EUROPEAN COMMISSION
Community Research

SEVENTH FRAMEWORK PROGRAMME

AMELI
Advanced Methodology for European Laeken Indicators

Deliverable 2.1

Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion

Version: 2011

Monique Graf, Desislava Nedyalkova,
Ralf Münnich, Jan Seger and Stefan Zins

The project **FP7–SSH–2007–217322 AMELI** is supported by European Commission funding from the Seventh Framework Programme for Research.

http://ameli.surveystatistics.net/

Contributors to Deliverable 2.1

Chapter 1: Monique Graf, Swiss Federal Statistical Office.

Chapter 2: Monique Graf, Swiss Federal Statistical Office.

Chapter 3: Monique Graf and Desislava Nedyalkova, Swiss Federal Statistical Office.

Chapter 4: Monique Graf and Desislava Nedyalkova, Swiss Federal Statistical Office.

Chapter 5: Monique Graf and Desislava Nedyalkova, Swiss Federal Statistical Office.

Chapter 6: Ralf Münnich and Jan Seger and Stefan Zins, University of Trier.

Chapter 7: Monique Graf and Desislava Nedyalkova, Swiss Federal Statistical Office; Jan Seger, University of Trier.

Main responsibility

Monique Graf and Desislava Nedyalkova, Swiss Federal Statistical Office.

Evaluators

Internal expert: Matthias Templ, Vienna University of Technology.

Aim and objectives of Deliverable 2.1

Present a state-of-the-art of the literature in parametric income distribution, which justifies the selection of an income distribution model that fits satisfyingly the equivalized income used in the EU-SILC survey and is, at the same time, easily applicable. From this study, it has become clear that a four-parameter size distribution called the Generalized Beta of the second kind (GB2) has been found as the best fitting distribution of income. Further we present and investigate the GB2 distribution and its properties and test different fitting methods for the GB2 (ML, Dagum's method of nonlinear regression on quantiles, moments) at the EU-SILC country level. The most promising methods are programmed and provide an input for simulation and robustness studies. Another goal is to get insight into the relationships between the characteristics of the theoretical distribution and a set of indicators, e.g. by sensitivity plots and to develop reliable variance estimation techniques for the fitted parameters and indicators. Also the use the mixture property of the GB2 distribution for fitting subgroup distributions by calibration and deduce by this method the subgroup indicators' estimates is investigated

Contents

Chapter 1

Introduction

In the context of the AMELI project, we aim at developing reliable and efficient methodologies for the estimation of a certain set of indicators of poverty and social exclusion computed within the EU-SILC survey, and in particular on the use of parametric estimation of the median, the at-risk-of-poverty rate (ARPR), the relative median poverty gap (RMPG), the quintile share ratio (QSR) and the Gini index (see e.g. Eurostat, 2009). This document investigates the use of parametric estimation in this context.

If we have income data, we can fit the theoretical distribution and compute the indicators from the parameters of the fitted 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.

Parametric income distributions have long been used for modeling income. The advantage of parametric estimation of income distributions is that there exist simple and explicit formulas for the inequality measures as functions of the parameters of the income distribution. Both modeling of the whole income range or of the tails of the distribution have been investigated in the literature.

Suppose we do not have the income micro data at disposal, but that the indicators, fitted on empirical data, are publicly available. 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 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 precision.

The deliverable is structured in the following way. Chapter 2 gives an overview of parametric estimation of income distributions. In Chapter 3, are given the basic properties of the generalized Beta distribution of the second kind. Chapter 4 gives a description of the methods used for fitting the GB2, both using the whole microdata information, or the set of empirical indicators only. Chapter 5 shows how the compounding property of the GB2 can be used to decompose the distribution and uses this model on subpopulations.

Chapter 6 is on the application of mixture distributions in the context of heterogeneous populations. Finally, Chapter 7 gives some conclusions.

Chapter 2

Review of Parametric Estimation in Income Distributions

2.1 Introduction

This bibliography collects seminal and recent papers in the field of parametric income distributions. The domain is so large and vivid that the collection is necessarily incomplete. The bibliography within each of the following references will give further information. We proceed by themes.
First publications on the mathematical properties of models for income distributions are described. They are followed by papers on international comparisons. Then some estimation procedures used in different contexts are reviewed. Next the Gini coefficient has given rise to much research. In particular the case of Gini with reference to negative incomes is considered. Finally we present the important subgroup decomposition of inequality indices in different contexts.

2.2 Statistical size distributions

Three books on income distributions and inequality indices are of great value:
Kleiber and Kotz (2003) is a reference book on statistical size distributions. It contains a encyclopedic bibliography on the derivation of the different types of distributions as well as on empirical applications. One huge difficulty that is overcome with the help of Kleiber and Kotz's book is the terminology they have unified and clarified. We propose to follow their terminological choices.

In the book of Chotikapanich (2008), seminal papers on size distributions and Lorenz curve are collected.

Atkinson and Bourguignon (2000) describe income distributions from a more econometric point of view. The book starts with a review of existing economic theories seeking to explain the distribution of income. Chap.1: relation between the idea of social justice and the analysis of income distribution(A.Sen); Chap.2: basis for comparing different

distributions and measuring inequality (F. Cowell); Chap. 3 and 4: historical perspectives; Chap. 5: empirical evidence on income inequality in industrialized countries (P. Gottschalk and T. Smeeding); Chap.6: income poverty in advanced countries, definitions of poverty and equivalence scales(M. Jäntti and S. Danziger); Chap. 7: theories of the distribution of earnings (D. Neal and S. Rosen); Chap. 14: income distribution, economic systems and transition (J. Flemming and J. Mickelwright). The rest of the book is of a less measurement nature (i.e. purely economic). This book is also relevant for WP1.

One prevailing family of income distributions is the Generalized Beta distribution of the Second Kind (GB2). Some recent papers about the GB2 are cited now:
McDonald (1984) gives a unified view of many income distributions, utilizing the generalized beta and gamma distributions family. This paper is the basis of Kleiber and Kotz (2003)'s chapter on the GB2.

Jenkins (2007) derives the generalized entropy class of inequality indices for the GB2 income distributions, thereby providing a full range of top-sensitive and bottom-sensitive measures. An examination of British income inequality in 1994/95 and 2004/05 illustrates the analysis. Jenkins (2008) is essentially the same paper.

Milgram (2006) is an electronic paper on the generalized hypergeometric function ${}_3F_2(1)$ (that appears in the Gini formula for GB2).

2.3 Estimation methods

Burkhauser et al. (2008) estimate trends in US income inequality with special emphasis on top income shares. On comparing with estimates from administrative data, they conclude that the trend is linked to the top-coding (for confidentiality reasons) of the CPS data. They show that their CPS estimates of trends in top income shares match the estimates of trends reported on the basis of administrative records, except for within the top 1% of the distribution. Thus, they argue that, if income inequality in the USA has increased substantially since 1993, such increases are confined to this very highest income group.

In the proceedings of the EU-SILC conference Eurostat (2007), Van Kerm (2007) considers extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. Social indicators are known to be sensitive to the presence of extreme incomes at either tail of the income distribution. It is therefore customary to make adjustments to extreme data before estimating such statistics. Thus it is important to evaluate the impact of such adjustments and assess how much resulting cross-country comparisons are affected by alternative adjustments. The paper presents the results of a large scale sensitivity analysis considering both simple, classical adjustments and a more sophisticated approach based on modeling parametrically the tails of the income distribution. A Pareto distribution was used as the parametric tail model. An inverse Pareto distribution was used for the lower tail.

In Biewen and Jenkins (2005), the decomposition of poverty differences is based on a parametric model of the income distribution and can be used to decompose differences in poverty rates across countries or years. The parameters of the GB2 family are modeled

with the help of covariates to account for population differences. The authors encountered sometimes convergence problems.

In the context of capital asset pricing model, McDonald (1989) estimates regression coefficient using partially adaptive techniques and a generalized t (GT) distribution for the error term. The idea is put further to any type of regression with positive variables in Butler et al. (1990) and McDonald and Butler (1990).

In Neocleous and Portnoy (2008), the partially linear Censored Regression Quantile (CRQ) model combines semiparametric estimation for censored data with quantile regression techniques, and uses B-splines for the estimation of the nonlinear term. An application to administrative unemployment data from the German Socio-Economic Panel Survey is presented. In a very interesting paper, McDonald and Butler (1987) apply generalized mixture distributions to unemployment duration.

Yu et al. (2004) present wage distributions via bayesian quantile regression.

Victoria-Feser and Ronchetti (1994) show that classical estimation methods are very sensitive to model deviations and set the scene for the optimal B-robust estimation (OBRE) in income distribution analysis for Gamma and Pareto models.

Victoria-Feser (2000) shows that robust techniques can play a useful role in income distribution analysis and should be used in conjunction with classical methods. The data available for estimating welfare indicators are often incomplete: they may be censored or truncated. Furthermore, for robustness reasons, researchers sometimes use trimmed samples. Cowell and Victoria-Feser (2003) derive distribution-free asymptotic variances for wide classes of welfare indicators not only in the complete data case, but also in the important cases where the data have been trimmed, censored or truncated.

2.4 The Gini coefficient

A huge literature exists on the Gini coefficient, and we do not pretend to be exhaustive. One interesting reference is Xu (2004)'s survey paper. Its aim is to help the reader to navigate through the major developments of the literature and to incorporate recent theoretical research results with a particular focus on different formulations and interpretations of the Gini index, its social welfare implication, and source or subgroup decomposition. One interesting question is the comparability of Gini indices between distributions without negative incomes and distributions that have some negative incomes. Chen et al. (1982) propose a normalized Gini coefficient that deals with the issue. Berberi and Silber (1985) point out a mistake in Chen and Saur's paper and propose an alternative formulation that is in turn criticised by Chen et al. (1985).

2.5 Subgroup decomposition

Another line of research is the decomposition of inequality measures, either by sub-groups or by source of income. The case of Gini and entropy is considered in Mussard and Terraza (2007) for both types of decomposition. The decomposition of inequality measures

by sub-groups is a subject of continuous interest. Dagum et al. (1984) compare male-female income distribution on the basis of an economic distance which is a normalized and dimensionless measure of inequality between distributions.

