the review of income and wealth

Review of Income and Wealth
Series 60, Number 4, December 2014
DOI: 10.1111/roiw.12031

MODELING OF INCOME AND INDICATORS OF POVERTY AND SOCIAL EXCLUSION USING THE GENERALIZED BETA DISTRIBUTION OF THE SECOND KIND

BY MONIQUE GRAF

Swiss Federal Statistical Office and Elpacos Statistics

AND

DESISLAVA NEDYALKOVA*

Swiss Federal Statistical Office and University of Neuchâtel

There are three reasons why estimation of parametric income distributions may be useful when empirical data and estimators are available: to stabilize estimation; to gain insight into the relationships between the characteristics of the theoretical distribution and a set of indicators, e.g. by sensitivity plots; and to deduce the whole distribution from known empirical indicators, when the raw data are not available. The European Union Statistics on Income and Living Conditions (EU-SILC) survey is used to address these issues. In order to model the income distribution, we consider the generalized beta distribution of the second kind (GB2). A pseudo-likelihood approach for fitting the distribution is considered, which takes into account the design features of the EU-SILC survey. An ad-hoc procedure for robustification of the sampling weights, which improves estimation, is presented. This method is compared to a non-linear fit from the indicators. Variance estimation within a complex survey setting of the maximum pseudo-likelihood estimates is done by linearization (a sandwich variance estimator), and a simplified formula for the sandwich variance, which accounts for clustering, is given. Performance of the fit and estimated indicators is evaluated graphically and numerically.

JEL Codes: C13, C46, D63

Keywords: income distribution, inequality, maximum pseudo-likelihood estimation, monetary indicators, sandwich variance estimator

1. INTRODUCTION

In December 2001, the European Council meeting took place in Laeken, Belgium. EU Heads of State and Government agreed on common objectives in the area of social inclusion, pensions, health, and long-term care. In order to compare practices in different countries and measure progress toward these common objectives, a set of common indicators was needed. These common indicators consist of an overall list of 14 indicators (Eurostat, 2009). The set of indicators called "Laeken indicators" was lately renamed to indicators of poverty and social exclusion. Our focus is placed, in particular, on the estimation of the at-risk-of-poverty rate (ARPR), the relative median poverty gap (RMPG), the quintile share ratio

Note: This research was conducted at the Swiss Federal Statistical Office and was partly funded by the FP7-SSH-2007-217322 AMELI Research Project.

*Correspondence to: Desislava Nedyalkova, Swiss Federal Statistical Office, Espace de l'Europe 10, 2010 Neuchâtel, Switzerland (desislava.nedyalkova@bfs.admin.ch).

(QSR), and the Gini index, as well as the median. Nevertheless, our results could be applied to other areas of indicator estimation as well.

Parametric income distributions have long been used for modeling income (see, e.g., Kleiber and Kotz, 2003; Chotikapanich, 2008). Modeling of both the whole income range or the tails of the distribution have been investigated in the literature. Here we concentrate on modeling the entire distribution. The advantage of parametric estimation of income distributions is that there are explicit formulas for poverty and inequality measures as functions of the parameters of the theoretical income distribution. The functional relationship between the indicators and the parameters under the assumed distribution gives insight into both: sensitivity of indicators to variations of shape can be assessed on the one hand, and on the other hand interpretation of shape parameters is deepened by the relationship to the indicators. Inequality measures are tightly linked to modeling income distributions (see, e.g., Cowell, 1995; Cowell and Flachaire, 2007).

The generalized beta distribution of the second kind (GB2), which is a four-parameter distribution, is acknowledged to give an excellent description of income distributions (see, e.g., McDonald, 1984; McDonald and Xu, 1995; Bordley *et al.*, 1996; McDonald and Ransom, 2008; Sepanski and Kong, 2008). Amongst its special cases are Fisk (or log-logistic), Dagum, and Singh–Maddala distributions. Interesting limiting cases also include the lognormal and the Generalized Gamma distributions. Empirical studies on income (see, e.g., Kleiber and Kotz, 2003, table B2; Dastrup *et al.*, 2007; Jenkins, 2009) tend to show that the GB2 outperforms other four-parameter distributions for modeling income. Moreover, generalizations to five-parameter distributions do not seem to improve the fit in general. Therefore the GB2 is sufficiently acceptable for a wide range of empirical distributions.

The main emphasis of this paper is to investigate different methods for fitting the GB2 model to the income distribution and to study income inequality at the country level in the context of the EU-SILC survey. The EU-SILC survey was developed in order to collect comparable cross-sectional and longitudinal microdata on income, poverty, social exclusion and, living conditions across participating EU countries. We fit the GB2 distribution to the income variable and compute the indicators from the parameters of the fitted distribution. We use the pseudo-loglikelihood: the population loglikelihood is approximated by the extrapolated sum of the scores. Once the parameter estimates have been obtained, the maximum pseudo-likelihood (PML) estimates of the indicators are obtained by plugging the parameter estimates into the functional expression for the indicators. We provide the design-based variance estimators of the GB2 parameters and of the derived indicators by linearization.

When measuring inequality, we are often confronted with extreme values (see, e.g., Cowell and Victoria-Feser, 1996; Cowell and Flachaire, 2007). In general, GB2 estimation and other maximum likelihood estimation from parametric distributions have robustness problems and are sensitive to extreme values and their specification (see, e.g., Victoria-Feser and Ronchetti, 1994; Victoria-Feser, 2000). Actions have been taken by the SILC data producers in order to limit the influence of very large incomes in the databases (Osier *et al.*, 2006). These consist in recoding of extreme weights to more acceptable values, but possibly less attention has been given to the left tail of the income distribution. In contrast to direct poverty estimates, parametric estimation is influenced by the whole left tail behavior of the income distribution. Small deviations to the GB2 hypothesis in the left tail result in biased indicators under the GB2. In our simulation study (Graf and Nedyalkova, 2011b), we have noticed that a certain bias in the estimates is induced. This led us to the idea of robustifying the sampling weights by creating an ad-hoc procedure for adjusting the sampling weights.

Suppose we do not have the income microdata at our disposal, but the indicators, fitted on empirical data, are publicly available (this is the situation of an external user of the Eurostat website). The indicators have been produced without any reference to a theoretical income distribution. It is then possible to go the other way round, that is to reconstruct the whole income distribution, knowing only the values of the empirical indicators and assuming that the theoretical distribution models the empirical distribution to an acceptable level of precision. This approach has been applied to EU-SILC data with success. This means that the set of indicators contains enough information to permit the reconstruction of the empirical distribution generally to an acceptable level of precision.

The article is structured as follows. In Section 2, the basic properties of the GB2 are briefly recalled and formulas for the above-mentioned indicators under the GB2 model are given. Section 3 focuses on fitting the GB2 by PML estimation, using the whole microdata information. In Section 3, a sandwich estimator of the variance of the PML estimates of the GB2 parameters is given. Section 4 presents an ad-hoc procedure for robustification of the sampling weights. In Section 5, a new method for fitting the parameters of the GB2, using only the set of empirical indicators, is presented. In Section 6, graphical and numerical results on comparison of the two methods of estimation are given. Finally, Section 7 gives some concluding remarks.

2. Indicators of Poverty and Social Exclusion in the EU-SILC Framework Under the GB2 Model

2.1. *The GB2 Model*

The GB2 is a four-parameter distribution and is denoted GB2(a, b, p, q). The GB2 can be obtained by a transformation of a standard beta random variable. Apart from the scale parameter $b > 0$, this distribution has three positive shape parameters a, p, and q. The parameter a represents the overall shape, p governs the left tail, and q the right tale. The GB2 density takes the form:

$$f(x;\theta) = \frac{a}{bB(p,q)} \frac{(x/b)^{ap-1}}{\left(1+(x/b)^{a}\right)^{p+q}}, \quad x \geq 0 \tag{1}$$

where $B(p, q)$ is the beta function, $\theta = (a, b, p, q)^T$ is the vector of parameters, and T stands for transposition.

The cumulative distribution function of a GB2 variable can be written as $F(x; \theta)$. It does not have an explicit form, but is easily obtainable in any statistical software, e.g. R package GB2 (R Development Core Team, 2008; Graf and Nedyalkova, 2011a). The derivation of moments and likelihood equations also necessitates the use of special mathematical functions, like the beta and gamma function and their derivatives. Note that there is a relationship between the beta (B) and gamma (Γ) functions:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

where $\Gamma(x) = (x - 1)!$, $x > 0$.

