Robust Pareto Tail Modeling for the Estimation of Indicators on Social Exclusion using the R Package laeken

Andreas Alfons[1], Matthias Templ[2], Peter Filzmoser[3], Josef Holzer[4]

Abstract In this vignette, robust semiparametric estimation of social exclusion indicators using the R package **laeken** is discussed. Special emphasis is thereby given to income inequality indicators, as the standard estimates for these indicators are highly influenced by outliers in the upper tail of the income distribution. This influence can be reduced by modeling the upper tail with a Pareto distribution in a robust manner. While the focus of the paper is to demonstrate the functionality of **laeken** beyond the standard estimation techniques, a brief mathematical description of the implemented procedures is given as well.

1 Introduction

From a robustness point of view, the standard estimators for some of the social exclusion indicators defined by Eurostat (2004, 2009) are problematic. In particular the income inequality indicators *quintile share ratio* (QSR) and *Gini coefficient* suffer from a lack of robustness. Consider, e.g., the QSR, which is estimated as the ratio of estimated totals or means (see Section 2.1 for an exact definition). It is well known that the classical estimates for totals or means have a breakdown point of 0, meaning that even a single outlier can distort the results to an arbitrary extent. In fact, the influence of a single observation in the upper tail of the income distribution on the estimation of the QSR is linear and therefore unbounded. For practical purposes, the standard QSR estimator thus cannot be recommended in many situations (cf. Hulliger and Schoch 2009). It is also important to note that the behavior of the Gini coefficient is similar to the behavior of the QSR.

The data basis for the estimation of the social exclusion indicators according to Eurostat (2004, 2009) is the *European Union Statistics on Income and Living Conditions* (EU-SILC), which is an annual panel survey conducted in EU member states and other European countries. On the one hand, EU-SILC data typically contain a considerable amount of *representative* outliers in the upper tail of the income distribution, i.e., correct observations that behave differently from the main part of the data, but that are not unique in the population and hence need to be considered for computing estimates of the indicators. On the other hand, EU-SILC data frequently contain some even more extreme *nonrepresentative* outliers, i.e., observations that are either incorrect or can be considered unique in the population. Consequently, such nonrepresentative outliers need to be excluded from the estimation process or downweighted.

[1] ORSTAT Research Center, Faculty of Business and Economics, KU Leuven
E-mail: andreas.alfons@econ.kuleuven.be

[2] Department of Statistics and Probability Theory, Vienna University of Technology
Methods Unit, Statistics Austria
E-mail: templ@tuwien.ac.at

[3] Department of Statistics and Probability Theory, Vienna University of Technology
E-mail: p.filzmoser@tuwien.ac.at

[4] Landesstatistik Steiermark
E-mail: josef.holzer@stmk.gv.at

As a remedy, the upper tail of the income distribution may be modeled with a *Pareto distribution* in order to recalibrate the sample weights or use fitted income values for observations in the upper tail when estimating the indicators (see Section 6). Nevertheless, classical estimators for the parameters of the Pareto distribution are highly influenced by the nonrepresentative outliers themselves. Using robust methods reduces the influence on fitting the Pareto distribution to the representative outliers and therefore on the estimation of the indicators.

Rather than evaluating these methods, the paper concentrates on showing how they can be applied in the statistical environment R (R Development Core Team 2011) with the add-on package **laeken** (Alfons et al. 2012). The basic design of the package, as well as standard estimation of the social exclusion indicators is discussed in detail in vignette `laeken-standard` (Templ and Alfons 2011a). Furthermore, the general framework for variance estimation is illustrated in vignette `laeken-variance` (Templ and Alfons 2011b). Those documents can be viewed from within R with the following commands:

```
R> vignette("laeken-standard")
R> vignette("laeken-variance")
```

Throughout the paper, the example data from package **laeken** is used. The data set is called `eusilc` and consists of 14 827 observations from 6 000 households. In addition, it was synthetically generated from Austrian EU-SILC survey data from 2006 using the data simulation methodology proposed by Alfons et al. (2011) and implemented in the R package **simPopulation** (Alfons and Kraft 2010). More information on the example data can be found in vignette `laeken-standard` or in the corresponding R help page.

```
R> library("laeken")
R> data("eusilc")
```

The rest of the paper is organized as follows. Section 2 gives a mathematical description of the Eurostat definitions of the social exclusion indicators QSR and Gini coefficient. In Section 3, the Pareto distribution is briefly discussed. Section 4 discusses a rule of thumb for estimating the threshold for the upper tail of the distribution, and illustrates graphical methods for exploring the data in order to find the threshold. Classical and robust estimators for the shape parameter of the Pareto distribution are described in Section 5. How to use Pareto tail modeling to estimate the social exclusion indicators is then shown in Section 6. Finally, Section 7 concludes.

