

GRAPHICAL REPRESENTATIONS AND ASSOCIATED GOODNESS-OF-FIT TESTS FOR PARETO AND LOG-NORMAL DISTRIBUTIONS BASED ON INEQUALITY CURVES

Emanuele Taufer^{a,c}, Flavio Santi^b, Giuseppe Espa^a, Maria Michela Dickson^a,

^a*Department of Economics and Management, University of Trento - Italy*

^b*Department of Economics, University of Verona - Italy*

^c*Corresponding author, email: emanuele.taufer@unitn.it*

Abstract

The Pareto and the Log-normal distributions can be characterized by having constant inequality curves. After providing formal proofs, these results will be exploited to obtain graphical and analytical tools for analysis and goodness-of-fit tests for these distributions. Order-statistics-based estimators of the curves will be presented and discussed. Examples with simulated and real-data will be used to show how the tools presented work in practice.

Keywords: Tail index, Pareto tail, graphical analysis, non-parametric estimation, parametric bootstrap.

1 Introduction

The choice between a distribution with light or heavy tails critically influences the behaviour of a probabilistic model used to analyse empirical data; this is especially important in areas such as growth theory, reliability, insurance risk, queueing theory, telecommunications network analysis, value-at-risk, asset pricing.

For example, Log-normal and type-I Pareto distributions are often competitors in problems of tail analysis and it turns out that, very often, it is quite hard to distinguish between the two cases. For example, in analysing the distribution of city sizes, the choice between a Log-normal or a Pareto distribution drives the growth theory for understanding the phenomenon; see Eeckhout (2004) and the connected literature for discussion.

In general, the distinction between power and exponential tails may appear to be easy but this is not often the case. Typically, graphical methods support the analysis; the most widely used plots are those based on estimates of the tail probability, the mean-excess function, the moment generating function, max-sum ratios or the Hill estimator. Plots may require to observe and compare linear trends, heavy jumps or blow-ups (towards infinity); for a deep discussion of weaknesses and advantages of these methods and references to their applications in various fields see Heyde and Kou (2004). One can also consult Clauset et al. (2009) for an overview of analytical methods for detecting power tails and examples of applications to several datasets.

Here we are mainly interested in two graphical tools of analysis which have the property of being constants for type-I Pareto and Log-normal distributions respectively.

The plots rely on two inequality curves, λ and Z , originally defined in Zenga (1984) in the context of concentration analysis; indeed the use of the index Z has been advocated by Cirillo (2013) in order to check for the Log-normal distribution (see however our Remark 1 below).

The original contribution of this paper with respect to previous literature is many-fold. First of all we study analytically the properties of the graphs; a) by proving that the curve $\lambda(p)$ is constant in $p \in (0, 1)$ only in the case of type-I Pareto distribution; b) discussing the behaviour of λ for distributions with a regularly varying tails; c) by proving that $Z(p)$ is constant in $p \in (0, 1)$ only in the case of Log-normal distributions.

Next, given that in practice, when using data, it is not always straightforward to judge on the actual "flatness" of the graphs, testing procedures for supporting the decision will be discussed. Our view on the resulting goodness-of-fit tests is that they should not be considered as standalone testing procedures, rather they should be used together with the graphical analysis, for which they have been developed. Notwithstanding, a comparison of the proposed tests statistics, with well established testing procedures will be carried on by means of simulations in order to show their performance.

Furthermore, examples of the graphical and inferential analyses to simulated and real data set will be presented.

Goodness-of-fit tests for Pareto and Log-normal distributions based on characterizations have been discussed in Afhami and Madadi (2017) which use entropy-based test statistics, Volkova (2016) which use an order-statistics-based characterization to define the test statistics, Obradović et al. (2015) which exploit a characterization based on the equivalence of distribution of random variables and maximum of their ratios; Rizzo (2009) which uses moment inequalities of random vector to obtain the test statistics, Meintanis (2009) which exploits the discrepancy between the empirical moment of arbitrary negative order and its theoretical counterpart in a weighted integral test statistic. As far as the log-normal distribution is concerned see Meintanis (2008) which uses a moment-based approach, Raqab et al. (2018), Dey and Kundu (2009), Kundu et al. (2005) which discuss likelihood-based techniques for discriminating the log-normal distribution from competing models.

Related works, which use graph-based methods in the context of tail estimation, are those of Kratz and Resnick (1996), which exploit properties of the QQ-plot and Grahovac et al. (2015) which use the asymptotic properties of the partition function, a moment statistics generally employed in the analysis of multi-fractality; see also Jia et al. (2018) which graphically analyses the real part of the characteristic function at the origin, Jammalamadaka and Gorla (2004), Jammalamadaka and Tian (2018) which discuss and review estimation and testing based on inequality measures.

The paper is organized as follows. Section 2 gives the theoretical results that support our analysis and Section 3 discusses accompanying bootstrap tests for goodness-of-fit. Section 4 contains examples of application and analysis with real and simulated data. An Appendix contains the proofs of the results of Section 2.

2 Main results

2.1 The curves λ and Z

Let X be a positive random variable with finite mean μ , absolutely continuous distribution function F , and probability density f . Two inequality curves, $\lambda(p)$ and $Z(p)$, $p \in (0, 1)$, defined in Zenga (1984), will be considered. The inequality curve λ is defined as:

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

where $F^{-1}(p) = \inf\{x: F(x) \geq p\}$ 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

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

In order to define Z , let the quantiles of F and Q be denoted respectively as x_p and x_p^* , i.e.

$$x_p = F^{-1}(p), \quad x_p^* = Q^{-1}(p), \quad 0 \leq p \leq 1.$$

Since $F(x) \geq Q(x) \forall x$, it follows that $x_p^* \geq x_p$. The Z curve is defined as

$$Z(p) = \frac{x_p^* - x_p}{x_p^*} = 1 - \frac{x_p}{x_p^*}, \quad 0 < p < 1. \quad (2)$$

For a type-I Pareto distribution with distribution function

$$F(x) = 1 - (x/x_0)^{-\alpha}, \quad x \geq x_0, \quad (3)$$

under the condition $\alpha > 1$, it holds that $\lambda(p) = 1/\alpha$, $p \in (0, 1)$, i.e. $\lambda(p)$ is constant in p . This is actually an if-and-only-if result, as we formalize in the following theorem.

