Robust semiparametric estimation of economic indicators from survey samples

A. Alfons, M. Templ, and P. Filzmoser
Robust semiparametric estimation of economic indicators from survey samples

Andreas Alfons\textsuperscript{a,b,}\textsuperscript{*}, Matthias Templ\textsuperscript{b,c}, Peter Filzmoser\textsuperscript{a}

\textsuperscript{a}Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 7, 1040 Vienna, Austria
\textsuperscript{b}ORSTAT Research Center, Faculty of Business and Economics, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium
\textsuperscript{c}Methods Unit, Statistics Austria, Guglgasse 13, 1100 Vienna, Austria

Abstract
Economic data typically contain variables with heavily tailed distributions. For analysis, the Pareto distribution is frequently used to model the tails of such variables. In the survey context, however, sample weights need to be considered so that the modeled distribution accurately reflects the true distribution on the population level. Therefore, commonly used estimators for the parameters of the Pareto distribution are adapted to take sample weights into account. Moreover, economic survey data often contain nonrepresentative outliers, i.e., extreme values that are either incorrectly recorded or can be considered unique in the population. Nonrepresentative outliers have a strong influence on economic indicators such as the Gini coefficient and thus need to be excluded from estimation or downweighted. While a semiparametric Pareto model can be used to detect such deviating data points, the parameters of the Pareto distribution need to be estimated in a robust manner to avoid corruption of the parameter estimates themselves. Hence the main contribution of this paper is to investigate the use of robust Pareto tail modeling in order to reduce the influence of nonrepresentative outliers on the estimation of economic indicators such as the Gini coefficient. In addition, all methods presented in this paper are available in the \textit{R} package \textit{laeken}.

Keywords: Robustness, semiparametric estimation, Pareto distribution, tail modeling, survey statistics

1. Introduction
Economic indicators monitor the economic performance of administrative units such as countries or regions for analysis and prediction purposes. In many cases, economic indicators are estimated from survey data due to a lack of availability of suitable population data. A well known example for such a survey is the \textit{European Union Statistics on Income and Living Conditions} (EU-SILC), which is an annual panel survey conducted in European Union member states and other European countries. This survey is used as data basis for a set of indicators to measure risk-of-poverty and social exclusion in Europe. One of these indicators is the \textit{Gini coefficient}, a measure of inequality originally proposed by Gini (1912). While the Gini coefficient is widely studied in the literature and applied in many fields of research, it is used to measure inequality of income in this context. However, it is important to note that the Gini coefficient and many other economic indicators are highly sensitive to outlying observations.

In economic data, the distributions of variables such as income or business turnover usually have heavy tails. Various theoretical distributions have been proposed in the literature to model such data. An overview of statistical distributions in economics is presented in, e.g., Kleiber and Kotz (2003). This paper is focused on using the Pareto distribution for semiparametric modeling of the upper tail of a distribution. More general Pareto-type distributions or more complex income distributions could be considered, but the Pareto distribution is chosen due to its simple form. In Pareto tail modeling, typically the shape of the Pareto distribution is estimated for points over a large threshold. Possibly the most widely known estimator was suggested by Hill (1975) and follows a maximum likelihood approach. Other classical estimators were introduced by Pickands (1975), Dekkers et al. (1989), and Kratz and Resnick (1996).

\textsuperscript{*}Corresponding author. Tel.: +32 16 326817.
Email addresses: andreas.alfons@econ.kuleuven.be (Andreas Alfons), templ@tuwien.ac.at (Matthias Templ), p.filzmoser@tuwien.ac.at (Peter Filzmoser)

Preprint submitted to Elsevier

October 11, 2011
Brazauskas and Serfling (2000a,b) examined various estimators with respect to their robustness properties. More advanced robust estimators were proposed by Victoria-Feser and Ronchetti (1994, 1997) following an optimal bias-robust approach, or by Dupuis and Morgenthaler (2002) and Dupuis and Victoria-Feser (2006) following a weighted maximum likelihood approach. Vandewalle et al. (2007), on the other hand, developed a promising robust estimator based on an integrated squared error criterion. An entirely different interpretation of robustness in the Pareto model is discussed in Beran and Schell (2010). There the proposed estimator is robust against model deviations at low quantiles.