Let X be a random variable following a GB2 distribution. Then the moment of order k is defined by

$$\text{(2)} \qquad \mathrm{E}\left(X^k\right) = b^k \frac{\Gamma(p + k/a)\Gamma(q - k/a)}{\Gamma(p)\Gamma(q)}.$$

Moments exist for $-ap < k < aq$.

The incomplete moment of order k (Butler and McDonald, 1989) is given by

$$\text{(3)} \qquad \frac{\mathrm{E}\left(X^k | X < x\right)}{\mathrm{E}\left(X^k\right)} = F_{(k)}(x; a, b, p, q) = F\left(x; a, b, p + \frac{k}{a}, q - \frac{k}{a}\right).$$

Thus it can be expressed with the help of a GB2 cumulative distribution function with special parameters.

The log density of the GB2 distribution is given by:

$$\log(f) = \log(a) - \log(b) - \log(\Gamma(p)) - \log(\Gamma(q)) + \log(\Gamma(p+q)) \\ + (ap - 1)\log(x/b) - (p+q)\log\left(1 + (x/b)^a\right).$$

2.2. *The Set of Indicators*

There are simple and explicit formulas for the inequality measures as functions of the parameters of the income distribution. McDonald (1984) gave the analytic form of the Gini index under the GB2 distribution, but the GB2 expressions for the other indicators are new and easily obtained through the cumulative distribution function, or the quantile function, or using the moments of the distribution.

The equivalized income is the main income variable in the EU-SILC survey and is equal to the total gross household income over the household equivalized size, where the household equivalized size is a weighted sum of the number of adults and children in the household. There is the particularity that the household's equivalized income equals the equivalized income of each member of the household. Another property of the survey is that all household members have the same sampling weight (Eurostat, 2009).

Here, we recall the definition of the ARPR, RMPG, QSR, and the Gini index and give their derived expressions (as functions of the parameters of the distribution) under the GB2 hypothesis. These inequality and poverty measures fulfill the well-known property of scale invariance (see, e.g., Atkinson and Bourguignon,

2000) and can be computed with a GB2 distribution for which the scale parameter b can be chosen arbitrarily, e.g. $b = 1$. Let F denote the cumulative distribution function of the equivalized income X and in particular the GB2 cumulative distribution function.

1. *At-risk-of-poverty rate (ARPR)*
   Let $m = x_{50} = F^{-1}(0.5)$ denote the median income, then $0.6m$ is called the at-risk-of-poverty threshold (ARPT) or "poverty line." The ARPR is the proportion of the population under the poverty line. Then, under the GB2 distribution,

$$ARPR(a, p, q) = \Pr(X < 0.6m) = F(0.6m; a, 1, p, q).$$

2. *Relative median poverty gap (RMPG)*
   Let $m_p = F^{-1}(ARPR/2)$ denote the median income of the poor (those under the poverty line). Then, $RMPG$ is the relative gap between the poverty line and the median income of the poor and is defined as one minus the ratio between the median income of the poor to 60 percent of the median income of the population.

$$RMPG = \frac{0.6m - m_p}{0.6m}.$$

   Under the GB2, if $A = ARPR(a, p, q)$, then:

$$RMPG(A, a, p, q) = 1 - F^{-1}(A/2, a, 1, p, q)/F^{-1}(A, a, 1, p, q),$$

   where F^{-1} stands for the GB2 quantile function.

3. *Quintile share ratio (QSR or S_{80}/S_{20})*
   Let x_{80} (resp. x_{20}) be the 80-th (resp. the 20-th) percentile of the GB2 distribution. The quintile share ratio is the ratio of the sum of the upper quintile incomes over the sum of the lower quintile incomes.

$$QSR = \frac{\mathrm{E}(X|X > x_{80})}{\mathrm{E}(X|X < x_{20})}.$$

   The quintile share ratio under the GB2 hypothesis can be expressed with the help of the incomplete moments of order 1 (equation (3), with $k = 1$):

$$QSR(a, p, q) = 1 - F_{(1)}(x_{80}; a, 1, p, q)/F_{(1)}(x_{20}; a, 1, p, q).$$

4. *Gini index*
   There are many definitions of the Gini index. One of them is: Let X and Y be two identically distributed independent positive random variables. Then, the Gini index is defined as:

$$Gini = \frac{\mathrm{E}(|X - Y|)}{2\mathrm{E}(X)}.$$

> The index is an inequality indicator measuring the expected absolute difference between two independently selected incomes relative to the mean income. The Gini index of the GB2 distribution is given by McDonald (1984). An efficient algorithm to compute the Gini index from its analytical expression has been described in Graf (2009), and implemented in the GB2 package in R.

As the four indicators are scale-invariant under the GB2, we can ask ourselves how these indicators behave in relation to the shape parameters *a*, *p*, and *q*. A sensitivity plot, implemented in the R package GB2, illustrates this. Figure 1 shows how the values of ARPR vary in relation to the parameters *p* and *q*, for different values of *a* which is kept fixed. We can see that for small values of *a*, ARPR depends on all three parameters, but when *a* increases, the dependence on *q* diminishes.

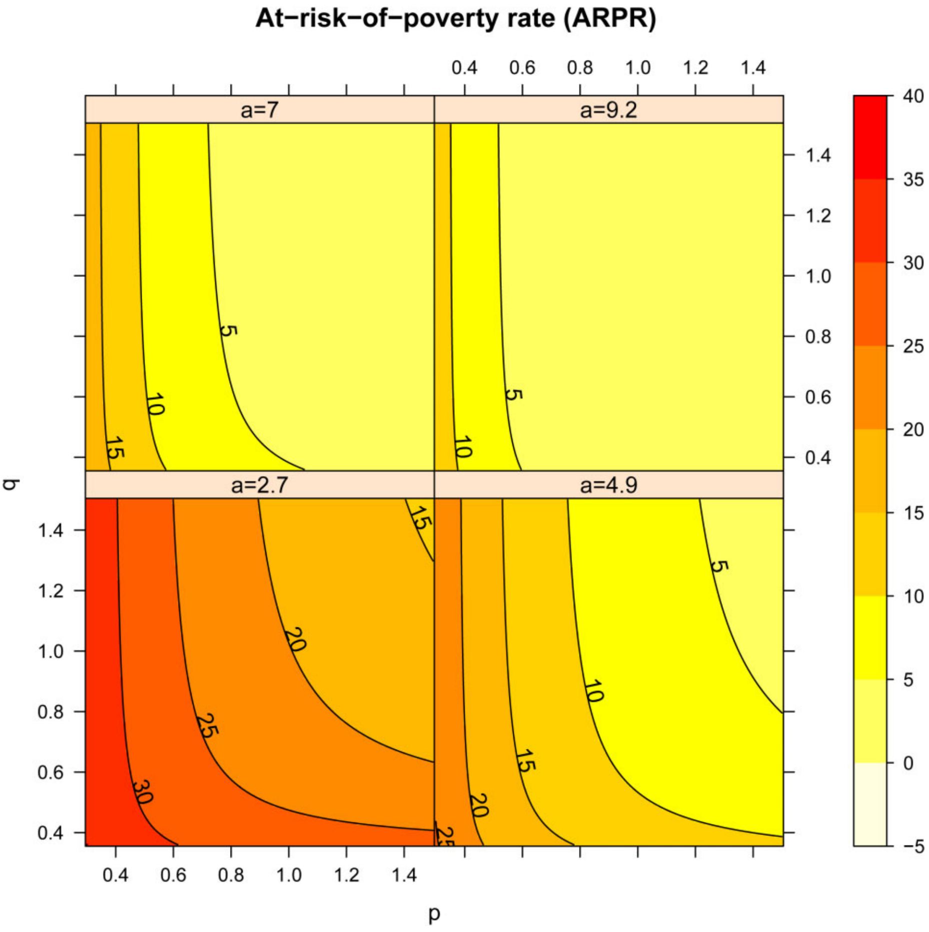

Figure 1. Sensitivity Plot of the ARPR

3. Maximum Pseudo-Likelihood Estimation of the Parameters of the GB2 Distribution Under Cluster Sampling

In the classical case of maximum likelihood estimation, the loglikelihood function is defined as a sum over the sample of the log density evaluated at the data points. The EU-SILC being a complex survey, estimating a model based on the EU-SILC microdata amounts to incorporating the sampling weights in the likelihood, producing a pseudo-likelihood (e.g., Skinner *et al.*, 1989; Chambers, 2003). The pseudo-loglikelihood is computed as a weighted sum over the sample of the log density of the distribution, where the weights are the sampling weights. It is a function of the parameters of the distribution. Maximizing it provides us with a set of parameters which fits the GB2 to the income variable by taking the sampling design into consideration.