Theorem 1. *The curve $\lambda(p)$ defined in (1) is constant in p , $p \in (0, 1)$, and equals $1/\alpha$ if, and only if, F satisfies (3) with $\alpha > 1$.*

A tail property of the type I Pareto distribution is worth of being noted: for X distributed according to (3), which we denote with $X \sim Pa(\alpha, x_0)$, 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 $Pa(\alpha, x_2)$. The implications of this property are twofold. Firstly, the truncated random variable is still distributed according to (3), thus Theorem 1 still applies. Secondly, the tail index α is the same both for original and for truncated random variable, thus λ will not change regardless of the truncation threshold x_2 . Consider the more general case of a positive random variable X with F satisfying

$$\bar{F}_X(x) = x^{-\alpha} H(x) \quad x \rightarrow \infty, \quad \alpha > 1, \quad (4)$$

where $\bar{F} = 1 - F$, and H is a slowly varying function, that is $H(tx)/H(x) \rightarrow 1$ as $x \rightarrow \infty$, 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 $\alpha = 1/\gamma$. Upon defining $Y = X|X > x_0$ one has that $F_Y(y) = 1 - \bar{F}_X(y)/\bar{F}_X(x_0)$. Under the assumption that $\bar{F} \in RV_{-\alpha}$ it holds that

$$\begin{aligned} F_Y(y) &= 1 - \left(\frac{y}{x_0}\right)^{-\alpha} \frac{H(\frac{y}{x_0}x_0)}{H(x_0)}, \quad y > x_0 \\ &= 1 - \left(\frac{y}{x_0}\right)^{-\alpha}, \quad y > x_0, \quad \text{as } x_0 \rightarrow \infty, \end{aligned} \quad (5)$$

by the properties of slowly varying functions. That is, a plot based on $(p, \lambda(p))$ will still be approximately constant for distributions satisfying (4) if only data larger than a sufficiently high threshold are used. This will make the plot based on λ a graphical tool for tail analysis; we will see that the remarkably smooth behaviour of estimated values of λ will make the graph based on λ quite effective in analyzing the tail of a distribution.

Turning now the attention to the curve Z , if we consider a Log-normal distribution, denoted by $Ln(\mu, \sigma^2)$, with density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma x}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, \quad x > 0. \quad (6)$$

Simple calculations show that in this case $Z(p) = 1 - e^{-\sigma^2}$, $p \in (0, 1)$. Again, we have an if-and-only-if result, formalized in the next theorem.

Theorem 2. *Assume that f has continuous derivatives up to the second order. The curve $Z(p)$ defined in (2) is constant in p , $p \in (0, 1)$, and equals $1 - e^{-\sigma^2}$ if, and only if, f satisfies (6).*

We refer the interested reader to Arcagni and Porro (2016), Zenga (1984), Zenga (1990) for a more general introduction, analysis and computations for other distributions of the curves λ and Z .

Remark 1. *We would like to point out that Z should not be confused with the Zenga (2007) $I(p)$, $p \in (0, 1)$, inequality curve which can still be defined via the Lorenz curve through the formula*

$$I(p) = 1 - \frac{L(p)}{p} \frac{1 - p}{1 - L(p)}, \quad 0 < p < 1. \quad (7)$$

Indeed, Cirillo (2013) states that $I(p)$ is constant in p for the Log-normal distribution, confusing the two curves Z and I .

Figures 1 and 2 report, respectively, the plots $(p, \lambda(p))$ and $(p, Z(p))$ for some common distributions. Figure 2, following Remark 1, provides a corrected version of Figure 7 in Cirillo (2013). Beyond the $Pa(\alpha, x_0)$ and the $Ln(\mu, \sigma)$ cases, defined respectively in (3) and (6), the following distributions over the positive domain are considered:

- 1) Fréchet distributions with $F(x) = \exp(-x^{-\alpha})$, $x \geq 0$, $Fr(\alpha)$ in short.

- 2) Gamma distributions with density $f(x) = (\theta^\alpha \Gamma(\alpha))^{-1} x^{\alpha-1} e^{-x/\theta}$, $\alpha, \theta > 0$, $x > 0$; denoted by $Ga(\alpha, \theta)$. The case $\alpha = 1$ will be denoted by $Exp(\theta)$.
- 3) Log-Weibull with $F(x) = 1 - \exp\{-(\log x)^{\theta+1}\}$, $\theta \in (0, 1)$, $x \geq 1$, denoted with $Lw(\theta)$.
- 4) Weibull distributions with $F(x) = 1 - \exp\{-x^\alpha\}$, denoted by $W(\alpha)$.

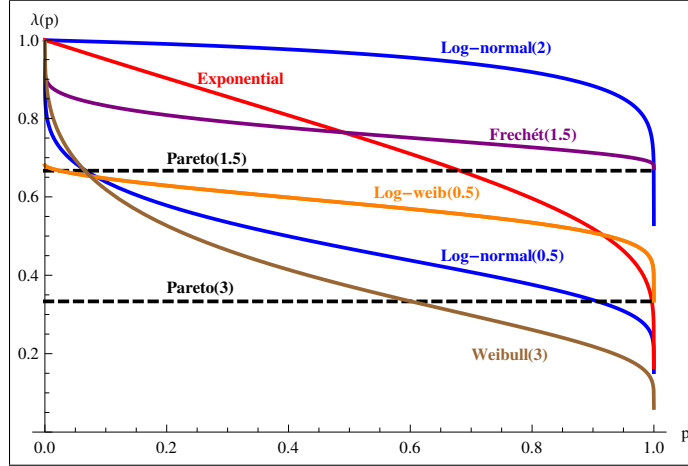


Figure 1: Theoretical $\lambda(p)$ curves for some common distributions.

Note that in Figure 1 most of the λ curves start in the upper left corner and end in the lower right corner, showing similar behaviours and, depending on parameter values, flatter or steeper curves, for the same distribution, could arise. This shows that graphs based on λ may not be optimal in order to determine the kind of distribution one is analysing. However, the Pareto case, characterized by a constant value for all p , stands out quite clearly and we will see a very powerful test for the hypothesis of paretianity can be constructed starting from the graph, unifying visual and analytical inspection.

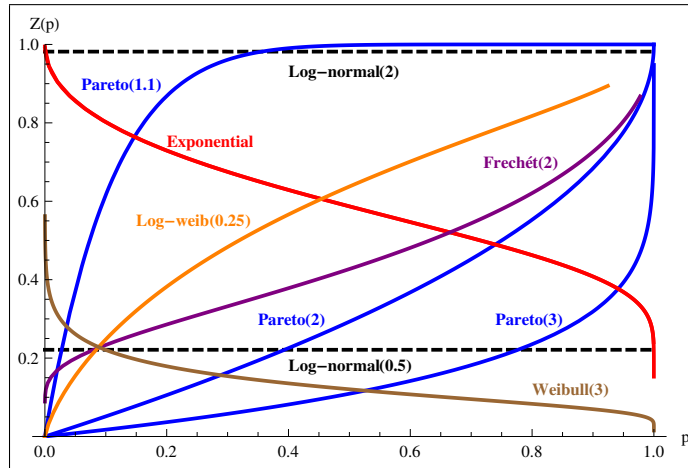


Figure 2: Theoretical $Z(p)$ curves for some common distributions.

Figure 2 shows a large variability in the behaviour of the curves for different distributions,

with the Log-normal, characterized by a constant value for all p , clearly standing out of the other distributions. A goodness-of-fit test will be derived in the next section.

