Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 339 (2004) 287–292
COMPTES RENDUS MATHEMATIQUE

Statistics/Probability Theory

Extreme quantiles estimation for actuarial applications

Emmanuel Delafosse, Armelle Guillou

Université Paris VI, L.S.T.A., boîte 158, 175, rue du Chevaleret, 75013 Paris, France

Received 11 March 2004; accepted after revision 7 June 2004

Available online 23 July 2004

Presented by Paul Deheuvels

Abstract

This paper is devoted to the estimation of tail index and extreme quantiles in actuarial applications. In this domain, the observations are often censored. Nevertheless, conversely to the classical randomly right-censored model, the censoring variables are always observed. Therefore, under this assumption, we introduce new estimators and we study their asymptotic properties. Their behaviour are illustrated in a small simulation study. ***To cite this article: E. Delafosse, A. Guillou, C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

Résumé

Estimation des quantiles extrêmes en actuariat. Ce papier concerne l'estimation des indices de queues et des quantiles extrêmes dans des applications actuarielles. Dans ce domaine, les observations sont souvent censurées. Néanmoins, contrairement au modèle classique de censure aléatoire à droite, les données censurantes sont toujours observées. Sous cette condition, nous introduisons de nouveaux estimateurs et nous étudions leurs propriétés asymptotiques. Leur comportement est illustré sur la base de simulations. ***Pour citer cet article : E. Delafosse, A. Guillou, C. R. Acad. Sci. Paris, Ser. I 339 (2004).***

Version française abrégée

Nous considérons des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) $X_1, \ldots, X_n$, de fonction de répartition F de type Pareto, c'est-à-dire telle qu'il existe une constante strictement positive γ_1, appelée indice des valeurs extrêmes, vérifiant

$$1 - F(x) = x^{-1/\gamma_1} \ell_1(x),$$

E-mail address: guillou@ccr.jussieu.fr (A. Guillou).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2004.06.005

où ℓ_1 est une fonction à variations lentes à l'infini. Nous supposons que ces données sont censurées par des variables aléatoires i.i.d. $Y_1, \ldots, Y_n$ de fonction de répartition G, également de type Pareto. Contrairement au modèle de censure aléatoire classique, ou seuls les couples $(Z_i = X_i \wedge Y_i, \delta_i = \mathbb{1}_{X_i \leqslant Y_i})$, sont observés, dans les applications actuarielles les variables Y_i le sont aussi. Par conséquent, nous proposons dans cette note, de tenir compte de cette information additionnelle, dans le but de construire un estimateur de γ_1, qui nous servira par la suite à obtenir un estimateur de quantiles extrêmes.

En s'inspirant du cas non censuré, où des estimateurs d'indice peuvent être interprétés comme des estimateurs de la pente d'un « Pareto quantile plot », nous proposons l'estimateur suivant de γ_1 :

$$H^{(c)}_{Z,Y,k,n} = \frac{\frac{1}{k}\sum_{j=1}^{k} \log Z_{n-j+1,n} - \log Z_{n-k,n}}{-\frac{1}{k}\sum_{j=1}^{k} \log(1-\widehat{F}_n(Z_{n-j+1,n})) + \log(1-\widehat{F}_n(Z_{n-k,n}))}, \quad k = 1, \ldots, n-1,$$

où $Z_{1,n}, \ldots, Z_{n,n}$ désigne la statistique d'ordre et, en notant G_n et H_n les fonctions de répartition empiriques des Y_i et des Z_i respectivement, $\widehat{F}_n(x) = 1 - \frac{1-H_n(x)}{1-G_n(x)}$.

Par extrapolation le long de la droite du « Pareto quantile plot », nous en déduisons un estimateur des quantiles extrêmes.

Puis, par le biais de simulations, nous comparons le comportement de notre nouvel estimateur $H^{(c)}_{Z,Y,k,n}$ à celui obtenu par l'approche POT (« Peaks-Over-Threshold ») basée sur les excès, adapté au préalable à la censure comme dans Davison et Smith [2]. Compte tenu des bonnes performances de ce nouvel estimateur de γ_1, nous nous concentrons sur celui-ci et établissons sa normalité asymptotique ainsi que le comportement limite de l'estimateur des quantiles extrêmes associé.