For the choice of the threshold, various proposals have been made in the literature as well. Beirlant et al. (1996a,b) and Danielsson et al. (2001) introduced different approaches to determine the optimal choice of the number of observations in the tail for the Hill estimator. Both approaches are thereby based on minimizing the asymptotic mean squared error (AMSE). While Beirlant et al. (1996a,b) developed an analytic procedure, Danielsson et al. (2001) proposed a bootstrap method with less analytical requirements. Nevertheless, those two procedures are not robust as they are designed for the non-robust Hill estimator. A robust prediction error criterion for simultaneously choosing the number of observations in the tail and estimating the shape parameter was introduced by Dupuis and Victoria-Feser (2006).

However, all the procedures mentioned above are designed for samples from infinite populations, survey samples are typically not considered in the literature on the Pareto model. Finite population survey sampling is in general based on complex sampling designs with unequal inclusion probabilities for the observations in the population, which leads to unequal weights for the observations in the sample (see, e.g., Tillé, 2006). The initial weights are also often further modified by techniques such as calibration (e.g., Deville et al., 1993) such that the sample weights of the observations in certain subsets sum up to known population totals. The idea behind Pareto tail modeling for survey data is that the upper tail of the population data follows a Pareto distribution. Hence sample weights need to be considered for fitting the distribution in order to avoid bias in the estimation of the parameters.

Concerning robustness in survey statistics, Chambers (1986) introduced the notion of representative and nonrepresentative outliers. Keep in mind that each observation in a survey sample represents a number of observations in the population as given by its sample weight. Representative outliers are observations whose values are correctly recorded and are not unique in the population. Therefore they contain relevant information and need to be considered in the estimation of quantities of interest. Nonrepresentative outliers are observations that either contain incorrect values or can in some sense be considered unique in the population. Consequently, they may corrupt the estimation of quantities of interest and need to be excluded or downweighted. In economic survey data, representative outliers are the observations forming the heavy tails, whereas nonrepresentative outliers are even more extreme observations that deviate from the observations in the tails.

The aim of this paper is to investigate the use of robust Pareto tail modeling in order to reduce the influence of nonrepresentative outliers on the estimation of economic indicators. All presented methods are available for the statistical environment R (R Development Core Team, 2011) in the contributed package laeken (Alfons et al., 2011a). The package functionality for semiparametric estimation based on Pareto tail modeling is described in detail in vignette "laeken-pareto" (Alfons et al., 2011c). If laeken is installed, the vignette can be viewed from within R with the following command:

R> vignette("laeken-pareto")

The rest of the paper is organized as follows. Brief descriptions of the Gini coefficient and the Pareto distribution are given in Sections 2 and 3, respectively. Section 4 presents the Pareto quantile plot for the case of survey data. Afterwards, selected estimators for the shape parameter of the Pareto distribution and their adaptations for sample weights are discussed in Section 5. How to use the semiparametric Pareto model for robust estimation of economic indicators is described in Section 6. In Section 7, the presented estimators are evaluated by means of simulation. Real data examples are then given in Section 8. Finally, Section 9 concludes.

2. Gini coefficient

The Gini coefficient (Gini, 1912) is a well known measure of inequality of a distribution and is widely applied in many fields of research. Most notably for this paper, it is often used to measure inequality of income as an economic
indicator. For instance, the Gini coefficient is part of a set of indicators defined by the European Union for measuring risk-of-poverty and social exclusion in its member states and other European countries. Many of these indicators, including the Gini coefficient, are estimated from the well known panel survey *European Union Statistics on Income and Living Conditions (EU-SILC)*. Since data from this survey are included as an example in Section 8, the Eurostat definition of the Gini coefficient is used in this paper.

Eurostat (2004, 2009) defines the Gini coefficient as the relationship of cumulative shares of the population arranged according to the level of income, to the cumulative share of the income received by them. All members of a household are thereby assigned the same *equivalized disposable income*. For details on the computation of the equivalized disposable income, the reader is referred to Eurostat (2004, 2009).

For a definition of the Gini coefficient in mathematical terms, let $x := (x_1, \ldots, x_n)'$ be the income with $x_1 \leq \ldots \leq x_n$ and let $w := (w_1, \ldots, w_n)'$ be the corresponding sample weights, where $n$ denotes the number of observations. Then the Gini coefficient according to Eurostat (2004, 2009) is estimated by

$$
\hat{Gini} := 100 \cdot \frac{2 \sum_{i=1}^{n} (w_i x_i - \sum_{i=1}^{n} w_i x_i^2)}{\left(\sum_{i=1}^{n} w_i \right) \left(\sum_{i=1}^{n} (w_i x_i)\right)} - 1.
$$

(1)

The Gini coefficient is closely related to the Lorenz curve (Lorenz, 1905), which plots the cumulative proportion of the total income against the corresponding proportion of the population. As for the Gini coefficient, the data are first sorted by income in non-decreasing order. An example for the Lorenz curve is shown in Figure 1 (left). The line at the angle of 45° thereby corresponds to perfect equality of incomes. The Gini coefficient can then be written as

$$
Gini = 100 \cdot 2A,
$$

(2)

where $A$ denotes the area between the Lorenz curve and the line of perfect equality. In the example in Figure 1 (left), the area $A$ is shaded in grey.

