A comparison of robust methods for Pareto tail modeling in the case of Laeken indicators

Andreas Alfons[1], Matthias Templ[1,2], Peter Filzmoser[1], and Josef Holzer[1,3]

This is a corrected reprint of a paper published in *Combining Soft Computing and Statistical Methods in Data Analysis*, volume 77 of *Advances in Intelligent and Soft Computing*.

Abstract The Laeken indicators are a set of indicators for measuring poverty and social cohesion in Europe. However, some of these indicators are highly influenced by outliers in the upper tail of the income distribution. This paper investigates the use of robust Pareto tail modeling to reduce the influence of outlying observations. In a simulation study, different methods are evaluated with respect to their effect on the quintile share ratio and the Gini coefficient.

1 Introduction

As a monitoring system for policy analysis purposes, the European Union introduced a set of indicators, called the *Laeken indicators*, to measure risk-of-poverty and social cohesion in Europe. The basis for most of these indicators is the EU-SILC *(European Union Statistics on Income and Living Conditions)* survey, which is an annual panel survey conducted in EU member states and other European countries. Most notably for this paper, EU-SILC data contain information on the income of the sampled households. Each person of a household is thereby assigned the same *equivalized disposable income* [9]. The subset of Laeken indicators based on EU-SILC is computed from this equivalized income, taking into account the sample weights.

In general the upper tail of an income distribution behaves differently than the rest of the data and may be modeled with a *Pareto* distribution. Moreover, EU-SILC data typically contain some extreme outliers that not only have a strong influence on some of the Laeken indicators, but also on fitting the

Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria, alfons@statistik.tuwien.ac.at, templ@tuwien.ac.at, p.filzmoser@tuwien.ac.at · Methods Unit, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria · now at Landesstatistik Steiermark, Hofgasse 13, 8010 Graz, Austria, josef.holzer@stmk.gv.at

Pareto distribution to the tail. Modeling the tail in a robust manner should therefore improve the estimates of the affected indicators.

The rest of the paper is organized as follows. Section 2 gives a brief description of selected Laeken indicators, while Section 3 discusses Pareto tail modeling. A simulation study is performed in Section 4 and Section 5 concludes.

2 Selected Laeken indicators

This paper investigates the influence of promising robust methods for Pareto tail modeling on the *quintile share ratio* and the *Gini coefficient*. Both indicators are measures of inequality and are highly influenced by outliers in the upper tail. Strictly following the Eurostat definitions [9], the indicators are implemented in the R package `laeken` [2].

For the following definitions, let $\boldsymbol{x} := (x_1, \ldots, x_n)'$ be the equivalized disposable income with $x_1 \leq \ldots \leq x_n$ and let $\boldsymbol{w} := (w_i, \ldots, w_n)'$ be the corresponding personal sample weights, where n denotes the number of observations.

2.1 Quintile share ratio

The income quintile share ratio is defined as the ratio of the sum of equivalized disposable income received by the 20% of the population with the highest equivalized disposable income to that received by the 20% of the population with the lowest equivalized disposable income [9]. Let $q_{0.2}$ and $q_{0.8}$ denote the weighted 20% and 80% quantiles of $\boldsymbol{x}$ with weights $\boldsymbol{w}$, respectively. With $I_{\leq q_{0.2}} := \{i \in \{1, \ldots, n\} : x_i \leq q_{0.2}\}$ and $I_{>q_{0.8}} := \{i \in \{1, \ldots, n\} : x_i > q_{0.8}\}$, the quintile share ratio is estimated by

$$QSR := \frac{\sum_{i \in I_{>q_{0.8}}} w_i x_i}{\sum_{i \in I_{\leq q_{0.2}}} w_i x_i}. \tag{1}$$

2.2 Gini coefficient

The Gini coefficient is defined as the relationship of cumulative shares of the population arranged according to the level of equivalized disposable income, to the cumulative share of the equivalized total disposable income received by them [9]. In mathematical terms, the Gini coefficient is estimated by

$$Gini := 100 \left[\frac{2 \sum_{i=1}^{n} \left(w_i x_i \sum_{j=1}^{i} w_j \right) - \sum_{i=1}^{n} {w_i}^2 x_i}{\left(\sum_{i=1}^{n} w_i\right) \sum_{i=1}^{n} \left(w_i x_i\right)} - 1 \right]. \tag{2}$$