2.2 Estimation of λ and Z

Let $X_{(1)}, \dots, X_{(n)}$ be the order statistics of the sample, $\mathbb{I}_{(A)}$ the indicator function of the event A . To estimate λ , define the empirical estimators

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

Under the Glivenko-Cantelli theorem it holds that $F_n(x) \rightarrow F(x)$ almost surely and uniformly in $0 < x < \infty$; under the assumption that $E(X) < \infty$, it holds that $Q_n(x) \rightarrow Q(x)$ almost surely and uniformly in $0 < x < \infty$ (Goldie (1977)). F_n and Q_n are both step functions with jumps at $X_{(1)}, \dots, X_{(n)}$. 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:

$$L_n(p) = Q_n(F_n^{-1}(p)) = \frac{\sum_{j=1}^i X_{(j)}}{T}, \quad \frac{i-1}{n} < p \leq \frac{i}{n}, \quad i = 1, 2, \dots, n, \quad (8)$$

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

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

The choice of $i = 1, \dots, n - \lfloor \sqrt{n} \rfloor$ guarantees that $\hat{\lambda}_i$ is consistent for λ_i for each $p_i = i/n$ as $n \rightarrow \infty$. To estimate Z , estimate first x_p and x_p^* by

$$\hat{x}_p = F_n^{-1}(p) = \inf\{x | F_n(x) \geq p\}, \quad \hat{x}_p^* = Q_n^{-1}(p) = \inf\{x | Q_n(x) \geq p\},$$

where

$$F_n^{-1}(p) = \begin{cases} x_{(1)} & \text{for } 0 < p \leq \frac{1}{n} \\ x_{(2)} & \text{for } \frac{1}{n} < p \leq \frac{2}{n} \\ \dots & \\ x_{(n-1)} & \text{for } \frac{n-2}{n} < p \leq \frac{n-1}{n} \\ x_{(n)} & \text{for } \frac{n-1}{n} < p \leq 1 \end{cases} \quad Q_n^{-1}(p) = \begin{cases} x_{(1)} & \text{for } 0 < p \leq \frac{T_1}{T} \\ x_{(2)} & \text{for } \frac{T_1}{T} < p \leq \frac{T_2}{T} \\ \dots & \\ x_{(n-1)} & \text{for } \frac{T_{n-2}}{T} < p \leq \frac{T_{n-1}}{T} \\ x_{(n)} & \text{for } \frac{T_{n-1}}{T} < p \leq 1 \end{cases}$$

and $T_j = \sum_{i=1}^j x_{(i)}$. Then set

$$\hat{z}_i = 1 - \frac{\hat{x}_{p_i}}{\hat{x}_{p_i}^*} \quad p_i = \frac{i}{n}, \quad i = \lfloor \sqrt{n} \rfloor, \lfloor \sqrt{n} \rfloor + 1, \dots, n - \lfloor \sqrt{n} \rfloor. \quad (9)$$

Again, the choice of the limits for i in (9) is in order to avoid highly irregular behaviour of the $\hat{z}_i$ at the limits of the domain for small sample sizes.

Remark 2. In Taufer et al. (2020) it has been shown that, under the assumption that μ be finite, $\hat{\lambda}(p)$ is uniformly consistent for $\lambda(p)$ over $p \in (0, 1)$. Moreover, uniform consistency of F_n and Q_n (assuming $\mu < \infty$) implies uniform consistency of $\hat{Z}(p)$ for $Z(p)$ over $p \in (0, 1)$.

2.3 Interpreting the graphical displays

From a practical point of view the estimated $\hat{\lambda}_i$ or, depending on the case at hand, $\hat{z}_i$, can be plotted against $p_i = i/n$ $i = 1, \dots, n - \lfloor \sqrt{n} \rfloor$ on a graph in order to inspect the whether the data follow a type-I Pareto distribution in the case of λ , of a Log-normal distribution in the case of Z .

One expects flat curves in the case the hypothesized model is correct. Note that in the case of the curve λ , from result (5), one would observe a flat curve also in the case of data following distributions defined by (4) if the sample has been appropriately truncated.

Since, when the curves are estimated, the situation may not be always clear cut, a testing procedure based on the graphical displays may help in taking a decision. For this reason two bootstrap test for the goodness-of-fit hypotheses will be discussed in the next section.

Note that if one is interested in distinguish between type-I Pareto (or Pareto type tail) and Log-normal distribution it might be useful to compare and analyse the two graphs. As we saw in Figures 1 and 2 the Log-normal distribution tends to flat-out in the case of the graph based on λ while the type-I Pareto distribution has a marked increasing behaviour in the graph based on Z .

3 Goodness-of-fit tests

Goodness-of-fit tests based on the inequality curves λ and Z can be defined by linear regression of λ_i , or z_i , on p_i . Let us discuss in detail the case of λ_i , the case for z_i can be carried on in an analogous way.

From Theorem 1, for a distribution F satisfying (3) with $\alpha > 1$, for any choice of p_i , $0 < p_i < 1$, $i = 1, \dots, m$, one has the linear equation

$$\lambda_i = \beta_0 + \beta_1 p_i,$$

where $\beta_0 = 1/\alpha$ and $\beta_1 = 0$. Given a random sample $X_1, \dots, X_n$, estimation and testing procedures can be defined through the regression:

$$\hat{\lambda}_i = \beta_0 + \beta_1 p_i + \varepsilon_i,$$

where $\varepsilon_i = \hat{\lambda}_i - \lambda_i$. Hence an estimator of β_0 can be used to estimate α while a test on the hypothesis $H_0 : \beta_1 = 0$ can be used to test that a distribution F satisfies (3).

Using least squares estimators and exploiting the knowledge that $\beta_1 = 0$ in the estimation of β_0 , define

$$\hat{\beta}_0 = \frac{1}{m} \sum_{i=1}^m \hat{\lambda}_i, \quad \hat{\beta}_{1,\lambda} = \sum_{i=1}^m \frac{\hat{\lambda}_i(p_i - \bar{p})}{S_p^2} = \sum_{i=1}^m \hat{\lambda}_i c(p_i) \quad (10)$$

where $\bar{p}$ is the mean of the p_i 's and $S_p^2 = \sum_{i=1}^m (p_i - \bar{p})^2$, $c(p_i) = (p_i - \bar{p})/S_p^2$. Note that since the sum of the first n natural numbers and their squares can be expressed as functions of n , one has that, for $p_i = i/n$, $i = 1, \dots, m$, $m = n - \lfloor \sqrt{n} \rfloor$,

$$\bar{p} = \frac{1}{2} \frac{m(m+1)}{n^2} = \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad S_p^2 = \frac{1}{12} \frac{m(m^2-1)}{n^2} = O(n).$$

Remark 3. From uniform consistency of $\hat{\lambda}_i$ (resp. $\hat{z}_i$), $i = 1, \dots, m$ it follows that $\hat{\beta}_{1,\lambda}$ (resp. $\hat{\beta}_{1,Z}$) converges in probability to 0 under the null hypothesis (12) (resp. (13)) discussed below.

Remark 4. Since λ does not depend on x_0 for the Pareto distribution and Z does not depend on μ for the Log-normal distribution, one can construct goodness-of-fit tests setting $x_0 = 1$ and $\mu = 0$ when needed.

Remark 5. A referee suggested using a Kolmogorov-Smirnov type statistics, namely

$$KS_\lambda = \max_{1 \leq i, j \leq m} |\hat{\lambda}_i - \hat{\lambda}_j| \quad \text{and} \quad KS_Z = \max_{1 \leq i, j \leq m} |\hat{z}_i - \hat{z}_j| \quad (11)$$

respectively for the Pareto test and for the Log-normal test. The approaches based on (10) and (11) will be compared in the next sections.