3. Pareto distribution

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

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

(3)
where \( x_0 > 0 \) is the scale parameter and \( \theta > 0 \) is the shape parameter (Kleiber and Kotz, 2003). The corresponding density function is given by

\[
f_\theta(x) = \frac{\theta x_0^\theta}{x^{\theta+1}}, \quad x \geq x_0.
\]  

Figure 1 (right) displays the density function of the Pareto distribution with scale parameter \( x_0 = 1 \) and different values of the shape parameter \( \theta \). The effect of changing the shape parameter \( \theta \) is thereby clearly visible: the lower \( \theta \), the lower the probability mass at \( x_0 \) and the longer the tail. In extreme value theory, the tail index \( \gamma \) is a measure of the tail heaviness of a distribution. For the Pareto distribution, the tail index is in fact given by \( \gamma = 1/\theta \).

In the semiparametric Pareto tail model, the cumulative distribution function on the whole range of \( x \) is modeled as

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

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

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

\[
\hat{x}_0 := x_{n-k}.
\]

If, on the other hand, an estimate \( \hat{x}_0 \) for the scale parameter of the Pareto distribution is available, \( k \) is given by the number of observations larger than \( \hat{x}_0 \). In this way, the estimation of \( x_0 \) and \( k \) directly corresponds with each other.

### 4. 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 by Beirlant et al. (1996a).

If the Pareto model holds, there exists a linear relationship between the logarithms 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

\[
- \log \left(1 - \frac{i}{n+1}\right), \quad i = 1, \ldots, n,
\]

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

\[
- \log \left(1 - \frac{\sum_{j=1}^{i} w_j n}{\sum_{j=1}^{n} w_j n + 1}\right), \quad i = 1, \ldots, n,
\]

where the correction factor \( \frac{n}{n+1} \) ensures that the quantiles reduce to (7) if all sample weights are equal.

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 \( 1/\theta \), the reciprocal of the shape parameter. Another advantage of the Pareto quantile plot is that nonrepresentative outliers, i.e., extreme observations in the upper tail that deviate from the Pareto model, are clearly visible.

Figure 2 shows Pareto quantile plots for Austrian and Belgian EU-SILC income survey data from 2005 and 2006. These data sets were provided by Eurostat and are used in Section 8 as real data examples for semiparametric estimation of the Gini coefficient. Note that the Austrian data are clean despite some irregularities in the upper tail of the 2005 data, whereas the Belgian data sets each contain one clear outlier from the Pareto tail model.
Figure 2: Pareto quantile plots of Austrian (top) and Belgian (bottom) EU-SILC income survey data from 2005 (left) and 2006 (right). The data were provided by Eurostat.

5. Estimation of the shape parameter

In this section, well-known estimators for the shape parameter of a Pareto distribution are described. Since none of the original proposals take sample weights into account, the estimators are adjusted for the case of survey samples.

5.1. Hill estimator

Hill (1975) introduced the maximum likelihood estimator for the shape parameter of the Pareto distribution, hence it is commonly referred to as the Hill estimator. The log-likelihood of the shape parameter $\theta$ in the Pareto model is given by

$$l(\theta) = \sum_{i=1}^{k} \log f_{\theta}(x_{n-k+1}) = \sum_{i=1}^{k} (\log \theta + \theta \log x_0 - (\theta + 1) \log x_{n-k+1})$$

(9)
By differentiating \( l(\theta) \) with respect to \( \theta \), equating the derivative to zero, and using \( \hat{x}_0 := x_{n-k} \) as an estimate for \( x_0 \), the Hill estimator for the case without sample weights is obtained by

\[
\hat{\theta}_{\text{Hill}} = \frac{k}{\sum_{i=1}^{k} \log x_{n-k+i} - k \log x_{n-k}}. \tag{10}
\]