In the EU-SILC framework, the data are observed at two levels—personal level and household level. Households (clusters) are sampled and then all persons in the selected households enter the sample. The personal disposable income is measured as an equivalized measure over all household members (Eurostat, 2009). All persons of a household have the same equivalized disposable income (x_i), which is also the household's equivalized disposable income, thus the observations are not independent (Clémenceau and Museux, 2007). The sampling weights (the sampling weight of a household equals the sampling weight of each person belonging to the household) are not simply the inverse of the inclusion probabilities but are obtained through calibration and adjusted for non-response (Osier *et al.*, 2006).

Let m, n_i and n denote, respectively, the number of households in the sample, the number of persons belonging to household i, and the number of persons in the sample. Then, the weighted (pseudo)-loglikelihood function, at the household level, is defined as

$$\ell_m(\theta) = \sum_{i=1}^{m} w_i n_i \log f(x_i;\theta),$$

where $f(\cdot)$ is the GB2 density, given in equation (1) and w_i are the sampling weights. In order to avoid large numerical values in the computation, we scale $\ell_m(\theta)$ by dividing by the sum of weights $\sum_{i=1}^{m} w_i n_i$. In classical likelihood theory, the weights are equal to one and thus sum to the sample size. To obtain the pseudo-loglikelihood at a similar scale one should multiply $\ell_m(\theta)$ by the sample size n.

The partial derivatives of the pseudo-loglikelihood function are readily obtained as weighted sums of the partial derivatives of $\log(f(x_i))$, evaluated at the data points. Thus, the first partial derivatives of ℓ_m with respect to θ are:

$$\ell'_m(\theta) = \sum_{i=1}^{m} w_i n_i u(x_i;\theta),$$

where

$$u(x_i;\theta) = [\log f(x_i;\theta)]' = \frac{\partial}{\partial\theta}\log f(x_i;\theta)$$

is the 4 by 1 vector of the first partial derivatives of log($f(x_i;\theta)$) with respect to θ, for a given observation i.

Similarly, the second partial derivatives of ℓ_m with respect to θ are:

$$\ell''_m(\theta) = \sum_{i=1}^{m} w_i n_i h(x_i;\theta),$$

where

$$h(x_i;\theta) = [\log f(x_i;\theta)]'' = \frac{\partial^2}{\partial\theta^2}\log f(x_i;\theta)$$

is a symmetric 4 by 4 matrix of the second partial derivatives of log($f(x_i;\theta)$) with respect to θ, for a given observation i.

The quantity

$$I(\theta) = -\mathrm{E}(\ell''_m(\theta))$$

is called the Fisher information matrix. For the GB2 distribution, the Fisher information matrix was computed by Prentice (1975) and more recently by Brazauskas (2002).

In classical maximum likelihood theory,

$$\mathrm{E}(\ell'_m(\theta)) = 0,$$

$$\mathrm{var}(\ell'_m(\theta)) = -\mathrm{E}(\ell''_m(\theta)).$$

The value of the parameter θ that maximizes the pseudo-loglikelihood is called the PML estimate $\hat{\theta}_m$ and is obtained by setting the first derivatives equal to zero. Thus we have

$$\ell'_m(\hat{\theta}_m) = 0. \tag{4}$$

When solving the likelihood equations, it is possible to express $\hat{p}$ and $\hat{q}$ as functions of a and b. The profile pseudo-loglikelihood logL_p has only two parameters $\theta' = (a, b)$. It is given by replacing θ in the full pseudo-loglikelihood by $\theta = (a, b, \hat{p}(a,b), \hat{q}(a,b)) = (\theta', \hat{p}(\theta'), \hat{q}(\theta'))$:

$$\log L_p = \sum w_i \log f(x_i;\theta') \Big/ \sum w_i.$$

Its advantages over the full pseudo-loglikelihood are that contour plots can be produced (see Figure 5) and that the fitting algorithm is faster.

Functions performing PML estimation based on the full and the profile pseudo-loglikelihoods are implemented in the R package GB2 (Graf and Nedyalkova, 2011a). PML estimation is obtained through methods for non-linear optimization like the BFGS method. Initial values for a and b come from the Fisk distribution, which is GB2 with $p = q = 1$. Moment estimators of a and b for this distribution are (see Graf, 2007):

$$\hat{m}_{log} = \sum w_i \log x_i \Big/ \sum w_i$$

$$\hat{v}_{log} = \sum w_i \left(\log x_i - \hat{m}_{log}\right)^2 \Big/ \sum w_i$$

(5) $$\hat{a} = \pi \Big/ \sqrt{3\hat{v}_{log}}$$

(6) $$\hat{b} = \exp\left(\hat{m}_{log}\right).$$

3.1. *Variance Estimation of the Parameters of the GB2 Distribution and the Derived Indicators*

A sandwich variance estimator is a common tool used in survey sampling for variance estimation of maximum likelihood estimates. It requires the vector of scores and the Fisher information matrix. However, if we do not know the true density, then we can use the so called robust sandwich variance estimator proposed by Huber (1967) and independently derived also by White (1980, 1982). An expository paper on this estimator is Freedman (2006). Pfeffermann and Sverchkov (2003) use this estimator in the survey sampling setting. Another name for this method of estimation is the Taylor-series linearization method.

We can approximate $\ell'_m\left(\hat{\theta}_m\right)$ by the first two terms of a Taylor series around θ. From equation (4), we have

$$\ell'_m\left(\hat{\theta}_m\right) \approx \ell'_m(\theta) + \ell''_m(\theta)\left(\hat{\theta}_m - \theta\right) = 0$$

$$\hat{\theta}_m - \theta \approx [-\ell''_m(\theta)]^{-1}\left(\ell'_m\left(\hat{\theta}_m\right) - \ell'_m(\theta)\right).$$

Then the variance of the maximum likelihood estimate is

$$\mathrm{var}\left(\hat{\theta}_m\right) = \mathrm{E}\left(\hat{\theta}_m - \theta\right)^2 \approx [-\ell''_m(\theta)]^{-1} V(\theta) [-\ell''_m(\theta)]^{-1},$$

where

$$V(\theta) = \mathrm{var}(\ell'_m(\theta)) = \mathrm{E}\left((\ell'_m(\theta))(\ell'_m(\theta))^T\right).$$

Thus the linearized sandwich variance estimator is

(7) $$\widehat{\mathrm{var}}\left(\hat{\theta}_m\right) \approx \left[-\ell''_m\left(\hat{\theta}_m\right)\right]^{-1} \hat{V}\left(\hat{\theta}_m\right) \left[-\ell''_m\left(\hat{\theta}_m\right)\right]^{-1},$$

where $\ell''_m(\theta)$ is estimated directly from the sample. Thus we have

$$\ell''_m\left(\hat{\theta}_m\right) = \sum_{i=1}^{m} w_i n_i h\left(x_i; \hat{\theta}_m\right)$$

and $\hat{V}\left(\hat{\theta}_m\right)$ can be calculated in the following way using all the available design information.

Let π_{Ii} denote the first-order inclusion probability of a household in the sample s_i and $\pi_{Ii_1i_2}$, respectively, the second-order inclusion probability for two different households i_1 and i_2 with respective sample sizes n_{i_1} and n_{i_2}. The Horvitz–Thompson estimator of $V(\theta)$ is:

$$\hat{V}\left(\hat{\theta}_m\right)=\sum_{i_1,i_2=1}^{m} w_{i_1}w_{i_2}u\left(x_{i_1};\hat{\theta}_m\right)u\left(x_{i_2};\hat{\theta}_m\right)^T\sum_{j=1}^{n_{i_1}}\sum_{k=1}^{n_{i_2}}\frac{\pi_{jk}-\pi_j\pi_k}{\pi_{jk}}, \tag{8}$$

where π_j and π_k are the inclusion probabilities for persons j and k. This gives us (Särndal *et al.*, 1992)

$$\pi_j=\pi_{Ii} \text{ if } j\in s_i. \tag{9}$$

$$\pi_{jk}=\begin{cases}\pi_{Ii} & \text{if } j,k\in s_i,\\ \pi_{Ii_1i_2} & \text{if } j\in s_{i_1} \text{ and } k\in s_{i_2},\end{cases} \tag{10}$$

where s_{i_1} and s_{i_2} denote, respectively the samples with respective sizes n_{i_1} and n_{i_2}.

