

A Robust Estimator for the Tail Index of Pareto-type Distributions

B. Vandewalle¹, M. Hubert², J. Beirlant² and A. Christmann³

¹ Department of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

² Department of Mathematics, Katholieke Universiteit Leuven, W. de Croylaan 54, B-3001 Leuven, Belgium

³ Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany.

1 Introduction

In extreme value statistics, the extreme value index is a well-known parameter to measure the tail heaviness of a distribution. This talk concentrates on Pareto-type distributions, with strictly positive extreme value index (or tail index). The most prominent extreme value methods are constructed on efficient maximum likelihood estimators based on specific parametric models which are fitted to excesses over large thresholds. Maximum likelihood estimators however are often not very robust, which makes them sensitive to few particular observations. Even in extreme value statistics, where the most extreme data usually receive most attention, this can constitute a serious problem. We illustrate the problem on a real data set from geopedology, in which a few abnormal soil measurements highly influence the estimates of the tail index. To overcome this problem, we propose a robust estimator of the tail index, by combining a refinement of the Pareto approximation for the conditional distribution of relative excesses over a large threshold with an integrated squared error approach on partial density component estimation. We show that the influence function of this newly proposed estimator is bounded and illustrate through several simulations that it performs reasonably well at contaminated as well as uncontaminated data.

2 Extreme Value Statistics

In extreme value statistics, the extreme value index, denoted by γ , is used to characterize the tail behavior of a distribution. This real-valued parameter helps to indicate the size and frequency of certain extreme events under a given probability distribution: the heavier the tail, the larger γ .

Consider X_1, \dots, X_n independent and identically distributed (i.i.d.) random variables with common distribution function F and quantile function Q . Denote the corresponding order statistics by $X_{1,n} \leq \dots \leq X_{n,n}$ and suppose there exist sequences of constants ($a_n > 0$) and ($b_n \in \mathbb{R}$) such that the properly centered and normed sample maxima $X_{n,n}$ converge in distribution to a non-degenerate limit H . Then, the limit distribution H is necessarily of generalized extreme value type:

$$H_\gamma(x) = \begin{cases} \exp\left(- (1 + \gamma x)^{-\frac{1}{\gamma}}\right), & 1 + \gamma x > 0, \gamma \neq 0, \\ \exp\left(- \exp(-x)\right), & x \in \mathbb{R}, \gamma = 0, \end{cases} \quad (1)$$

and F is said to belong to the maximum domain of attraction of H_γ , denoted as $F \in \mathcal{D}(H_\gamma)$. Most common continuous distribution functions satisfy this weak condition quite naturally.

The real-valued parameter γ is referred to as the tail index (or extreme value index) of F . Distributions for which $\gamma > 0$ are called Pareto-type (or heavy tailed) distributions, as the tail typically decays polynomially, i.e. $\bar{F}(x) = x^{-1/\gamma} l_F(x)$ with l_F a slowly varying function. Examples in this class are the Fréchet, Pareto and Burr distributions.

Recently, it was shown (Beirlant, Joossens and Segers, 2004) that the conditional distribution F_{Y_t} of relative excesses $Y_t = \left(\frac{X}{t} \mid X > t\right)$ over a threshold t can be approximated with

$$F_{\theta_t}(y) = (1 - \delta)[1 - y^{-1/\gamma}] + \delta[1 - y^{-1/\gamma + \rho}], \quad (2)$$

where $\theta_t = (\gamma, \delta, \rho)$ and the range of parameters is $\frac{1}{\gamma\rho} \leq \delta < 1$, $\gamma > 0$ and $\rho < 0$. Assuming this Pareto-mixture approximation to hold exactly as a model for the conditional distribution of the relative excesses Y_{jk} ($j = 1, \dots, k$) above a high threshold $t = X_{n-k,n}$ quite naturally leads to a maximum likelihood estimator for γ . This recently proposed mixture model and corresponding maximum likelihood estimator lead to potentially less biased and more stable estimates for γ than the Hill estimator.

3 Robust tail index estimator

In order to obtain robust estimates for the tail index, we adopt a minimum distance criterion based on integrated squared errors, which was discussed, among others, by Scott (2001, 2004). More exactly, we consider the partial density component approach of Scott (2004) and apply it to the mixture approximation of equation (2) for the conditional distribution of relative excesses over a threshold.

With t a suitable threshold and $Y_{jk} = X_{n-j+1,n}/t$ ($j = 1, \dots, k$) the k corresponding relative excesses, this leads to

$$\hat{\theta}_k^w = \arg \min_{\theta, w} \left[w^2 \int_1^\infty f_\theta^2(y) dy - \frac{2w}{k} \sum_{j=1}^k f_\theta(Y_{jk}) \right] \quad (3)$$

with

$$f_\theta(y) = (1 - \delta) \left[\frac{1}{\gamma} y^{-(1+1/\gamma)} \right] + \delta \left[\left(\frac{1}{\gamma} - \rho \right) y^{-(1+1/\gamma-\rho)} \right], \quad (4)$$

where $\theta = (\gamma, \delta, \rho)$, $\gamma > 0$, $\rho < 0$ and $\frac{1}{\gamma\rho} \leq \delta < 1$. Here, k is easily seen to be equal to $n(1 - F_n(t))$ if $t = X_{n-k,n}$, where F_n is the empirical distribution function of the original data.

We show that the influence function of the resulting tail index estimator is bounded. Finally we illustrate with simulations and real examples its robustness towards contamination in real data.

References

- J. Beirlant, E. Joossens and J. Segers (2004). Discussion of "Generalized Pareto Fit to the Society of Actuaries' Large Claims Database" by A. Cebrian, M. Denuit and P. Lambert, *North American Actuarial Journal*, 8, 108-111.
- D.W. Scott (2001). Parametric Statistical Modeling by Minimum Integrated Squared Error. *Technometrics*, 43, 274-285.
- D.W. Scott (2004). Partial Mixture Estimation and Outlier Detection in Data and Regression. In: M. Hubert et al., editors, *Theory and Applications of Recent Robust Methods*, pp. 297-306, Birkhäuser Verlag Basel, Switzerland.