Statistics of Extremes under Censoring Schemes

M. Ivette Gomes
Universidade de Lisboa, Faculdade de Ciências, DEIO e CEAUL
e-mail: ivette.gomes@fc.ul.pt

Abstract

Statistics of Extremes deals essentially with the estimation of parameters of extreme events like a *high quantile* situated in the border or even beyond the range of the available data. The most common assumption on any set of univariate data, $(X_1, X_2, \cdots, X_n)$, is to consider that we are in the presence of a *complete* sample of size n, either indepently, identically distributed or weakly dependent and stationary, from an unknown distribution function $F = F_X$. In all applications of extreme value theory, the estimation of the *extreme value index* is of primordial importance. Such a parameter measures the heaviness of the tail and has been widely studied in the literature. It is indeed the real parameter γ in the extreme value distribution function $EV_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}),\ 1+\gamma x > 0$. This distribution function appears as the limiting distribution of the maximum, linearly normalized, whenever such a non-degenerate limit does exist. For heavy tails, i.e. whenever $\gamma > 0$, we mention the classical *Hill* estimator (Hill, 1975) and one of the most recent *minimum-variance reduced-bias* (MVRB) estimators of the extreme value index (Caeiro et al., 2005) and of extreme quantiles (Gomes and Pestana, 2007). For a general extreme value index estimation, we mention the *moment* (Dekkers et al., 1989), the "*maximum likelihood*" (Smith, 1987; Drees et al., 2004), the *generalized Hill* (Beirlant et al., 1996) and the *mixed moment* (Fraga Alves et al., 2007) estimators. In all these papers the available sample is complete. However, in the analysis of lifetime data or reliability data, observations are usually censored. For simplicity we shall assume first the case of right censorship, where no difficulties appear, following immediately to the case of random censorship, where apart from a recent paper by Einmahl et al. (2008), there is only, as far as we know, a brief reference to the topic in Reiss and Thomas (1997, Section 6.1) and a paper by Beirlant et al. (2007). We shall give here special attention to the estimation of γ, as well as associated high quantile and right endpoint estimation under random censoring, i.e. we shall assume that there is a random variable Y such that only $Z = X \wedge Y$ and $\delta = I_{\{X \le Y\}}$ are observed. The indicator variable δ determines whether X has been censored or not. Consequently, we have access to the random sample (Z_i, δ_i), $1 \le i \le n$, of independent copies of (Z, δ), but our goal is to make inference on the right tail of the unknown lifetime distribution, i.e. on $\overline{F}_X(x) := P(X > x) = 1 - F_X(x)$, while F_Y, the distribution function of Y, is considered to be a nonparametric nuisance parameter. As mentioned in Einmahl *et al.* (2008), all the extreme value index estimators need to be

slightly modified in order to be consistent for the estimation of γ. A possible and simple modification is suggested in Einmahl *et al.* (2008): replace $\widehat{\gamma}(k|Z)$ by $\widehat{\gamma}(k|X) = \widehat{\gamma}(k|Z)/\widehat{p}$, with $\widehat{p} = \frac{1}{k}\sum_{j=1}^{k}\delta_{[n-j+1]}$, with $\delta_{[j]}$ the concomitant value of δ associated with $Z_{j:n}$, the j-th ascending order statistic, $1 \leq j \leq n$, associated with the observed sample $(Z_1, Z_2, \cdots, Z_n)$. We shall here apply the methodology in Einmahl *et al.* (2008) to a few sets of survival data, available in the literature, as well as simulated data, providing some hints for the adequate estimation of the extreme value index, high quantiles and right endpoint of X.

Keywords and phrases. Extreme value index, censoring, extreme quantiles

References

[1] Beirlant, J., Guillou, A., Dierckx, G. and Fils-Viletard, A. *Estimation of the extreme value index and extreme quantiles under random censoring.* Extremes, **10**(3) (2007), 151-174

[2] Beirlant, J., Vynckier, P. and Teugels, J. *Tail index estimation, Pareto quantile plots, and regression diagnostics.* J. Amer. Statist. Assoc., **91** (1996), 1659-1667

[3] Caeiro, F., Gomes, M. I. and Pestana, D. *Direct reduction of bias of the classical Hill estimator.* Revstat, **3**(2) (2005), 113-136

[4] Dekkers, A., Einmahl, J. and de Haan, L. *A moment estimator for the index of an extreme-value distribution.* Ann. Statist., **17** (1989), 1833-1855

[5] Drees, H., Ferreira, A. and de Haan, L. *On maximum likelihood estimation of the extreme value index.* Ann. Appl. Probab., **14** (2004), 1179-1201

[6] Einmahl, J.H.J., Fils-Villetard, A. and Guillou, A. *Statistics of extremes under random censoring.* Bernoulli, **14**(1) (2007), 207-227

[7] Fraga Alves, M.I., Gomes, M.I., de Haan, L. and Neves, C. *Mixed Moment Estimator and Location Invariant Alternatives.* Notas e Comunicações CEAUL, **14** (2007). Available at http://www.ceaul.fc.ul.pt/notas.html?ano=2007

[8] Gomes, M.I. and Pestana, D. *A sturdy reduced bias extreme quantile (VaR) estimator.* J. Amer. Statist. Assoc., **102**(477) (2007), 280-292

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

[10] Reiss. R.-D. and Thomas, M. *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and other Fields* (1997), Birkhäuser Verlag, Basel

[11] Smith, R.L. *Estimating tails of probability distributions.* Ann. Statist., **15** (1987), 1174-1207