3.1 Pareto distribution

Following the discussion in the previous section, to test the hypothesis

$$H_0 : F \text{ is } Pa(\alpha, x_0), \quad \alpha > 1, \quad x_0 > 0, \quad (12)$$

reject the null if $|\hat{\beta}_{1,\lambda}|$ (defined in (10)), or alternatively, if KS_λ defined in (11), is large. A normal approximation to the distribution of $\hat{\beta}_{1,\lambda}$, not discussed here, holds under the assumption that $\alpha > 2$, leaving uncovered the case $1 < \alpha \leq 2$. Since the test statistic is invariant under a change in x_0 , a simple parametric bootstrap procedure can be carried on as follows:

1. Given a random sample of size n , estimate $\hat{\alpha} = 1/\hat{\beta}_0$ and $\hat{\beta}_{1,\lambda}$ as in (10) (or KS_λ as in (11)).
2. Generate a sample of size n from a $Pa(\hat{\alpha}, 1)$ and estimate $\hat{\beta}_{1,\lambda}$ (or KS_λ).
3. Repeat step 2 M times.
4. Get an estimated p -value of $\hat{\beta}_{1,\lambda}$ (or KS_λ) from the bootstrap distribution.

In a simulation experiment, the test statistics $\hat{\beta}_{1,\lambda}$ and KS_λ will be compared to classical Kolmogorov-Smirnov and Cramer-von-Mises tests statistics (see, e.g. D'agostino and Stephens (1986)), indicated respectively as KS and CM. A test based on a characterization discussed by Obradović et al. (2015) and Volkova (2016) (which generalizes the results of Obradović et al. (2015)), denote as V will be also considered.

Table 1 contains the power estimates for tests of critical level 0.05. Results are based on 1000 samples of size n . The null and alternative distributions considered have been presented in Section 2.1. For implementing the test V , which is based on a characterization valid only for the case $x_0 = 1$ the sample data $X_1, \dots, X_n$ were normalized by the sample minimum value.

Looking at the results in the table, note that the performance of $\hat{\beta}_{1,\lambda}$ and KS_λ is very similar in all cases. Also, these test statistics tend to over-reject under the null for small sample size, implying a possibly inflated power in such cases. There are ways to fix this, see for instance Lloyd (2005). Since the tests achieve the correct significance level for higher

	$\hat{\beta}_{1,\lambda}$			KS_λ			KS			CM			V		
$n \rightarrow$	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100
Pa(2,1)	.11	.07	.05	.09	.08	.07	.05	.06	.04	.05	.05	.04	.05	.05	.05
Pa(3,2)	.09	.07	.05	.08	.07	.06	.05	.06	.04	.05	.05	.04	.05	.05	.05
Ln(0,1)	.91	1.0	1.0	.92	.99	1.0	.76	.99	1.0	.81	1.0	1.0	.90	1.0	1.0
Ln(0,2)	.28	.50	.72	.29	.47	.70	.76	.99	1.0	.81	1.0	1.0	.90	1.0	1.0
Lw(0.25)	.23	.38	.59	.26	.42	.57	.08	.23	.47	.07	.26	.56	.17	.31	.54
Lw(0.5)	.46	.81	.96	.50	.83	.96	.16	.61	.95	.16	.68	.98	.33	.71	.95
Fr(2)	.61	.88	.95	.67	.89	.97	.40	.93	.99	.42	.95	.99	.61	.97	1.0
Ga(2)	.99	1.0	1.0	.99	1.0	1.0	.89	1.0	1.0	.93	1.0	1.0	.98	1.0	1.0
W(2)	.99	1.0	1.0	1.0	1.0	1.0	.94	1.0	1.0	.97	1.0	1.0	.98	1.0	1.0

Table 1: Estimated power for tests of size 0.05 for the hypothesis (12). Results based on 1000 replications, p-value estimates were obtained by parametric bootstrap with $M = 500$.

sample sizes, we believe that the results in the table, given the objectives of the paper discussed in the introduction, depict quite clearly the overall situation.

Reading the figures, note that the proposed test statistics, in general, compare well with the competitors considered, showing definitely a better performance in the case of Log-weibull distributions. It appears that $\hat{\beta}_{1,\lambda}$ and KS_λ has a worst performance in the case of Log-normal distributions with large value of the parameter σ . This was expected since, as noted in Section 2.1, the curve λ for the Log-normal distribution becomes flatter when σ increases. As pointed out, the proposed tests should be carried on together with the graphical analysis and if a direct comparison Pareto versus Log-normal is of interest, a test for Log-normality should be carried on; see also the discussion in the next sections.

Finally note that all tests show very high power for Gamma and Weibull alternatives. In the case of the Exponential distribution (not shown in the table because very similar to the case Ga(2)) we have the peculiar result that λ does not depend on θ : for this distribution the test always attains the same power irrespectively of the scale.

3.2 Log-normal distribution

A formal test for the general hypothesis that the data come from a distribution F with density (6), i.e.

$$H_0 : F \text{ is } Ln(\mu, \sigma), \quad \mu \in \mathbb{R}, \quad \sigma > 0 \quad (13)$$

can be carried on by computing the regression coefficient on the data $(p_i, \hat{z}_i)$ and rejecting the null if $|\hat{\beta}_{1,Z}|$ is large or, alternatively, if KS_Z as in (11)) is large; in this case, following (10), $\hat{\beta}_{1,Z} = \sum_i^m \hat{z}_i c(p_i)$. In order to practically implement the test one can rely on a parametric bootstrap procedure as follows:

1. Given a random sample of size n , estimate $\hat{\sigma}$ using the standard deviation of the log-data (the estimator $\hat{\sigma} = \sqrt{-\log(1 - \hat{\beta}_0)}$ obtained by Lemma 2 turned out to be inefficient) and $\hat{\beta}_{1,Z}$ (or KS_Z as in (11)).
2. Generate a sample of size n from a $\text{Ln}(0, \hat{\sigma})$ (see Remark 4) and estimate $\hat{\beta}_1$ (or KS_Z).

	$\hat{\beta}_{1,Z}$			KS_Z			KS			CM			SW		
n	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100
Ln(0,1)	.05	.06	.05	.05	.06	.06	.05	.05	.06	.05	.06	.07	.05	.05	.05
Ln(0,2)	.05	.06	.04	.04	.05	.04	.05	.05	.06	.05	.06	.06	.05	.06	.05
Pa(2,1)	.30	.90	.99	.23	.84	1.0	.59	.96	1.0	.73	.98	1.0	.84	.99	1.0
Pa(3,2)	.11	.86	1.0	.12	.65	.98	.59	.96	1.0	.73	.98	1.0	.84	.99	1.0
Exp(1)	.23	.45	.77	.24	.45	.74	.22	.45	.74	.24	.59	.86	.33	.72	.95
Ga(2)	.14	.23	.45	.15	.23	.44	.11	.25	.44	.14	.31	.55	.18	.44	.72
W(2)	.15	.30	.66	.17	.30	.57	.25	.45	.76	.27	.57	.87	.34	.72	.95

Table 2: Estimated power for tests of size 0.05 for the hypothesis (13). Results based on 1000 replications, p-value estimates were obtained by parametric bootstrap with $M = 500$.

