AJAS / SPAS
ISSN 2316-0861

African Journal of Applied Statistics
Vol. 4 (1), 2017, pages 157–164.
DOI: http://dx.doi.org/10.16929/ajas/2017.157.207

ISSN 2316-0861

Measuring inequality: application of semi-parametric methods to real life data

Tchilabalo Abozou Kpanzou[1,*], Tertius de Wet [2] and Gane Samb Lo.[3]

[1]Faculty of Science and Technology, University of Kara, Kara, Togo
[2]Faculty of Economics and Management Science, University of Stellenbosch, Stellenbosch, South Africa
[3]LERSTAD, Gaston Berger University, Saint-Louis, Senegal; Affiliated LASTA, UPMC, France; Associated Professor, African University of Science and Technologies (AUST), Nigeria.

Received May 8, 2017; Accepted June 10, 2017; Published Online July 10, 2017

Abstract. A number of methods have been introduced in order to measure the inequality in various situations such as income and expenditure. In order to curry out statistical inference, one often needs to estimate the available measures of inequality. Many estimators are available in the literature, the most used ones being the non parametric estimators. Kpanzou (2011) has developed semi-parametric estimators for measures of inequality and showed that these are very appropriate especially for heavy tailed distributions. In this paper we apply such semi-parametric methods to a practical data set and show how they compare to the non parametric estimators. A guidance is also given on the choice of parametric distributions to fit in the tails of the data.

Key words: income distribution; inequality measures; confidence intervals; extreme value theory.
AMS 2010 Mathematics Subject Classification : 62F10; 62G05; 62P05.

Presented by Dr Diakarya Barro, University of Ouagadougou, Burkina Faso
Corresponding Member of the Editors Board.

*Corresponding author Tchilabalo Abozou Kpanzou: kpanzout@gmail.com
Tertius de Wet : tdewet@sun.ac.za
Gane Samb Lo : gane-samb.lo@edu.ugb.sn/gslo@aust.edu.ng

Résumé. Des méthodes ont été introduites en vue de mesurer l'inégalité dans diverses situations comme, par exemple, dans la répartition du revenu. Pour faire de l'inférence statistique, l'on a souvent besoin d'estimer les mesures d'inégalité disponibles. Des estimateurs ont été developpés, les plus utilisés étant les estimateurs non paramétriques. Kpanzou (2011) a developpé des estimateurs semi-paramétriques, estimateurs qui sont beaucoup plus appropriés surtout quand l'on a à traiter avec les distributions à queue épaisse. Dans cet article, nous appliquons ces estimateurs semi-paramétriques à des données de la vie réelle et les comparons à leurs équivalents non paramétriques. Une indication est aussi donnée sur le choix des distributions paramétriques à ajuster à la queue.

1. Introduction

An area of application of the statistical methods is that of measures of inequality. These are very popular in economics and have applications in many other branches of Science, see e.g. EU-SILC (2004); Hullinger and Schoch (2009); Allison (1978). In order to curry out statistical inference, one often needs to estimate the available measures of inequality, namely the Gini, the Generalised entropy, the Atkinson, the quintile share ratio, just to mention a few. Many estimators are available in the literature, the most used ones being the non parametric estimators. Details on the definitions of inequality measures as well as non parametric estimators can be found in, e.g., Cowell and Flachaire (2007), Langel and Tillé (2011), Kpanzou (2011), Kpanzou (2014) and Kpanzou (2015). Further on inequality and poverty measures can be found in Lo (2013) and Lo and Mergane (2013).

Kpanzou (2011) has developed semi-parametric estimators for measures of inequality and showed that these are very appropriate especially for heavy tailed distributions. In this paper we apply such semi-parametric methods to a practical data set and show how they compare to the non parametric estimators. A guidance is also given on the choice of parametric distributions to fit in the tails of the data. The data used are claims data from a South African short term insurer.

The remainder of the paper is organized as follows. In Section 2, we briefly describe the semi-parametric methods used. Results on application to the aforementioned data set are given in Section 3. We give some concluding remarks in Section 4.