2 Social exclusion indicators

This paper is focused on the inequality indicators *quintile share ratio* (QSR) and *Gini coefficient*, which are both highly influenced by outliers in the upper tail of the distribution. Note that for the estimation of the social exclusion indicators, each person in a household is assigned the same *eqivalized disposable income*. See vignette `laeken-standard` (Templ and Alfons 2011a) for the computation of the equivalized disposable income with the R package **laeken**.

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 (QSR)

The income *quintile share ratio* (QSR) is defined as the ratio of the sum of the 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 (Eurostat 2004, 2009).

For the estimation of the quintile share ratio from a sample, let $\hat{q}_{0.2}$ and $\hat{q}_{0.8}$ denote the weighted 20% and 80% quantiles, respectively. With $0 \leq p \leq 1$, these weighted quantiles are given by

$$\hat{q}_p = \hat{q}_p(\boldsymbol{x}, \boldsymbol{w}) := \begin{cases} \frac{1}{2}(x_j + x_{j+1}), & \text{if } \sum_{i=1}^{j} w_i = p \sum_{i=1}^{n} w_i, \\ x_{j+1}, & \text{if } \sum_{i=1}^{j} w_i < p \sum_{i=1}^{n} w_i < \sum_{i=1}^{j+1} w_i. \end{cases} \tag{1}$$

Using index sets $I_{\leq \hat{q}_{0.2}} := \{i \in \{1, \ldots, n\} : x_i \leq \hat{q}_{0.2}\}$ and $I_{>\hat{q}_{0.8}} := \{i \in \{1, \ldots, n\} : x_i > \hat{q}_{0.8}\}$, the quintile share ratio is estimated by

$$\widehat{QSR} := \frac{\sum_{i \in I_{>\hat{q}_{0.8}}} w_i x_i}{\sum_{i \in I_{\leq \hat{q}_{0.2}}} w_i x_i}. \tag{2}$$

With package **laeken**, the quintile share ratio can be estimated using the function `qsr()`. Sample weights can thereby be supplied via the `weights` argument.

```
R> qsr("eqIncome", weights = "rb050", data = eusilc)

Value:
[1] 3.971415
```

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 (Eurostat 2004, 2009).

For the estimation of the Gini coefficient from a sample, the sample weights need to be taken into account. In mathematical terms, the Gini coefficient is estimated by

$$\widehat{Gini} := 100 \left[\frac{2 \sum_{i=1}^{n} \left(w_i x_i \sum_{j=1}^{i} w_j \right) - \sum_{i=1}^{n} {w_i}^2 x_i}{\left(\sum_{i=1}^{n} w_i \right) \sum_{i=1}^{n} \left(w_i x_i \right)} - 1 \right]. \tag{3}$$

The function `gini()` is available in **laeken** to estimate the Gini coefficient. As before, sample weights can be specified with the `weights` argument.

```
R> gini("eqIncome", weights = "rb050", data = eusilc)

Value:
[1] 26.48962
```

3 The Pareto distribution

The *Pareto distribution* is well studied in the literature and is defined in terms of its cumulative distribution function

$$F_\theta(x) = 1 - \left(\frac{x}{x_0} \right)^{-\theta}, \qquad x \geq x_0, \tag{4}$$

where $x_0 > 0$ is the scale parameter and $\theta > 0$ is the shape parameter (Kleiber and Kotz 2003). Furthermore, its density function is given by

$$f_\theta(x) = \frac{\theta x_0^\theta}{x^{\theta+1}}, \qquad x \geq x_0. \tag{5}$$

Figure 1 visualizes the Pareto probability density function with scale parameter $x_0 = 1$ and different values of the shape parameter θ. Clearly, the Pareto distribution is a highly right-skewed distribution with a heavy tail. It is therefore reasonable to assume that a random variable following a Pareto distribution contains extreme values. The effect of changing the shape parameter θ is visible in the probability mass at the scale parameter x_0: the higher θ, the higher the probability mass at x_0.

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

$$F(x) = \begin{cases} G(x), & \text{if } x \leq x_0, \\ G(x_0) + (1 - G(x_0)) F_\theta(x), & \text{if } x > x_0, \end{cases} \tag{6}$$

where G is an unknown distribution function (Dupuis and Victoria-Feser 2006).