3. Repeat step 2 M times.

4. Get an estimated p -value of $\hat{\beta}_{1,Z}$ (or KS_Z) from the bootstrap distribution.

The test statistics $\hat{\beta}_{1,Z}$ and KS_Z will be compared to classical Kolmogorov-Smirnov and Cramer-von-Mises tests statistics (see, e.g. D’agostino and Stephens (1986)), indicated respectively as KS and CM. The Shapiro-Wilk test for normality (see, e.g. D’agostino and Stephens (1986)), previous log-transformation of the data will be also considered (indicated as SW); this test has proven quite powerful in several situations and it will provide a hard comparison to our test statistics.

Table 2 contains the power estimates for tests of critical level 0.05. Results are based on 1000 samples of size n . The null and alternative distributions considered have been presented in Section 2.1.

Looking at the results of the table, one can note that, similarly to the case of the test for the Pareto distribution, $\hat{\beta}_{1,Z}$ and KS_Z have a similar performance. Inspection of the graphs during the simulations (not shown here) show that the estimated curve Z shows high variability for small sample sizes; this reflects in the low power of the tests based on $\hat{\beta}_{1,Z}$ and KS_Z for very small sample sizes. When sample size increases, the performance of the tests proposed is similar to the KS and CM statistics. Above all the results stands out the performance of the SW test which is already indicated in the literature as a very powerful test for normality and here, previous log-transformation of the data, proves quite good also for the log-normal case.

Remark 6. *Remark 2 and remark 3 imply consistency of the tests proposed for hypotheses 12 and 13. Moreover, under the null hypothesis, the proposed parametric bootstrap will asymptotically attain the actual significance level.*

4 Examples

4.1 Simulated data

As a first example of application, consider some simulated data. A sample of size $n = 1000$ from each of the following distributions is considered: Pa(2,1), Ln(0,1), Fré(2) and

Exponential ($\theta = 1$).

Figure 3 reports the empirical curve $\hat{\lambda}(p)$ as a function of p at different levels of truncation of the sample. Note the remarkably smooth and regular behaviour of the curves. As the level of truncation increases, as expected, the curve λ for the Fréchet distribution gets closer and closer to the Pareto one, which is always remarkably constant. It appears that the tails of the two distributions are the same for data above the 50th percentile.

The graphical analysis seems able to effectively distinguish between a Pareto distribution and other cases for each level of truncation. In order to have an analytical result, a bootstrap test for the hypothesis (12) is carried on for each case. Table 3 reports the estimated p-values obtained. Note that in all cases, the p-values confirm the expected results.

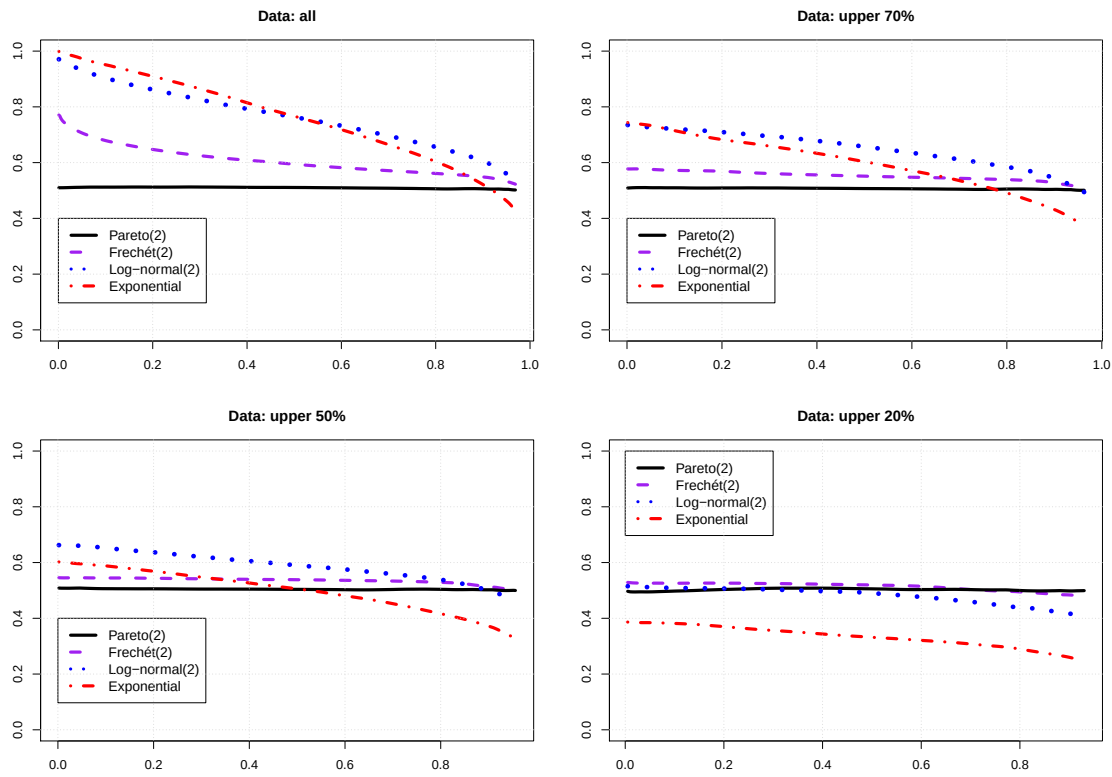


Figure 3: Plot of $\hat{\lambda}(p)$ and p for Pareto, Fréchet, Log-normal and Exponential distributions at various levels of truncation. Sample size $n = 1000$.

Data	All	Upper 70%	Upper 50%	Upper 20%
Pa(2,1)	0.696	0.738	0.828	0.988
Fre(2)	0.000	0.000	0.226	0.196
Ln(0,1)	0.000	0.000	0.000	0.002
Exp	0.000	0.000	0.000	0.000

Table 3: P-values for hypothesis (12), bootstrap tests for Pareto, Fréchet, Log-normal and Exponential samples. Sample size $n = 1000$ (all data).

Note that for truncated samples the graphs in figure 3 together with the bootstrap tests are able to distinguish effectively distributions satisfying (4) from distributions with exponential type decay in the tail such as the Log-normal and the Exponential.

To compare these results with those of some largely used graphical tools, consider the tail probability plot (also known as Zipf's plot or log-log-plot), which plots $(\log x, \log \bar{F}(x))$. This kind of plot should display a straight line for power type tails, but not for exponential type tails (see Heyde and Kou (2004) for details). Figure 4 reports the log-log plot for the entire sample and the upper 50% of the observation; note that in both displays, all curves show an essentially linear behaviour, making it difficult to distinguish exponential type from power type distributions.