3 Pareto tail modeling

The *Pareto* distribution is defined in terms of its cumulative distribution function

$$F_\theta(x) = 1 - \left(\frac{x}{x_0} \right)^{-\theta}, \qquad x \geq x_0, \tag{3}$$

where $x_0 > 0$ is the scale parameter and $\theta > 0$ is the shape parameter [12]. Furthermore, the density is given by

$$f_\theta(x) = \frac{\theta x_0^\theta}{x^{\theta+1}}, \qquad x \geq x_0. \tag{4}$$

In Pareto tail modeling, the cumulative distribution function on the whole range of x is modeled as

$$F(x) = \begin{cases} G(x), & \text{if } x \leq x_0, \\ G(x_0) + (1 - G(x_0)) F_\theta(x), & \text{if } x > x_0, \end{cases} \tag{5}$$

where G is an unknown distribution function [8].

Let n be the number of observations and let $\boldsymbol{x} = (x_1, \ldots, x_n)'$ denote the observed values with $x_1 \leq \ldots \leq x_n$. In addition, let k be the number of observations to be used for tail modeling. In this scenario, the threshold x_0 is estimated by

$$\hat{x}_0 := x_{n-k}. \tag{6}$$

On the other hand, if an estimate $\hat{x}_0$ for the scale parameter of the Pareto distribution has been obtained, k is given by the number of observations larger than $\hat{x}_0$. Thus estimating x_0 and k directly corresponds with each other. Various methods for the estimation of x_0 or k have been proposed [5, 6, 8, 17]. However, this paper is focused on evaluating robust methods for estimating the shape parameter θ (with respect to their influence on the selected Laeken indicators) once the threshold is fixed.

3.1 Hill estimator

The maximum likelihood estimator for the shape parameter of the Pareto distribution was introduced by [10] and is referred to as the *Hill* estimator. It is given by

$$\hat{\theta} = \frac{k}{\sum_{i=1}^{k} \log x_{n-k+i} - k \log x_{n-k}}. \tag{7}$$

Note that the Hill estimator is non-robust, therefore it is included for benchmarking purposes.

3.2 Weighted maximum likelihood (WML) estimator

The weighted maximum likelihood (WML) estimator [7, 8] falls into the class of M-estimators and is given by the solution $\hat{\theta}$ of

$$\sum_{i=1}^{k} \Psi(x_{n-k+i}, \theta) = 0 \tag{8}$$

with

$$\Psi(x, \theta) := w(x, \theta) \frac{\partial}{\partial \theta} \log f(x, \theta) = w(x, \theta) \left(\frac{1}{\theta} - \log \frac{x}{x_0} \right), \tag{9}$$

where $w(x, \theta)$ is a weight function with values in $[0, 1]$. In this paper, a Huber type weight function is used, as proposed in [8]. Let the logarithms of the relative excesses be denoted by

$$y_i := \log \left(\frac{x_{n-k+i}}{x_{n-k}} \right), \qquad i = 1, \ldots, k. \tag{10}$$

In the Pareto model, these can be predicted by

$$\hat{y}_i := -\frac{1}{\theta} \log \left(\frac{k+1-i}{k+1} \right), \qquad i = 1, \ldots, k. \tag{11}$$

The variance of y_i is given by

$${\sigma_i}^2 := \sum_{j=1}^{i} \frac{1}{\theta^2 (k-i+j)^2}, \qquad i = 1, \ldots, k. \tag{12}$$

Using the standardized residuals

$$r_i := \frac{y_i - \hat{y}_i}{\sigma_i}, \tag{13}$$

the Huber type weight function with tuning constant c is defined as

$$w(x_{n-k+i}, \theta) := \begin{cases} 1, & \text{if } |r_i| \leq c, \\ \frac{c}{|r_i|}, & \text{if } |r_i| > c. \end{cases} \tag{14}$$

For this choice of weight function, the bias of $\hat{\theta}$ is approximated by

$$\hat{B}(\hat{\theta}) = -\frac{\sum_{i=1}^{k} \left(w_i \frac{\partial}{\partial\theta} \log f_i \right) |_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-k+i}) - F_{\hat{\theta}}(x_{n-k+i-1}) \right)}{\sum_{i=1}^{k} \left(\frac{\partial}{\partial\theta} w_i \frac{\partial}{\partial\theta} \log f_i + w_i \frac{\partial^2}{\partial\theta^2} \log f_i \right) |_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-k+i}) - F_{\hat{\theta}}(x_{n-k+i-1}) \right)}, \tag{15}$$

where $w_i := w(x_{n-k+i}, \theta)$ and $f_i := f(x_{n-k+i}, \theta)$. This term is used to obtain a bias-corrected estimator

$$\tilde{\theta} := \hat{\theta} - \hat{B}(\hat{\theta}). \tag{16}$$

For details and proofs of the above statements, the reader is referred to [7, 8].