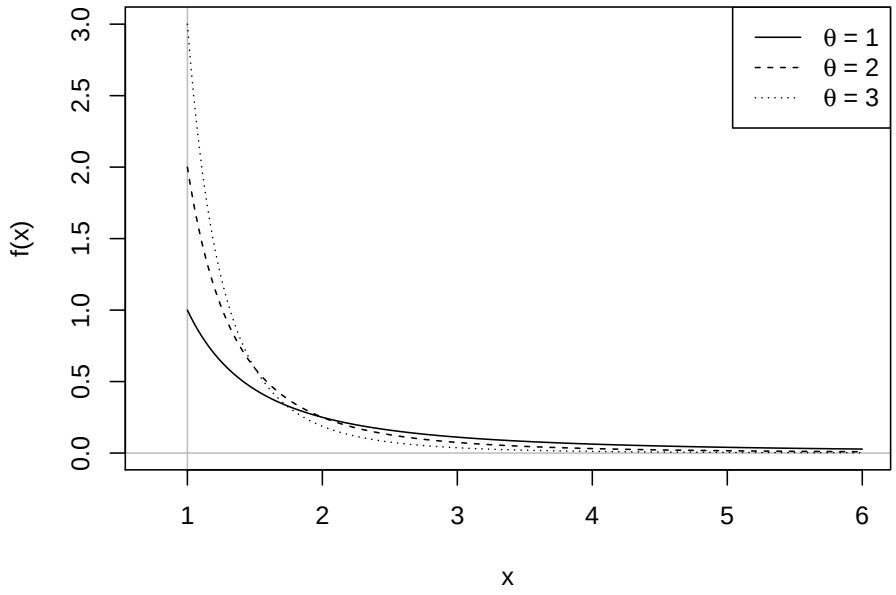

Figure 1: Pareto probability density functions with parameters $x_0 = 1$ and $\theta = 1, 2, 3$.

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

$$\hat{x}_0 := x_{n-k}. \tag{7}$$

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.

In the remainder of this package vignette, the equivalized disposable income of the EU-SILC example data is of main interest. Consequently, the Pareto distribution will be modeled at the household level rather than the individual level. Moreover, the focus of this vignette is on robust estimation of the social exclusion indicators. Hence the equivalized disposable income of the household with the largest income is replaced by a large outlier.

```
R> hID <- eusilc$db030[which.max(eusilc$eqIncome)]
R> eusilc[eusilc$db030 == hID, "eqIncome"] <- 10000000
```

Since the aim is to model a Pareto distribution at the household level, the following command creates a data set that contains only the equivalized disposable income and the sample weights on the household level. This data set will be used in Sections 4 and 5 to estimate the parameters of the Pareto distribution.

```
R> eusilcH <- eusilc[!duplicated(eusilc$db030), c("eqIncome", "db090")]
```

4 Finding the threshold

The aim of the methods presented in this sections is to find the threshold x_0 for modeling the Pareto distribution. Several methods for the estimation of the threshold x_0 or the number of observations k in the tail have been proposed in the literature, but those proposals typically do not consider sample weights.

Beirlant et al. (1996a,b) developed a procedure that analytically determines the optimal choice of k for the Hill estimator of the shape parameter (Hill 1975, see also Section 5.1 of this paper)

by minimizing the asymptotic mean squared error (AMSE). In package **laeken**, this approach is implemented in the function `minAMSE()`. However, the procedure is designed for the non-robust Hill estimator and is therefore not further discussed in this paper. Furthermore, Danielsson et al. (2001) proposed a bootstrap method to find the optimal k for the Hill estimator with respect to the AMSE, which has less analytical requirements than the approach by Beirlant et al. (1996a,b). Please note that this method is not robust either and that it is currently not available in package **laeken**. A robust prediction error criterion for choosing the number of observations k in the tail and estimating the shape parameter θ was developed by Dupuis and Victoria-Feser (2006). Nevertheless, our implementation of this robust criterion was unstable and is therefore not included in **laeken**.

In any case, Holzer (2009) concludes that graphical methods for finding the threshold outperform those analytical approaches in the case of EU-SILC data. While this section is thus focused graphical methods, a simple rule of thumb designed specifically for the equivalized disposable income in EU-SILC data is described in the following as well.

4.1 Van Kerm's rule of thumb

Van Kerm (2007) presented a formula that is more of a rule of thumb for the threshold of the equivalized disposable income in EU-SILC data. Is is given by

$$\hat{x}_0 := \min(\max(2.5\bar{x}, q_{0.98}), q_{0.97}), \tag{8}$$

where $\bar{x}$ is the weighted mean, and $q_{0.98}$ and $q_{0.97}$ are weighted quantiles as defined in Equation (1).

In package **laeken**, the function `paretoScale()` provides functionality for computing the threshold with van Kerm's rule of thumb. The argument `w` is available to supply sample weights.

```
R> ts <- paretoScale(eusilcH$eqIncome, w = eusilcH$db090)
R> ts

Threshold: 48459.43
Number of observations in the tail: 119
```

It should be noted that the function returns an object of class `"paretoScale"`, which consists of a component `x0` for the threshold (scale parameter) and a component `k` for the number of observations in the tail of the distribution, i.e., that are larger than the threshold.

4.2 Pareto quantile plot

The *Pareto quantile plot* is a graphical method for inspecting the parameters of a Pareto distribution. For the case without sample weights, it is described in detail in Beirlant et al. (1996a).

If the Pareto model holds, there exists a linear relationship between the lograrithms of the observed values and the quantiles of the standard exponential distribution, since the logarithm of a Pareto distributed random variable follows an exponential distribution. Hence the logarithms of the observed values, $\log(x_i)$, $i = 1, \ldots, n$, are plotted against the theoretical quantiles.

