Tail analysis of a distribution by means of an inequality curve

Analisi della coda di una distribuzione attraverso una curva di concentrazione

E. Taufer, F. Santi, G. Espa and M. M. Dickson

Abstract The Zenga (1984) inequality curve $\lambda(p)$ is constant in p for Type I Pareto distributions. We show that this property holds exactly only for the Pareto distribution and, asymptotically, for distributions with power tail with index α, with $\alpha > 1$. Exploiting these properties one can develop powerful tools to analyze and estimate the tail of a distribution. An estimator for α is discussed. Inference is based on an estimator of $\lambda(p)$ which utilizes all sample information for all values of p. The properties of the proposed estimation strategy is analyzed theoretically and by means of simulations.

Abstract *La curva di concentrazione $\lambda(p)$ di Zenga (1984) è costante in p per le distribuzioni di Pareto di tipo I. Questa proprietà vale esattamente solo per la distribuzione di Pareto e, asintoticamente, per le distribuzioni con indice di coda $-\alpha$, con $\alpha > 1$. Sfruttando queste proprietà si possono sviluppare metodi molto efficaci per analizzare e stimare la coda di una distribuzione. Si discute uno stimatore per α. L'inferenza si basa su uno stimatore di $\lambda(p)$. Le proprietà della strategia di stima proposta sono analizzate teoricamente e mediante simulazioni*

Key words: Tail index, inequality curve, non-parametric estimation

E. Taufer
University of Trento, Trento e-mail: emanuele.taufer@unitn.it

F. Santi
University of Trento, Trento e-mail: flavio.santi@unitn.it

G. Espa
University of Trento, Trento e-mail: giuseppe.espa@unitn.it

M. M. Dickson
University of Padua, e-mail: dickson@stat.unipd.it

1 Introduction

Consider an iid random sample $X_1, X_2, \ldots, X_n$ drawn from a random variable with distribution function F satisfying

$$\bar{F}(x) = x^{-\alpha} L(x), \tag{1}$$

where $\bar{F} = 1 - F$, and $L(x)$ is a slowly varying function, that is $L(tx)/L(x) \to 1$ as $x \to \infty$, for any $t > 0$. We will say that $\bar{F}$ is regularly varying (RV) at infinity with index $-\alpha$, denoted as $\bar{F} \in RV_{-\alpha}$. The parameter $\alpha > 0$ is usually referred to as *tail index*; alternatively, in the extreme value (EV) literature it is typical to refer to the EV index $\gamma > 0$ with $\alpha = 1/\gamma$ (see e.g. [10]).

The paper proposes an estimator of the tail index α which relies on Zenga inequality curve $\lambda(p)$, $p \in (0,1)$ [12]. The curve $\lambda(p)$ has the property of being constant for Type I Pareto distributions and, as it will be shown, this property holds for distributions satisfying (1). See [1], [12] , [13] for a general introduction and analysis of $\lambda(p)$.

Probably the most well-known estimator of the tail index is the Hill [6] estimator, which exploits the k upper order statistics. The Hill estimator may suffer from high bias and is heavily dependent on the choice of k (see e.g. [2]). It has been thoroughly studied and several generalization have appeared in the literature. For recent review of estimation procedures for the tail index of a distribution see [3].

The approach to estimation proposed here, directly connected to the inequality curve $\lambda(p)$ has a nice graphical interpretation and could be used to develop graphical tools for tail analysis. Another graph-based method is to be found in [9], which exploits properties of the QQ-plot; while a recent approach based on the asymptotic properties of the partition function, a moment statistic generally employed in the analysis of multi-fractality, has been introduced by [5]; see also [8] which analyzes the real part of the characteristic function at the origin.