If we plug equations (9) and (10) into equation (8) we find the general design-based formula for the midterm of the sandwich variance estimator:

$$\hat{V}\left(\hat{\theta}_m\right)=\sum_{i_1,i_2=1}^{m} w_{i_1}n_{i_1}w_{i_2}n_{i_2}u\left(x_{i_1};\hat{\theta}_m\right)u\left(x_{i_2};\hat{\theta}_m\right)^T\frac{\pi_{Ii_1i_2}-\pi_{Ii_1}\pi_{Ii_2}}{\pi_{Ii_1i_2}}, \tag{11}$$

where n_{i_1} and n_{i_2} are, respectively, the number of persons in samples s_{i_1} and s_{i_2}.

If we suppose that the households are independently (but not identically due to the different n_i) distributed, then

$$\pi_{Ii_1i_2}=\pi_{Ii_1}\pi_{Ii_2} \text{ if } i_1\neq i_2.$$

In this case, equation (11) becomes

$$\hat{V}\left(\hat{\theta}_m\right)=\sum_{i=1}^{m} w_i^2n_i^2u\left(x_i;\hat{\theta}_m\right)u\left(x_i;\hat{\theta}_m\right)^T(1-\pi_{Ii}).$$

Moreover, if we neglect the finite population correction, thus supposing $1-\pi_{Ii}\simeq 1$ we obtain a simplified formula for the midterm of the sandwich variance estimator:

$$\hat{V}\left(\hat{\theta}_m\right)=\sum_{i=1}^{m} n_i^2w_i^2u\left(x_i;\hat{\theta}_m\right)u\left(x_i;\hat{\theta}_m\right)^T. \tag{12}$$

The midterm of the sandwich variance estimator (11) can be calculated numerically, for example using the R package survey (see Lumley, 2010). In this case, inclusion probabilities, sample strata sizes, etc. are considered when calculating the variance of the scores. We have successfully implemented this in our simulation study (Graf and Nedyalkova, 2011b). We have seen that our variance estimate by linearization using the simplified formula for the variance of the scores (equation 12) is almost equal to the design variance calculated with the package survey for the one-stage sampling designs.

Now we would like to estimate the variance of the fitted indicators, to construct confidence intervals and to compare with their empirical estimates. We know that the median, ARPR, RMPG, QSR, and Gini can all be expressed as functions of the GB2 parameters a, b, p, and q (see Section 2). Thus in order to obtain a variance estimator for a given indicator, we can apply the delta method (see, e.g., Davison, 2003). If we denote, for example, $\hat{A} = A(\hat{\theta}_m)$, the PML estimate of the ARPR, then by the delta method, we have:

$$\widehat{\text{var}}(\hat{A}) = \frac{\partial \hat{A}}{\partial \hat{\theta}_m}' \widehat{\text{var}}(\hat{\theta}_m) \frac{\partial \hat{A}}{\partial \hat{\theta}_m},$$

where $\widehat{\text{var}}(\hat{\theta}_m)$ is given in equation (7). The derivatives of the indicators with respect to the vector of parameters are calculated numerically. Next, we can easily compute confidence intervals and confidence domains.

4. Robustification of the Sampling Weights

Due to the fact that pseudo-maximum likelihood estimators are often sensitive to extreme weights we came to the idea of robustifying the sampling weights. Our procedure is inspired by, but does not directly follow, the MAD-rule (see Luzi *et al.*, 2007. We start from the Fisk distribution, which is a GB2 with $p = q = 1$. Its cumulative distribution function (see Kleiber and Kotz, 2003, p. 222) is given by:

$$F(x; a, b, 1, 1) = \frac{(x/b)^a}{1 + (x/b)^a}.$$

The α–th quantile of the Fisk(a, b) is given by:

$$x_\alpha = b\left(\frac{\alpha}{1-\alpha}\right)^{1/a}. \tag{13}$$

From equation (13), it follows that:

$$\frac{x_\alpha}{b} \frac{x_{1-\alpha}}{b} = 1.$$

Thus the geometric mean between the two symmetric quantiles x_α and $x_{1-\alpha}$ is equal to b, the median under the Fisk distribution.

Let x denote the observed value, in our case, the equivalized income. Our procedure is as follows:

1. First, we define our scale as:

$$d = \frac{x_{1-\alpha}}{b} - \frac{x_\alpha}{b}, \tag{14}$$

   where α takes a small value, e.g. 0.001.
2. Next, the correction factor is calculated as follows:

$$corr = \max\left\{c, \min\left(1, \frac{d}{|b/x-1|}, \frac{d}{|x/b-1|}\right)\right\}, \tag{15}$$

   where c is a constant that can take different values, e.g. 0.2, and can be used to limit the correction factor. The correction factor is of Huber-type (Huber, 1981). One can easily find that the correction factor *corr* is given by

$$corr = \begin{cases} c & \text{if} & & & x/b & \le & c/(d+c), \\ dx/(b-x) & \text{if} & c/(d+c) & \le & x/b & \le & 1/(d+1), \\ 1 & \text{if} & 1/(d+1) & \le & x/b & \le & d+1, \\ db/(x-b) & \text{if} & d+1 & \le & x/b & \le & (d+c)/c, \\ c & \text{if} & (d+c)/c & \le & x/b. & & \end{cases}$$

3. The sampling weights are multiplied by the correction factor *corr*.
4. The weights are multiplied by the ratio of the sum of the unadjusted weights and the sum of the adjusted weights, in order to keep the sum of weights constant.

This robust procedure tends to make the fitted GB2 parameters p and q closer. For example, in our simulation study with the AMELIA dataset (Alfons *et al.*, 2011), if this adjustment is processed, we downweight about 0.2 percent of the observations, essentially on the left tail. Figure 2 shows the correction of the weights obtained with Fisk parameter $a = 1.78$ and tuning parameter $\alpha = 0.01$ (which implies that $d \approx 13$), and $c = 0.1$. These parameters are similar to those used with the AMELIA dataset.

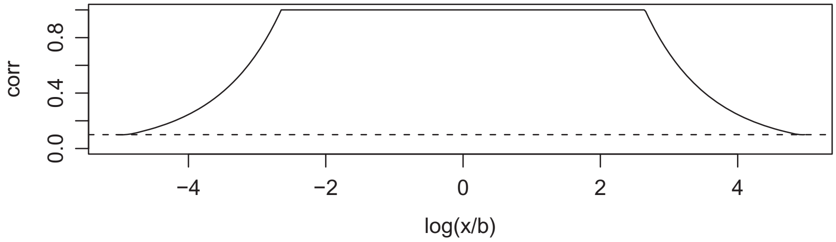

Figure 2. Correction Factor for the Robustification of Weights (Huber-Type Function). Dotted line corresponds to limit c. x is the income and b the median income

5. A New Method of Estimation of Income Data from a Set of Indicators

5.1. *Nonlinear Fit for Indicators*

Let us suppose that we have access to indicators fitted on the empirical data. It is then possible to reconstruct the whole income distribution, assuming the GB2 model fits acceptably the empirical data. Consider a set of indicators A = (median, ARPR, RMPG, QSR, Gini) and their corresponding GB2 expressions $A_{GB2}(a, b, p, q)$. The method of estimation we developed (hereafter referred to as method of nonlinear fit for indicators) consists of finding the set of GB2 parameters a, b, p, and q that minimizes the distance between the empirical estimates of the indicators A_{empir} and their GB2 representations $A_{GB2}(a, b, p, q)$:

$$\sum_{i=1}^{5} c_i \left\{A_{empir,i} - A_{GB2,i}(a, b, p, q)\right\}^2,$$

where the weights c_i take the differing scales into account.

This idea was successfully implemented in the package GB2 in R using a non-linear regression model. Instead of fitting the GB2 parameters all together, we process in two consecutive steps, which appears to be more efficient:

- In the first step, we use the set of indicators A, excluding the median. These indicators are scale-invariant and their corresponding expressions are given as functions of a, ap, and aq (we set $b = 1$), where ap denotes the product of the two parameters a and p and aq the product of a and q. We choose ap and aq instead of simply choosing p and q because constraints on the moments (see equation 2) imply bounds on ap and aq. We have chosen $ap > 1$ and $aq > 2$, so that at least $E(1/X) < \infty$ and $E(X^2) < \infty$ under the GB2. The bounds for the parameter a can be defined in function of the coefficient of variation of the PML estimate of parameter a or simply as $0.2 \cdot a_0$ and $1.8 \cdot a_0$, where a_0 denotes the initial value of the parameter a. Thus, from the first step we deduce the fitted parameters $\hat{a}$, $\hat{p}$ and $\hat{q}$.
- In the second step, only the parameter b is estimated, fitting a non-linear regression model on the empirical median in function of the GB2 median calculated with the NLS estimates of the parameters a, p, and q obtained in the first step. Thus we obtain the fitted parameter $\hat{b}$.