In the case without sample weights, the theoretical quantiles of the standard exponential distribution are given by

$$-\log\left(1 - \frac{i}{n+1}\right), \qquad i = 1, \ldots, n, \tag{9}$$

i.e., by dividing the range into $n+1$ equally sized subsets and using the resulting n inner gridpoints as probabilities for the quantiles. If the data contain sample weights, the range of the exponential distribution needs to be divided according to the weights of the n observations. The Pareto quantile plot is thus generalized by using the theoretical quantiles

$$-\log\left(1 - \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{n} w_j} \frac{n}{n+1}\right), \qquad i = 1, \ldots, n, \tag{10}$$

where the correction factor $\frac{n}{n+1}$ ensures that the quantiles reduce to (9) if all sample weights are equal.

```
R> paretoQPlot(eusilcH$eqIncome, w = eusilcH$db090)
```

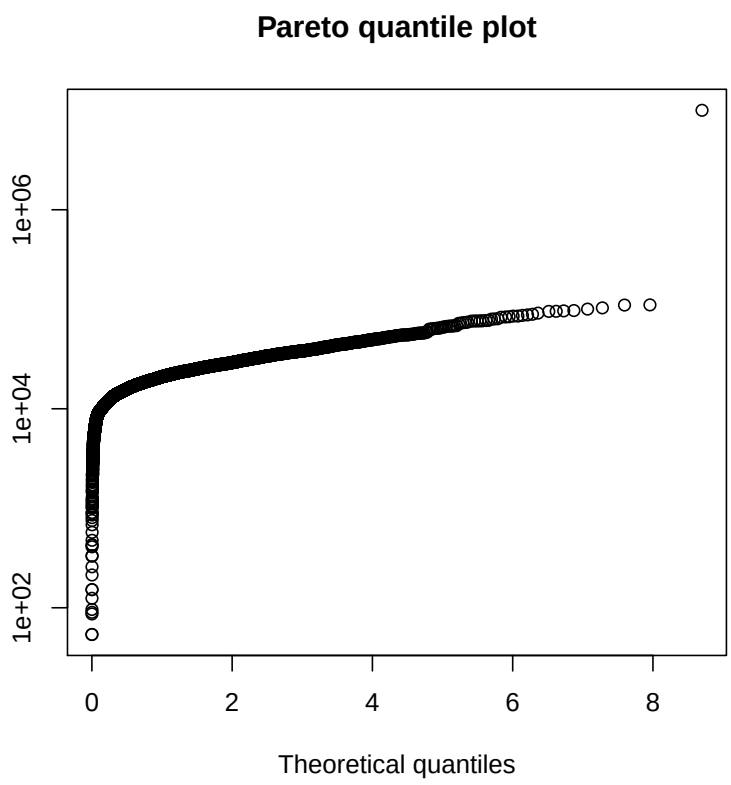

Figure 2: Pareto Quantile plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier.

If the tail of the data follows a Pareto distribution, those observations form almost a straight line. The leftmost point of a fitted line can thus be used as an estimate of the threshold x_0, the scale parameter. All values starting from the point after the threshold may be modeled by a Pareto distribution, but this point cannot be determined exactly. Furthermore, the slope of the fitted line is in turn an estimate of $\frac{1}{\theta}$, the reciprocal of the shape parameter.

Figure 2 displays the Pareto quantile plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier. The plot is generated using the function `paretoQPlot()`, which allows to supply sample weights via the argument `w`. In addition, the threshold can be selected interactively by clicking on a data point. Information on the selected threshold is then printed on the R console. When the interactive selection is terminated, which is typically done by a secondary mouse click, the selected threshold is returned as an object of class `"paretoScale"`.

Another advantage of the Pareto quantile plot is also illustrated in Figure 2. Nonrepresentative outliers such as the large income introduced into the example data in Section 3, i.e., extreme observations in the upper tail that deviate from the Pareto model, are clearly visible.

4.3 Mean excess plot

The *mean excess plot* is another graphical method for inspecting the threshold for Pareto tail modeling, but it does not provide information on the shape parameter. It is based on the excess function

$$e(x_0) := \mathbb{E}(x - x_0 | x > x_0), \qquad x_0 \geq 0. \tag{11}$$

```
R> meanExcessPlot(eusilcH$eqIncome, w = eusilcH$db090)
```