2 The curve $\lambda(p)$ and estimation strategy

Let X be a positive random variable with finite mean μ, distribution function F, and probability density f. The inequality curve $\lambda(p)$ is defined as:

$$\lambda(p) = 1 - \frac{\log(1 - Q(F^{-1}(p)))}{\log(1-p)}, \quad 0 < p < 1, \tag{2}$$

where $F^{-1}(p) = \inf\{x\colon F(x) \geq p\}$ is the generalized inverse of F and $Q(x) = \int_0^x t f(t) dt / \mu$ is the first incomplete moment. Q can be defined as a function of p via the Lorenz curve

$$L(p) = Q(F^{-1}(p)) = \frac{1}{\mu} \int_0^p F^{-1}(t) dt. \tag{3}$$

For a Type I Pareto distribution [7, 573 ff.] with

$$F(x) = 1 - (x/x_0)^{-\alpha}, \quad x \geq x_0 \tag{4}$$

it holds that $\lambda(p) = 1/\alpha$, i.e. $\lambda(p)$ is constant in p. This is actually an if-and-only-if result, as we formalize in the following lemma:

Lemma 1. *The curve* $\lambda(p)$ *defined in* (2) *is constant in* p *if, and only if,* F *satisfies* (4).

The following result can also be stated, asymptotically for the case where $\bar{F}$ satisfies (1) as it is stated in the next lemma. For this purpose write

$$\lambda(x) = 1 - \frac{\log(1-Q(x))}{\log(1-F(x))}, \tag{5}$$

Lemma 2. *If* $\bar{F}$ *satisfies* (1)*, then* $\lim_{x\to\infty} \lambda(x) = 1/\alpha$.

A tail property of Pareto type I distribution is worth of being noted. Let X be a random variable distributed according to (4) – that is, $X \sim \text{Pareto}(\alpha, x_0)$ –, the following property holds for any $x_1 > x_2 > x_0$:

$$\mathbb{P}[X > x_1 | X > x_2] = \left(\frac{x_1}{x_2}\right)^{-\alpha},$$

hence, the truncated random variable $(X|X > x_2)$ is distributed as $\text{Pareto}(\alpha, x_2)$.

The implications of this property are twofold. Firstly, the truncated random variable is still distributed according to (4), thus Lemma 1 still applies. Secondly, the tail index α is the same both for original and for truncated random variable, thus function $\lambda(p)$ can be used for the estimation of α regardless of the truncation threshold x_2.

The same property we have just outlined holds asymptotically for distribution functions satisfying (1).

Figure 1 reports the empirical curve $\hat{\lambda}(p)$ as a function of p for a Pareto distribution defined by (4) with $\alpha = 2$ and $x_0 = 1$, denoted with $\text{Pareto}(2,1)$ and a Fréchet distribution with $F(x) = \exp(-x^{-\alpha})$ for $x \geq 0$ and $\alpha = 2$, denoted by Fréchet(2) at different truncation thresholds. Note the remarkably regular behavior or the curves and the closeness to the theoretical form for the Fréchet case already for low levels of truncation.

In this paper the above properties are exploited for devising an estimation method of the tail index α for distributions of class (1).

Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of the sample, $\mathbb{I}_{(A)}$ the indicator function of the event A. To estimate $\lambda(p)$, define the preliminary estimates

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(X_i \leq x)} \qquad Q_n(x) = \frac{\sum_{i=1}^{n} X_i \mathbb{I}_{(X_i \leq x)}}{\sum_{i=1}^{n} X_i} \tag{6}$$