1. Introduction and statement of results

Let $X_1, \ldots, X_n$ be a sequence of nonnegative i.i.d. random variables with distribution function F. Along with the X-sequence, let $Y_1, \ldots, Y_n$ be a sample of i.i.d. censoring random variables with arbitrary distribution function G and also being independent of the X's. In the randomly right-censored model, we only observe the variables $Z_i = X_i \wedge Y_i$ together with $\delta_i = \mathbb{1}_{X_i \leqslant Y_i}$, $i \in \mathbb{N}$, the indicator of no-censoring.

Nevertheless, on many occasions in insurance, the censoring values Y are also known. Indeed, in insurance, the reported payments cannot be larger than the maximum payment value of the contract. When the reported payment equals the maximum payment, this real payment can indeed be equal to the maximum or can be censored. Therefore, we propose here to take into account this additional information, and consequently, we introduce new estimators in the case (Z, δ, Y) are observed and we study their asymptotic properties.

Being motivated by actuarial applications we confine ourselves to the case where sample maxima from X samples are in the domain of attraction of the Fréchet law. That means that we suppose that there exists a strictly positive constant γ_1 for which

$$1 - F(x) = x^{-1/\gamma_1}\ell_1(x),$$

where ℓ_1 is a slowly varying function at infinity satisfying

$$\frac{\ell_1(\lambda x)}{\ell_1(x)} \to 1 \quad \text{as } x \to \infty, \text{ for all } \lambda > 0.$$

The present model is well-known to be equivalent to the model $U_F(x) = x^{\gamma_1}\ell_{1,U}(x)$, where $U_F(x) = \inf\{y\colon F(y) \geqslant 1 - 1/x\}$, $x > 1$, and $\ell_{1,U}(x)$ again a slowly varying function at infinity.

Moreover, in order for the censoring to be not too heavy, it appears natural to assume that the censoring distribution G is also heavy tailed

$$1 - G(x) = x^{-1/\gamma_2}\ell_2(x),$$

for some $\gamma_2 > 0$ and ℓ_2 a slowly varying function. Under the assumption that X and Y are independent, we have

$$1 - H(x) = x^{-(1/\gamma_1+1/\gamma_2)}\ell_H(x) \quad \text{or equivalently} \quad U_H(x) = x^{\gamma_1\gamma_2/(\gamma_1+\gamma_2)}\ell_U(x), \tag{1}$$

with ℓ_H and ℓ_U also slowly varying functions at infinity.

From (1), it is clear that the classical Hill estimator $H_{Z,k,n}$ (see Hill [4]) defined as

$$H_{Z,k,n} = \frac{1}{k}\sum_{j=1}^{k} \log Z_{n-j+1,n} - \log Z_{n-k,n}, \quad k = 1, \ldots, n-1, \tag{2}$$

is not consistent for γ_1 but for $\gamma_1\gamma_2/(\gamma_1+\gamma_2)$.

Therefore, we have to modify this estimator in the censoring context. Recall that one of the interpretation of the Hill estimator is the fact that it can be viewed as a trivial slope estimator in the Pareto quantile plot, defined, in the case of no-censoring, as the plot with coordinates $(\log\frac{n+1}{j}, \log X_{n-j+1,n})$, which is an approximation of $(-\log(1 - F_n(X_{n-j+1,n})), \log X_{n-j+1,n})$, where F_n is the classical empirical distribution function. Therefore, taking into account the fact that Y are observed, we propose to replace $1 - F_n$ (which is unknown in our case) by $1 - \widehat{F}_n(x) := \frac{1-H_n(x)}{1-G_n(x)}$, where H_n and G_n are also the empirical distribution functions associated to the Z and Y samples. Then, the trivial slope estimator can be expressed as

$$H^{(c)}_{Z,Y,k,n} = \frac{\frac{1}{k}\sum_{j=1}^{k} \log Z_{n-j+1,n} - \log Z_{n-k,n}}{-\frac{1}{k}\sum_{j=1}^{k} \log(1 - \widehat{F}_n(Z_{n-j+1,n})) + \log(1 - \widehat{F}_n(Z_{n-k,n}))}. \tag{3}$$