Initial values for the parameters can be taken as the moment estimators of the Fisk distribution in equations (6) and (5), and $p = q = 1$. A natural choice can also be an initial value for b given by the empirical median, and for a by the inverse of the empirical Gini coefficient. This is in accordance with the information the user is supposed to have, namely the set of indicators A. If the PML estimates of the GB2 parameters are known, they give a third choice for the initial values.

The weights c_i, $i = 2, \ldots, 5$ should be selected in order to take the different scales of the indicators into account. A natural choice can be the inverse of the empirical variances of the indicators. Another choice of weights, which we have used in our procedure with success, is the set of weights $c_i = 0.1, 0.1, 1, 1$. In this case more weight is given to QSR and Gini. The user is however free to select his own set of weights.

6. An Application

In this section we present different plots, produced with R, for the case of the EU-SILC survey, in which the method of non-linear fit for indicators is compared with the method of PML estimation. Some numerical results of the fitted GB2 parameters and indicators are shown.

6.1. *Distribution Plots*

We present two different types of plots. The first plot is a cumulative distribution plot in which the GB2 cumulative distribution function is plotted against the empirical distribution function. The second plot is a density plot in which are plotted a kernel density estimate (Epanechnikov) and a GB2 density estimate of the income variable. The Epanechnikov kernel is given by a quadratic weight function within an interval around each observed value. The length of the interval is called the bandwidth and N is the sample size.

Figure 3 shows an example of the fitted GB2 distribution by PML estimation and the method of non-linear fit for indicators with the Austrian EU-SILC data, 2006. From the plot it can be seen that the GB2 fits the empirical distribution well and that the fit by PML is slightly better than the non-linear fit for the indicators.

Figure 4 shows the cumulative distribution plots of the fitted GB2 distribution by the methods of maximum pseudo-likelihood estimation and non-linear fit for indicators for the remaining 25 participating countries in the EU-SILC survey, 2006. We can again conclude that the GB2 correctly fits the empirical income distribution and that the NLS method guesses the income distribution quite well.

6.2. *Contour Plot of the Profile Loglikelihood*

Figure 5 presents a contour plot of the profile pseudo-loglikelihood (see equation 3) for the Austrian EU-SILC sample, 2006. The Fisk, PML prof, and NLS estimates of the parameters a and b are denoted respectively as “F,” “P,” and “N.” Each contour represents a value of the profile pseudo-loglikelihood. We can see that the value of the PML estimate based on the profile pseudo-loglikelihood (“P”) is close to the small quadrangle on the figure, which is the graphical representation of the maximum value of the pseudo-loglikelihood. We can notice that the profile pseudo-loglikelihood is rather flat around the maximum. Thus, different sets of parameters produce close values of the profile pseudo-loglikelihood. The values of the parameters a, b and the profile pseudo-loglikelihood for Fisk, PML prof, and NLS estimates are given in Table 1.

6.3. *Estimated Parameters and Indicators, EU-SILC Participating Countries 2006*

Tables 2 and 3 present the fitted GB2 parameters, the estimated median, ARPR, RMPG, QSR and Gini index for the 26 participating countries in the EU-SILC 2006 survey, i.e. AT, BE, CY, CZ, DE, DK, EE, ES, FI, FR, GR, HU, IE, IS, IT, LT, LU, LV, NL, NO, PL, PT, SE, SI, SK, UK. Estimates are based on PML estimation using the full and profile pseudo-loglikelihoods with adjusted sampling weights following the ad-hoc procedure described in Section 4, and the

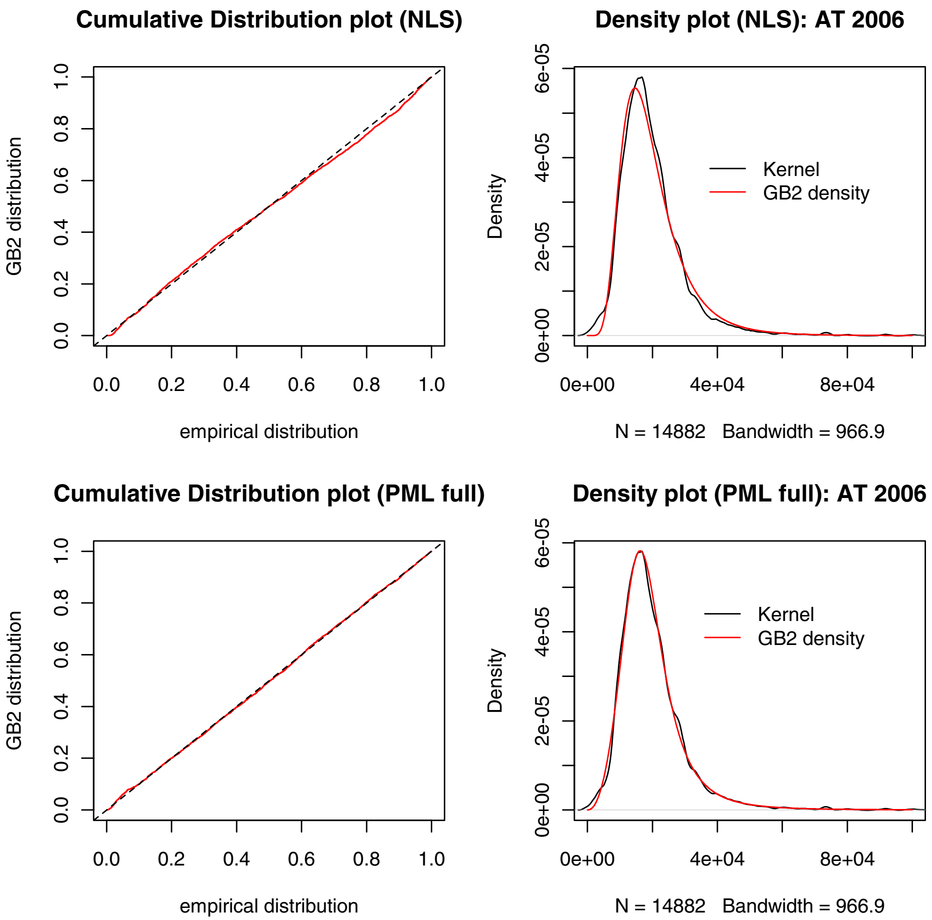

Figure 3. Distribution and Density Plots, Austria 2006. Top Panels: Non-Linear Fit for Indicators. Bottom Panels: PML Estimation Using the Full Pseudo-Loglikelihood

method of non-linear fit for indicators applying as initial values for a and b, respectively, the inverse of the Gini coefficient and the empirical median, and $p = q = 1$.

We can see from the table that PML estimation tends to overestimate ARPR and RMPG. However the robustification of the sampling weights has improved the point estimates. We do not provide tables with the results using the non-adjusted sampling weights. Comparative tables of PML estimation with and without adjusting the sampling weights, based on the simulated universe AMELIA are provided in Graf and Nedyalkova (2011b). We can also see that the methods of PML estimation using full and profile pseudo-likelihoods give similar results, which is expected as analytically the solutions to the optimization problem are similar. At the optimum of the full PML, the parameters p and q should verify the constraints utilized in the profile PML. The method of non-linear fit for indicators succeeds in reproducing the empirical indicators.