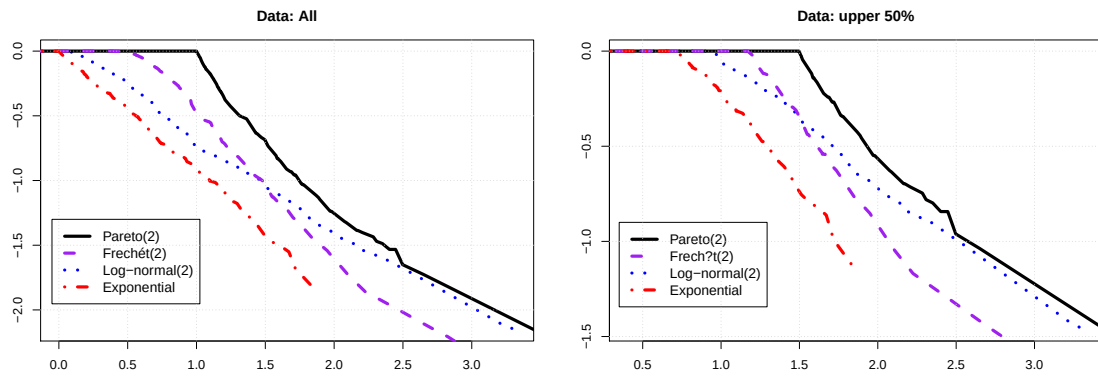


Figure 4: Zipf's plot for Pareto, Fréchet, Log-normal and Exponential distributions at various levels of truncation. Sample size $n = 1000$.

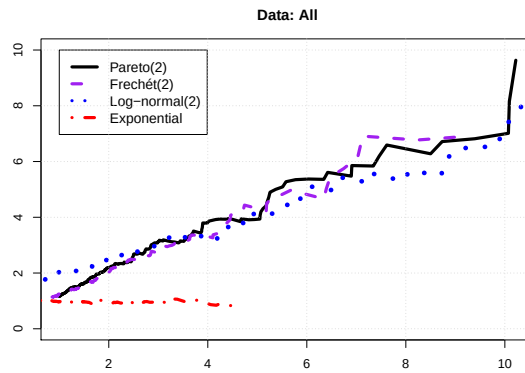


Figure 5: Mean Excess function plot for Pareto, Fréchet, Log-normal and Exponential distributions. Sample size $n = 1000$.

Another popular plot used for tail analysis is the mean-excess function plot based on a sample estimate of $e(t) = E(X - t|X > t)$. The plot of $(t, e(t))$ suggest a power type distribution if it is linear with positive slope and null intercept (see Heyde and Kou (2004)

for details). Figure 5 reports the mean-excess function plot for a large range of t values; as we see, in this case it turns out to be quite hard to distinguish a Log-normal distribution from Pareto or Fréchet ones.

Turning to the case of the curve Z , Figure 6 reports the estimated Z curves from simulated data for the following distributions: $\text{Ln}(0,1)$, $\text{Pa}(2,1)$, $W(2)$ and $\text{Ga}(2)$. Since for this graph it does not make sense to consider truncated samples, we report the graphs of the estimated Z for sample sizes $n = 250$ on the left and $n = 500$ on the right. The estimated Z curves show a more erratic behaviour if compared to the estimates of the curve λ ; however, the graphs are still effective in discriminating the Log-normal distribution from the other cases.

To support the graphical analysis with analytical results, a bootstrap test for the hypothesis (13) has been carried out for each case. The results are reported in Table 4. Again, for each single case, the results are those one would expect to obtain under the full knowledge of the data.

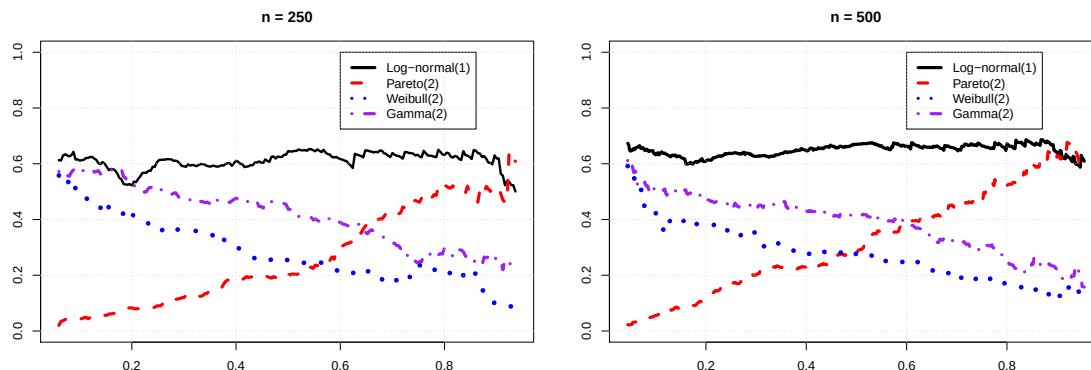


Figure 6: Plot of estimated $Z(p)$ against p for Log-normal, Pareto, Weibull and Gamma distributions. Left: sample size $n = 250$. Right: sample size $n = 500$.

n	$\text{Ln}(0,1)$	$\text{Pa}(2,1)$	$W(2)$	$\text{Ga}(2)$
250	0.764	0.000	0.000	0.000
500	0.686	0.000	0.000	0.000

Table 4: P-values for hypothesis (13), bootstrap tests for Log-normal, Pareto, Weibull and Gamma samples. Sample size: 250 and 500.

4.2 City sizes

There is an ongoing debate on the specific regularity on the distribution of city sizes: Zipf's law (a distribution with a Pareto tail with $\alpha = 1$) or a Log-normal distribution. Early studies show that Zipf's law generally holds well for data on the largest cities in a country (see Gabaix and Ioannides (2004) and references therein). However, a discussion was started by Eeckhout (2004) which extended the analysis to the entire range of city sizes and argued that the size of all populated places in the US follows a Log-normal distribution. Later, Levy (2009) argued that significant deviations from a Log-normal distribution exist in the top

range of the largest cities ($\log\text{-size} > 11$) by using graphical analysis. A reply of Eeckhout (2009) is followed by several other papers, not recalled here, discussing the issue. The data set used by Eeckhout (2004) and Levy (2009) containing the entire distribution with population ranging from 1 to several millions is analysed here. The “US Census 2000 places” dataset is based on 25,358 units of “city, town, borough, or village”. For a preliminary analysis using log-log plots and other analytical tools we refer the reader to the above mentioned papers.

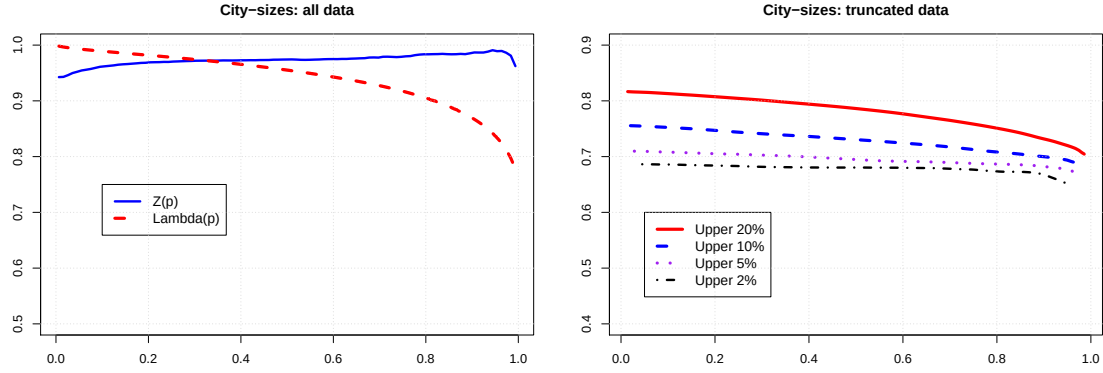


Figure 7: City sizes data. Left: estimated λ and Z curves on all data. Right: λ curves for upper order statistics.