Chiappero-Martinetti and Civardi (2006) propose a decomposition of the Foster, Greer, Thorbecke (FGT) class of poverty indexes into two additive components (namely, poverty within groups and poverty between groups) when both a community-wide threshold and a specific poverty line for each subgroup of population is used. For any given order of stochastic dominance, Makdissi and Mussard (2006) decompose standard concentration curves into contribution curves corresponding to within-group inequalities, between-group inequalities, and transvariational inequalities. The latter gauges between-group inequalities issued from the groups with lower mean incomes and thus brings out the intensity with which the groups are polarized.

Mussard (2007) first introduces between-group and within-group transfers, then axiomatically derives Gini's mean difference (Gini (1912)) and Dagum's Gini index between two populations (Dagum (1987)). An application is performed with the Gini decomposition in order to understand the impact of within- and between-group transfers on the variations of the overall Gini index. A conclusion follows to highlight the debate between the use of entropy and Gini measures throughout the prism of decomposition techniques.

Dastrup et al. (2007) extend the analysis using the generalized beta distributions to include the impact of transfer payments and taxes on the distribution of income.

The paper by Lilla (2007) attempts to measure income inequality and its changes over the period 1993-2000 for a set of 13 Countries in ECHP. Focusing on wages and incomes of workers in general, inequality is mainly analyzed with respect to educational levels as proxy of individual abilities. Estimation of education premia is performed by quantile regressions to stress differences in income distribution and questioning the true impact of education. The same estimates are used to decompose income inequality and show the rise in residual inequality.

Chapter 3

Basic Properties of the Generalized Beta Distribution of the Second Kind

3.1 Introduction

The Generalized Beta Distribution of the Second Kind is a four-parameter distribution and is denoted by GB2(a, b, p, q). It has been derived by McDonald (1984). Most of the following formulas are collected in Kleiber and Kotz (2003). The GB2 distribution encompasses Fisk's ($p = q = 1$), Dagum's ($q = 1$) and Singh - Maddala's ($p = 1$) distributions. Empirical studies on income (see e.g. Jenkins, 2007; Dastrup et al., 2007; Kleiber and Kotz, 2003, Table B2), tend to show that the GB2 outperforms other 4-parameter distributions.

The GB2 can be obtained by a transformation of a standard Beta random variable. The derivation of moments and likelihood equations also necessitates the use of special mathematical functions, like the beta function and the gamma function and its derivatives. The Fisher information matrix has been obtained by Brazauskas (2002). Formulas for the indicators are new, except for the Gini index that was derived by McDonald (1984). An efficient method for the computation of the Gini index is described in Graf (2009). The paper is given in Annex C of this document.

3.2 Density and distribution function

The GB2 density takes the form:

$$f(x; a, b, p, q) = \frac{a}{bB(p, q)} \frac{(x/b)^{ap-1}}{(1 + (x/b)^a)^{p+q}}, \tag{3.1}$$

where $B(p, q)$ is the beta function, $b > 0$ is a scale parameter, $p > 0, q > 0$ and $a > 0$ are shape parameters. The parameter a represents the overall shape, p governs the left tail and q - the right tale.

Let $I_z(p,q)$ be the incomplete beta function ratio, given by

$$I_z(p,q) = \frac{1}{B(p,q)} \int_0^z u^{p-1}(1-u)^{q-1} du, \qquad 0 \le z \le 1 \tag{3.2}$$

The distribution function of a GB2 variable can be written as

$$F(x;a,b,p,q) = I_z(p,q), \tag{3.3}$$

where $z = y/(1+y)$ and $y = (x/b)^a$.

Let z_α be the $\alpha-$th quantile of the Beta(p,q) distribution, and $y_\alpha = z_\alpha/(1-z_\alpha)$, then the $\alpha-$th quantile of the GB2(a,b,p,q) is given by:

$$x_\alpha = b\, y_\alpha^{1/a}. \tag{3.4}$$

If Z is a random variate of a standard Beta(p,q) distribution, and $Y = Z/(1-Z)$, then the GB2(a,b,p,q) random variate is given by

$$X = b\, Y^{1/a}. \tag{3.5}$$

3.3 Log density

If we set $y = (x/b)^a$, then $\partial y/\partial a = (1/a)y\log(y)$ and $\partial y/\partial b = (-a/b)y$. Then the log density of the GB2 distribution is given by:

$$\begin{aligned} \log(f) &= \log(a) - \log(b) - \log(\Gamma(p)) - \log(\Gamma(q)) + \log(\Gamma(p+q)) \\ &+ (ap-1)\log(x/b) - (p+q)\log(1+y), \end{aligned}$$

where $\log(f)$ stands for $\ln(f)$ and Γ is the gamma function.

Let us set $r = p/(p+q)$ and $s = p+q$. Then $p = rs$ and $q = (1-r)s$. Then, we obtain a new formula for the log density, given by

$$\begin{aligned} \log f &= \log a - \log b - \log\Gamma(rs) - \log\Gamma((1-r)s) + \log\Gamma(s) \\ &+ (ars-1)\log(x/b) - s\log(1+y) \end{aligned}$$

The partial derivatives of the log density with respect to a, b, r and s are given by

$$\begin{aligned} \frac{\partial \log f}{\partial a} &= \frac{1}{a} + rs\log(x/b) - s\frac{(1/a)y\log(y)}{1+y} = \frac{1}{a}\left[1 + rs\log(y) - s\frac{y\log(y)}{1+y}\right] \\ \frac{\partial \log f}{\partial b} &= -\frac{1}{b} - \frac{ars-1}{b} + \frac{as}{b}\frac{y}{1+y} = \frac{as}{b}\left[\frac{y}{1+y} - r\right] \\ \frac{\partial \log f}{\partial r} &= -s\psi(rs) + s\psi((1-r)s) + s\log(y) = s\left[\log(y) - \psi(rs) + \psi((1-r)s)\right] \\ \frac{\partial \log f}{\partial s} &= -r\psi(rs) - (1-r)\psi((1-r)s) + \psi(s) + r\log(y) - \log(1+y) \end{aligned}$$

In Appendix A are given the first and second partial derivatives of the log density with respect to a, b, p and q.

3.4 GB2 Log-likelihood Equations

We express the log-likelihood as a weighted mean of the log density evaluated at the data points.

$$\log L = \sum w_i \log f(x_i; a, b, rs, (1-r)s) / \sum w_i,$$

where $f(.)$ is the GB2 density in Equation (3.1). Next, we can calculate the score functions, which are readily obtained as weighted sums of the partial derivatives of $\log f$ evaluated at data points.

It is easy to solve the ML equations for r and s in function of a and b: $\partial \log L / \partial b = 0 <=>$

$$\hat{r} = \sum w_i \frac{y_i}{1+y_i} / \sum w_i \tag{3.6}$$

$\partial \log L / \partial a = 0 <=>$

$$\begin{aligned} \hat{s}^{-1} &= \sum w_i \log(y_i) \left(\frac{y_i}{1+y_i} - \hat{r} \right) / \sum w_i & (3.7) \\ &= \sum w_i \frac{y_i}{1+y_i} (\log(y_i) - m) / \sum w_i, & (3.8) \end{aligned}$$

where

$$m = \sum w_i \log(y_i) / \sum w_i. \tag{3.9}$$

We see that $\hat{s}^{-1}$ is the empirical covariance between $\log y_i$ and $y_i/(1+y_i)$.

Introducing these solutions into the likelihood leads to the profile log-likelihood $\log L_p$ which has two parameters a and b,

$$\log L_p = \sum w_i \log f(x_i; a, b, \hat{r}\hat{s}, (1-\hat{r})\hat{s}) / \sum w_i \tag{3.10}$$

The advantage over the full log-likelihood is that contour plots can be produced (see Figure 4.3).

3.5 Moments and other properties

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

$$\mathrm{E}(X^k) = b^k \frac{\Gamma(p + k/a)\Gamma(q - k/a)}{\Gamma(p)\Gamma(q)}, \tag{3.11}$$

where moments exist for $-ap < k < aq$.

The incomplete moment of order k is given by

$$\frac{\mathrm{E}(X^k|X<x)}{\mathrm{E}(X^k)} = F_{(k)}(x;a,b,p,q) = F(x;a,b,p+\frac{k}{a},q-\frac{k}{a}). \tag{3.12}$$

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

Equation (3.11) can be viewed as the moment generating function of $\log(X)$. Thus the moments of $\log(X)$ can be easily obtained by differentiation. Let denote by ψ the digamma function (the logarithmic derivative of the gamma function). The polygamma function of order n, $\psi^{(n)}$, is the n-th order derivative of the digamma function. The expectation, variance, skewness and kurtosis coefficients of $\log(X)$ are given by:

$$\begin{aligned}
\mathrm{E}(\log X) &= (\psi(p)-\psi(q))/a + \log(b) \\
\mathrm{Var}(\log X) &= (\psi^{(1)}(p)+\psi^{(1)}(q))/a^2 \\
\mathrm{Skew}(\log X) &= (\psi^{(2)}(p)-\psi^{(2)}(q))/(\psi^{(1)}(p)+\psi^{(1)}(q))^{3/2} \\
\mathrm{Kurt}(\log X) &= (\psi^{(3)}(p)+\psi^{(3)}(q))/(\psi^{(1)}(p)+\psi^{(1)}(q))^{2}
\end{aligned}$$

The four parameters have a direct interpretation in terms of the distribution of $\log(X)$. The location parameter is $\log(b)$, a is the scale parameter, and p and q determine the asymmetry and the skewness of the distribution. One can easily prove that when $p=q$, all odd moments vanish (except the first), thus the distribution of $\log X$ is symmetric around $\log(b)$ in this case; in general, it is skewed to the right if $p>q$, and to the left if $p<q$. Let us also remark that, contrary to X, $\log(X)$ has moments of all orders.

3.6 Indicators of poverty and social exclusion in the EU-SILC framework

The advantage of parametric estimation of income distributions, and in particular the GB2, is that there exist 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. An efficient algorithm to compute the Gini index from its analytical expression has been described in GRAF (2009), see Annex C, and implemented in R.

The following inequality measures are defined in EUROSTAT (2009). Robust methods for the direct estimates are addressed in Deliverable 4.2. The implementation in EU-SILC is described in Deliverable 5.1. Here we derive the indicators under the GB2 hypothesis.