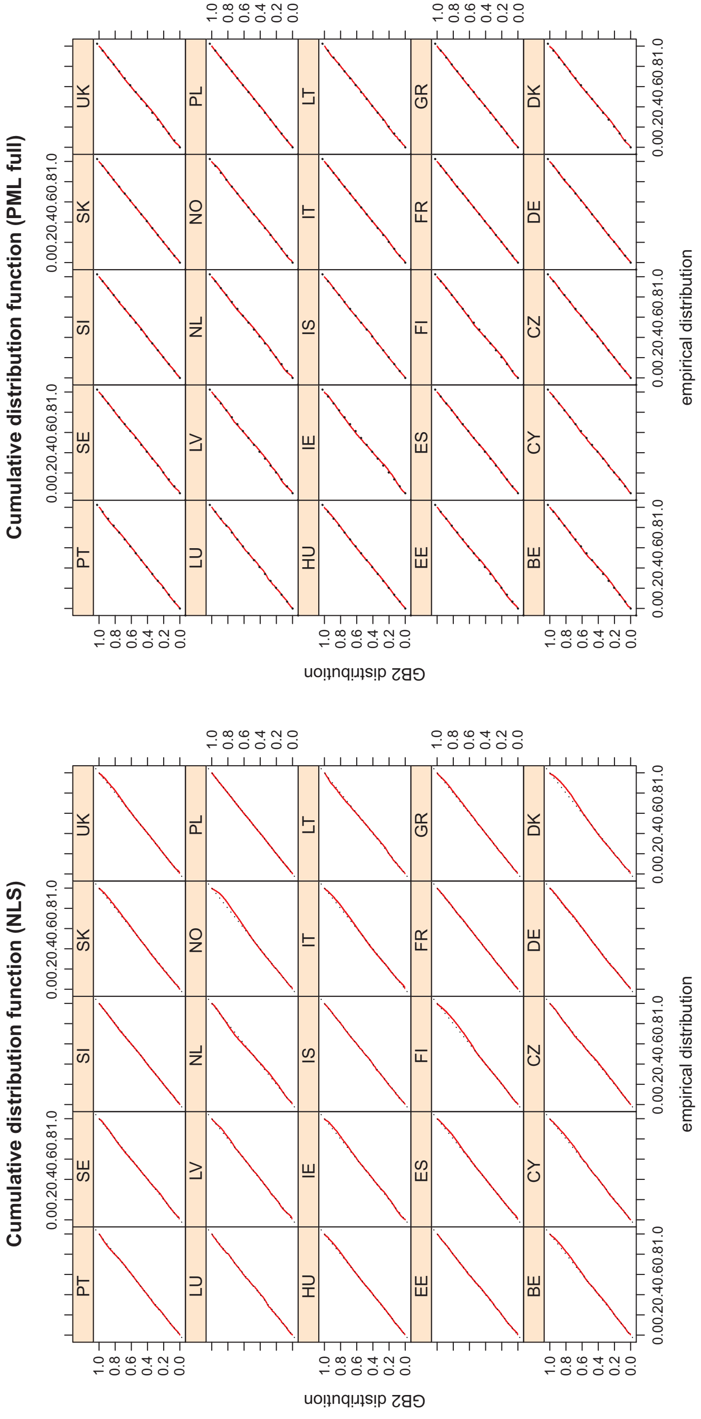

Figure 4. Cumulative Distribution Plots for the Remaining 25 Countries in the EU-SILC Survey, 2006. Left: Non-Linear Fit for Indicators. Right: PML Estimation Using the Full Pseudo-Loglikelihood

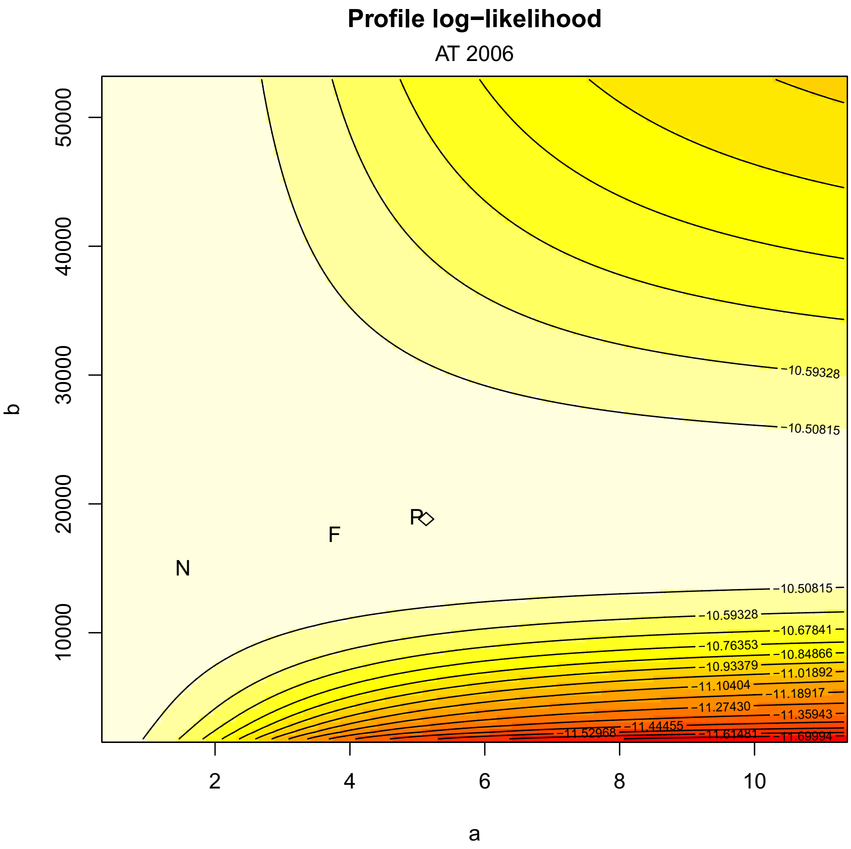

Figure 5. Contour Plot of the Profile Loglikelihood, Austria 2006

TABLE 1
Profile Pseudo-Loglikelihood and Parameter Values Corresponding to the Points Depicted in Figure 5

	a	b	pseudo-loglikelihood
F	3.747	17,642.39	−10.42652
P	4.990	18,995.88	−10.42299
N	1.523	15,049.95	−10.44769
Graphical ML	5.131	18,818.55	−10.42302

7. Discussion

We have seen that parametric distributions may be useful in a survey setting, i.e. the EU-SILC survey. We have chosen the GB2 as a parametric model for the empirical income distribution. This approach offers the advantage of presenting monetary indicators as functions of the parameters of the chosen distribution. Thus parametric modeling allows us to stabilize estimation and gain insight into the relationship between indicators and the distribution of the parameters. The

TABLE 2

GB2 Fitted Parameters and Indicators, Countries 1–13

Country	Type	a	b	p	q	Median	ARPR	RMPG	QSR	GINI
AT	Direct	–	–	–	–	17,854	12.547	15.425	3.647	0.253
AT	NLS	1.523	15,050	5.195	4.079	17,854	12.547	15.425	3.646	0.257
AT	PML full	4.964	19,005	0.654	0.790	17,911	12.716	19.833	3.661	0.253
AT	PML prof	4.990	18,996	0.650	0.784	17,911	12.710	19.840	3.662	0.253
BE	Direct	–	–	–	–	17,225	14.547	19.034	3.960	0.272
BE	NLS	1.941	18,719	2.474	2.853	17,225	14.547	19.034	3.960	0.270
BE	PML full	3.367	18,643	1.050	1.319	17,043	13.707	19.740	3.791	0.260
BE	PML prof	3.279	18,706	1.090	1.376	17,041	13.729	19.698	3.787	0.260
CY	Direct	–	–	–	–	14,532	15.747	18.965	4.268	0.288
CY	NLS	1.132	13,919	6.487	6.194	14,532	15.747	18.965	4.268	0.285
CY	PML full	2.642	14,245	1.564	1.536	14,366	14.343	18.922	4.128	0.280
CY	PML prof	2.551	14,192	1.658	1.617	14,361	14.362	18.829	4.124	0.280
CZ	Direct	–	–	–	–	4,797	9.796	16.967	3.516	0.253
CZ	NLS	7.017	4,619	0.537	0.465	4,797	9.796	16.967	3.516	0.252
CZ	PML full	4.846	4,609	0.854	0.751	4,796	10.213	16.187	3.457	0.248
CZ	PML prof	4.869	4,610	0.849	0.746	4,796	10.208	16.198	3.457	0.248
DE	Direct	–	–	–	–	15,646	12.339	19.625	3.800	0.260
DE	NLS	5.831	15,902	0.555	0.586	15,646	12.339	19.625	3.800	0.263
DE	PML full	7.481	16,351	0.400	0.468	15,680	12.458	20.791	3.703	0.255
DE	PML prof	7.530	16,348	0.397	0.465	15,680	12.448	20.796	3.701	0.255
DK	Direct	–	–	–	–	22,718	11.326	15.159	3.241	0.230
DK	NLS	0.870	26,747	14.380	16.525	22,718	11.289	15.169	3.257	0.233
DK	PML full	6.332	24,834	0.517	0.732	22,665	11.275	19.302	3.174	0.223
DK	PML prof	6.261	24,840	0.525	0.743	22,661	11.262	19.255	3.172	0.223
EE	Direct	–	–	–	–	3,645	18.141	21.841	5.361	0.328
EE	NLS	1.878	3,354	2.203	1.929	3,645	18.141	21.841	5.360	0.331
EE	PML full	2.597	3,972	1.116	1.298	3,679	18.804	24.781	5.517	0.331
EE	PML prof	2.557	3,984	1.140	1.331	3,680	18.834	24.788	5.511	0.331
ES	Direct	–	–	–	–	11,493	19.760	25.399	5.109	0.308
ES	NLS	0.912	22,321	5.824	10.393	11,493	19.474	25.464	5.164	0.316
ES	PML full	2.691	15,675	0.914	1.738	11,476	19.465	27.468	5.108	0.307
ES	PML prof	2.722	15,628	0.900	1.707	11,477	19.461	27.491	5.109	0.306
FI	Direct	–	–	–	–	18,317	12.523	14.459	3.631	0.258
FI	NLS	1.078	12,092	11.079	7.198	18,317	12.523	14.459	3.632	0.257
FI	PML full	3.803	18,083	1.091	1.101	18,024	11.279	16.886	3.466	0.246
FI	PML prof	3.769	18,074	1.106	1.115	18,023	11.279	16.854	3.465	0.246
FR	Direct	–	–	–	–	16,197	13.050	18.361	3.936	0.272
FR	NLS	3.561	15,957	1.075	1.034	16,197	13.050	18.361	3.936	0.271
FR	PML full	4.000	16,251	0.900	0.911	16,179	12.947	18.817	3.894	0.269
FR	PML prof	3.991	16,248	0.903	0.914	16,178	12.946	18.806	3.893	0.269
GR	Direct	–	–	–	–	9,880	20.137	25.049	5.698	0.337
GR	NLS	1.270	11,471	3.395	4.037	9,880	20.137	25.049	5.698	0.337
GR	PML full	2.433	10,794	1.176	1.410	9,803	19.391	25.478	5.695	0.336
GR	PML prof	2.425	10,800	1.182	1.418	9,803	19.397	25.477	5.694	0.336
HU	Direct	–	–	–	–	3,854	15.650	23.309	5.165	0.327
HU	NLS	5.862	3,842	0.453	0.448	3,854	15.650	23.309	5.165	0.325
HU	PML full	6.163	3,906	0.424	0.446	3,841	15.538	23.560	4.951	0.315
HU	PML prof	6.283	3,906	0.414	0.436	3,841	15.526	23.617	4.957	0.315
IE	Direct	–	–	–	–	19,679	18.464	16.358	4.870	0.319
IE	NLS	0.714	5,356	21.948	8.870	19,679	17.688	16.584	5.051	0.321
IE	PML full	2.047	16,587	2.379	1.816	19,372	15.810	18.841	4.688	0.307
IE	PML prof	1.822	16,037	2.945	2.179	19,372	15.924	18.616	4.666	0.306

