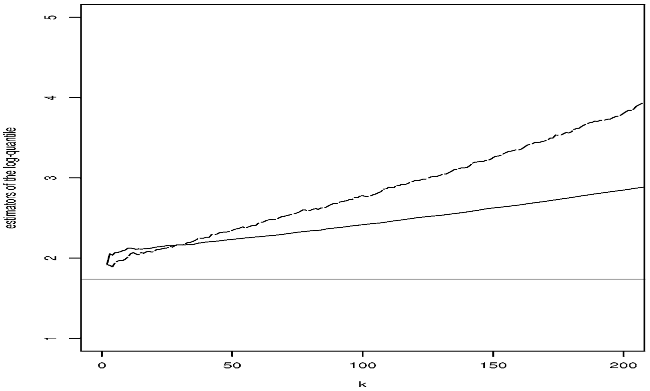

Fig. 1. Median of $\log(\hat{x}^H_{p,k})$ (dashed line) and $\log(\hat{x}_{p,k})$ (full line) based on 100 samples of size 500 from a $|t(10)|$ distribution with $p = \frac{1}{10n}$. The horizontal line indicates the true value of the parameter.

Note that it is equivalent to the assumption (3.2) in Kratz and Resnick [5]. The second condition restricts the growth of $k = k(n)$:

$$k = k(n) \to \infty, \quad \frac{k}{n} \to 0, \quad \text{and} \quad \sqrt{k}\, b\left(\frac{n+1}{k+1}\right) \to 0 \quad \text{as } n \to \infty, \tag{7}$$

with b as defined in Assumption (R_{ℓ_U}), and corresponds to the assumption (3.4) in Kratz and Resnick [5]. The third condition, relative to the parameter p is

$$p = p_n \to 0 \text{ such that } np \to 0 \text{ as } n \to \infty. \tag{8}$$

Now we can state the main result of this paper.

Theorem 1.1. *Under Assumption* (R_{ℓ_U}), (7) *and* (8), *we have*

$$\frac{\sqrt{k}}{\log a_n}(\log \hat{x}_{p,k} - \log x_p) \longrightarrow^d \mathcal{N}\big(0, 2\gamma^2\big), \quad \textit{as } n \to \infty,$$

where $a_n = \frac{k+1}{(n+1)p}$.

Note that the logarithm of the Weissman estimator (5) has a similar asymptotic behavior, but an asymptotic variance of γ^2. However, it exhibits considerable bias in certain circumstances; thus asymptotic variance cannot be the only criterion to compare the quality of estimators. Especially, as can be shown from Assumption (R_{ℓ_U}), estimation problems will arise if ρ is close to 0. This is illustrated in Fig. 1, where we compare the behaviour of $\log(\hat{x}^H_{p,k})$ and $\log(\hat{x}_{p,k})$ in case of a $|t(10)|$ distribution, whose parameter ρ, equal to $-1/5$, is very close to 0.

As can be observed in Fig. 1, the bias of $\log(\hat{x}_{p,k})$ is less important than the one of the Weissman estimator. Moreover, one interesting feature of $\hat{x}_{p,k}$ is the smoothness of the realizations as a function of k, which alleviates the problem of choosing k to some extent. This problem will be considered in some future work.

2. Proof of Theorem 1.1

Let $(U_{j,n})$, respectively $(E_{j,n})$, be the order statistics from an i.i.d. uniform (0,1), respectively standard exponential, sample of size n. By using (2), the definition (4) of x_p, and the fact that $E_{n-j+1,n} =^d \log(U^{-1}_{j,n})$, for $j \leqslant n$, we can write:

$$\begin{aligned}\log \hat{x}_{p,k} - \log x_p &=^d \frac{\gamma}{k}\sum_{j=1}^{k}\left(E_{n-j+1,n} - \log\frac{n+1}{j}\right) - (\hat{\gamma}_{k,n} - \gamma)\left(\frac{1}{k}\sum_{j=1}^{k}\log\frac{n+1}{j} + \log p\right)\\ &\quad + \frac{1}{k}\sum_{j=1}^{k}\log\frac{\ell_U(U_{j,n}^{-1})}{\ell_U(p^{-1})}\\ &=: B_{1,k} + B_{2,k} + B_{3,k}.\end{aligned}$$

We successively discuss each of the terms $B_{j,k}$, for $j = 1, 2, 3$, separately.