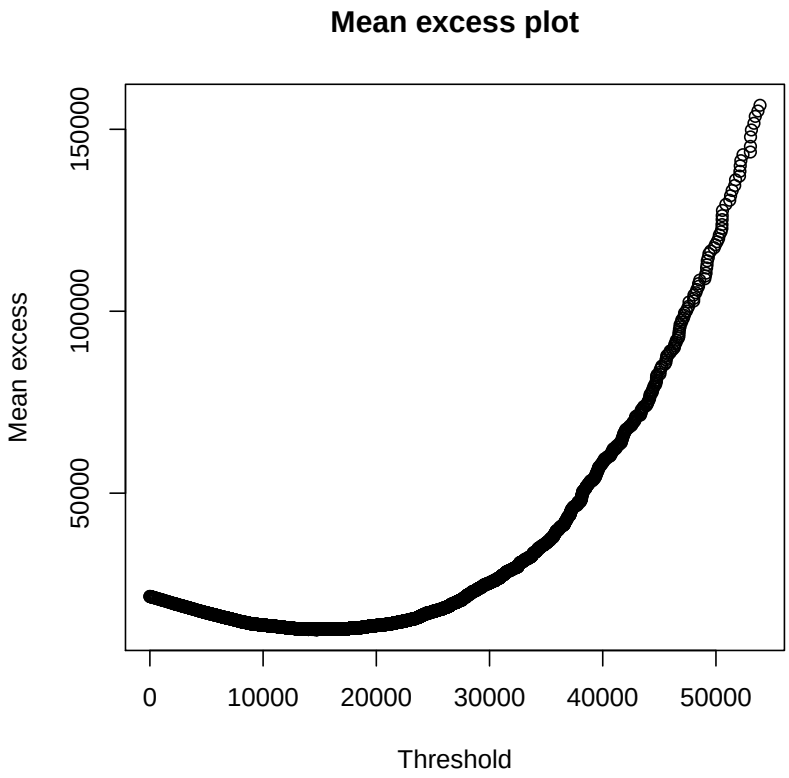

Figure 3: Mean excess plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier.

A detailed description for the case without sample weights can be found in Borkovec and Klüppelberg (2000).

For the following definition of the mean excess plot, keep in mind that the observations are sorted such that $x_1 \leq \ldots \leq x_n$. For each observation x_i, $i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor$, the empirical excess function e_n is computed. In the case without sample weights, the expectation in Equation (11) is replaced by the arithmetic mean, and the empirical excess function is given by

$$e_n(x_i) := \frac{1}{n-i} \sum_{j=i+1}^{n} (x_j - x_i), \qquad i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor. \tag{12}$$

The values of the empirical excess function $e_n(x_i)$ are then plotted against the corresponding x_i, $i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor$. If sample weights are available in the data, the mean excess plot is simply generalized by using the weighted mean for the empirical excess function:

$$e_n(x_i) := \frac{1}{\sum_{j=i+1}^{n} w_j} \sum_{j=i+1}^{n} w_j (x_j - x_i), \qquad i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor. \tag{13}$$

If the tail of the data follows a Pareto distribution, those observations show a positive linear trend. The leftmost point of a fitted line can thus be used as an estimate of the threshold x_0, the scale parameter. As for the Pareto quantile plot, a disadvantage of the mean excess plot is that the threshold cannot be determined exactly.

Figure 3 shows the mean excess plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier. The function `meanExcessPlot()` is thereby used to

produce the plot. Sample weights can be supplied via the argument `w`. Interactive selection of the threshold works just like for the Pareto quantile plot. Again, the selected threshold is returned as an object of class `"paretoScale"`.

5 Estimation of the shape parameter

This section is focused on methods for estimating the shape parameter θ once the threshold x_0 is fixed. It should be noted that none of the original proposals takes sample weights into account. Most estimators presented in the following were therefore adjusted for the case of sample weights.

5.1 Hill estimator

The maximum likelihood estimator for the shape parameter of the Pareto distribution was introduced by Hill (1975) and is referred to as the *Hill* estimator. If the data do not contain sample weights, it is given by

$$\hat{\theta}_{\text{Hill}} = \frac{k}{\sum_{i=1}^{k} \log x_{n-k+i} - k \log x_{n-k}}. \tag{14}$$

In the case of sample weights, the *weighted Hill* (wHill) estimator is given by generalizing Equation (14) to

$$\hat{\theta}_{\text{wHill}} = \frac{\sum_{i=1}^{k} w_{n-k+i}}{\sum_{i=1}^{k} w_{n-k+i} \left(\log x_{n-k+i} - \log x_{n-k}\right)}. \tag{15}$$

Package **laeken** provides the function `thetaHill()` to compute the Hill estimator. It requires to specify either the number of observations in the tail via the argument `k`, or the threshold via the argument `x0`. Furthermore, the argument `w` can be used to supply sample weights. In the following example, the shape parameter is estimated using the largest observations (first command) and the threshold (second command) as computed with van Kerm's rule of thumb in Section 4.1.

```
R> thetaHill(eusilcH$eqIncome, k = ts$k, w = eusilcH$db090)

[1] 3.437979

R> thetaHill(eusilcH$eqIncome, x0 = ts$x0, w = eusilcH$db090)

[1] 3.437979
```