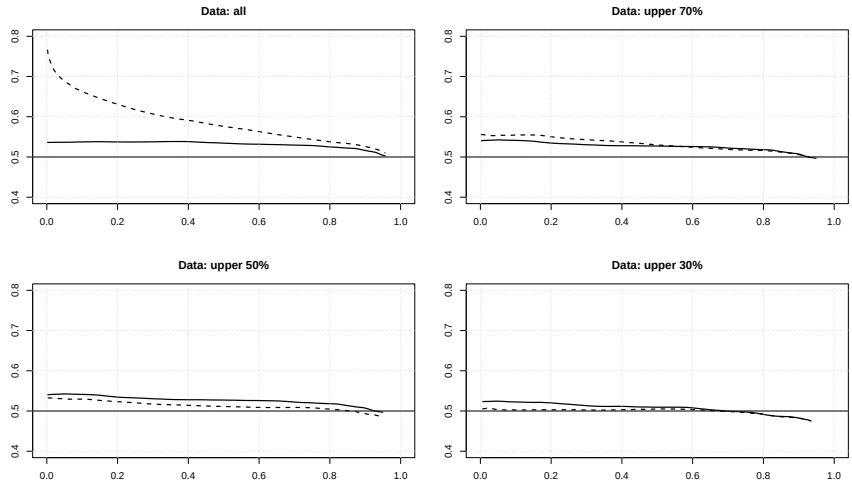

Fig. 1 Plot of $\hat{\lambda}(p)$ and p for Pareto$(2,1)$ (solid line) and Fréchet(2) (dashed line) at various levels of truncation. Sample size $n = 500$. Horizontal line at $1/\alpha = 0.5$

Under the Glivenko-Cantelli theorem (see e.g. [11]) it holds that $F_n(x) \to F(x)$ almost surely and uniformly in $0 < x < \infty$; under the assumption that $E(X) < \infty$, it holds that $Q_n(x) \to Q(x)$ almost surely and uniformly in $0 < x < \infty$. F_n and Q_n are both step functions with jumps at $X_{(1)}, \ldots, X_{(n)}$. The jumps of F_n are of size $1/n$ while the jumps of Q_n are of size $X_{(i)}/T$ where $T = \sum_{i=1}^{n} X_{(i)}$. Define the empirical counterpart of L as follows:

$$L_n(p) = Q_n(F_n^{-1}(p)) = \frac{\sum_{j=1}^{i} X_{(j)}}{T}, \quad \frac{i}{n} \leq p < \frac{i+1}{n}, \quad i = 1,2,\ldots,n-1, \tag{7}$$

where $F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$. To estimate α define

$$\hat{\lambda}_i = 1 - \frac{\log(1 - L_n(p_i))}{\log(1 - p_i)}, \quad p_i = \frac{i}{n}, \quad i = 1,2,\ldots n - \left\lfloor \sqrt{n} \right\rfloor. \tag{8}$$

and let $\hat{\alpha} = 1/\bar{\lambda}$ where $\bar{\lambda}$ is the mean of the $\hat{\lambda}_i$'s. The choice of $i = 1,\ldots,n - \lfloor\sqrt{n}\rfloor$ guarantees that $\hat{\lambda}_i$ is consistent for λ_i for each $p_i = i/n$ as $n \to \infty$.

3 Simulations

To evaluate the performance of $\hat{\alpha}$, some numerical comparisons are carried out with respect to some reduced bias competitors optimized with respect to the choice of k, the number of largest order statistics used in estimation, as discussed in [4]. The

class of moment of order p estimators [4], which reduce to the Hill estimator when $p = 0$ is considered; in the tables they are indicated as $Mop(p)$.

As far as the estimator $\hat{\alpha}$ is concerned, different levels of truncation of the data are considered. In the tables $Ze(q)$ indicates the estimator $\hat{\alpha}$ with q indicating the fraction of upper order statistics used in estimation.

For the comparisons, the Pareto and the Fréchet distributions, as defined in the previous section, are used. Random numbers for the Pareto distribution are simply generated in `R` using the function `runif()` and inversion of F; random numbers from the Fréchet are simulated using the function `rfrechet()` from the library `evd` with shape parameter set equal to α.