- *At-risk-of-poverty threshold (ARPT)*

  Let x_{50} be the median of the GB2(a,b,p,q), computed from Equation (3.4) with $\alpha = 50\%$. Then $ARPT$ is given by

  $$ARPT(a,b,p,q) = 0.6\, x_{50} \tag{3.13}$$

- *At-risk-of-poverty rate (ARPR)*

  The at-risk-of-poverty rate being scale-free, b can be chosen arbitrarily and can be fixed to the value of 1.

  $$ARPR(a,p,q) = F(ARPT(a,1,p,q);a,1,p,q), \tag{3.14}$$

  where F is the GB2 distribution function given in Equation (3.3).

- *Relative median poverty gap*

  RMPG is defined as one minus the ratio between the median income of the poor to 60% of the median income of the population.
  If $A = ARPR(a,p,q)$,

  $$RMPG(A,a,p,q) = 1 - qGB2(A/2,a,1,p,q)/qGB2(A,a,1,p,q), \tag{3.15}$$

  where $qGB2$ is the GB2 quantile function.

- *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 (see Equation (3.4)). The quintile share ratio can be expressed with the help of the incomplete moments of order 1 (Equation 3.12, with $k = 1$):

  $$QSR(a,p,q) = (\,1 - F_{(1)}(x_{80};a,1,p,q)\,)/F_{(1)}(x_{20};a,1,p,q) \tag{3.16}$$

- *Gini index*

  The Gini index of the GB2 distribution is given by (McDonald, 1984):

  $$GINI(a,p,q) = \frac{B(2p+1/a,2q-1/a)}{B(p,q)B(p+1/a,q-1/a)}\left\{\frac{1}{p}G_1 - \frac{1}{p+1/a}G_2\right\}, \tag{3.17}$$

  where

  $$G_1 = {}_3F_2\left[\begin{array}{lll} 1, & p+q, & 2p+1/a \\ & p+1, & 2(p+q) \end{array};1\right] \tag{3.18}$$

  and

  $$G_2 = {}_3F_2\left[\begin{array}{lll} 1, & p+q, & 2p+1/a \\ & p+1+1/a, & 2(p+q) \end{array};1\right], \tag{3.19}$$

  where ${}_3F_2$ is the generalized hypergeometric series. A direct application of Equation (3.17) can lead to convergence problems.

- *Gini: Particular cases*

  In some special cases, the Gini takes a simpler form:

  B2 distribution ($a = 1$):

  $$GINI(p,q) = \frac{B(2p,2q-1)}{2pB^2(p,q)} \tag{3.20}$$

  Dagum distribution ($q = 1$):

  $$GINI(a,p) = \frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(2p)\Gamma(p+1/a)} - 1 \tag{3.21}$$

  Singh-Maddalah distribution ($p = 1$):

  $$GINI(a,q) = 1 - \frac{\Gamma(q)\Gamma(2q-1/a)}{\Gamma(2q)\Gamma(q-1/a)} \tag{3.22}$$

3.7 Sensitivity plots

As ARPR, RMPG, QSR and Gini do not depend on the scale parameter b, we can ask ourselves how do these indicators behave in function of the shape parameters a, p and q. A sensitivity plot, implemented in the R package GB2 GRAF and NEDYALKOVA (2010), illustrates this.

Figure 3.1 shows how the values of ARPR vary in function of 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.

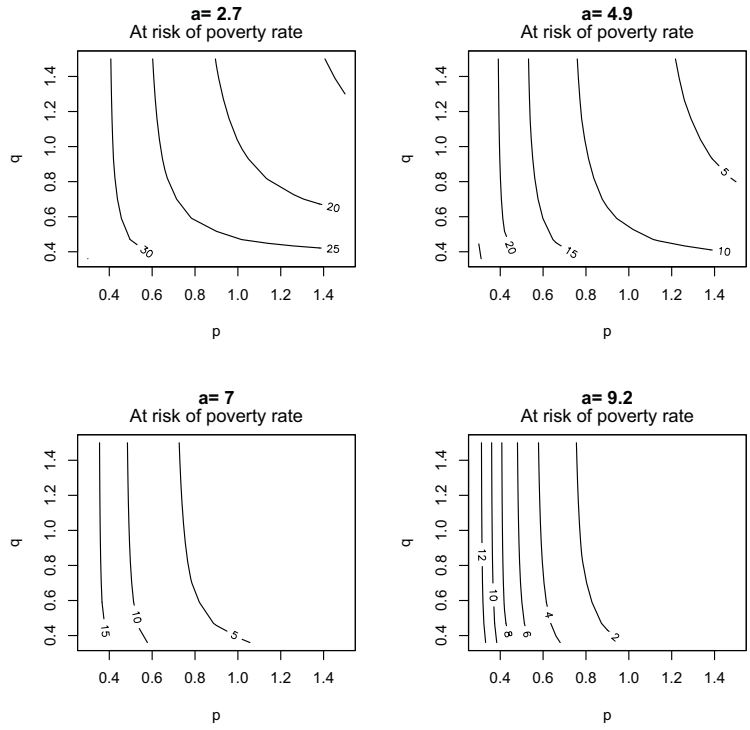

Figure 3.1: Sensitivity plot of the ARPR

Sensitivity plots can also be produced for RMPG, QSR and Gini.

Chapter 4

Methods of estimation of the parameters of the GB2

In this section we consider several methods of estimation of the GB2 parameters a, b, p and q. Amongst them, the pseudo maximum likelihood, nonlinear least squares on the quantile function (Dagum (1977)), nonlinear fit for indicators. In our experience, the pseudo maximum likelihood estimation has proven to be the most suitable, giving the best fit of the distribution and allowing for easy calculation of variance estimates (by linearization) of the fitted parameters and indicators. Variance estimation takes the sampling design into account. The pseudo log-likelihood is computed as a weighted sum over the sample of the log density of the distribution, where the weights are the sample weights. It is a function of the parameters of the distribution. Optimizing the pseudo-likelihood provides us with a set of parameters which fits the GB2 to the income variable by taking the sampling design into consideration.

4.1 Dagum's Method

Let $\hat{F}(x)$ be the empirical distribution function estimated at x and $F_{GB2}(x; a, b, p, q)$ the GB2 cumulative distribution function.
Dagum's method (Dagum (1977)) consists of finding a, b, p, q that minimise the following objective function:

$$\sum w_i \left[\hat{F}(x_i) - F_{GB2}(x_i; a, b, p, q)\right]^2, \tag{4.1}$$

where w_i is the sampling weight of x_i.

We start with initial values from the Fisk distribution, which is GB2 with $p = q = 1$. Moment estimators of a and b for this distribution are, (see Graf, 2007):

$$m_\ell = \sum w_i \log x_i / \sum w_i \tag{4.2}$$
$$v_\ell = \sum w_i (\log x_i - m_\ell)^2 / \sum w_i$$
$$\hat{b} = \exp(m_\ell) \tag{4.3}$$
$$\hat{a} = \pi / \sqrt{3 v_\ell} \tag{4.4}$$

4.2 Pseudo maximum likelihood estimation of the parameters of the GB2 distribution under cluster sampling

In the classical case of maximum likelihood estimation the log-likelihood function is defined as a sum over the sample of the log density evaluated at the data points. However, in the framework of EU-SILC, we are in the case where the data is observed at two levels - personal level and household level. Households (clusters) are sampled and then all persons of the selected households enter in the sample. All persons of a household have the same equivalised disposable income (x_i), which is also the household's equivalised disposable income, thus the observations are not independent. Let m, n_i and n denote, respectively, the number of selected households, the number of persons belonging to household i and the number of selected persons. Then, the weighted pseudo log-likelihood function (see e.g. SKINNER et al., 1989, Chapter 3.4.4), at the household level, is defined as

$$l_m(\theta) = \sum_{i=1}^{m} w_i n_i \log f(x_i;\theta), \tag{4.5}$$

where $f(\cdot)$ is the GB2 density, given in Equation (3.1), $\theta = (a, b, p, q)^T$ is the vector of parameters and w_i are the sampling weights (the sampling weight of a household equals the sampling weight of each person belonging to the household). We can scale $l_m(\theta)$ by dividing by the mean of weights over the sample of households $\bar{w}_m = \sum_{i=1}^{m} w_i/m$ in order to avoid large numerical values in the computation.

The partial derivatives of the log-likelihood function are readily obtained as weighted sums of the partial derivatives of $\log(f(x_i))$ (see Section 3.3), evaluated at the data points. Thus, the first and second partial derivatives of l_m with respect to θ are:

$$l'_m(\theta) = \sum_{i=1}^{m} w_i n_i u_i(x_i;\theta), \tag{4.6}$$

where

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

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

Similarly, we have

$$l''_m(\theta) = \sum_{i=1}^{m} w_i n_i h_i(x_i;\theta), \tag{4.7}$$

where

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

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

The quantity

$$I(\theta) = -\,\mathrm{E}(l''_m(\theta)).$$

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

In classical maximum likelihood theory, when the assumed model is correct, it can be proved that

$$\mathrm{E}(l'_m(\theta)) = 0 \tag{4.8}$$
$$\mathrm{Var}(l'_m(\theta)) = -\,\mathrm{E}(l''_m(\theta)) \tag{4.9}$$

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

$$l'_m(\hat{\theta}_m) = 0. \tag{4.10}$$

Functions performing pseudo maximum likelihood estimation based on the full and the profile log-likelihoods are implemented in Graf and Nedyalkova (2010). Maximum likelihood estimation is obtained through methods for non-linear optimization like the BFGS method. As for Dagum's method, the same initial values for a and b, given in Equations (4.4) and (4.3), are chosen.

4.3 Robustification of the sampling weights

In general, GB2 estimation and other ML estimation from parametric distributions have robustness problems and are sensitive to extremes (see e.g. Victoria-Feser and Ronchetti, 1994; Victoria-Feser, 2000). Actions have been taken by the SILC data producers in order to avoid very large incomes in the databases, but less attention has been given to the left tail of the income distribution. In our simulation study (see Chapter 7 of Deliverable D7.1), we have noticed that a certain bias in the estimates is induced. This led us to the idea to robustify the sampling weights in creating an ad hoc procedure for correcting the sampling weights. Our procedure is inspired, but not following directly, by 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:

$$F(x; a, b, 1, 1) = \frac{(x/b)^a}{1 + (x/b)^a} = \frac{y}{1+y}, \tag{4.11}$$

where $y = (x/b)^a$.