In the case of survey samples, the log-likelihood is replaced by the weighted log-likelihood

\[
I_w(\theta) = \sum_{i=1}^{k} w_{n-k+i} \log f_0(x_{n-k+i}) = \sum_{i=1}^{k} w_{n-k+i} (\log \theta + \theta \log x_0 - (\theta + 1) \log x_{n-k+i}). \tag{11}
\]

For the weighted Hill (wHill) estimator, Equation (10) thus generalizes to

\[
\hat{\theta}_{\text{wHill}} = \frac{\sum_{i=1}^{k} w_{n-k+i}}{\sum_{i=1}^{k} w_{n-k+i} (\log x_{n-k+i} - \log x_{n-k})}. \tag{12}
\]

However, it is important to note that the Hill estimator is non-robust, therefore it is included for benchmarking purposes.

5.2. Integrated squared error (ISE) estimator

Terrell (1990) first proposed estimation based on an integrated squared error minimum distance criterion as an alternative to the maximum likelihood framework. Intuitively speaking, this estimation method reduces the influence of outliers by trying to find largest proportion of the data that matches the assumed parametric model (Vandewalle et al., 2007). A detailed discussion on this behavior can be found in Scott (2001).

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

\[
y_i := \frac{s_{n-k+i}}{x_{n-k}}, \quad i = 1, \ldots, k. \tag{13}
\]

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

\[
f_0(y) = \theta y^{-(1+\theta)}. \tag{14}
\]

With this density, the integrated squared error criterion to find an estimate of the parameter \( \theta \) is given by

\[
\hat{\theta} = \arg \min_{\theta} \left[ \int (f_0^2(y)dy - f(y)^2 dy) \right] \tag{15}
\]

\[
= \arg \min_{\theta} \left[ \int f_0^2(y)dy - \int f_0(y)f(y)dy + \int f_0^{-2}(y)dy \right]. \tag{16}
\]

where \( f(y) \) denotes the unknown true density. Since the last term is constant with respect to \( \theta \), it can be omitted. Furthermore, the middle term denotes the expected value of the model density. Hence Equation (16) can be rewritten as

\[
\hat{\theta} = \arg \min_{\theta} \left[ \int f_0^2(y)dy - 2\mathbb{E}(f_0(Y)) \right]. \tag{17}
\]

If there are no sample weights in the data, the ISE estimator is obtained by using the mean as an unbiased estimator of \( \mathbb{E}(f_0(Y)) \):

\[
\hat{\theta}_{\text{ISE}} = \arg \min_{\theta} \left[ \int f_0^2(y)dy - \frac{2}{k} \sum_{i=1}^{k} f_0(y_i) \right]. \tag{18}
\]

For survey samples, the mean in Equation (18) is simply replaced by a weighted mean. This leads to the weighted integrated squared error (wISE) estimator

\[
\hat{\theta}_{\text{wISE}} = \arg \min_{\theta} \left[ \int f_0^2(y)dy - \frac{2}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_0(y_i) \right]. \tag{19}
\]
5.3. Partial density component (PDC) estimator

In an application of the integrated squared error criterion to outlier detection and regression, Scott (2004) noticed that this criterion only requires the true density \( f \) to be a real density, but not \( f_\theta \). Vandewalle et al. (2007) use this result to define the partial density component (PDC) estimator for the Pareto model. This estimator minimizes the integrated squared error criterion based on an incomplete density mixture model \( u f_\theta \). If the data do not contain sample weights, the PDC estimator is thus given by

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

(20)

In order to obtain an estimate for the parameter \( u \), the expression between brackets in Equation (20) is differentiated with respect to \( \theta \) and evaluated at \( \hat{\theta}_{PDC} \). Equating to zero and solving the resulting equation then leads to the estimate

\[
\hat{u} = \frac{1}{k} \sum_{i=1}^{k} f_\theta(y_i) \left/ \int f_\theta^2(y)dy. \right.
\]

(21)

If the uncontaminated part of the data follows a distribution \( F_\theta \) and the contaminated part follows a distribution taking most of its mass outside the range of the model distribution or in a region where the density \( f_\theta \) is almost zero, \( \hat{u} \) can be interpreted as a measure of the proportion of the uncontaminated part of the data. However, this interpretation is not suitable if the assumptions described above do not hold, which includes the case of the uncontaminated part only approximately following a distribution \( F_\theta \). For a more detailed discussion on this issue, the reader is referred to Vandewalle et al. (2007).