Tables 1 and 2 contain the results of simulations. For each sample size $n = 100, 200, 500, 1000, 2000$, $M = 1000$ Monte-Carlo replicates were generated. Computations have been carried out with `R` version 3.3.1 and each experiment, i.e. given a chosen distribution and a chosen n, has been initialized using `set.seed(1)`.

n	*Hill*	*Mop*(0.5)	*Mop*(1)	*Ze*(1)	*Ze*(0.7)	*Ze*(0.5)	*Ze*(0.3)
100	3.41	1.01	1.02	6.44	5.93	5.46	4.68
200	3.97	1.01	1.02	8.67	8.11	7.38	6.53
500	1.99	1.00	0.99	5.33	4.84	4.58	4.07
1000	2.75	1.00	1.00	8.31	7.64	7.11	6.49
2000	1.10	1.00	0.99	3.67	3.48	3.18	2.84

Table 1 Hill estimator: RMSE. Other estimators: relative RMSE *wrt* to the Hill estimator. Pareto$(2,1)$ distribution. Results based on 1000 replications.

n	*Hill*	*Mop*(0.5)	*Mop*(1)	*Ze*(1)	*Ze*(0.7)	*Ze*(0.5)	*Ze*(0.3)
100	0.72	0.99	0.97	1.19	1.41	1.34	1.11
200	0.62	0.99	0.96	1.01	1.26	1.29	1.13
500	0.50	0.98	0.94	0.81	1.07	1.15	1.09
1000	0.44	1.00	0.93	0.71	0.93	1.06	1.06
2000	0.37	1.00	0.91	0.61	0.82	0.94	1.00

Table 2 Hill estimator: RMSE. Other estimators: relative RMSE *wrt* to the Hill estimator. Fréchet(2) distribution. Results based on 1000 replications.

From the tables one can note that the performance of $\hat{\alpha}$ is brilliant for the Pareto and slightly better of Mop estimators for the Fréchet. Truncation seems to have only a small effect on the performance of the estimator.

References

1. Arcagni A., Porro F. (2016) . A comparison of income distributions models throught inequality curves. Statistica & Applicazioni. XIV (2), 123–144
2. Embrechts, P., C. Klüppelberg, T. Mikosch (1997). Modelling Extremal Events. Springer.

3. Gomes, M. I., & Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: a review. *International Statistical Review*, 83(2), 263–292.
4. Gomes, M. I., Brilhante, M. F., & Pestana, D. (2016). New reduced-bias estimators of a positive extreme value index. *Communications in Statistics-Simulation and Computation*, 45(3), 833–862.
5. Grahovac, D., Jia, M., Leonenko, N. N., Taufer, E. (2015) Asymptotic properties of the partition function and applications in tail index inference of heavy-tailed data. *Statistics: A Journal of Theoretical and Applied Statistics* 49, 1221–1242.
6. Hill, B. M. (1975) A simple general approach to inference about the tail of a distribution. *The Annuals of Statistics* 3(5), 1163–1174.
7. Johnson N. L., S. Kotz, N. Balakrishnan (1995) *Continuous Univariate Distributions, Vol. 2*, 2nd ed, Wiley.
8. Jia, M., Taufer, E., Dickson, M. (2018). Semi-parametric regression estimation of the tail index. *Electronic Journal of Statistics* 12, 224–248.
9. Kratz, M. F., Resnick, S. I. (1996) The QQ-estimator and heavy tails. *Comm. Statist. Stochastic Models* 12 (4), 699–724.
10. McNeil, A. J., R. Frey, P. Embrechts (2005) *Quantitative Risk Management*, Princeton University Press.
11. Resnik, S. I. (1999) *A probability path*, Birkhäuser.
12. Zenga, M. (1984). Proposta per un indice di concentrazione basato sui rapporti fra quantili di popolazione e quantili di reddito. *Giornale degli Economisti e Annali di Economia* 5/6, 301–326
13. Zenga M.(1990). Concentration curves and Concentration indexes derived from them. In *Income and Wealth Distribution, Inequality and Poverty*, Dagum, C., Zenga, M. (Ed.), Springher -Verlag, 94–110.