The $\alpha-$th quantile of the Fisk(a, b) is given by:

$$x_\alpha = b\left(\frac{\alpha}{1-\alpha}\right)^{1/a}. \tag{4.12}$$

From Equation 4.12, it follows that:

$$\frac{x_\alpha}{b}\frac{x_{1-\alpha}}{b} = 1. \tag{4.13}$$

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:

$$d = |\frac{x_\alpha}{b} - \frac{x_{1-\alpha}}{b}|, \tag{4.14}$$

   where α can take different values, e.g. $0.001, 0.002$, etc.

2. Next, the correction factor is calculated as follows:

$$corr = \max(c, \min(1, \frac{d}{|b/x - 1|}, \frac{d}{|x/b - 1|})), \tag{4.15}$$

   where c is a constant, that can take different values, e.g. $0.1, 0.2$, etc. and that allows 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

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

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 data set (created by Kolb et al., 2011), if this adjustment is processed, we downweight about 0.2% of the observations, essentially on the left tail. Figure 4.1 shows the correction of the weights obtained with $a = 1.78$ and $\alpha = 0.01$ (which implies that $d \approx 13$), and $c = 0.1$. These parameters are similar to those used with the AMELIA dataset.

4.4 Variance estimation

We fit the GB2 by pseudo-maximum likelihood and derive the design variance of both the parameters and the indicators by linearization. Simulations with the AMELIA artificial dataset show a bias of the linearization variances relative to the simulation variance of around 10% (see Chapter 7 of Deliverable D7.1).

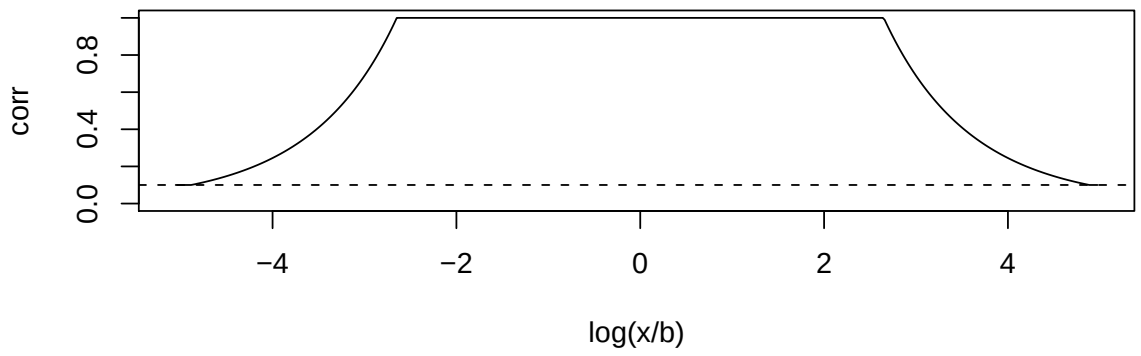

Figure 4.1: Correction factor for the robustification of weights (Huber-type function). Dotted line corresponds to limit c.

4.4.1 Variance estimation of the parameters of the GB2 distribution

We can approximate $l'_m(\hat{\theta}_m)$ by the first two terms of a Taylor series around θ. Thus we have

$$l'_m(\hat{\theta}_m) \approx l'_m(\theta) + l''_m(\theta)(\hat{\theta}_m - \theta),$$
$$\hat{\theta}_m - \theta \approx [-l''_m(\theta)]^{-1} l'_m(\theta).$$

Then

$$\mathrm{Var}(\hat{\theta}_m) = \mathrm{E}(\hat{\theta}_m - \theta)^2 \approx [l''_m(\theta)]^{-1} V(\theta) [-l''_m(\theta)]^{-1}, \tag{4.16}$$

where

$$V(\theta) = \mathrm{Var}(l'_m(\theta)),$$
$$= \mathrm{E}((l'_m(\theta))(l'_m(\theta))').$$

This formula leads to the so called sandwich variance estimator (Freedman (2006); Huber (1967); Pfeffermann and Sverchkov (2003)):

$$\hat{\mathrm{Var}}(\hat{\theta}_m) \approx [l''_m(\hat{\theta}_m)]^{-1} \hat{V}(\hat{\theta}_m) [l''_m(\hat{\theta}_m)]^{-1}, \tag{4.17}$$

where $l''_m(\theta)$ and $V(\theta)$ are estimated directly from the sample. Thus we have

$$l''_m(\hat{\theta}_m) = \sum_{i=1}^{m} w_i n_i h_i(x_i; \hat{\theta}_m) \tag{4.18}$$

and $\hat{V}(\hat{\theta}_m)$ can be calculated in two different ways. If we do not consider the cluster effect, thus supposing that the persons are independently distributed within a household, then

the variance of $l'_m(\theta)$ is readily obtained as the sum of variances of the scores weighted at the personal level, so

$$\begin{aligned}\hat{V}(\hat{\theta}_m) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_i^2 u_i(x_i;\hat{\theta}_m)u_i(x_i;\hat{\theta}_m)' \\ &= \sum_{i=1}^{m} n_i w_i^2 u_i(x_i;\theta)u_i(x_i;\theta)'. \end{aligned} \tag{4.19}$$

In our case, we use a variance estimator which takes into account the cluster effect, in supposing that the households are independently (but not identically due to the different n_i) distributed. Thus we sum the squared sums at he household level of the weighted scores, i.e.

$$\begin{aligned}\hat{V}(\hat{\theta}_m) &= \sum_{i=1}^{m}\left(\sum_{j=1}^{n_i} w_i u_i(x_i;\hat{\theta}_m)\right)\left(\sum_{j=1}^{n_i} w_i u_i(x_i;\hat{\theta}_m)\right)' \\ &= \sum_{i=1}^{m} n_i^2 w_i^2 u_i(x_i;\hat{\theta}_m)u_i(x_i;\hat{\theta}_m)'. \end{aligned} \tag{4.20}$$

Note that, in the case of a correctly specified model, the variance of the MLE is given by the inverse of the Fisher information matrix $\left(I(\hat{\theta}_m)\right)^{-1}$.

We can also calculate the midterm of the sandwich variance estimator numerically, using the full design information, e.g. 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 implemented this in our simulation study with success. We have seen that our variance estimate by linearization is almost equal to the design variance calculated with the package survey for the one-stage sampling designs. Results and comments will be given in Deliverable of WP7 (see Chapter 7 of Deliverable D7.1).

4.4.2 Variance estimation of the aggregate indicators

Now we would like to estimate the variance of the estimated Laeken indicators, to construct confidence intervals and to compare with the empirical estimates of the indicators. We know that the median, ARPR, RMPG, QSR and Gini all can be expressed as functions of the GB2 parameters a, b, p and q (see Section 3.6) . 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 ML estimate of the ARPR, then by the delta method, we have:

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

$\hat{V}(\hat{\theta}_m)$ is given in Equation 4.20. 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.

4.5 Estimation of income data from a set of indicators

Suppose we do not have the income micro data at disposal, but that the indicators, fitted on empirical data, are publicly available. 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 value of the empirical indicators and assuming that the theoretical distribution models the empirical distribution to an acceptable precision. This approach has been applied to EU-SILC data with success. This means that the set of indicators contains quite a lot of information about the empirical distribution.

Consider a set of indicators A = (median, ARPR, RMPG, QSR, Gini) and their corresponding GB2 expressions $A_{\text{GB2}}(a, b, p, q)$. 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.

Instead of fitting the GB2 parameters all together, we can also 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 do not depend on the parameter b, thus we set $b = 1$ and their corresponding expressions are given in function of a, ap and aq. This is done in order to be able to bound the parameters ap and aq in the algorithm, so that the constraints for the existence of the moments of order at least 2 ($aq > 1$) and the existence of the excess for the calculation of the Gini ($ap > 1$) are fulfilled. The bounds for the parameter a can be defined in function of the coefficient of variation of the ML estimate of the parameter a.

- In the second step, only the parameter b is estimated, optimizing the weighted square difference between the empirical median and the GB2 median in function of the already obtained NLS estimates of the parameters a, p and q.

Initial values 1. Initial values for the parameters can be taken as the moment estimators of the Fisk distribution in Equations (4.3) and (4.4), and $p = q = 1$; 2. Alternatively, the initial value for b can be given by the empirical median, and for a by the inverse of the Gini coefficient, which is in accordance with the information the user is supposed to have, namely the set of indicators A; 3. If the ML estimates of the GB2 parameters are known, they give a third choice for the initial values.

4.6 Graphical representations and evaluation of the GB2 fit

In order to visualize the various results of fitting the GB2 distribution we present examples of different plots, programmed in R, for the case of the EU-SILC survey.

4.6.1 Distribution plots

- Cumulative distribution plot: presents the GB2 versus the empirical distribution function.
- Density plot: presents a kernel density estimate (Epanechnikov) of the income variable and the fitted GB2 density.

The Epanechnikov kernel is 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 4.2 shows an example of the GB2 fitted distribution by maximum likelihood estimation and the method of non-linear fit for indicators with the Austrian EU-SILC data, 2006. We can see that the fit by the pseudo maximum likelihood is better.

