A goodness of fit test for the Pareto distribution

Emanuele Taufer\textsuperscript{a}, Flavio Santi\textsuperscript{a}, Giuseppe Espa\textsuperscript{a}, Maria Michela Dickson\textsuperscript{b},

\textsuperscript{a}Department of Economics and Management, University of Trento - Italy  \textsuperscript{a}Department of Statistics, University of Padua - Italy

Abstract

The Zenga (1984) inequality curve \( \lambda(p) \) is constant in \( p \) for Type I Pareto distributions. This characterizing behavior will be exploited to obtain graphical and analytical tools for tail analysis and goodness of fit tests. A testing procedure for Pareto-type behavior based on a regression of \( \lambda(p) \) against \( p \) will be introduced.

Keywords: Tail index, inequality curve, non-parametric estimation, goodness-of-fit.

1 Introduction

Let \( X \) be a positive random variable with finite mean \( \mu \), distribution function \( F \), and probability density \( f \). The inequality curve, \( \lambda(p) \), defined in \cite{11} is defined as:

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

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

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

\( \lambda(p) \) can be used to define a concentration measure as it has been done in \cite{11}. Here we exploit the curve in order to define goodness-of-test for the Pareto. In fact as it will be more formally shown below \( \lambda(p) \) is constant in \( p \) for type I Pareto distributions. Indeed the above properties can also be exploited in order to define graphical tools for the analysis of distributions and their tails. For related works see \cite{8, 10, 2, 4, 5, 6, 7, 9}.

For a Type I Pareto distribution \cite{3} 573 ff.] with

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

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 \cite{11} is constant in \( p \) if, and only if, \( F \) satisfies \cite{3}.


2 Goodness-of-fit tests

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

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

Under the Glivenko-Cantelli theorem (see e.g. [9]) 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:

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

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

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

The choice of $i = 1, \ldots, n - \lfloor \sqrt{n} \rfloor$ guarantees that $\hat{\lambda}_i$ is consistent for $\lambda_i$ for each $p_i = i/n$ as $n \to \infty$.

Goodness-of-fit tests can be defined by linear regression of $\lambda_i$, on $p_i$. From Lemma 1 for a distribution $F$ satisfying (3) with $\lambda > 1$, for any choice of $p_i, 0 < p_i < 1, i = 1, \ldots, m$, 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 $X_1, \ldots, X_n$, 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 $\beta_0$ can be used to estimate $\alpha$ 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 $\beta_0$, define

$$ \hat{\beta}_0 = \frac{1}{m} \sum_{i=1}^{m} \hat{\lambda}_i, \quad \hat{\beta}_1 = \sum_{i=1}^{m} \frac{\hat{\lambda}_i (p_i - \bar{p})}{S_p^2} = \sum_{i} \frac{\hat{\lambda}_i c(p_i)}{S_p^2} $$

where $\bar{p}$ is the mean of the $p_i$’s and $S_p^2 = \sum_{i} (p_i - \bar{p})^2, c(p_i) = (p_i - \bar{p})/S_p^2$. 


Note that since
\[
\sum_{i=1}^{n} i = \frac{n(n + 1)}{2} \quad \text{and} \quad \sum_{i=1}^{n} i^2 = \frac{n(n + 1)(2n + 1)}{6}
\]
then, for \( p_i = i/n, i = 1, \ldots, m, m = n - \lfloor \sqrt{n} \rfloor, \)
\[
\bar{p} = \frac{1}{2} \frac{m(m+1)}{n^2} = \frac{1}{2} + O \left( \frac{1}{\sqrt{n}} \right) \quad \text{and} \quad S_p^2 = \frac{1}{12} \frac{m(m^2-1)}{n^2} = O(n).
\]

**Remark 1.** Since \( \lambda(p) \) does not depend on location parameters, one can construct goodness-of-fit tests free of \( x_0 \) for the Pareto distribution.