3.3 Partial density component (PDC) estimator

For the partial density component (PDC) estimator [16], the Pareto distribution is modeled in terms of the relative excesses

$$y_i := \frac{x_{n-k+i}}{x_{n-k}}, \qquad i = 1, \ldots, k. \tag{17}$$

The density function of the Pareto distribution for the relative excesses is approximated by

$$f_\theta(y) = \theta y^{-(1+\theta)}. \tag{18}$$

The PDC estimator is then given by

$$\hat{\theta} = \arg\min_{\theta} \left[w^2 \int f_\theta^2(y) dy - \frac{2w}{k} \sum_{i=1}^{k} f_\theta(y_i) \right], \tag{19}$$

i.e., by minimizing the integrated squared error criterion [15] using an incomplete density mixture model $w f_\theta$. The parameter w can be interpreted as a measure of the uncontaminated part of the sample and is estimated by

$$\hat{w} = \frac{\frac{1}{k} \sum_{i=1}^{k} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y) dy}. \tag{20}$$

See [16] and references therein for more information on the PDC estimator.

4 Simulation study

Various robust methods for the estimation of poverty and inequality indicators, mostly non-parametric, have been investigated in [17], but neither the WML nor the PDC estimator for Pareto tail modeling are considered

there. Preliminary results with income generated from theoretical distributions [11] are an indication that both estimators are promising in the context of Laeken indicators. This is further investigated in this section. However, variance estimation is not yet considered in this paper.

The simulations are carried out in R [14] using the package `simFrame` [1, 4], which is a general framework for statistical simulation studies. A synthetic data set consisting of 35 041 households and 81 814 individuals is used as population data in the simulation study. This data set has been generated with the R package `simPopulation` [3, 13] based on Austrian EU-SILC survey data from 2006 and is about 1% of the size of the real Austrian population. A thorough investigation in a close-to-reality environment using real-life sized synthetic Austrian population data is future work.

From the synthetic data, 500 samples are drawn using simple random sampling. Each sample consists of 6 000 households, which is roughly the sample size used in the real-life survey. With these samples, two scenarios are investigated. In the first scenario, no contamination is added. In the second, the equivalized disposable income of 0.25% of the households is contaminated. The contamination is thereby drawn from a normal distribution with mean $\mu = 1\,000\,000$ and standard deviation $\sigma = 10\,000$. Note that the *cluster effect* is considered, i.e., all persons in a contaminated household receive the same income. The threshold for Pareto tail modeling is in both cases set to $k = 275$ based on graphical exploration of the original EU-SILC sample with a Pareto quantile plot [5]. Furthermore, the tuning constant $c = 2.5$ is used for the bias-corrected WML estimator due to favorable robustness properties [11].

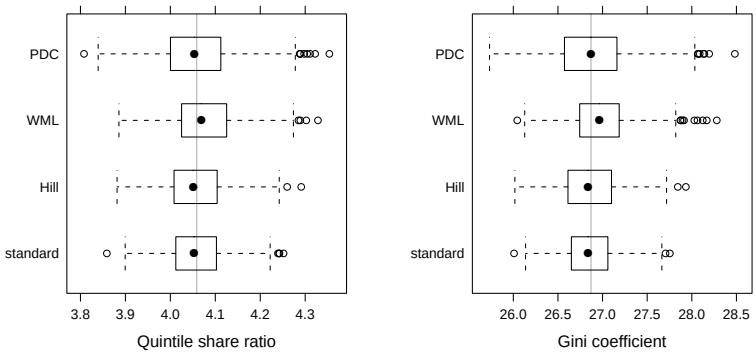

Fig. 1 Simulation results for the quintile share ratio *(left)* and the Gini coefficient *(right)* without contamination.

Figure 1 shows the results of the simulations without contamination for the quintile share ratio *(left)* and the Gini coefficient *(right)*. The three methods for tail modeling as well as the standard estimation method without tail

modeling behave very similarly and are very close to the true values, which are represented by the grey lines. This is also an indication that the choice of k is suitable.