4.6.2 Contour plot of the profile log-likelihood

On Figure 4.3, we can see a contour plot of the profile log-likelihood for the Austrian EU-SILC sample, 2006. With **F**, **M** and **N** are given the Fisk, ML and NLS estimates of the parameters a and b, respectively. The value of the log-likelihood at these points can be read on the plot. We can see that the value of the estimated maximum log-likelihood (**M**) is close to the small quadrangle on the figure, which is the graphical representation of the maximum value of the log-likelihood. The values of the parameters and the log-likelihood are given in Table 4.1. We can also notice that the profile log-likelihood is really flat.

	a	b	log-likelihood
F	3.45	17494	-10.44958
M	5.89	19410	-10.43806
N	1.56	15039	-10.48561
graphical ML	5.91	19410	-10.42302

Table 4.1: Log-likelihood and parameter values corresponding to the points depicted in Figure 4.3

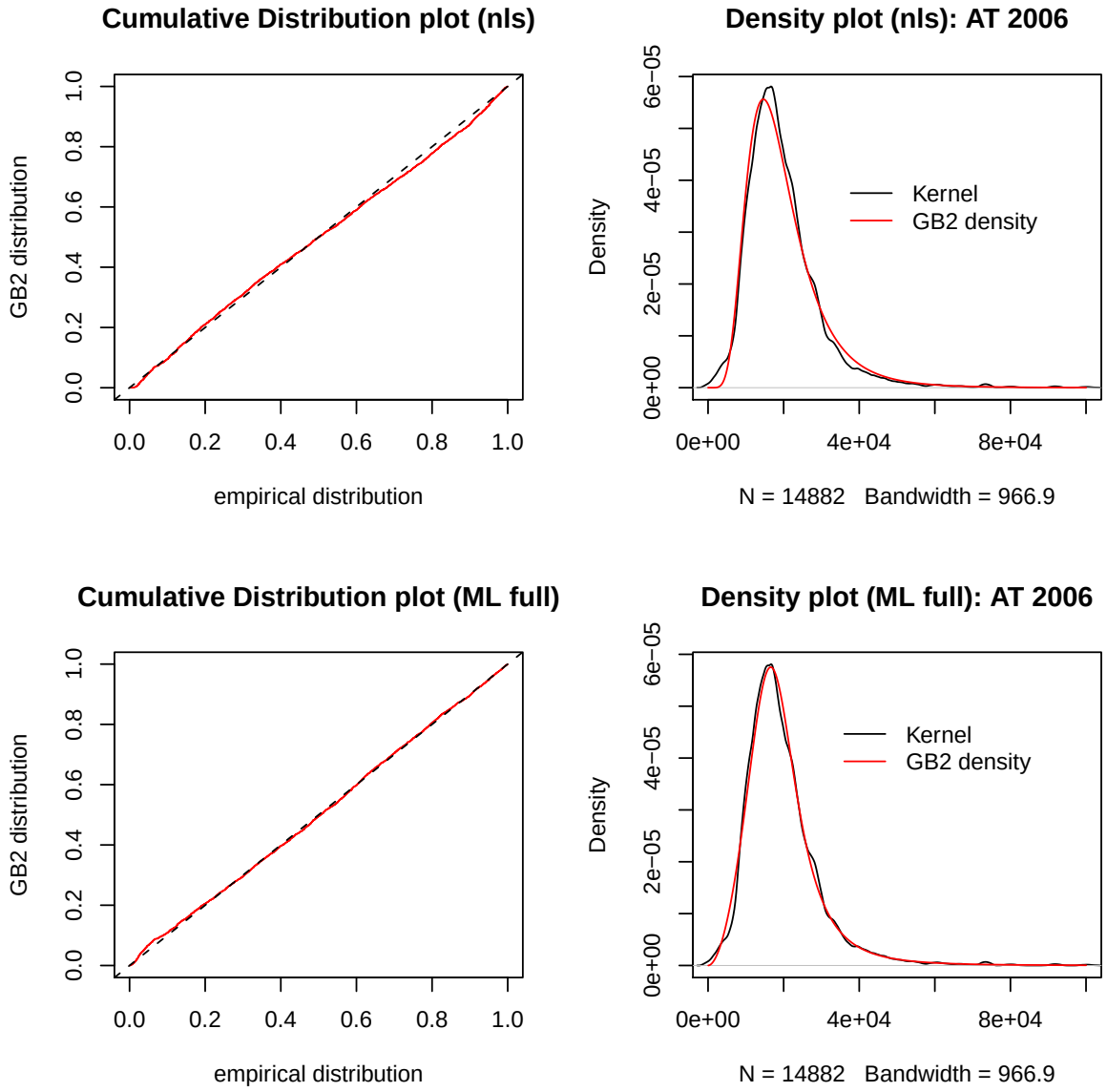

Figure 4.2: Distribution and density plots, AT 2006

4.6.3 Estimated parameters and indicators, EU-SILC participating countries 2006

In Tables 4.2 and 4.3 are presented the fitted GB2 parameters, the estimated median, ARPR, RMPG, QSR and Gini index for the 26 participating countries in the EU-SILC 2006 survey. The used methods of estimation are maximum likelihood using the full and profile log-likelihoods with adjusted sampling weights using the ad hoc procedure described in Section 4.3 and the method of nonlinear fit for indicators using the third approach.