A formal test for the general hypothesis that the data comes from a distribution \( F \) satisfying \( \lambda(p) \) i.e.
\[
H_0 : F \text{ is } Pa(\alpha, x_0), \quad \alpha > 1, \quad x_0 > 0
\]
The null hypotheses is rejected if |\( \hat{\beta}_1 \)| is large. In order to carry on practically the test we have two possibilities. The first is to use a normal approximation to \( \hat{\beta}_1 \), properly normalized. This is feasible only for the cases \( \alpha > 2 \).

The second way is to carry on a parametric bootstrap procedure as follows:

1. Given a random sample of size \( n \), estimate \( \hat{\alpha} = 1/\hat{\beta}_0 \) and \( \hat{\beta}_1 \).
2. Generate a sample of size \( n \) from a \( Pa(\hat{\alpha}, 1) \)and estimate \( \hat{\beta}_1 \). Note that since \( \lambda(p) \) does not depend on \( x_0 \), we do not need to estimate it and use, for example, always the same value \( 1 \).
3. Repeat step 2 \( M \) times.
4. Get an estimated \( p \)-value of \( \hat{\beta}_1 \) from the bootstrap distribution.

To compare the performance of the test proposed here, consider \( \text{[10]} \). For the distributions and sample sizes considered in Table 8 of \( \text{[10]} \), Table 1 contains the power estimates obtained with the parametric bootstrap for tests of level 0.05. Results are based on 500 samples of size \( n \) from null and alternative distributions; for each of them a parametric bootstrap with \( M = 500 \) was carried on.

We see that the test proposed here performs better than its competitors in several cases. The Log-normal distribution looks like a hard alternative for large values of the standard deviation. Further simulation and analyses will be carried on in a subsequent work.
Table 1: Estimated power for tests of size 0.05. Results based on 500 replications, $p$-value estimates were obtained by parametric bootstrap with $M = 500$.

| n  | Pa(2) | LN(1) | LN(2.5) | LN(3) | Exp | Ga(2) | LW(0.25) | LW(0.5) |
|----|-------|-------|---------|-------|-----|-------|----------|--------|
| 20 | 0.19  | 0.99  | 0.99    | 0.23  | 0.46|
| 50 | 0.054 | 1.00  | 0.19    | 0.04  | 1.00| 0.38   | 0.81     |
| 100| 0.048 | 1.00  | 0.61    | 0.03  | 1.00| 0.59   | 0.96     |
| 500| 0.038 | 1.00  | 0.81    | 0.08  | 1.00| 0.96   | 1.00     |
| 1000|0.044 | 1.00  | 0.95    | 0.12  | 1.00| 0.98   | 1.00     |

References

[1] Embrechts, P., C. Klüppelberg, T. Mikosch (1997). Modelling Extremal Events. Springer.
[2] 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.
[3] Johnson N. L., S. Kotz, N. Balakrishnan (1995) *Continuous Univariate Distributions, Vol. 2*, 2nd ed, Wiley.
[4] Jia, M., Taufer, E., Dickson, M. M. (2018). Semi-parametric regression estimation of the tail index. *Electronic Journal of Statistics* 12, 224–248.
[5] Leonenko, N. N., & Taufer, E. (2006). Weak convergence of functionals of stationary long memory processes to Rosenblatt-type distributions. *Journal of statistical planning and inference*, 136(4), 1220–1236.
[6] Leonenko, N., Petherick, S., & Taufer, E. (2013). Multifractal models via products of geometric OU-processes: Review and applications. *Physica A: Statistical Mechanics and its Applications*, 392(1), 7–16.
[7] McNeil, A. J., R. Frey, P. Embrechts (2005) *Quantitative Risk Management*, Princeton University Press.
[8] Meintanis, S. G., Ngatchou-Wandji, J., & Taufer, E. (2015). Goodness-of-fit tests for multivariate stable distributions based on the empirical characteristic function. *Journal of Multivariate Analysis*, 140, 171–192.
[9] Resnik, S. I. (1999) *A probability path*, Birkhäuser.
[10] Volkova, K. (2016). Goodness-of-fit tests for the Pareto distribution based on its characterization. *Statistical Methods & Applications*, 25(3), 351-373.
[11] 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