2. Methodology

In this section we present the semi-parametric estimation method. The procedure relies on the estimation of the underlying distribution in a semi-parametric setting.

Define a semi-parametric distribution function by

$$\tilde{F}(x) = \begin{cases} F(x), & x \leq x_0, \\ F(x_0) + (1 - F(x_0))F_\theta(x), & x > x_0, \end{cases} \quad (1)$$

for a given x_0, where F_θ is a parametric distribution satisfying the condition $F_\theta(x_0) = 0$, and F is an unknown distribution. Note that θ can be a vector parameter. Choose

$x_0 = Q(F, 1-\alpha),\ \alpha \in [0,1]$, where Q denotes the quantile function associated with F.

We then have

$$\begin{aligned} F(x_0) + (1-F(x_0))F_\theta(x) &= 1-\alpha+(1-(1-\alpha))F_\theta(x) \\ &= 1-\alpha+\alpha F_\theta(x) \\ &= 1-\alpha(1-F_\theta(x)). \end{aligned}$$

It follows that

$$\tilde{F}(x) = \begin{cases} F(x), & x \leq Q(F,1-\alpha), \\ 1-\alpha(1-F_\theta(x)), & x > Q(F,1-\alpha), \end{cases} \tag{2}$$

where F_θ satisfies the condition $F_\theta(Q(F,1-\alpha)) = 0$.

Estimating θ by $\widehat{\theta}$, and estimating F by the empirical distribution function F_n, we estimate the underlying distribution semi-parametrically as

$$\tilde{F}_n(x) = \begin{cases} F_n(x), & x \leq Q(F_n,1-\alpha), \\ 1-\alpha(1-F_{\widehat{\theta}}(x)), & x > Q(F_n,1-\alpha). \end{cases} \tag{3}$$

Equation (3) is very important as it estimates the underlying distribution. However, a choice of the parametric distribution F_θ is required in order to make the estimation process possible. We address that issue by making use of results from Extreme Value Theory (EVT). See e.g. Beirlant et al. Beirlant et al. (2004) for these results.

Given a certain threshold u, we consider the conditional distribution of the exceedance of u given that u was exceeded. We consider two types of exceedances:

1. $X-u$ given $X > u$ (absolute exceedance).
2. X/u given $X > u$ (relative exceedance).

From EVT, if F belongs to the domain of attraction of the generalized extreme value distribution, the following limiting results hold when $u \to \infty$.

1. The distribution of $X-u|X>u$ converges to the generalized Pareto distribution (GPD)

$$G(x;\sigma,\gamma) = 1-\left(1+\frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}},\ x>0, \tag{4}$$

   where $\sigma > 0$ is the scale parameter and $\gamma > 0$ is the extreme value index (EVI).
2. The distribution of $X/u|X>u$ converges to the strict Pareto (Pa) distribution

$$F_P(x) = 1-x^{-\frac{1}{\gamma}},\ x>1,\ \gamma>0, \tag{5}$$

   where γ is the EVI.
3. A second order approximation of the distribution of the relative exceedance is the perturbed Pareto distribution (PPD) defined by the survival function

$$\overline{G}(x;\gamma,c,\tau) = 1-G(x;\gamma,c,\tau) = (1-c)x^{-1/\gamma} + cx^{-1/\gamma-\tau}, \tag{6}$$

   where $x > 1$, $\gamma > 0$, $\tau > 0$ and $c \in (-1/\tau, 1)$. The idea here is to fit such a PPD to the relative exceedance, aiming for a more accurate estimation of the unknown tail. See Beirlant et al. (2004) for more details.

Once the semi-parametric estimators for the distribution function are obtained, we plug them in the functional form of the inequality measures to obtain their SP estimators. Here we only give the estimator for Gini when fitting the GPD in order to illustrate the procedures we are aiming to apply.