Taking sample weights into account, the weighted partial density component (wPDC) estimator is obtained by replacing the mean as an estimator of \( \mathbb{E}(f_\theta(Y)) \) by the weighted mean. Thus Equations (20) and (21) are generalized to

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

(22)

\[
\hat{u} = \frac{1}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \left/ \int f_\theta^2(y)dy. \right.
\]

(23)

6. Semiparametric estimation based on Pareto tail modeling

This section introduces two approaches for reducing the influence of outliers on economic indicators. Both methods are based on the semiparametric Pareto model for the upper tail of a distribution. In particular, they have in common that they first declare values larger than \( F_{\hat{\theta}}^{-1}(1-\alpha) \), the \((1-\alpha)\)-quantile of the fitted distribution, as nonrepresentative outliers. From experience, \( \alpha = 0.005 \) or \( \alpha = 0.01 \) seem to be suitable choices for this tuning parameter. It should be noted that \( \alpha = 0.005 \) is used throughout this paper. In any case, the treatment of the outliers is described in the following:

**Calibration for nonrepresentative outliers (CN):** Since nonrepresentative outliers are considered to be unique to the population data in some sense, 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):** The nonrepresentative outliers are replaced by values drawn from the fitted distribution, thereby preserving the order of the original values.

In addition, preliminary simulation results for a simpler approach based on replacing all values above the threshold can be found in Alfons et al. (2010b). That approach was first proposed by Vandewalle et al. (2007). Nevertheless, it introduces a considerable amount of additional uncertainty and is therefore not considered in this paper.
7. Simulation studies

The simulation studies presented in this section are performed in R (R Development Core Team, 2011) using package simFrame (Alfons et al., 2010a; Alfons, 2011a), which is a general framework for simulations in statistics research.

7.1. Estimation of the shape parameter

The first simulation experiment compares the weighted and unweighted estimators for the shape parameter of the Pareto distribution presented in Section 5. Its aim is to demonstrate the importance of considering the sample weights in the finite population sampling context.

First, 100 population data sets of size $N = 10,000$ are generated. Values in the variable of interest are drawn from a Pareto distribution with scale parameter $x_0 = 1$ and shape parameter $\theta = 1$. The scale parameter $x_0$ is thereby assumed to be known throughout the simulation study. In addition, an auxiliary variable giving probability weights for sampling is then created for each population in the following way. Let $x = (x_1, \ldots, x_N)'$ denote the Pareto distributed variable of interest. Then the probability weights $p = (p_1, \ldots, p_N)'$ are given by

$$p_i = \begin{cases} 
1, & x_i > F_{\theta}^{-1}(\frac{2}{3}) \\
2, & F_{\theta}^{-1}(\frac{1}{3}) < x_i \leq F_{\theta}^{-1}(\frac{2}{3}) \\
3, & x_i \leq F_{\theta}^{-1}(\frac{1}{3}) 
\end{cases}, \quad i = 1, \ldots, N,$$

where $F_{\theta}$ is the cumulative distribution function of the Pareto distribution from (3). Second, 100 samples of size $n = 200$ observations are drawn from each of the population data sets, resulting in a total number of 10,000 simulation runs. The samples are thereby taken using Midzuno’s method for unequal probability sampling (Midzuno, 1952) with inclusion probabilities determined by the probability weights $p$. Hence observations with lower values in the variable of interest have higher inclusion probabilities, which in turn results in lower sample weights.

In order to investigate the behavior of the estimators under contamination, a proportion $\varepsilon$ of randomly selected observations in the samples are replaced by outliers. The contamination level $\varepsilon$ is thereby varied from 0 to 0.5 in steps of 0.05, and the values of the selected observations are drawn from a normal distribution $N(\mu, \sigma)$ with mean $\mu = 1000$ and standard deviation $\sigma = 10$.

![Figure 3: Average simulation results for the estimation of the shape parameter $\theta$ with contamination level $\varepsilon$ varying between 0 and 50%.](image)
Figure 3 displays the average simulation results for varying contamination level ε, where the true shape parameter \( \theta = 1 \) is indicated by the grey horizontal line. Clearly, the unweighted methods overestimate the shape parameter if there is no contamination in the data. With increasing contamination level, the large outliers have a decreasing effect on the shape parameter, resulting in underestimation for higher contamination levels. Note that the outliers thereby have the strongest influence on the Hill estimator, whereas the PDC estimator only exhibits this effect for more than 10% contamination. The weighted estimators, on the other hand, are very close to the true shape parameter in the case of no contamination. As contamination increases, the wHill estimator and (to a lesser extent) the wISE estimator immediately move away from the true value. However, the robust wPDC remains accurate until about 20% contamination. To summarize, the sample weights need to be taken into account in the finite population sampling context, and the wPDC estimator is clearly favorable.