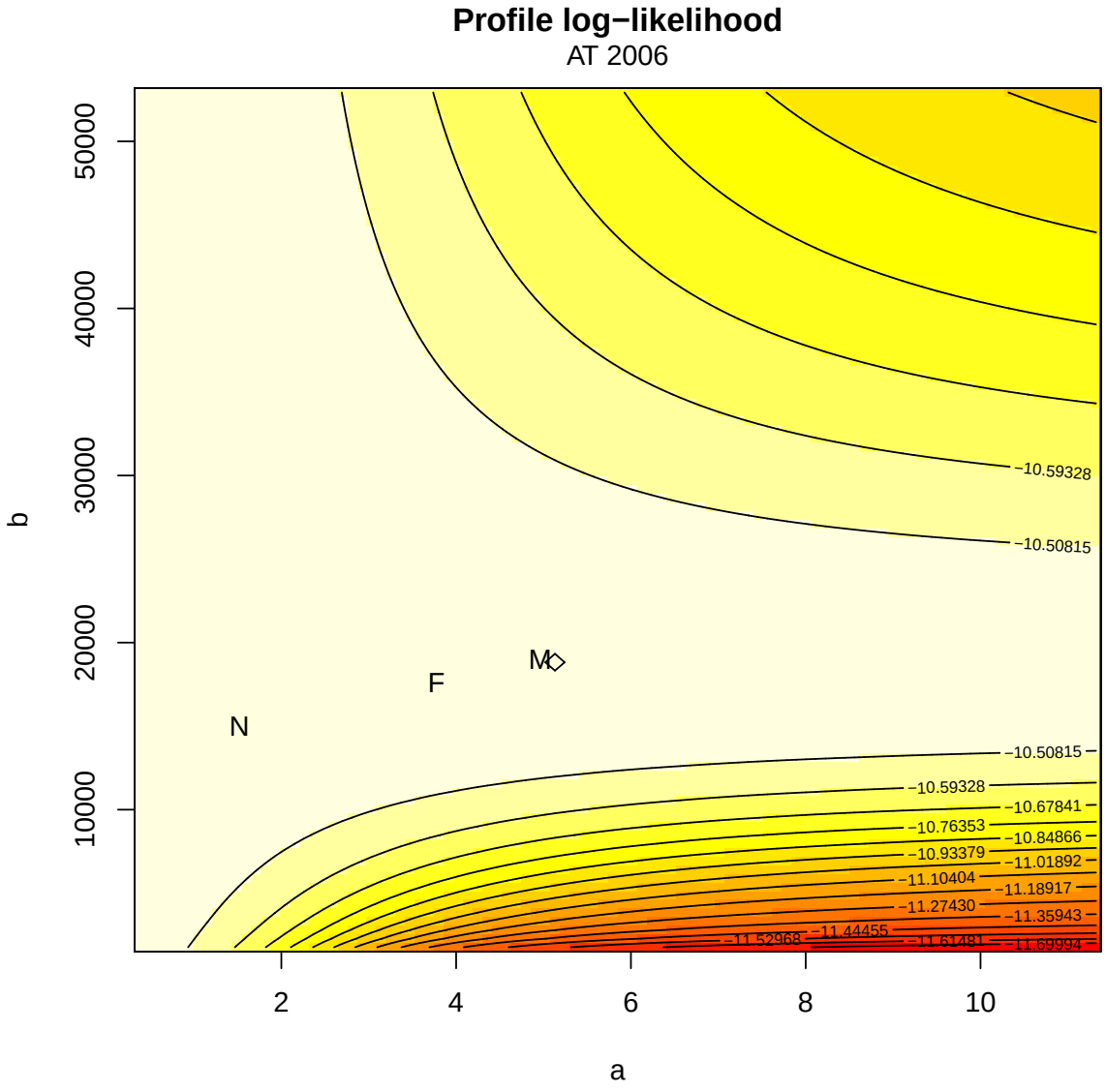

Figure 4.3: Contour plot of the profile log-likelihood, AT 2006

country	type	a	b	p	q	median	ARPR	RMPG	QSR	GINI
AT	Direct	–	–	–	–	17854	12.547	15.425	3.647	0.253
AT	NLS	1.523	15050	5.195	4.079	17854	12.547	15.425	3.646	0.257
AT	ML full	4.964	19005	0.654	0.790	17911	12.716	19.833	3.661	0.253
AT	ML prof	4.990	18996	0.650	0.784	17911	12.710	19.840	3.662	0.253
BE	Direct	–	–	–	–	17225	14.547	19.034	3.960	0.272
BE	NLS	1.941	18719	2.474	2.853	17225	14.547	19.034	3.960	0.270
BE	ML full	3.367	18643	1.050	1.319	17043	13.707	19.740	3.791	0.260
BE	ML prof	3.279	18706	1.090	1.376	17041	13.729	19.698	3.787	0.260
CY	Direct	–	–	–	–	14532	15.747	18.965	4.268	0.288
CY	NLS	1.132	13919	6.487	6.194	14532	15.747	18.965	4.268	0.285
CY	ML full	2.642	14245	1.564	1.536	14366	14.343	18.922	4.128	0.280
CY	ML prof	2.551	14192	1.658	1.617	14361	14.362	18.829	4.124	0.280
CZ	Direct	–	–	–	–	4797	9.796	16.967	3.516	0.253
CZ	NLS	7.017	4619	0.537	0.465	4797	9.796	16.967	3.516	0.252
CZ	ML full	4.846	4609	0.854	0.751	4796	10.213	16.187	3.457	0.248
CZ	ML prof	4.869	4610	0.849	0.746	4796	10.208	16.198	3.457	0.248
DE	Direct	–	–	–	–	15646	12.339	19.625	3.800	0.260
DE	NLS	5.831	15902	0.555	0.586	15646	12.339	19.625	3.800	0.263
DE	ML full	7.481	16351	0.400	0.468	15680	12.458	20.791	3.703	0.255
DE	ML prof	7.530	16348	0.397	0.465	15680	12.448	20.796	3.701	0.255
DK	Direct	–	–	–	–	22718	11.326	15.159	3.241	0.230
DK	NLS	0.870	26747	14.380	16.525	22718	11.289	15.169	3.257	0.233
DK	ML full	6.332	24834	0.517	0.732	22665	11.275	19.302	3.174	0.223
DK	ML prof	6.261	24840	0.525	0.743	22661	11.262	19.255	3.172	0.223
EE	Direct	–	–	–	–	3645	18.141	21.841	5.361	0.328
EE	NLS	1.878	3354	2.203	1.929	3645	18.141	21.841	5.360	0.331
EE	ML full	2.597	3972	1.116	1.298	3679	18.804	24.781	5.517	0.331
EE	ML prof	2.557	3984	1.140	1.331	3680	18.834	24.788	5.511	0.331
ES	Direct	–	–	–	–	11493	19.760	25.399	5.109	0.308
ES	NLS	0.912	22321	5.824	10.393	11493	19.474	25.464	5.164	0.316
ES	ML full	2.691	15675	0.914	1.738	11476	19.465	27.468	5.108	0.307
ES	ML prof	2.722	15628	0.900	1.707	11477	19.461	27.491	5.109	0.306
FI	Direct	–	–	–	–	18317	12.523	14.459	3.631	0.258
FI	NLS	1.078	12092	11.079	7.198	18317	12.523	14.459	3.632	0.257
FI	ML full	3.803	18083	1.091	1.101	18024	11.279	16.886	3.466	0.246
FI	ML prof	3.769	18074	1.106	1.115	18023	11.279	16.854	3.465	0.246
FR	Direct	–	–	–	–	16197	13.050	18.361	3.936	0.272
FR	NLS	3.561	15957	1.075	1.034	16197	13.050	18.361	3.936	0.271
FR	ML full	4.000	16251	0.900	0.911	16179	12.947	18.817	3.894	0.269
FR	ML prof	3.991	16248	0.903	0.914	16178	12.946	18.806	3.893	0.269
GR	Direct	–	–	–	–	9880	20.137	25.049	5.698	0.337
GR	NLS	1.270	11471	3.395	4.037	9880	20.137	25.049	5.698	0.337
GR	ML full	2.433	10794	1.176	1.410	9803	19.391	25.478	5.695	0.336
GR	ML prof	2.425	10800	1.182	1.418	9803	19.397	25.477	5.694	0.336
HU	Direct	–	–	–	–	3854	15.650	23.309	5.165	0.327
HU	NLS	5.862	3842	0.453	0.448	3854	15.650	23.309	5.165	0.325
HU	ML full	6.163	3906	0.424	0.446	3841	15.538	23.560	4.951	0.315
HU	ML prof	6.283	3906	0.414	0.436	3841	15.526	23.617	4.957	0.315
IE	Direct	–	–	–	–	19679	18.464	16.358	4.870	0.319
IE	NLS	0.714	5356	21.948	8.870	19679	17.688	16.584	5.051	0.321
IE	ML full	2.047	16587	2.379	1.816	19372	15.810	18.841	4.688	0.307
IE	ML prof	1.822	16037	2.945	2.179	19372	15.924	18.616	4.666	0.306

Table 4.2: GB2 fitted parameters and indicators, countries 1-13

Table 4.3: GB2 fitted parameters and indicators, countries 14-26

country	type	a	b	p	q	median	ARPR	RMPG	QSR	GINI
IS	Direct	–	–	–	–	28015	9.540	18.480	3.578	0.257
IS	NLS	7.794	27600	0.451	0.425	28015	10.247	17.982	3.514	0.250
IS	ML full	8.162	27573	0.436	0.406	28065	9.949	17.764	3.470	0.248
IS	ML prof	8.283	27566	0.429	0.399	28063	9.938	17.791	3.472	0.248
IT	Direct	–	–	–	–	14559	19.216	23.210	5.233	0.316
IT	NLS	0.632	17728	14.071	15.893	14559	19.214	23.211	5.234	0.322
IT	ML full	3.396	17318	0.711	1.062	14584	18.816	26.652	5.226	0.314
IT	ML prof	3.390	17333	0.713	1.066	14584	18.822	26.659	5.225	0.314
LT	Direct	–	–	–	–	2536	19.927	28.852	6.163	0.347
LT	NLS	4.317	2857	0.488	0.657	2536	19.927	28.852	6.163	0.346
LT	ML full	2.883	2942	0.807	1.077	2552	20.717	28.369	6.336	0.352
LT	ML prof	2.946	2926	0.786	1.041	2551	20.679	28.366	6.349	0.353
LU	Direct	–	–	–	–	29683	13.925	19.403	4.082	0.278
LU	NLS	3.428	29996	1.054	1.082	29683	13.925	19.403	4.082	0.277
LU	ML full	3.278	28902	1.185	1.106	29727	13.603	18.571	4.087	0.279
LU	ML prof	3.198	28869	1.230	1.145	29728	13.633	18.519	4.084	0.279
LV	Direct	–	–	–	–	2546	22.731	24.315	7.303	0.386
LV	NLS	0.645	1170	13.574	8.351	2546	22.731	24.315	7.303	0.387
LV	ML full	2.521	2763	0.931	1.076	2551	22.039	29.206	7.502	0.388
LV	ML prof	2.468	2770	0.959	1.111	2551	22.074	29.195	7.485	0.387
NL	Direct	–	–	–	–	17293	9.399	16.601	3.571	0.255
NL	NLS	7.586	16367	0.508	0.409	17293	9.399	16.601	3.571	0.257
NL	ML full	5.214	17499	0.695	0.698	17479	11.311	17.968	3.574	0.252
NL	ML prof	5.240	17495	0.691	0.693	17478	11.304	17.977	3.574	0.252
NO	Direct	–	–	–	–	27806	11.001	18.117	3.967	0.280
NO	NLS	7.050	26401	0.497	0.411	27806	11.001	18.117	3.967	0.278
NO	ML full	10.552	28955	0.288	0.346	27770	11.414	20.424	3.411	0.238
NO	ML prof	10.270	28953	0.297	0.358	27751	11.393	20.353	3.403	0.238
PL	Direct	–	–	–	–	3112	19.018	24.977	5.605	0.332
PL	NLS	2.539	3359	1.140	1.322	3112	19.018	24.977	5.605	0.334
PL	ML full	2.744	3505	0.970	1.221	3129	19.319	25.976	5.661	0.334
PL	ML prof	2.746	3505	0.969	1.220	3129	19.319	25.977	5.661	0.334
PT	Direct	–	–	–	–	7311	18.466	23.468	6.726	0.377
PT	NLS	3.368	6605	0.859	0.686	7311	18.467	23.469	6.726	0.383
PT	ML full	4.443	6858	0.569	0.481	7339	18.422	24.905	7.170	0.396
PT	ML prof	4.362	6861	0.582	0.492	7342	18.467	24.887	7.151	0.396
SE	Direct	–	–	–	–	17795	11.609	20.097	3.334	0.231
SE	NLS	7.747	19003	0.401	0.522	17795	11.609	20.097	3.334	0.233
SE	ML full	6.948	20412	0.416	0.690	17920	12.742	21.468	3.300	0.227
SE	ML prof	6.858	20433	0.422	0.702	17919	12.728	21.422	3.298	0.227
SI	Direct	–	–	–	–	9316	11.677	18.539	3.388	0.238
SI	NLS	4.697	9954	0.753	0.930	9316	11.677	18.539	3.388	0.238
SI	ML full	4.342	10220	0.817	1.070	9360	11.919	18.682	3.377	0.237
SI	ML prof	4.336	10221	0.819	1.072	9360	11.920	18.678	3.377	0.237
SK	Direct	–	–	–	–	3313	11.608	19.918	4.034	0.280
SK	NLS	7.139	3260	0.448	0.422	3313	12.018	19.643	4.001	0.276
SK	ML full	8.545	3372	0.362	0.389	3312	11.718	20.135	3.682	0.256
SK	ML prof	8.325	3372	0.373	0.401	3312	11.716	20.066	3.677	0.256
UK	Direct	–	–	–	–	19375	18.976	22.395	5.208	0.320
UK	NLS	0.741	19096	11.153	11.037	19375	18.976	22.396	5.208	0.322
UK	ML full	2.803	22495	0.976	1.329	19412	18.517	25.260	5.176	0.316
UK	ML prof	2.758	22487	1.001	1.359	19406	18.516	25.195	5.173	0.316

Chapter 5

The Generalized Beta Distribution of the Second Kind as a Compound Distribution

Authors: Monique Graf and Desislava Nedyalkova

5.1 Introduction

The GB2 distribution can be expressed as an infinite mixture of distributions with varying scale parameters, that is as a compound distribution, (see Kleiber and Kotz, 2003, Table 6.1). Thus, as quoted by Kleiber and Kotz (2003), the GB2 distribution and its subfamilies can be given a theoretical justification as a representation of incomes arising from a heterogeneous population of income receivers. The compounding property will be used to derive a decomposition of the GB2 into a finite mixture of components.

The GB2 parameters a, b, p, q need a large sample size (a few thousands) in order to be estimated with an acceptable precision. The GB2 model is thus hardly applicable to domains, even of moderate size. The compounding property of the GB2 distribution will allow us to exploit the model fitted at the national level, also for small sub-populations. The idea behind is that the population consists of heterogeneous groups with respect to the scale of income and that this heterogeneity is well represented by the GB2. The aim is to set up a model that estimates the heterogeneity of subgroups and is consistent with the overall fit. Once the distribution of incomes in the subgroup is determined, any subgroup characteristic (e.g. an indicator of poverty and social exclusion) can be computed.

In Section 5.2 we give a theoretical justification for the decomposition of the GB2. Two different decompositions are presented: with respect to the right or the left tail of the distribution. Next, an example that illustrates both approaches is given.

In Section 5.3 we explain how the compounding property of the GB2 can be used in a survey context and in the context of small sub-populations. We define two models, with

or without auxiliary information. The pseudo log-likelihood, using the survey weights is defined and the method of estimation is presented.