TABLE 3

GB2 Fitted Parameters and Indicators, Countries 14–26

Country	Type	a	b	p	q	Median	ARPR	RMPG	QSR	GINI
IS	Direct	–	–	–	–	28,015	9.540	18.480	3.578	0.257
IS	NLS	7.794	27,600	0.451	0.425	28,015	10.247	17.982	3.514	0.250
IS	PML full	8.162	27,573	0.436	0.406	28,065	9.949	17.764	3.470	0.248
IS	PML prof	8.283	27,566	0.429	0.399	28,063	9.938	17.791	3.472	0.248
IT	Direct	–	–	–	–	14,559	19.216	23.210	5.233	0.316
IT	NLS	0.632	17,728	14.071	15.893	14,559	19.214	23.211	5.234	0.322
IT	PML full	3.396	17,318	0.711	1.062	14,584	18.816	26.652	5.226	0.314
IT	PML prof	3.390	17,333	0.713	1.066	14,584	18.822	26.659	5.225	0.314
LT	Direct	–	–	–	–	2,536	19.927	28.852	6.163	0.347
LT	NLS	4.317	2,857	0.488	0.657	2,536	19.927	28.852	6.163	0.346
LT	PML full	2.883	2,942	0.807	1.077	2,552	20.717	28.369	6.336	0.352
LT	PML prof	2.946	2,926	0.786	1.041	2,551	20.679	28.366	6.349	0.353
LU	Direct	–	–	–	–	29,683	13.925	19.403	4.082	0.278
LU	NLS	3.428	29,996	1.054	1.082	29,683	13.925	19.403	4.082	0.277
LU	PML full	3.278	28,902	1.185	1.106	29,727	13.603	18.571	4.087	0.279
LU	PML prof	3.198	28,869	1.230	1.145	29,728	13.633	18.519	4.084	0.279
LV	Direct	–	–	–	–	2,546	22.731	24.315	7.303	0.386
LV	NLS	0.645	1,170	13.574	8.351	2,546	22.731	24.315	7.303	0.387
LV	PML full	2.521	2,763	0.931	1.076	2,551	22.039	29.206	7.502	0.388
LV	PML prof	2.468	2,770	0.959	1.111	2,551	22.074	29.195	7.485	0.387
NL	Direct	–	–	–	–	17,293	9.399	16.601	3.571	0.255
NL	NLS	7.586	16,367	0.508	0.409	17,293	9.399	16.601	3.571	0.257
NL	PML full	5.214	17,499	0.695	0.698	17,479	11.311	17.968	3.574	0.252
NL	PML prof	5.240	17,495	0.691	0.693	17,478	11.304	17.977	3.574	0.252
NO	Direct	–	–	–	–	27,806	11.001	18.117	3.967	0.280
NO	NLS	7.050	26,401	0.497	0.411	27,806	11.001	18.117	3.967	0.278
NO	PML full	10.552	28,955	0.288	0.346	27,770	11.414	20.424	3.411	0.238
NO	PML prof	10.270	28,953	0.297	0.358	27,751	11.393	20.353	3.403	0.238
PL	Direct	–	–	–	–	3,112	19.018	24.977	5.605	0.332
PL	NLS	2.539	3,359	1.140	1.322	3,112	19.018	24.977	5.605	0.334
PL	PML full	2.744	3,505	0.970	1.221	3,129	19.319	25.976	5.661	0.334
PL	PML prof	2.746	3,505	0.969	1.220	3,129	19.319	25.977	5.661	0.334
PT	Direct	–	–	–	–	7,311	18.466	23.468	6.726	0.377
PT	NLS	3.368	6,605	0.859	0.686	7,311	18.467	23.469	6.726	0.383
PT	PML full	4.443	6,858	0.569	0.481	7,339	18.422	24.905	7.170	0.396
PT	PML prof	4.362	6,861	0.582	0.492	7,342	18.467	24.887	7.151	0.396
SE	Direct	–	–	–	–	17,795	11.609	20.097	3.334	0.231
SE	NLS	7.747	19,003	0.401	0.522	17,795	11.609	20.097	3.334	0.233
SE	PML full	6.948	20,412	0.416	0.690	17,920	12.742	21.468	3.300	0.227
SE	PML prof	6.858	20,433	0.422	0.702	17,919	12.728	21.422	3.298	0.227
SI	Direct	–	–	–	–	9,316	11.677	18.539	3.388	0.238
SI	NLS	4.697	9,954	0.753	0.930	9,316	11.677	18.539	3.388	0.238
SI	PML full	4.342	10,220	0.817	1.070	9,360	11.919	18.682	3.377	0.237
SI	PML prof	4.336	10,221	0.819	1.072	9,360	11.920	18.678	3.377	0.237
SK	Direct	–	–	–	–	3,313	11.608	19.918	4.034	0.280
SK	NLS	7.139	3,260	0.448	0.422	3,313	12.018	19.643	4.001	0.276
SK	PML full	8.545	3,372	0.362	0.389	3,312	11.718	20.135	3.682	0.256
SK	PML prof	8.325	3,372	0.373	0.401	3,312	11.716	20.066	3.677	0.256
UK	Direct	–	–	–	–	19,375	18.976	22.395	5.208	0.320
UK	NLS	0.741	19,096	11.153	11.037	19,375	18.976	22.396	5.208	0.322
UK	PML full	2.803	22,495	0.976	1.329	19,412	18.517	25.260	5.176	0.316
UK	PML prof	2.758	22,487	1.001	1.359	19,406	18.516	25.195	5.173	0.316

GB2 is highly flexible due to its four parameters and the presented methodology could be easily extended to other contexts and other estimators, e.g. finance, insurance, biostatistics.