5.2 Weighted maximum likelihood estimator

The *weighted maximum likelihood* (WML) estimator (Dupuis and Morgenthaler 2002, Dupuis and Victoria-Feser 2006) falls into the class of M-estimators and is given by the solution $\hat{\theta}$ of

$$\sum_{i=1}^{k} \Psi(x_{n-k+i}, \theta) = 0 \tag{16}$$

with

$$\Psi(x, \theta) := u(x, \theta) \frac{\partial}{\partial \theta} \log f(x, \theta) = u(x, \theta) \left(\frac{1}{\theta} - \log \frac{x}{x_0}\right), \tag{17}$$

where $u(x, \theta)$ is a weight function with values in $[0, 1]$. In the implementation in package **laeken**, a Huber type weight function is used by default, as proposed by Dupuis and Victoria-Feser (2006). Let the logarithms of the relative excesses be denoted by

$$z_i := \log \left(\frac{x_{n-k+i}}{x_{n-k}}\right), \qquad i = 1, \ldots, k. \tag{18}$$

In the Pareto model, these can be predicted by

$$\hat{z}_i := -\frac{1}{\theta} \log \left(\frac{k+1-i}{k+1}\right), \qquad i = 1, \ldots, k. \tag{19}$$

The variance of z_i is given by

$$\sigma_i{}^2 := \sum_{j=1}^{i} \frac{1}{\theta^2(k-i+j)^2}, \qquad i = 1, \ldots, k. \tag{20}$$

Using the standardized residuals

$$r_i := \frac{z_i - \hat{z}_i}{\sigma_i}, \tag{21}$$

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

$$u(x_{n-k+i}, \theta) := \begin{cases} 1, & \text{if } |r_i| \leq c, \\ \frac{c}{|r_i|}, & \text{if } |r_i| > c. \end{cases} \tag{22}$$

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

$$\hat{B}(\hat{\theta}) = -\frac{\sum_{i=1}^{k} \left(u_i \frac{\partial}{\partial\theta} \log f_i\right)|_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-k+i}) - F_{\hat{\theta}}(x_{n-k+i-1})\right)}{\sum_{i=1}^{k} \left(\frac{\partial}{\partial\theta} u_i \frac{\partial}{\partial\theta} \log f_i + u_i \frac{\partial^2}{\partial\theta^2} \log f_i\right)|_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-k+i}) - F_{\hat{\theta}}(x_{n-k+i-1})\right)}, \tag{23}$$

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

$$\tilde{\theta} := \hat{\theta} - \hat{B}(\hat{\theta}). \tag{24}$$

For details and proofs of the above statements, as well as for information on a probability-based weight function $u(x, \theta)$, the reader is referred to Dupuis and Morgenthaler (2002) and Dupuis and Victoria-Feser (2006). However, note the WML estimator does not consider sample weights. An adjustment of the estimator to take sample weights into account is currently not available due to its complexity. For sampling designs that lead to equal sample weights, the WML estimator may still be useful, though.

The function `thetaWML()` is available in **laeken** to compute the WML estimator. Again, either the argument `k` or `x0` needs to be used to specify the number of observations in the tail or the threshold. Since the sample weights in the example data are not equal, the following example is only included to demonstrate the use of the function.

```
R> thetaWML(eusilcH$eqIncome, k = ts$k)

[1] 4.226204

R> thetaWML(eusilcH$eqIncome, x0 = ts$x0)

[1] 4.226204
```

5.3 Integrated squared error estimator

For the *integrated squared error* (ISE) estimator (Vandewalle et al. 2007), the Pareto distribution is modeled in terms of the relative excesses

$$y_i := \frac{x_{n-k+i}}{x_{n-k}}, \qquad i = 1, \ldots, k. \tag{25}$$

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

$$f_\theta(y) = \theta y^{-(1+\theta)}. \tag{26}$$

The ISE estimator is then given by minimizing the integrated squared error criterion (Terrell 1990):

$$\hat{\theta} = \arg\min_{\theta} \left[\int f_\theta^2(y)dy - 2\mathbb{E}(f_\theta(Y))\right]. \tag{27}$$

If there are no sample weights in the data, the mean is used as an unbiased estimator of $\mathbb{E}(f_\theta(Y))$ in order to obtain the ISE estimate

$$\hat{\theta}_{\text{ISE}} = \arg\min_\theta \left[\int f_\theta^2(y)dy - \frac{2}{k} \sum_{i=1}^{k} f_\theta(y_i) \right]. \tag{28}$$

See Vandewalle et al. (2007) for more information on the ISE estimator for the case without sample weights.