First, concerning $B_{1,k}$, we need the Rényi representation of standard exponential order statistics, which states that:

$$E_{n-j+1,n} =^d \sum_{i=j}^{n}\frac{f_{n-i+1}}{i} \quad \text{for } j \leqslant n,$$

where f_i are i.i.d. standard exponential random variables.

From this, we deduce the equality

$$B_{1,k} =^d \frac{\gamma}{k}\sum_{j=1}^{k}\left(\sum_{i=j}^{n}\frac{f_{n-i+1}}{i} - \log\frac{n+1}{j}\right).$$

Note that

$$\sum_{j=1}^{k}\sum_{i=j}^{n}\frac{f_{n-i+1}}{i} = \sum_{i=n-k+1}^{n} f_i + k\sum_{i=k+1}^{n}\frac{f_{n-i+1}}{i},$$

which leads to

$$B_{1,k} =^d \gamma\left(E_{n-k,n} - \log\frac{n+1}{k+1}\right) - \gamma\left(\frac{1}{k}\sum_{j=1}^{k}\log\frac{n+1}{j} - \log\frac{n+1}{k+1}\right) + \frac{\gamma}{k}\sum_{i=1}^{k} f_i. \tag{9}$$

Now by Lemma 2.2 in Kratz and Resnick [5], we get

$$B_{1,k} =^d \gamma\left(E_{n-k,n} - \log\frac{n+1}{k+1}\right) + \gamma\left(\frac{1}{k}\sum_{i=1}^{k} f_i - 1\right) + \mathrm{o}(1). \tag{10}$$

It can be easily proved that

$$\sqrt{k}\left(E_{n-k,n} - \log\frac{n+1}{k+1}\right) \to^d \mathcal{N}(0,1), \tag{11}$$

which implies

$$\frac{\sqrt{k}}{\log a_n} B_{1,k} \longrightarrow^P 0 \tag{12}$$

since $a_n \to \infty$ (a_n being defined in Theorem 1.1).

Next, the third term $B_{3,k}$ can be rewritten as follows:

$$B_{3,k} = \frac{1}{k}\sum_{j=1}^{k}\left(\log\frac{\ell_U\left(\frac{n+1}{k+1}\right)}{\ell_U\left(\frac{k+1}{(n+1)p}\frac{n+1}{k+1}\right)} + \log\frac{\ell_U\left(\frac{n+1}{k+1}U_{j,n}^{-1}\frac{k+1}{n+1}\right)}{\ell_U\left(\frac{n+1}{k+1}\right)}\right).$$

Using Assumption (R_{ℓ_U}) together with the fact that $\frac{n+1}{k+1}U_{k+1,n} \to^P 1$, and $\sqrt{k}\, b\big(\frac{n+1}{k+1}\big) \to 0$ as $n \to \infty$, we derive that

$$\frac{\sqrt{k}}{\log a_n} B_{3,k} \longrightarrow^P 0. \tag{13}$$

Finally, both Lemma 2.2 and Theorem 3.1 in Kratz and Resnick [5] imply

$$\frac{\sqrt{k}}{\log a_n} B_{2,k} \longrightarrow \mathcal{N}\big(0, 2\gamma^2\big), \tag{14}$$

under our assumptions. Combining (12)–(14) provides Theorem 1.1. □

Acknowledgements

The authors are very grateful to the referees for careful reading of the first version of this Note; these comments lead to significant improvements.

References

[1] S. Csörgő, L. Viharos, Estimating the Tail Index, Asymptotic Methods in Probability and Statistics, Elsevier, Amsterdam, 1998.

[2] A.L.M. Dekkers, L. de Haan, On the estimation of the extreme-value index and large quantile estimation, Ann. Statist. 17 (1989) 1795–1832.

[3] L. de Haan, H. Rootzén, On the estimation of high quantiles, J. Statist. Plann. Inference 35 (1993) 1–13.

[4] B.M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Statist. 3 (1975) 1163–1174.

[5] M. Kratz, S. Resnick, The qq-estimator and heavy tails, Comm. Statist. Stochastic Models 12 (1996) 699–724.

[6] J. Schultze, J. Steinebach, On least squares estimates of an exponential tail coefficient, Statist. Decisions 14 (1996) 353–372.

[7] I. Weissman, Estimation of parameters and large quantiles based on the k largest observations, J. Amer. Statist. Assoc. 73 (1978) 812–815.