5.2 Decomposition of the GB2 distribution

Starting with a generalized gamma distribution $GG(a, \theta, p)$ with scale parameter θ, the compound representation of the GB2 distribution is obtained by assigning a inverse generalized gamma distribution $InvGG(a, b, q)$ to θ (see, e.g. JOHNSON et al., 1995).

Let us recall that the probability density of the GB2 with parameters a, b, p, q is given by:

$$f(x; a, b, p, q) = \frac{a}{b\, B(p, q)} \frac{(x/b)^{ap-1}}{((x/b)^a + 1)^{p+q}} \tag{5.1}$$

with $a, b, p, q > 0$.

The density $g(.; a, \theta, p)$ of $GG(a, \theta, p)$ is given by

$$g(x; a, \theta, p) = \frac{a}{\theta\, \Gamma(p)} (x/\theta)^{ap-1} \exp -(x/\theta)^a \tag{5.2}$$

and the density $h(.; a, b, q)$ of the distribution $InvGG(a, b, q)$ is

$$h(\theta; a, b, q) = \frac{a}{b\, \Gamma(q)} (\theta/b)^{-aq-1} \exp -(\theta/b)^{-a} \tag{5.3}$$

The GB2 density is obtained by integration over θ:

$$f(x; a, b, p, q) = \int_0^\infty h(\theta; a, b, q)\, g(x; a, \theta, p)\, d\theta \tag{5.4}$$

This is the compounding property of the GB2. The proof is recalled in Appendix B.1.

5.2.1 Decomposition with respect to the right or the left tail

Notice that the distribution of the random scale parameter θ does not depend on the shape parameter p governing the left tail. For this reason, we denote the decomposition in Equation (5.4) a decomposition with respect to the right tail.

A similar decomposition with respect to the left tail can be obtained using the following property of the GB2:
Let $y = 1/x$ denote the inverse of the income variable x. Then y also follows a GB2 distribution and its density can be written as

$$f(y; a', b', p', q'),$$

where $a' = a$, $b' = b^{-1}$, $p' = q$ and $q' = p$ (see Kleiber and Kotz, 2003).

We have, using Equation (5.4):

$$f(y; a', b', p', q') = \int_0^\infty h(\theta; a', b', q')\, g(y; a', \theta, p')\, d\theta \tag{5.5}$$

By a change of variable ($x = 1/y$) in Equation (5.5), we obtain the left tail decomposition of the GB2 density in Equation (5.1):

$$f(x; a, b, p, q) = \int_0^\infty h(\theta; a, b^{-1}, p)\, (1/x^2) g(1/x; a, \theta, q)\, d\theta \tag{5.6}$$

The decomposition with respect to the left tail emphasizes the variability of the poor and gives better results for the poverty indicators.

5.2.2 Right tail discretization

For simplicity, let us drop the explicit reference to the fixed parameters a, b, p, q in Equation (5.4).

We propose to use the decomposition in the following way: Discretize the random scale parameter θ by partitioning its domain of integration into L intervals, with limits

$$\theta_0 = 0 < \theta_1 < ... < \theta_L = \infty.$$

Then the GB2 density can be written as a mixture:

$$\begin{aligned} f(x) &= \sum_{\ell=1}^{L} \int_{\theta_{\ell-1}}^{\theta_\ell} h(\theta) g(x, \theta)\, d\theta \\ &= \sum_{\ell=1}^{L} \left[\int_{\theta_{\ell-1}}^{\theta_\ell} h(\theta)\, d\theta \right] \frac{\int_{\theta_{\ell-1}}^{\theta_\ell} h(\theta) g(x, \theta)\, d\theta}{\int_{\theta_{\ell-1}}^{\theta_\ell} h(\theta)\, d\theta} = \sum_{\ell=1}^{L} p_{L-\ell}\, f_{L-\ell}(x) \end{aligned} \tag{5.7}$$

The conditional density $f_{L-\ell}(x)$ given that the scale parameter is in $(\theta_{\ell-1}, \theta_\ell)$ is defined by the fraction in Equation (5.7). The term in brackets is the probability $p_{L-\ell}$ giving the weight of the density $f_{L-\ell}(x)$ in the mixture. (The numbering with $L - \ell$ instead of ℓ is such that densities with more mass towards zero have a larger index.)

Evaluation of $f_\ell(x)$ and p_ℓ

With $u = (\theta/b)^{-a}$ (see Equation 5.3), the integration bounds are changed to

$$u_\ell = (\theta_{L-\ell}/b)^{-a}, \ell = 0, ..., L, \qquad (u_{\ell-1} < u_\ell).$$

Denoting by $P(\cdot, q)$ the cumulative distribution function of the standard gamma distribution with shape parameter q, we obtain

$$p_\ell = P(u_\ell, q) - P(u_{\ell-1}, q) \tag{5.8}$$

In practice, the p_ℓ are chosen and determine the u_ℓ.
Set $t = (x/b)^a + 1$. The component density is given by:

$$f_\ell(x) = f(x)\frac{P(tu_\ell, p+q) - P(tu_{\ell-1}, p+q)}{P(u_\ell, q) - P(u_{\ell-1}, q)} \tag{5.9}$$

where $f(x)$ is the GB2 density in Equation (5.1). Proofs are given in Appendix B.2.

5.2.3 Left tail discretization

The principle is to apply the right tail discretization to the inverse income y, and obtain the decomposition in the original income scale by a change of variables $x = 1/y$.
For the inverse income, we have: $u' = (\theta'/b')^{-a} = (\theta^{-1}/b^{-1})^{-a} = (\theta/b)^a$ and

$$u'_\ell = (\theta_\ell/b)^a, \ell = 0, ..., L, \qquad (u'_{\ell-1} < u'_\ell).$$

Knowing that $q' = p$, we see that u'_ℓ is determined by:

$$\tilde{p}_\ell = P(u'_\ell, p) - P(u'_{\ell-1}, p).$$

With $t' = (y/b')^{a'} + 1 = (x/b)^{-a} + 1$, and changing to the variable $x = 1/y$, we obtain new component densities $\tilde{f}_\ell(x)$:

$$\tilde{f}_\ell(x) = f(x)\frac{P((t'u'_\ell, p+q) - P((t'u'_{\ell-1}, p+q)}{P(u'_\ell, p) - P(u'_{\ell-1}, p)} \tag{5.10}$$

The proof is in Appendix B.3. Finally we have, that:

$$f(x) = \sum_{\ell=1}^{L} \tilde{p}_\ell \, \tilde{f}_\ell(x) \tag{5.11}$$

Notice that in this representation, densities with more mass towards zero have a smaller index. Now, we can fit the compound GB2 distribution using this new decomposition of the GB2 density function.

Figure 5.1 shows the right and left tail decomposition of the GB2 for AT2006, with $p_\ell = \tilde{p}_\ell = 1/3$, $\ell = 1, 2, 3$. One sees clearly that the very poor are totally in f_1 for the left tail decomposition (bottom pane), but are scattered between all 3 components in the right tail decomposition (upper pane).

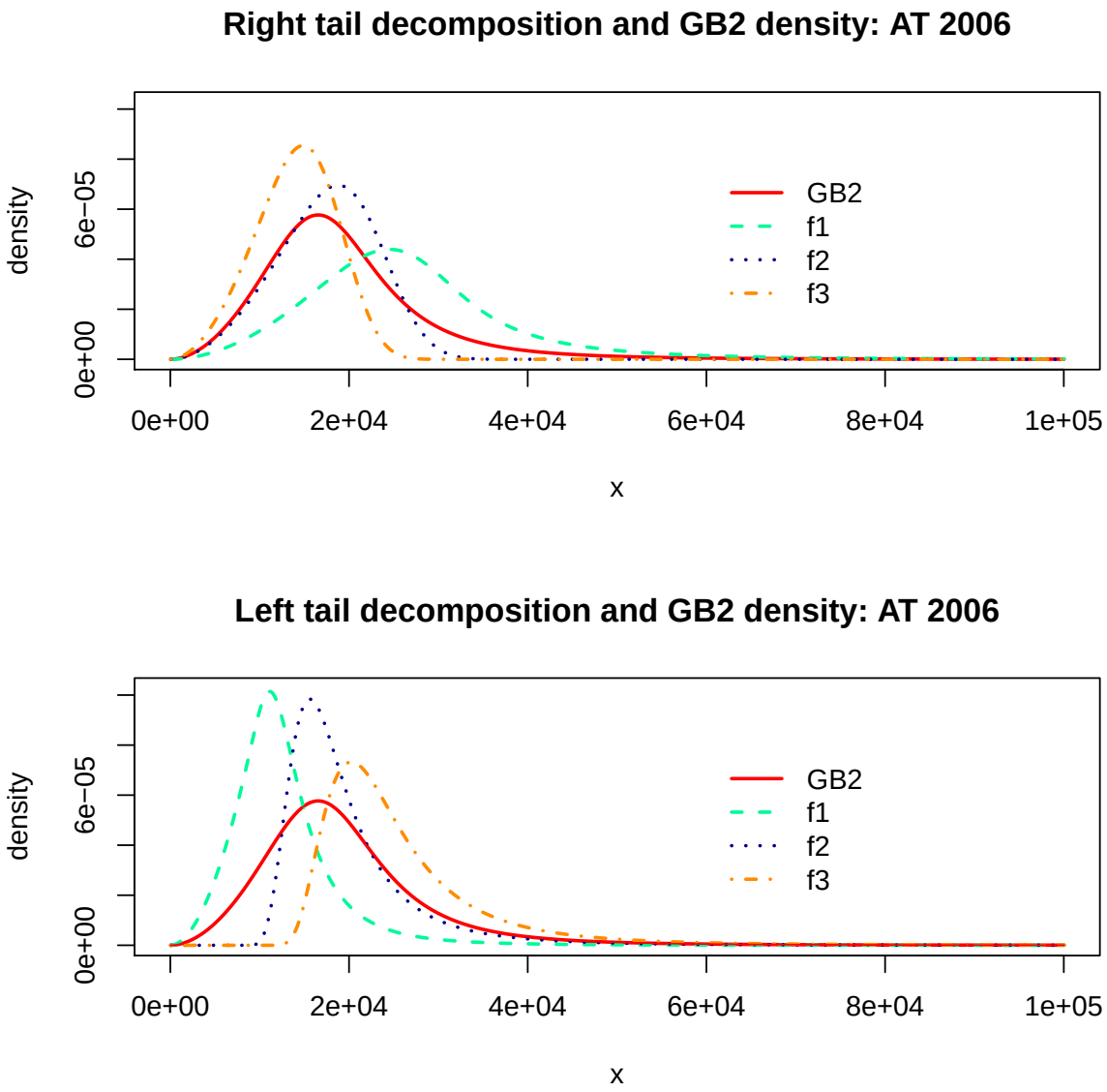

Figure 5.1: Right and left tail decomposition and the parent GB2 density

5.2.4 Sensitivity plot to the mixture probabilities

Consider a GB2 fit, determined by $a = 5.89$, $p = 0.49$, $q = 0.65$, close to the AMELIA fitted parameters. Because we are only interested in scale-free indicators, b can be given an arbitrary value, e.g. 1. The component densities of the right tail decomposition and the left tail decomposition with $p_\ell = 1/3$, $\ell = 1, \ldots, 3$ are computed. The break points in Equation 5.8 are $(0, u_1, u_2, \infty)$ and $u = (u_1, u_2) = (0.17, 0.65)$ $((0.09, 0.65))$, for the right (left) tail decomposition, respectively. In Figure 5.2 and Figure 5.3, we let the mixture probabilities pp_1 and pp_2 of f_1 and f_2 respectively vary. The probability of f_3 is thus $1 - pp_1 - pp_2$.