Figure 7 reports, on the left, the λ and Z curves estimated on all data; on the right, the λ curves for upper order statistics at various levels of truncation. One can use the left graph to check for Log-normality and the right in order to analyse the tail of the distribution of city-sizes.

On the left, Z is rather flat, while λ is decreasing, both supporting, at first sight, the Log-normal model for the data. A bootstrap test for the hypothesis (13) however provides a p-value equal to 0.016, i.e. rejecting the hypothesis of Log-normality. This result is actually confirmed by a Kolmogorov-Smirnov test for hypothesis (13) carried on by Eeckhout (2009) which resulted in a p-value equal to 0.0189 (Eeckhout maintained however his argument for the Log-normal distribution). An inspection of standard PP and QQ-plots reveals that the fit of the Log-normal model might have some problems on the right tail.

On the right part of Figure 7, λ can be used to analyse the upper tail of the empirical distribution. Note that the curves flatten out by reducing the number of upper order statistics considered. A bootstrap test for the hypothesis (12), does not reject the null hypothesis for samples based on the upper 5% and 2% of the data. The upper 2% of the order statistics, limit from which, according to Levy (2009), a Pareto tail should start, considers cities with size larger than 59,874 (or 11 in log-terms), which are 510 in total. Considering the upper 5% of the sample, one actually considers cities with more than 29,314 inhabitants (1,269 cities in total).

In definitive, from a statistical point of view, the support of the inequality curves λ and Z in the analysis provides relevant information by discarding the Log-normal model for the whole empirical distribution and strongly supporting a Pareto tail in the data starting already at the 95th percentile.

5 Conclusions

Characterizations of type-I Pareto and Log-normal distributions in terms of inequality curves have been proven and an extension to the case of distributions with regularly varying tails has been discussed. Associated testing procedures, aimed at helping on evaluation of the graphical results have been developed. Beyond the task of providing support in individuating Pareto or Log-normal models, the graphical analysis has been compared to well known graphical methods used for analysing the tail behaviour of distributions. The estimated curve λ shows quite a smooth and regular behaviour already for small sample sizes, while the estimated curve Z needs moderate sample sizes in order to be effective. These characteristics reflects on the performance of the testing procedure discussed, quite powerful in most cases the test based on λ , with performance depending on sample size in the case of the test based on Z . However it is important to point out that these testing procedures are not proposed as goodness-of-fit strategies *per se*, rather they should support interpretation of the graphs.

Indeed, as shown in the examples, the *joint* use of the curves λ and Z do actually provide an effective tool for separating cases which other graphical displays typically do not. In particular, the analysis proposed in Section 4, either on simulated and real data, shows the effectiveness in distinguishing the difficult case of Log-normal versus Pareto.

6 Proofs

Proof of Theorem 1. It is trivially verified that if F satisfies (3) then $\lambda(p) = 1/\alpha$. Suppose now that $\lambda(p) = k$, $p \in (0, 1)$, where k is some constant. Then it must hold that $1 - L(p) = (1 - p)^k$ or equivalently, after some algebraic manipulation,

$$\int_0^p F^{-1}(u) du = \mu[1 - (1 - p)^k]$$

Taking derivatives on both sides we have that

$$\frac{d}{dp} \int_0^p F^{-1}(u) du = \frac{d}{dp} \mu[1 - (1 - p)^k],$$

which gets

$$F^{-1}(p) = \mu k(1 - p)^{k-1}$$

from which, setting $x_p = F^{-1}(p)$, which implies $p = F(x_p)$, it follows that, after some further elementary manipulations,

$$\left(\frac{x_p}{\mu k}\right)^{1/(k-1)} = 1 - F(x_p).$$

Setting $1/(k - 1) = -\alpha$, properly normalized, the above F follows (3). □

Proof of Theorem 2. Definitions of F and Q imply that for any $p \in (0, 1)$:

$$\int_0^{F^{-1}(p)} f(x) dx = \frac{1}{E(X)} \int_0^{Q^{-1}(p)} x f(x) dx. \quad (14)$$

If $Z(p) = 1 - c^{-1}$, then $Q^{-1}(p) = cF^{-1}(p)$ for any $p \in (0, 1)$, whereas $c \geq 1$, and Equation (14) becomes:

$$\int_0^{F^{-1}(p)} f(x) dx = \frac{1}{E(X)} \int_0^{cF^{-1}(p)} xf(x) dx. \quad (15)$$

If both sides of (15) are differentiated with respect to $x_p = F^{-1}(p)$, it follows that:

$$f(x_p) = \frac{1}{E(X)} c^2 x_p f(cx_p). \quad (16)$$

Under theorem assumptions, Equation (16) is equivalent to (15), hence if it is proved that Equation (16) is satisfied only by log-normal random variables whose log-variance parameter equals $\ln c$, the proof is complete. In the rest of the proof the functional equation (16) in f will be solved and it will be proved that the solution is unique.

The following functional equation has to be solved:

$$f(cx) = \frac{1}{c^2 x} \left(\int_0^{+\infty} wf(w) dw \right) f(x), \quad (17)$$

with respect to a function $f: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that $\int_{\mathbb{R}} f(w) dw = 1$, and where $c \in [1, +\infty)$. Equation (17) can be restated as it follows:

$$f(cx) = \frac{E(X)}{c^2 x} f(x). \quad (18)$$

Let $g: \mathbb{R} \rightarrow \mathbb{R}$ be a function defined as:

$$g(x) = \ln f(e^x). \quad (19)$$

From (19) it follows that functional equation (18) can be restated as:

$$g(\omega + \ln c) = \ln E(X) - 2 \ln c - \omega + g(\omega). \quad (20)$$

If both sides of Equation (20) are differentiated twice with respect to ω , we obtain the following second order differential equation:

$$g''(\omega + T) = g''(\omega),$$

which characterises a function of period $T = \ln c$. From the assumption in the lemma, g'' is expandable in Fourier series, thus:

$$g''(\omega) = \sum_{n=-\infty}^{+\infty} s_n e^{\frac{2\pi}{T} in\omega}, \quad (21)$$

where $s_n \in \mathbb{C}$ are the coefficients of the Fourier series expansion, and are such that, for any $n \in \mathbb{N}$, s_n and s_{-n} are complex conjugates. From (21) it follows that:

$$g(\omega) = -\frac{T^2}{4\pi^2} \left(\sum_{n \in \mathbb{Z}^*} \frac{s_n}{n^2} e^{\frac{2\pi}{T} in\omega} \right) + s_0 \omega^2 + k_1 \omega + k_2, \quad (22)$$

where $k_1, k_2 \in \mathbb{R}$ resulted from integrations, and $\mathbb{Z}^*$ is the set of integers except zero.