### 7.2. Estimation of the Gini coefficient

In this simulation study, semiparametric estimation of the Gini coefficient is investigated in a close-to-reality setting for EU-SILC data. The basis for the simulations are synthetic population data generated from Austrian EU-SILC data from 2006, the latter of which were provided by Statistics Austria. The population data are thereby simulated with the methodology described in Alfons et al. (2011b) and implemented in the R package simPopulation (Alfons and Kraft, 2010). In total, the population consists of 8 182 222 individuals from 3 505 145 households. For each individual, information on demographics and income is available. It is important to note that the synthetic population data do not contain outliers in the income data, as these are generated in the samples for full control over the amount of contamination (cf. Alfons, 2011b).

Concerning the sampling design, 1 000 samples of 6 000 households are drawn via stratified cluster sampling. While the strata are given by the nine federal states of Austria, the primary sampling units (PSUs) correspond to the households. Within each stratum, households are drawn with unequal probabilities using Midzuno’s method (Midzuno, 1952). The inclusion probabilities are thereby based on the household sizes such that the numbers of sampled households within the strata are proportional to the corresponding numbers of households in the population.

Keep in mind that the Gini coefficient in the case of EU-SILC is estimated from an equivalized household income (see Section 2). Contamination is therefore inserted on the household level in this simulation study. To be more precise, households are first selected by simple random sampling. Then the equivalized household income of the selected households is drawn from a normal distribution \( N(\mu, \sigma) \) with \( \mu = 1 000 000 \) and \( \sigma = 20 000 \). Note that all individuals within a contaminated household are assigned the same value. Moreover, EU-SILC data typically contain only a small amount of outliers. For a realistic scenario, the contamination level is set to \( \varepsilon = 0.0025 \), which in this case corresponds to 15 contaminated households. Additionally, the contamination level \( \varepsilon = 0.01 \) is investigated. This corresponds to 60 contaminated households out of a total 6 000 households, which in official statistics would be considered rather poor data quality. More information on modeling contamination in simulation studies for survey statistics can be found in, e.g., Béguin and Hulliger (2008) and Alfons (2011b).

In this simulation setting, the semiparametric methods from Section 6 are evaluated for varying number of households \( k \) used for Pareto tail modeling. Only the weighted estimators of the shape parameter from Section 5 are thereby considered. Since the Gini coefficient is estimated from an equivalized household income, households are used as observations for fitting the Pareto distribution and detecting outliers. In the CN approach for calibration of nonrepresentative outliers, the sample weight of each individual in an outlying household is set to 1. Calibration on the remaining individuals is then done by raking (see, e.g., Deville et al., 1993) on the subsets given by the strata of the sampling procedure (the nine federal states). In the RN approach for replacement of nonrepresentative outliers, values from the fitted distribution are drawn for the outlying households and all individuals within the same household receive the same replacement value. It is important to note that this simulation study is focused on exploring the behavior of the semiparametric methods for different choices of the threshold for tail modeling. The estimation of the threshold is not addressed in this paper. Furthermore, standard estimation of the Gini coefficient according to Eurostat (see Section 2) is investigated for comparison.

Figure 4 presents the simulation results for the case without contamination. The semiparametric methods are evaluated by the average over the simulation runs (top) and the root mean squared error (RMSE; bottom) for varying number of households \( k \) used for tail modeling. While left panels correspond to the CN approach, the results for the RN procedure are shown in the right panels. In addition, reference lines are drawn for the standard estimation according to Eurostat (dash-dotted grey lines), as well as for the true population value of the Gini coefficient (solid
The results for the scenario with contamination level $\varepsilon = 0.0025$ are displayed in Figure 5. For low values of $k$, the behavior of the semiparametric methods is similar to the standard estimation until they reach a point where they start to reduce the influence of the outliers. Afterwards, there is a smooth transition towards the true population value. The approaches based on the robust wPDC and wISE estimators thereby reach this changepoint very quickly at $k \approx 20$. In particular for the wPDC estimator, there is a large drop in bias and RMSE until $k \approx 30$, after which point the curves continue with a slower rate of reduction. With the ISE estimator, the transition is smoother, hence not as steep in the beginning. Both methods stabilize at $k \approx 130$ and then remain very close to the true population value. Note that the CN approach performs slightly better with respect to bias and RMSE than the RN approach, albeit not enough to state a clear preference. Furthermore, the non-robust wHill estimator needs much more households in the tail to move the estimates of the Gini coefficient towards the true population value. Nevertheless, the estimates are very close to the true value in the upper half of the investigated range of $k$. It is quite surprising that the influence of the outliers on the Gini coefficient in this example can be reduced via a non-robust estimator of the shape parameter at all.