If sample weights are available in the data, the mean in Equation (28) is simply replaced by a weighted mean to obtain the *weighted integrated squared error* (wISE) estimator:

$$\hat{\theta}_{\text{wISE}} = \arg\min_\theta \left[\int f_\theta^2(y)dy - \frac{2}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \right]. \tag{29}$$

With package **laeken**, the ISE estimator can be computed using the function `thetaISE()`. The arguments `k` and `x0` are available to specify either the number of observations in the tail or the threshold, and sample weights can be supplied via the argument `w`.

```
R> thetaISE(eusilcH$eqIncome, k = ts$k, w = eusilcH$db090)
[1] 3.993801
R> thetaISE(eusilcH$eqIncome, x0 = ts$x0, w = eusilcH$db090)
[1] 3.993801
```

5.4 Partial density component estimator

For the *partial density component* (PDC) estimator Vandewalle et al. (2007) minimizes the integrated squared error criterion using an incomplete density mixture model uf_θ. If the data do not contain sample weights, the PDC estimator in is thus given by

$$\hat{\theta}_{\text{PDC}} = \arg\min_\theta \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{k} \sum_{i=1}^{k} f_\theta(y_i) \right]. \tag{30}$$

The parameter u can be interpreted as a measure of the uncontaminated part of the sample and is estimated by

$$\hat{u} = \frac{\frac{1}{k} \sum_{i=1}^{k} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y)dy}. \tag{31}$$

See Vandewalle et al. (2007) and references therein for more information on the PDC estimator for the case without sample weights.

Taking sample weights into account, the *weighted partial density component* (wPDC) estimator is obtained by generalizing Equations (30) and (31) to

$$\hat{\theta}_{\text{wPDC}} = \arg\min_\theta \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \right], \tag{32}$$

$$\hat{u} = \frac{\frac{1}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y)dy}. \tag{33}$$

The function `thetaPDC()` is implemented in package **laeken** to compute the PDC estimator. As for the other estimators, it is necessary to specify either the number of observations in the tail via the argument `k`, or the threshold via the argument `x0`. Sample weights can be supplied using the argument `w`.

```
R> thetaPDC(eusilcH$eqIncome, k = ts$k, w = eusilcH$db090)
[1] 4.132596
R> thetaPDC(eusilcH$eqIncome, x0 = ts$x0, w = eusilcH$db090)
[1] 4.132596
```

6 Estimation of the indicators using Pareto tail modeling

Three approaches based on Pareto tail modeling for reducing the influence of outliers on the social exclusion indicators are implemented in the R package **laeken**:

Calibration for nonrepresentative outliers (CN): Values larger than a certain quantile of the fitted distribution are declared as nonrepresentative outliers. Since these are considered to be unique to the population data, the sample weights of the corresponding observations are set to 1 and the weights of the remaining observations are adjusted accordingly by calibration.

Replacement of nonrepresentative outliers (RN): Values larger than a certain quantile of the fitted distribution are declared as nonrepresentative outliers. Only these nonrepresentative outliers are replaced by values drawn from the fitted distribution, thereby preserving the order of the original values.

Replacement of the tail (RT): All values above the threshold are replaced by values drawn from the fitted distribution. The order of the original values is preserved.

An evaluation of the RT approach by means of a simulation study can be found in Alfons et al. (2010).

Keep in mind that the largest observation in the example data `eusilc` was replaced by a large outlier in Section 3. With the following command, the Gini coefficient is estimated according to the Eurostat definition to show that even a single outlier can completely distort the results for the standard estimation (see Section 2.2 for the original value).

```
R> gini("eqIncome", weights = "rb050", data = eusilc)
```

```
Value:
[1] 29.24333
```

For Pareto tail modeling, the function `paretoTail()` is implemented in **laeken**. It returns an object of class `"paretoTail"`, which contains all the necessary information for further analysis using the three approaches described above. Note that the household IDs are supplied via the argument `groups` such that the Pareto distribution is fitted on the household level rather than the individual level. In addition, the PDC is used by default to estimate the shape parameter. Other estimators can be specified via the `method` argument.

```
R> fit <- paretoTail(eusilc$eqIncome, k = ts$k,
+     w = eusilc$db090, groups = eusilc$db030)
```

The function `reweightOut()` is available for semiparametric estimation with the CN approach. It returns a vector of the recalibrated weights. In this example, regional information is used as auxiliary variables for calibration. The function `calibVars()` thereby transforms a factor into a matrix of binary variables, as required by the calibration function `calibWeights()`, which is called internally. These recalibrated weights are then simply used to estimate the Gini coefficient with function `gini()`.

```
R> w <- reweightOut(fit, calibVars(eusilc$db040))
R> gini(eusilc$eqIncome, w)
```

```
Value:
[1] 26.45973
```

For the RN approach, the function `replaceOut()` is implemented. Since values are drawn from the fitted distribution to replace the observations flagged as outliers, the seed of the random number generator is set first for reproducibility of the results. The returned vector of incomes is then supplied to `gini()` to estimate the Gini coefficient.

```
R> set.seed(1234)
R> eqIncome <- replaceOut(fit)
R> gini(eqIncome, weights = eusilc$rb050)
```

```
Value:
[1] 26.46924
```

Similarly, the function `replaceTail()` is available for the RT approach. Again, the seed of the random number generator is set beforehand.

```
R> set.seed(1234)
R> eqIncome <- replaceTail(fit)
R> gini(eqIncome, weights = eusilc$rb050)

Value:
[1] 26.64921
```

It should be noted that `replaceTail()` can also be used for the RN approach by setting the argument `all` to `FALSE`. In fact, `replaceOut(x, ...)` is a simple wrapper for `replaceTail(x, all = FALSE, ...)`.