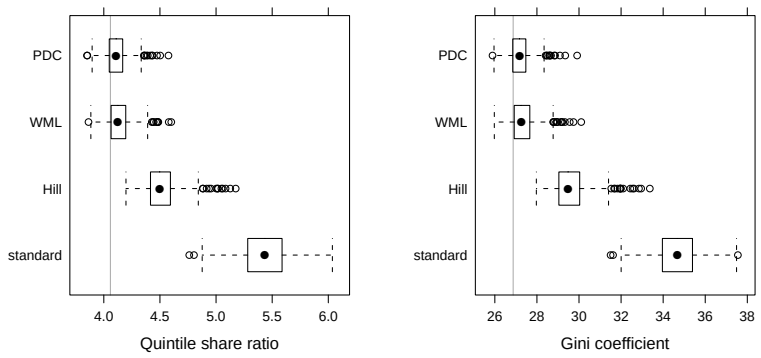

Fig. 2 Simulation results for the quintile share ratio *(left)* and the Gini coefficient *(right)* with 0.25% contamination.

Figure 2 shows the results of the simulations with 0.25% contamination for the quintile share ratio *(left)* and the Gini coefficient *(right)*. Even such a small amount of contamination completely corrupts the standard estimation of these inequality indicators. Fitting the Pareto distribution with the Hill estimator is still highly influenced by the outliers. The best results are obtained with the PDC estimator, while the WML estimator shows a slightly larger bias.

5 Conclusions and outlook

The quintile share ratio and the Gini coefficient, which are inequality indicators belonging to the set of Laeken indicators, are highly influenced by outliers. A simulation study for the case of simple random sampling showed that robust Pareto tail modeling can be used to reduce the influence of the outlying observations. The partial density component (PDC) estimator thereby performed best.

The simulation study in this paper is limited to simple random sampling because the estimators for Pareto tail modeling do not account for sample weights. Future work is to modify the estimators such that sample weights are taken into account, to investigate variance estimation, and to perform simulations using real-life sized synthetic population data.

Acknowledgements This work was partly funded by the European Union (represented by the European Commission) within the 7th framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322). Visit `http://ameli.surveystatistics.net` for more information on the project.

References

1. Alfons A (2009). `simFrame`: Simulation Framework. R package version 0.1.2.
2. Alfons A, Holzer J, Templ M (2010). `laeken`: Laeken indicators for measuring social cohesion. R package version 0.1.
3. Alfons A, Kraft S, Templ M, Filzmoser P (2010) Simulation of synthetic population data for household surveys with application to EU-SILC. *Research Report CS-2010-1*, Department of Statistics and Probability Theory, Vienna University of Technology. `http://www.statistik.tuwien.ac.at/forschung/CS/CS-2010-1complete.pdf`
4. Alfons A, Templ M, Filzmoser P (2009) `simFrame`: An object-oriented framework for statistical simulation. *Research Report CS-2009-1*, Department of Statistics and Probability Theory, Vienna University of Technology. `http://www.statistik.tuwien.ac.at/forschung/CS/CS-2009-1complete.pdf`
5. Beirlant J, Vynckier P, Teugels JL (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association*, **31**(436), 1659–1667.
6. Beirlant J, Vynckier P, Teugels JL (1996). Excess functions and estimation of the extreme-value index. *Bernoulli*, **2**(4), 293–318.
7. Dupuis DJ, Morgenthaler S (2002). Robust weighted likelihood estimators with an application to bivariate extreme value problems. *The Canadian Journal of Statistics*, **30**(1), 17–36.
8. Dupuis DJ, Victoria-Feser M-P (2006). A robust prediction error criterion for Pareto modelling of upper tails. *The Canadian Journal of Statistics*, **34**(4), 639–658.
9. EU-SILC (2004). Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. *EU-SILC 131-rev/04*, Eurostat, Luxembourg.
10. Hill BM (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**(5), 1163–1174.
11. Holzer J (2009). Robust methods for the estimation of selected Laeken indicators. Master's Thesis, Vienna University of Technology.
12. Kleiber C, Kotz S (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley, Hoboken, ISBN 0-471-15064-9.
13. Kraft S, Alfons A (2010). `simPopulation`: Simulation of synthetic populations for surveys based on sample data. R package version 0.1.1.
14. R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. `http://www.R-project.org`.
15. Terrell G (1990). Linear density estimates. In: *Proceedings of the Statistical Computing Section*, American Statistical Association, pp. 297–302.
16. Vandewalle B, Beirlant J, Christmann A, Hubert M (2007). A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics & Data Analysis*, **51**(12), 6252–6268.
17. Van Kerm P (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. *IRISS Working Paper Series 2007-01*, CEPS/INSTEAD.