Recall that given a distribution function F of a random variable X with mean μ, the ordinary Gini coefficient can be defined as

$$I_G = \frac{1}{\mu}\int_0^{\infty} F(x)(1-F(x))dx. \tag{7}$$

Consider a random sample $X_1, X_2, \dots, X_n$ from F, with associated order statistics $X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$, and suppose the threshold above which a parametric distribution is fitted is $x_0 = Q(F, 1-\alpha)$. The Gini coefficient is then estimated semi-parametrically as

$$\widehat{I}_{SPG} = \frac{1}{\widehat{\mu}}\int_0^{\infty} \tilde{F}_n(x)(1-\tilde{F}_n(x))dx, \tag{8}$$

where $\tilde{F}_n$ is given in Equation (3), and $\widehat{\mu}$ is the estimator of μ using $\tilde{F}_n$, that is

$$\widehat{\mu} = \int_0^{\infty} x d\tilde{F}_n(x). \tag{9}$$

Estimate the threshold u by $X_{n-k,n}$, and assume that for a large n, the GPD is a reasonable approximation to the distribution of the exceedances $X_{n-k+1,n} - X_{n-k,n}, X_{n-k+2,n} - X_{n-k,n}, \dots, X_{n,n} - X_{n-k,n}$ for a given k. A semi-parametric estimator for I_G is then given by

$$\widehat{I}_{SPG} = \frac{1}{\widehat{\mu}}\sum_{i=1}^{n-k-1}\frac{i}{n}(1-\frac{i}{n})(X_{i+1,n} - X_{i,n}) + \frac{k\widehat{\sigma}\left[2n-k-\widehat{\gamma}(n-k)\right]}{n^2\widehat{\mu}(1-\widehat{\gamma})(2-\widehat{\gamma})}, \tag{10}$$

where $\widehat{\sigma}$ and $\widehat{\gamma}$ are estimators for the unknown scale and shape parameters σ and γ of the GPD using the exceedances, with $\widehat{\gamma} < 1$, and

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n-k} X_{i,n} + \frac{k}{n}\left(X_{n-k,n} + \frac{\widehat{\sigma}}{1-\widehat{\gamma}}\right) \tag{11}$$

is an estimator for μ.

Theoretical details and other estimators can be found in Kpanzou (2011). We remind that the aim of this work is to show how the semi-parametric methods can be applied to real life data, and so we direct the reader to the previous reference for all the theoretical derivations.

3. Numerical applications and interpretations

Here we consider claims data from a South African short term insurer. These consist of a portfolio of claims from 1 July 2004 to 21 July 2006. The dates used, were the dates the claims occurred, and not the dates the claims were registered. The claim amounts were the total claim amounts and any excesses paid by the client were ignored. The claim amounts

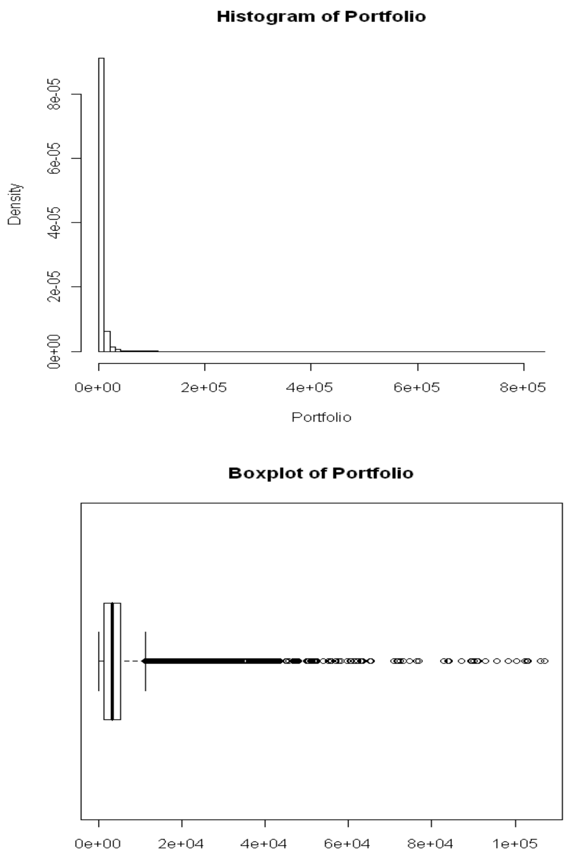

Fig. 1. Histogram and Boxplot for Portfolio data

were adjusted for inflation to July 2006 as base month. Finally, any negative or zero claim amounts were deleted from the data set. Negative amounts occur if, for instance, the value of the items salvaged from the wreck exceeds the claim amount in value. The final sample size is 16104. We will refer to the data set as Portfolio.