In any case, the estimates for the semiparametric approaches based on Pareto tail modeling are very close to the original value before the outlier has been introduced (see Section 2.2), whereas the standard estimation is corrupted by the outlier. Furthermore, the estimation of other indicators such as the quintile share ratio (see Section 2.1) using the semiparametric approaches is straightforward and hence not shown here.

7 Conclusions

This vignette shows the functionality of package **laeken** for robust semiparametric estimation of social exclusion indicators based on Pareto tail modeling. Most notably, it demonstrates that the functions are easy to use and that the implementation follows an object-oriented design. While the focus of the paper lies on the use of the package, a mathematical description of the methods is given as well.

Furthermore, it is shown that the standard estimation of the inequality indicators can be corrupted by a single outlier, thus underlining the need for robust alternatives. Three approaches for robust semiparametric estimation based on Pareto tail modeling are thereby implemented such that the corresponding functions share a common interface for ease of use.

Acknowledgments

This work was partly funded by the European Union (represented by the European Commission) within the 7$^{\text{th}}$ 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

A. Alfons and S. Kraft. ***simPopulation**: Simulation of synthetic populations for surveys based on sample data*, 2010. URL http://CRAN.R-project.org/package=simPopulation. R package version 0.2.1.

A. Alfons, M. Templ, P. Filzmoser, and J. Holzer. A comparison of robust methods for Pareto tail modeling in the case of Laeken indicators. In C. Borgelt, G. González-Rodríguez, W. Trutschnig, M.A. Lubiano, M.A. Gil, P. Grzegorzewski, and O. Hryniewicz, editors, *Combining Soft Computing and Statistical Methods in Data Analysis*, volume 77 of *Advances in Intelligent and Soft Computing*, pages 17–24. Springer, Heidelberg, 2010. ISBN 978-3-642-14745-6.

A. Alfons, S. Kraft, M. Templ, and P. Filzmoser. Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20(3): 383–407, 2011.

A. Alfons, J. Holzer, and M. Templ. ***laeken**: Laeken indicators for measuring social cohesion*, 2012. URL http://CRAN.R-project.org/package=laeken. R package version 0.3.2.

J. Beirlant, P. Vynckier, and J.L. Teugels. Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association*, 31(436):1659–1667, 1996a.

J. Beirlant, P. Vynckier, and J.L. Teugels. Excess functions and estimation of the extreme-value index. *Bernoulli*, 2(4):293–318, 1996b.

M. Borkovec and C. Klüppelberg. Extremwerttheorie für Finanzzeitreihen – ein unverzichtbares Werkzeug im Risikomanagement. In L. Johanning and B. Rudolph, editors, *Handbuch Risikomanagement*, pages 219–241. Uhlenbruch, Bad Soden, 2000. ISBN 3933207150.

J. Danielsson, L. de Haan, L. Peng, and C.G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76(2):226–248, 2001.

D.J. Dupuis and S. Morgenthaler. Robust weighted likelihood estimators with an application to bivariate extreme value problems. *The Canadian Journal of Statistics*, 30(1):17–36, 2002.

D.J. Dupuis and M.-P. Victoria-Feser. A robust prediction error criterion for Pareto modelling of upper tails. *The Canadian Journal of Statistics*, 34(4):639–658, 2006.

Eurostat. Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. EU-SILC 131-rev/04, Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, Eurostat, Luxembourg, 2004.

Eurostat. Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). Doc. LC-ILC/39/09/EN-rev.1, Unit F-3: Living conditions and social protection, Directorate F: Social and information society statistics, Eurostat, Luxembourg, 2009.

B.M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

J. Holzer. Robust methods for the estimation of selected Laeken indicators. Master's thesis, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria, 2009.

B. Hulliger and T. Schoch. Robustification of the quintile share ratio. New Techniques and Technologies for Statistics, Brussels, 2009.

C. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons, Hoboken, 2003. ISBN 0-471-15064-9.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org. ISBN 3-900051-07-0.

M. Templ and A. Alfons. Standard methods for point estimation of social inclusion indicators using the R package **laeken**. Research Report CS-2011-1, Department of Statistics and Probability Theory, Vienna University of Technology, 2011a. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-1complete.pdf.

M. Templ and A. Alfons. Variance estimation of social inclusion indicators using the R package **laeken**. Research Report CS-2011-3, Department of Statistics and Probability Theory, Vienna University of Technology, 2011b. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-3complete.pdf.

G. Terrell. Linear density estimates. In *Proceedings of the Statistical Computing Section*, pages 297–302. American Statistical Association, 1990.

P. Van Kerm. Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. IRISS Working Paper Series 2007-01, CEPS/INSTEAD, 2007.

B. Vandewalle, J. Beirlant, A. Christmann, and M. Hubert. A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics & Data Analysis*, 51(12):6252–6268, 2007.