In Figure 6, the simulation results for the configuration with contamination level $\varepsilon = 0.01$ are shown. Due to the increased number of outlying households, a larger number of households in the tail is necessary before the wPDC and wISE estimators are able to reduce the influence of the outliers ($k \approx 100$ for wPDC, $k \approx 110$ for wISE). The wPDC estimator in this case also stabilizes much earlier than the wISE estimator ($k \approx 170$ for wPDC, $k \approx 240$ for wISE). In addition, the Gini estimates based on the wHill estimator remain highly influenced by outliers throughout the entire investigated range of $k$. Thus the semiparametric methods based on the wPDC estimator clearly perform best in this scenario. However, the CN approach leads to very accurate results for the more robust estimators of the shape parameter, whereas there remains some bias and a slightly larger RMSE for the RN method.

To summarize the simulation results for the Gini coefficient, semiparametric estimation based on the wPDC and wISE estimators produces excellent results for a wide range of the number of households $k$ used for tail modeling. In particular under heavier contamination in the upper tail, the wPDC estimator is preferable. Furthermore, the CN approach is favorable over RN since it leads to more accurate results and does not require drawing random values from the fitted distribution.
Figure 5: Simulation results for semiparametric estimation of the Gini coefficient with contamination level $\varepsilon = 0.0025$: average results (top) and RMSE (bottom). Reference lines are drawn for the standard estimation according to Eurostat (dash-dotted grey lines), as well as for the true population value of the Gini coefficient (solid grey line in the top plot).

Figure 6: Simulation results for semiparametric estimation of the Gini coefficient with contamination level $\varepsilon = 0.01$: average results (top) and RMSE (bottom). Reference lines are drawn for the standard estimation according to Eurostat (dash-dotted grey lines), as well as for the true population value of the Gini coefficient (solid grey line in the top plot).
8. Application to real data

In this section, the semiparametric methods for the estimation of the Gini coefficient are applied to real Austrian and Belgian EU-SILC survey data from 2005 and 2006. The data were provided by Eurostat and have already been used for the Pareto quantile plots in Figure 2. Furthermore, the threshold for Pareto tail modeling is determined by Van Kerm’s formula (Van Kerm, 2007). It is given by

\[ \hat{x}_0 := \min(\max(2.5\bar{x}, Q(0.98)), Q(0.97)), \] (24)

where \( \bar{x} \) is the weighted mean and \( Q(\cdot) \) denotes weighted quantiles. Note that this formula was developed specifically for the equivalized disposable income in EU-SILC data and has more of a rule-of-thumb nature. A drawback of the formula is that it is designed to handle only very few outliers, but that is not a problem in this application. Semiparametric estimation of the Gini coefficient is then done in the same manner as in the simulation study from the previous section, except that the CN approach uses raking according to regional information on a more aggregated level, as information on the strata used for sampling is not available in the data provided by Eurostat.

Table 1 shows the results for standard and semiparametric estimation of the Gini coefficient for the Austrian and Belgian EU-SILC data from 2005 and 2006. Standard deviations estimated via stratified bootstrap with calibration are thereby given in parenthesis. Details on the computation of this bootstrap estimator are out of scope for this paper and can be found in Templ and Alfons (2011). For the semiparametric methods, all steps are performed for each bootstrap sample to account for the additional uncertainty. As mentioned above, the data provided by Eurostat do not contain regional information on the strata used for sampling in practice, only more aggregated information is available. Hence estimates of the standard deviation may not be the most accurate since stratification cannot be replicated exactly in the bootstrap samples. Nevertheless, they illustrate the importance of robust estimation of the Gini coefficient.

Regarding the Austrian data, it can easily be seen in Figure 2 (top) that there are no clear outliers from the Pareto model in the upper tail of either the 2005 or 2006 data. There are, however, some irregularities in the upper tail for 2005 (see Figure 2, top left). Since the 2005 and 2006 data are very similar otherwise, those irregularities may be responsible for the differences in the corresponding Gini coefficient estimates. In any case, the semiparametric methods do not detect any outliers in either of the two years, therefore the estimates are identical to the standard estimation of the Gini coefficient.