According to the definition (19) of g , Equation (22) implies that:

$$f(x) = e^{-\frac{T^2}{4\pi^2} \sum_{n \in \mathbb{Z}^*} \frac{s_n}{n^2} e^{\frac{2\pi}{T} i n \ln x}} e^{s_0 \ln^2 x + k_1 \ln x + k_2}. \quad (23)$$

Solution (23) satisfies (21), however specific constraints on constants s_0 , k_1 , k_2 and on s_n (for $n \in \mathbb{Z}^*$) are implicitly stated by (18). Consider the following equation:

$$\begin{aligned} \ln \left(\frac{x f(cx)}{f(x)} \right) &= \ln x - \frac{T^2}{4\pi^2} \sum_{n \in \mathbb{Z}^*} \left(e^{\frac{2\pi}{T} i n \ln c} - 1 \right) \frac{s_n}{n^2} e^{\frac{2\pi}{T} i n \ln x} + \\ &+ s_0 \ln^2 c + 2s_0 \ln x \ln c + k_1 \ln c + k_2 \end{aligned} \quad (24)$$

and note that, according to Equation (18), $x f(cx)/f(x)$ must be constant. It follows that the right-hand side of Equation (24) should be constant as well and should equal $\ln E(X) - 2 \ln c$. This implies that:

$$s_0 = -\frac{1}{2 \ln c} \quad s_n = 0 \quad \forall n \in \mathbb{Z}^*, \quad (25)$$

since, for any $n \in \mathbb{Z}^*$, $(e^{\frac{2\pi}{T} i n \ln c} - 1)s_n$ and $(e^{-\frac{2\pi}{T} i n \ln c} - 1)s_{-n}$ are complex conjugates. It follows that the following condition must hold:

$$-\frac{1}{2} \ln c + k_1 \ln c + k_2 = \ln E(X) - 2 \ln c. \quad (26)$$

Restrictions

$$\int_0^{+\infty} f(x) dx = 1, \quad \int_0^{+\infty} x f(x) dx = E(X)$$

implies that

$$k_1 = \frac{\ln E(X)}{\ln c} - \frac{3}{2}, \quad k_2 = 0. \quad (27)$$

If values (25) and (27) are substituted into (23), the probability density function of a log-normal distribution with log-variance $\ln c$ is found. Uniqueness of the solution stems from uniqueness theorems in differential equations theory. $\square$

7 Acknowledgement

The authors would like to thank two anonymous referees whose constructive comments lead to an improved version of the paper.

References

- Afhami, B., & Madadi, M. (2017). Entropy-based goodness-of-fit tests for the Pareto I distribution. *Communications in Statistics-Theory and Methods*, 46(8), 3649–3666.
- Arcagni A., Porro F. (2016) . A comparison of income distributions models throught inequality curves. *Statistica & Applicazioni*. XIV (2), 123–144

- Clauset, A.n; Shalizi, C. R.; Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.* 51(4), 661–703.
- Cirillo, P. (2013). Are your data really Pareto distributed? *Physica A: Statistical Mechanics and its Applications*, 392(23), 5947–5962.
- D’Agostino, R. B., Stephens, M. A., (1986). *Goodness-of-fit-techniques*. CRC Press.
- Dey, A. K., & Kundu, D. (2009). Discriminating between the log-normal and log-logistic distributions. *Communications in Statistics-Theory and Methods*, 39(2), 280–292.
- Eeckhout, J. (2004). Gibrat’s law for (all) cities. *The American Economic Review* 94(5), 1429–1451.
- Eeckhout, J. (2009). Gibrat’s law for (all) cities: reply. *The American Economic Review* 99(4), 1676–1683.
- Gabaix, X., Ioannides Y.M. (2004). The evolution of city size distributions. *Handbook of Regional and Urban Economics* 4, 2341–2378.
- Goldie, C. M. (1977). Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*, 9(4), 765–791.
- 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.
- Heyde, C. C.; Kou, S. G. (2004). On the controversy over tailweight of distributions. *Oper. Res. Lett.* 32(5), 399–408.
- Jammalamadaka, S. R., & Gorla, M. N. (2004). A test of goodness-of-fit based on Gini’s index of spacings. *Statistics & probability letters*, 68(2), 177–187.
- Jammalamadaka, S. R., & Tian, G. (2018). Indirect Inference for Fitting Income Distributions. *Journal of the Indian Society for Probability and Statistics*, 19(2), 345–358.
- Johnson N. L., S. Kotz, N. Balakrishnan (1995) *Continuous Univariate Distributions*, Vol. 2, 2nd ed, Wiley.
- Jia, M., Taufer, E., Dickson, M. (2018). Semi-parametric regression estimation of the tail index. *Electronic Journal of Statistics* 12, 224–248.
- Kratz, M. F., Resnick, S. I. (1996). The QQ-estimator and heavy tails. *Comm. Statist. Stochastic Models* 12 (4), 699–724.
- Kundu, D., Gupta, R. D., & Manglick, A. (2005). Discriminating between the log-normal and generalized exponential distributions. *Journal of Statistical Planning and Inference*, 127(1-2), 213–227.

- Levy, M. (2009). Gibrat's law for (all) cities: comment. *The American Economic Review* 99(4), 1672–1675.
- Lloyd, C. J. (2005). Estimating test power adjusted for size. *Journal of Statistical Computation and Simulation*, 75(11), 921–933.
- Meintanis, S. G. (2008) A powerful method of assessing the fit of the lognormal distribution. *Comm. Statist. Theory Methods* 37(11-12), 1948–1958.
- Meintanis, S. G. (2009). A unified approach of testing for discrete and continuous Pareto laws. *Statistical Papers*, 50(3), 569–580.
- Obradović, M., Jovanović, M., & Milošević, B. (2015). Goodness-of-fit tests for Pareto distribution based on a characterization and their asymptotics. *Statistics*, 49(5), 1026–1041.
- Raqab, Mohammad Z.; Al-Awadhi, Shafiqah A.; Kundu, Debasis; (2018). Discriminating among Weibull, log-normal, and log-logistic distributions. *Comm. Statist. Simulation Comput.* 47(5), 1397–1419.
- Rizzo, Maria L. (2009). New goodness-of-fit tests for Pareto distributions. *ASTIN Bulletin: The Journal of the IAA* 39(2), 691–715.
- Taufer, E., Santi, F., Novi Inverardi, P. L., Espa, G., Dickson, M. M. (2020). Extreme Value Index Estimation by Means of an Inequality Curve. *Mathematics* 8, 1834.
- Volkova, K. (2016). Goodness-of-fit tests for the Pareto distribution based on its characterization. *Statistical Methods & Applications*, 25(3), 351–373.
- 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
- Zenga M.(1990). Concentration curves and Concentration indexes derived from them. In *Income and Wealth Distribution, Inequality and Poverty*, Dagum, C., Zenga, M. (Ed.), Springer -Verlag, 94–110.
- Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica e Applicazioni* 1, 3–27.