The histogram and the boxplot for the data are shown in Fig. 1., giving an idea of the tails of the distribution.

Table 1. Descriptive statistics for the Portfolio data set

	Sample size n	Median	MAD	Maximum
Portfolio	16104	3268.5050	2951.3525	835567.7000

Table 2. Non parametric estimates of inequality measures for Portfolio data set

	Gini	GE0	A1	QSR
Portfolio	0.5767	0.7070	0.5069	34.8292

Table 3. Non parametric and semi-parametric estimates of inequality measures for Portfolio data set

Portfolio	NP	SPGPD	SPPa	SPPPD
Gini	0.5767	0.5923	0.5145	0.5850
GE0	0.7070	0.7499	0.6999	0.7453
A1	0.5069	0.5276	0.5034	0.5072
QSR	34.8292	36.6990	34.4200	34.9122

We see from Fig. 1. that the data have heavy tails and so we can use them to illustrate our methods. Table 1 gives some descriptive statistics (Median, Median Absolute Deviation (MAD) and maximum value).

We next give the estimates of the inequality measures, starting with the non parametric ones summarised in Table 2.

The semi-parametric estimates for the inequality measures using the methods described in Section 2 (fitting the GPD, the strict Pareto and the PPD in the tails) are now calculated. We use 10% of the data (upper order statistics) in each case both to estimate the parameters in the tail distribution and to fit that distribution. For comparison purposes we put together both the non parametric and semi-parametric estimates. The results are given in Table 3. We only show these for the Gini coefficient, the generalised entropy with parameter 0 (GE0), the Atkinson with parameter 1 (A1) and the quintile share ratio (QSR). The non parametric estimators are denoted by NP, the semi-parametric (SP) ones by SPGPD (when fitting the GPD in the tails), SPPa (when fitting the Pa) and SPPPD (when fitting the PPD).

We see from these tables that the SP estimates are not far away from the NP estimates. The interpretation is that for the portfolio data considered, we have a high level of inequality in the distribution. In fact, Gini is around 60%, GE0 around 70%, A1 around 50%, and the QSR shows that the 20% upper claims are worth 35 times the 20% lowers claims. For the interpretation of inequality measures, see e.g. Creedy (2014).

However, the advantage of using the semi-parametric estimators relies on the fact that they are more resistant to extreme values in the data, as shown by Kpanzou (2011). In addition, although we do not show confidence intervals in this paper, note that the coverage probabilities are very satisfactory for confidence intervals based on SP methods. This is also demonstrated in Kpanzou (2011).

Table 4. Measure of representativeness

$\underline{X}$	$R(\underline{X}, GPD)$	$R(\underline{X}, Pa)$	$R(\underline{X}, PPD)$
Portfolio	0.97362	0.99987	0.99988

The question that is now raised is how one can choose the right parametric distribution to fit in the tails. Such a choice is done among the GPD, the Pa and the PPD. In order to decide which one to use in a specific application, we suggest the use of a measure of representativeness of the sample to each distribution. In our applications, we use the one given by Bertino (2006). Below is the description of that measure.

Let $\underline{X} = (X_1, X_2, \ldots, X_n)$ be a simple random sample of size n from a distribution F and denote by $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ its associated order statistics. A representativeness index is given by

$$R(\underline{X}, F) = 1 - \frac{12n}{4n^2-1} \sum_{i=1}^{n} \left(F(X_{i.n}) - \frac{2i-1}{2n} \right)^2 . \tag{12}$$

This index is used to measure how well the sample $\underline{X}$ represents its parent distribution. Therefore, it can be used to select the model one should use when having to choose between a number of different models.