In many applications, however, this estimation is only an intermediate goal. Indeed, what we would like to do, is, for instance, to estimate an extreme quantile. With this aim, the classical approach consists to extrapolate the Pareto quantile plot along the fitted line, which leads to the following extreme quantile estimator for $x_{F,p} = U_F(1/p)$, $p < 1/n$:

$$\hat{x}^{(c)}_{p,Y,k,n} = Z_{n-k,n}\left(\frac{1 - H_n(Z_{n-k,n})}{p(1 - G_n(Z_{n-k,n}))}\right)^{H^{(c)}_{Z,Y,k,n}}. \tag{4}$$

This estimator is similar to the one proposed by Weissman (see Weissman [6]) in the case of no-censoring.

Another approach, for the estimation of γ_1, consists to use the *maximum likelihood method* based on POT's and on the results given by Balkema and de Haan [1] and Pickands [5], stating that the limit distribution of the exceedances over a threshold t when $t \to \infty$ is given by a generalized Pareto distribution (GPD). In the case of censoring, we can easily adapt the likelihood as follows (where N_t denotes the number of excesses over t)

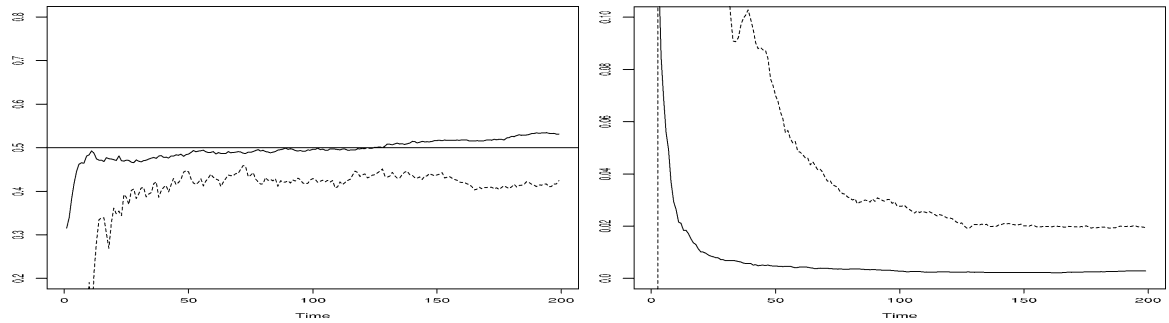
$$\prod_{j=1}^{N_t} \big[f_{\mathrm{GPD}}(E_j)\big]^{\delta_j}\big[1 - F_{\mathrm{GPD}}(E_j)\big]^{1-\delta_j},$$

where $E_j = Z_j - t$ if $Z_j > t$ and $1 - F_{\mathrm{GPD}}(x) = (1 + \frac{\gamma_1 x}{\sigma})^{-1/\gamma_1}$ (see Davison and Smith [2] for further details). Then the maximization of this expression leads to a POT estimator for γ_1 which we further denote by $\hat{\gamma}^{(c)}_{t,ML}$.

The aim of the next section is to compare, in a small simulation study, the behaviour of the two estimators $H^{(c)}_{Z,Y,k,n}$ and $\hat{\gamma}^{(c)}_{Z_{n-k,n},ML}$. Then, we state our main theoretical results. The proofs have been deposited with the Service des Comptes rendus, and maybe consulted by request.