A maximum pseudo-likelihood approach of estimation of the parameters of the GB2 was considered and compared to a new method of "non-linear fit from the indicators" which tends to reconstruct the whole income distribution, knowing only the values of the empirical indicators. The pseudolikelihood methodology for the survey setting was already well known in the literature (see, e.g., Pfeffermann and Sverchkov, 2003). We have fitted the GB2 model using this approach in the special context of the EU-SILC survey, where households are sampled according to a country-dependent sampling design and all persons belonging to a household are included in the sample, thus the sample is clustered.

When fitting the GB2 to real income data and simulated data (AMELIA), we noticed that maximum pseudo-likelihood estimation sometimes produces a very large difference between the fitted and empirical estimates of ARPR and RMPG. We have developed an ad-hoc procedure for robustification of the sampling weights, which markedly improves the quality of the point estimates of the fitted indicators. In Tables 2 and 3 we have provided results on PML estimation for the case of the EU-SILC survey, 2006. In our simulation report (Graf and Nedyalkova, 2011b) we provide results based on estimation with and without using robustified weights. From our study, we have concluded that another advantage of the robustification of the sampling weights is that, in all cases, it significantly reduces the relative root mean squared error (RRMSE) of our estimates.

The contribution of this paper is to show that PML estimation can successfully be used to fit the GB2 model and estimate the derived indicators in a complex survey context. It also provides insight into the data. This approach offers the advantage of easy calculation of variance estimates through linearization of the parameters and the derived indicators. The proposed sandwich variance estimator accounts for clustering and we have seen in our simulation study with the AMELIA universe that the variance of the estimated indicators is smaller than the variance of the direct estimates. For single-stage sampling designs, we do also not require the whole design information (Graf and Nedyalkova, 2011b).

The novelty of the paper is the new method we have developed, i.e. the method of non-linear fit for indicators. We have seen that the five indicators of poverty and inequality (ARPR, RMPG, QSR, Gini, and median income) provide enough information about the underlying income distribution to permit the reconstruction of this distribution under the GB2 hypothesis. While variance estimation for this methodology has not yet been developed, this is possible if we have the variance–covariance matrix of the empirical indicators.

All described methods are programmed in the open source software R and are accessible through the GB2 R package (Graf and Nedyalkova, 2011a), which is part of the output of the AMELI project.

References

Alfons, A., M. Templ, P. Filzmoser, S. Kraft, B. Hulliger, J.-P. Kolb, and R. Münnich, "Synthetic Data Generation of SILC Data," Deliverable 6.2. of the Ameli project, 2011.

Atkinson, A. B. and F. Bourguignon (eds), *Handbook of Income Distribution*, North-Holland, Amsterdam, 2000.

Bordley, R. F., J. B. McDonald, and A. Mantrala, "Something New, Something Old: Parametric Models for the Size Distribution of Income," *Journal of Income Distribution*, 6(1), 91–103, 1996.

Brazauskas, V., "Fisher Information Matrix for the Feller-Pareto Distribution," *Statistics and Probability Letters*, 59(2), 159–67, 2002.

Butler, R. and J. B. McDonald, "Using Incomplete Moments to Measure Inequality," *Journal of Econometrics*, 42(1), 109–19, 1989.

Chambers, R. L., "Introduction to Part A," in R. L. Chambers and C. J. Skinner (eds), *Analysis of Survey Data*, John Wiley & Sons, Chichester, 13–27, 2003.

Chotikapanich, D. (ed.), *Modeling Income Distributions and Lorenz Curves*, Springer, New York, 2008.

Clémenceau, A. and J.-M. Museux, "EU-SILC: An EU Statistical Instrument Collecting Cross National Comparable Data on Income and Living Conditions and the Measure of Well Being," in *Perspectives of Improving Economic Welfare Measurement in a Changing Europe*, 34th CEIES Seminar, Helsinki, 2007.

Cowell, F. A., *Measuring Inequality*, 2nd edition, Harvester Wheatsheaf, Hemel Hempstead, 1995.

Cowell, F. A. and E. Flachaire, "Income Distribution and Inequality Measurement: The Problem of Extreme Values," *Journal of Econometrics*, 141(2), 1044–72, 2007.

Cowell, F. A. and M.-P. Victoria-Feser, "Robustness Properties of Inequality Measures," *Econometrica*, 64(1), 77–101, 1996.

Dastrup, S. R., R. Hartshorn, and J. B. McDonald, "The Impact of Taxes and Transfer Payments on the Distribution of Income: A Parametric Comparison," *Journal of Economic Inequality*, 5(3), 353–69, 2007.

Davison, A. C., *Statistical Models*, Cambridge University Press, Cambridge, 2003.

Eurostat, "Algorithms to Compute Overarching Indicators Based on EU-SILC and Adopted Under the Open Method of Coordination (OMC)," Technical Report, European Commission, Directorate F: Social Statistics and Information Society, 2009.

Freedman, D. A., "On the So-Called 'Huber Sandwich Estimator' and 'Robust Standard Errors'," *The American Statistician*, 60, 299–302, 2006.

Graf, M., "Use of Distributional Assumptions for the Comparison of Four Laeken Indicators on EU-SILC Data," in *56th Session of the ISI*, Lisbon, 2007.

———, "An Efficient Algorithm for the Computation of the Gini Coffcient of the Generalized Beta Distribution of the Second Kind, in *JSM Proceedings, Business and Economic Statistics Section*, American Statistical Association, Alexandria, VA, 4835–43, 2009.

Graf, M. and D. Nedyalkova, *GB2: Generalized Beta Distribution of the Second Kind: Properties, Likelihood, Estimation*, R Package Version 1.0, 2011a.

———, "Parametric Estimation of Income Distributions and Derived Indicators Using the GB2 Distribution," in B. Hulliger (ed.), *Report on the Simulation Results*, Deliverable 7.1 of the AMELI project, chapter 7.1, 2011b.

Huber, P. J., "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221–33, 1967.

———, *Robust Statistics*, John Wiley & Sons, New York, 1981.

Jenkins, S. P., "Distributionally-Sensitive Inequality Indices and the GB2 Income Distribution," *Review of Income and Wealth*, 55(2), 392–8, 2009.

Kleiber, C. and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, John Wiley & Sons, Hoboken, 2003.

Lumley, T., "Survey: Analysis of Complex Survey Samples," R Package Version 3.23–3, 2010.

Luzi, O., T. De Waal, and B. Hulliger, *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*, EDIMBUS Project, 2007.

McDonald, J. B., "Some Generalized Functions for the Size Distribution of Income," *Econometrica*, 52(3), 647–63, 1984.

McDonald, J. B. and M. Ransom, "The Generalized Beta Distribution as a Model for the Distribution of Income: Estimation and Related Measures of Inequality," in D. Chotikapanich, (ed.), *Modeling Income Distributions and Lorenz Curves*, Springer, New York, 147–66, 2008.

McDonald, J. B. and Y. J. Xu, "A Generalization of the Beta Distribution with Applications," *Journal of Econometrics*, 66(1–2), 133–52 (Erratum: *Journal of Econometrics*, 69, 427–8, 1995.

Osier, G., J. M. Museux, P. Seoane and V. Verma, "Cross-Sectional and Longitudinal Weighting for the EU-SILC Rotational Design," in *Methodology of Longitudinal Surveys*, Essex, UK, 2006.

Pfeffermann, D. and M. Yu Sverchkov, "Fitting Generalized Linear Model Under Informative Sampling," in R. L. Chambers and C. J. Skinner (eds), *Analysis of Survey Data*, John Wiley & Sons, New York, 175–95, 2003.

Prentice, R. L., "Discrimination Among Some Parametric Models," *Biometrika*, 62(3), 607–14, 1975.

R Development Core Team, *R: A Language and Environment for Statistical Computing*, http://www.R-project.org, R Foundation for Statistical Computing, Vienna, 2008.

Särndal, C.-E., B. Swensson, and J. H. Wretman, *Model Assisted Survey Sampling*, Spinger, New York, 1992.

Sepanski, J. H. and L. Kong, "A Family of Generalized Beta Distributions For Income," *Advances and Applications in Statistics*, 10(1), 75–84, 2008.

Skinner, C. J., D. Holt, and T. M. F. Smith (eds), *Analysis of Complex Surveys*, John Wiley & Sons, New York, 1989.

Victoria-Feser, M.-P., "Robust Methods for the Analysis of Income Distribution, Inequality and Poverty," *International Statistical Review*, 68(3), 277–93, 2000.

Victoria-Feser, M.-P. and H. Ronchetti, "Robust Methods for Personal-Income Distribution Models," *Canadian Journal of Statistics*, 22(2), 247–58, 1994.

White, H., "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(4), 817–38, 1980.

———, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1), 1–25, 1982.