In our case, we use the measure to decide between the three parametric choices, GPD, strict Pareto and PPD. To do this we calculate $R(\underline{X}, GPD)$, $R(\underline{X}, Pa)$ and $R(\underline{X}, PPD)$, and the largest value determines the preferred distribution to use.

Applying the measure of representativeness to 10% of the upper values of the Portfolio data considered, we obtain the results in Table 4.
These results show that the strict Pareto and the PPD are well represented by the data in the tails and so either of them can be used. The corresponding estimators are the most reliable, and should be considered for assessing the inequality in the given situation.

Given a practical data set, in order to use a semi-parametric method, a two-step approach would be to first determine which of the three parametric distributions to use in the tail estimation. The corresponding semi-parametric procedure can then be used as a preferred choice to estimate the desired measures.

4. Conclusions

In this paper we have illustrated the application of semi-parametric estimators to real life data. Such estimators indeed perform very well, especially in the case of heavy tailed distributions. They are based on semi-parametric estimators of the underlying distribution using results from Extreme Value Theory. As the name indicates, a part of a semi-parametric estimator is made up with non parametric estimation and the other part (the tails) with parametric estimation. Three options have been considered for fitting such parametric distributions, namely the generalized Pareto, the strict Pareto and the perturbed Pareto. Given a data set, one therefore needs to decide which one of the three he should use for the estimation. We have addressed this question by suggesting the use of the measure of representativeness proposed by Bertino (2006). In order to illustrate the way one should use it, this measure has also been applied to the portfolio data set considered.

From the analysis done in this paper, we make the following recommendations. Suppose one has a data set from a heavy tailed distribution. Then:

(i) Choose a threshold above which a particular distribution fits well (we have the three possibilities mentioned above);
(ii) Use the measure of representativeness to choose the best fitting distribution;
(iii) Calculate the estimates of the inequality measures using the estimators corresponding to the distribution obtained in (ii).

Applying this procedure guarantees, to some extent, that the conclusions drawn from the analysis will be more reliable.

References

Allison P.D. (1978). Measures of inequality. *American Sociological Review,* 43, 865–880.

Beirlant J., Goegebeur Y., Segers J. and Teugels J. (2004). *Statistics of Extremes: Theory and Applications.* Wiley, New York.

Bertino S. (2006). A Measure of Representativeness of a Sample for Inferential Purposes. *International Statistical Review,* 74, 149–159.

Cowell F.A. and Flachaire E. (2007). Income Distribution and Inequality Measurement: The Problem of Extreme Values. *Journal of Econometrics,* 141, 1044-1072.

Creedy J. (2014). Interpreting Inequality Measures and Changes in Inequality. *Working Papers in Public Finance, Victoria Business School,* 11, 1-21.

EU-SILC 131/04 (2004). *Common Cross-sectional EU Indicators Based on EU-SILC; The Gender Pay Gap.* Eurostat, Luxembourg.

Hulliger B. and Schoch T. (2009). The Robustification of the Quintile Share Ratio, *in the Proceedings of New Technologies and Techniques in Statistics (NTTS).* Eurostat, Brussel.

Kpanzou T.A. (2011). *Statistical Inference for Inequality Measures based on Semi-Parametric Estimators.* Ph.D. dissertation, Stellenbosch University.

Kpanzou, T.A. (2014). Asymptotic distribution of the quintile share ratio estimator. *Afrika Statistika,* 9 (1), 659-670.

Kpanzou, T.A. (2015). On the Influence Function of the Quintile Share Ratio. *Communications in Statistics – Simulation and Computation,* 44, 2492–2499.

Langel M. and Tillé Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference,* 141, 2976-2985.

Lo, G. S. (2013). The Generalized Poverty Index. *Far East Journal of Theoretical Statistics,* 42(1), 1-22.

Lo, G. S. and Mergane P. D. (2013). On the influence of the Theil-like inequality measure on the growth. *Applied Mathematics,* 4, 986-1000. doi:10.4236/am.2013.47136, arXiv:1210.3190.