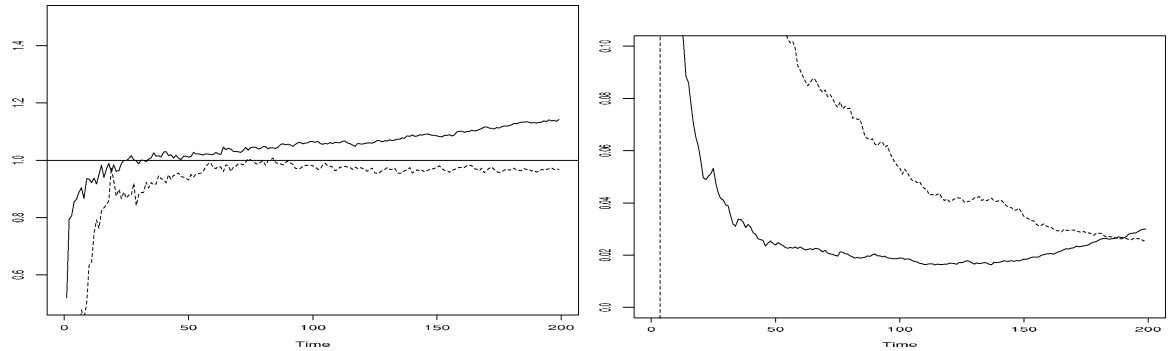
2. Simulation study

To illustrate the behaviour of the new estimators of γ_1, we report the results of a small simulation study. The distributions which we included are:



(a) (b)

Fig. 1. (a) Medians and (b) empirical MSE's of the estimators $H^{(c)}_{Z,Y,k,n}$ (full line) and $\hat{\gamma}^{(c)}_{Z_{n-k,n},ML}$ (dashed line) as a function of k obtained from 100 simulated samples of length 500 from a Burr$(1, 4, \frac{1}{2})$ distribution censored by a Burr$(10, 1, \frac{1}{2})$ distribution.



(a) (b)

Fig. 2. (a) Medians and (b) empirical MSE's of the estimators $H^{(c)}_{Z,Y,k,n}$ (full line) and $\hat{\gamma}^{(c)}_{Z_{n-k,n},ML}$ (dashed line) as a function of k obtained from 100 simulated samples of length 500 from a Fréchet(1) distribution censored by a Burr$(10, \frac{1}{2}, 1)$ distribution.

– a Burr$(1, 4, \frac{1}{2})$ distribution censored by a Burr$(10, 1, \frac{1}{2})$ distribution (Fig. 1),
– a Fréchet(1) distribution censored by a Burr$(10, \frac{1}{2}$,1) distribution (Fig. 2).

Note that a Burr(β, τ, λ) distribution is defined by $F(x) = 1 - [\beta/(\beta + x^{\tau})]^{\lambda}$, in which case the extreme value index γ_1 is equal to $1/(\lambda\tau)$ and a Fréchet(1) distribution is defined as $F(x) = \exp(-x^{-1})$ with an extreme value index 1.

From each of the above distributions, 100 random samples of length $n = 500$ were generated and for each sample, we compare the estimator $H^{(c)}_{Z,Y,k,n}$ (full line) with the estimator $\hat{\gamma}^{(c)}_{Z_{n-k,n},ML}$ (dashed line). Figs. 1 and 2(a) represent the medians as a function of k and Figs. 1 and 2(b) the empirical mean squared errors (MSE). The horizontal line indicates always the true value of the parameter.

We can observe the good behaviour of the estimator $H^{(c)}_{Z,Y,k,n}$ in these two examples, but also in all the other simulations that we tried. Consequently, we focus, in the next section, on this estimator, for which we establish its asymptotic normality. From this result, we then deduce the asymptotic behaviour of $\hat{x}^{(c)}_{p,Y,k,n}$.

3. Asymptotic results

Since the model (1) cannot be studied in its full generality, we impose the following classical condition on the slowly varying function ℓ_U:

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

$$\log \frac{\ell_U(\lambda x)}{\ell_U(x)} \sim b_U(x) \int_1^\lambda x^{\rho_U - 1} \mathrm{d}x.$$