The situation is different for the Belgian data. Figure 2 (bottom left) reveals one clear outlier in the 2005 data, otherwise the upper tail follows the Pareto model quite nicely. For 2006 (Figure 2, bottom right), the largest observation is also a clear outlier. In addition, the second largest observation slightly deviates from the Pareto model. Nevertheless, it is debatable whether the latter constitutes an outlier. Except for the outliers, the 2005 and 2006 data are very similar. In this example, all semiparametric methods detect the largest observation as the only outlier in 2005, and the two largest observations in 2006. Thus the semiparametric estimates are all very similar and significantly lower than the corresponding standard estimates. It should be noted that the outliers are also detected via the non-robust wHill

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Austria 2005</th>
<th>Austria 2006</th>
<th>Belgium 2005</th>
<th>Belgium 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td></td>
<td>26.13 (0.26)</td>
<td>25.33 (0.22)</td>
<td>28.53 (1.59)</td>
<td>27.82 (1.02)</td>
</tr>
<tr>
<td>wPDC CN</td>
<td></td>
<td>26.13 (0.26)</td>
<td>25.33 (0.22)</td>
<td>26.91 (0.39)</td>
<td>26.78 (0.26)</td>
</tr>
<tr>
<td>wPDC RN</td>
<td></td>
<td>26.13 (0.26)</td>
<td>25.33 (0.22)</td>
<td>26.92 (0.38)</td>
<td>26.79 (0.26)</td>
</tr>
<tr>
<td>wISE CN</td>
<td></td>
<td>26.13 (0.26)</td>
<td>25.33 (0.22)</td>
<td>26.91 (0.40)</td>
<td>26.78 (0.26)</td>
</tr>
<tr>
<td>wISE RN</td>
<td></td>
<td>26.13 (0.26)</td>
<td>25.33 (0.22)</td>
<td>26.91 (0.39)</td>
<td>26.79 (0.26)</td>
</tr>
<tr>
<td>wHill CN</td>
<td></td>
<td>26.13 (0.26)</td>
<td>25.33 (0.22)</td>
<td>26.91 (0.37)</td>
<td>26.78 (0.26)</td>
</tr>
<tr>
<td>wHill RN</td>
<td></td>
<td>26.13 (0.26)</td>
<td>25.33 (0.22)</td>
<td>26.92 (0.38)</td>
<td>26.79 (0.26)</td>
</tr>
</tbody>
</table>

Table 1: Gini coefficient estimated from Austrian and Belgian EU-SILC survey data from 2005 and 2006. Standard deviations estimated via stratified bootstrap with calibration are given in parenthesis. The data were provided by Eurostat.
estimator. This is in accordance with the simulation results from Section 7.2, where semiparametric estimation via the wHill estimator was able to reduce the influence of a small amount of outliers given that enough households are used for tail modeling. However, the estimated standard deviations are much larger for the standard estimation of the Gini coefficient than for the semiparametric methods. Considering that the sample size is 5,137 households for 2005 and 5,860 households for 2006, this suggests a disproportionally high influence of only 1 or 2 households with extremely large equivalized income. The semiparametric estimates may therefore be considered more reliable.

9. Conclusions

This paper introduced robust semiparametric methods for the estimation of economic indicators such as the Gini coefficient from survey samples. More specifically, two approaches based on Pareto tail modeling are considered. For this purpose, commonly used estimators for the shape parameter of the Pareto distribution were adapted to allow for sample weights. The importance of taking sample weights into account was demonstrated by a small simulation experiment. Moreover, a close-to-reality simulation study and an application to real EU-SILC data demonstrated the excellent performance of robust semiparametric estimation in the case of the Gini coefficient. In particular the weighted partial density component (wPDC) estimator for the shape parameter of the Pareto distribution is favorable, as it still lead to excellent results in the simulations with an unrealistically high amount of contamination. Even though the Gini coefficient is used as an example in this paper, the developed semiparametric approaches can be applied to other economic indicators as well. For instance, an extensive collection of results for the Gini coefficient and the income quintile share ratio (see Eurostat, 2004, 2009) for a wide range of simulation settings can be found in a technical report (Hulliger et al., 2011). Last but not least, all methods proposed in this paper are implemented in the R package laeken. The use of the package is thereby demonstrated in a separate package vignette.

Acknowledgments

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). For more information on the project, visit http://ameli.surveystatistics.net.

References

URL http://ameili.surveystatistics.net
URL http://www.R-project.org
URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-3complete.pdf