Under this assumption, which holds for most common heavy-tailed distributions, we establish the asymptotic normality of all our estimators. Before stating these theorems, we would like to mention that in Hall [3] this assumption was specified further for a subclass of Pareto-type distributions, often referred to as the Hall class. For this class, ℓ_1 and ℓ_2 are determined in such a way that for some constants C_i and $D_i \in \mathbb{R}$, $\rho_i < 0$, $i = 1, 2$,

$$\ell_i(x) = C_i\big[1 + D_i x^{\rho_i/\gamma_i}\big(1 + \mathrm{o}(1)\big)\big], \quad \text{as } x \to \infty,$$

which implies

$$\ell_U(x) = (C_1 C_2)^{\gamma_1\gamma_2/(\gamma_1+\gamma_2)}\left[1 + \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} D (C_1 C_2 x)^{(\max(\gamma_1\rho_2, \gamma_2\rho_1))/(\gamma_1+\gamma_2)}\big(1 + \mathrm{o}(1)\big)\right],$$

for a suitable constant D. Note that, under such a model, $b(x) = Cx^\rho$ with $\rho = \frac{\max(\gamma_1\rho_2, \gamma_2\rho_1)}{\gamma_1+\gamma_2}$.

Well-known examples in the Hall class are the Student t, the Fréchet and the Burr distributions, among others.

The asymptotic normality result of our estimator $H^{(c)}_{Z,Y,k,n}$ now reads as follows:

Theorem 3.1. *Suppose assumption* $(R_\ell(\rho, b))$ *holds for all the slowly varying functions* ℓ *considered. Whenever* $k \to \infty$, $\max(\frac{1}{k}, (\frac{k}{n})^{\gamma_2/(\gamma_1+\gamma_2)}) \log\log n \to 0$ *and* $\sqrt{k}\, b_U(\frac{n+1}{k+1}) \to 0$ *as* $n \to \infty$, *we have*

$$\sqrt{k}\big(H^{(c)}_{Z,Y,k,n} - \gamma_1\big) \longrightarrow^d N\left(0, \gamma_1^2\left(\frac{\gamma_1+\gamma_2}{\gamma_2}\right)^2\right).$$

Remark that the asymptotic variance is always larger than that of the original Hill estimator which equals, in the case of no-censoring, to γ_1^2. This case now appears as a limit case taking $\gamma_2 \to \infty$. Remark that the smaller γ_2 (i.e. the heavier the censoring) the higher the price to pay in variance.

The estimator $H^{(c)}_{Z,Y,k,n}$ figures in the power of the expression of the quantile estimator defined in (4), and hence the finite sample distribution of $\hat{x}^{(c)}_{p,Y,k,n}$ will rather be lognormal than normal for small samples. Hence we always work below with $\log \hat{x}^{(c)}_{p,Y,k,n}$. The asymptotic behaviour of the quantile estimator $\log \hat{x}^{(c)}_{p,Y,k,n}$ is stated in the following theorem under the assumptions of Theorem 3.1 and

$$p = p_n \quad \text{such that} \quad \frac{k}{np} \to \infty. \tag{5}$$

In the sequel, we use the notation $a_{k,n} = \frac{k}{np(1-G_n(Z_{n-k,n}))}$. Note that under the assumption (5), $a_{k,n} \to \infty$.

Theorem 3.2. *Suppose that the assumptions of Theorem* 3.1 *hold. Then, if* (5) *is satisfied, we have*

$$\frac{\sqrt{k}}{\log a_{k,n}}\big(\log \hat{x}^{(c)}_{p,Y,k,n} - \log x_{F,p}\big) \longrightarrow^d N\left(0, \gamma_1^2\left(\frac{\gamma_1+\gamma_2}{\gamma_2}\right)^2\right).$$

Acknowledgement

The authors would like to thank Jan Beirlant and Patrick Leveillard (actuaries responsible for AGF) for valuable suggestions.

References

[1] A. Balkema, L. de Haan, Residual life time at great age, Ann. Probab. 2 (1974) 792–804.
[2] A.C. Davison, R.L. Smith, Models for exceedances over high thresholds, J. Roy. Statist. Soc. Ser. B 52 (1990) 393–442.
[3] P. Hall, On some simple estimates of an exponent of regular variation, J. Roy. Statist. Soc. Ser. B 44 (1982) 37–42.
[4] B.M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Statist. 3 (1975) 1163–1174.
[5] J. Pickands III, Statistical inference using extreme order statistics, Ann. Statist. 3 (1975) 119–131.
[6] I. Weissman, Estimation of parameters and large quantiles based on the k largest observations, J. Am. Statist. Assoc. 73 (1978) 812–815.