For the right tail decomposition, f_1 is the component having the less mass towards zero

(Figure 5.2), whereas for the left tail decomposition f_1 is the component having the most mass towards zero (Figure 5.3). The dot in each panel shows the position of the indicator in the original GB2 distribution, here corresponding to $pp_1 = pp_2 = 1/3$.

With varying probabilities (pp_1, pp_2) the left tail decomposition (Figure 5.3) generates a much larger range for the indicators of poverty ARPR and RMPG. This is the reason why this approach proves to be more efficient in our context than the right tail decomposition. The instabilities of RMPG in Figure 5.3, when $pp_2 \approx 0$ for small values of pp_1, has to be noticed. If $pp_3 \to 1$, a large variance in RMPG has to be expected; on the contrary, if $pp_3 \to 0$ (diagonal in the graph), RMPG is almost insensitive to the shares of pp_1 and pp_2.

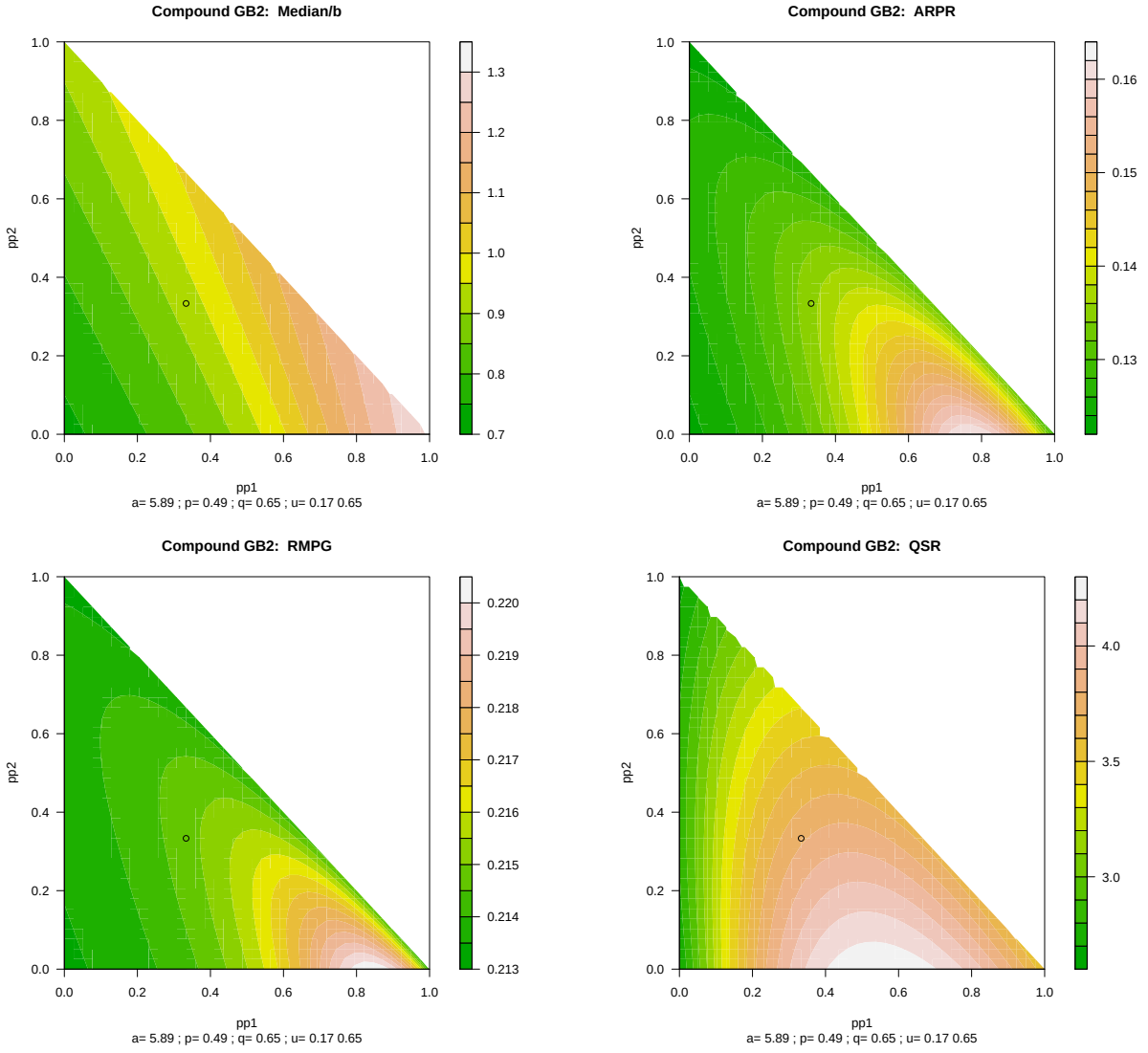

Figure 5.2: Right tail decomposition: sensitivity plots

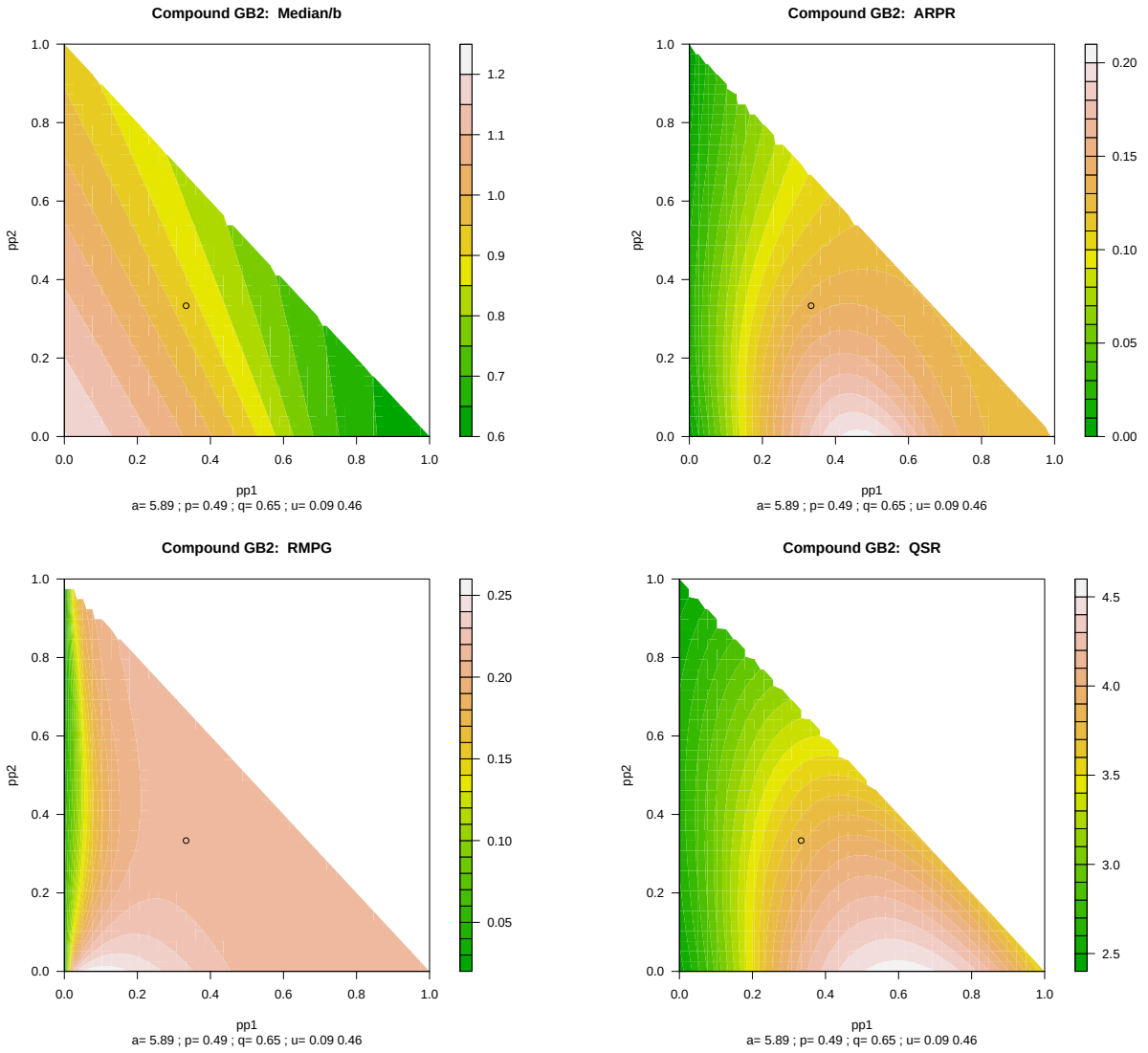

Figure 5.3: Left tail decomposition: sensitivity plots

5.3 Use of the decomposition

5.3.1 New model

The GB2 parameters a, b, p, q are determined at the global (national) level.

Now, given a partition into L intervals for the scale parameter θ of incomes, we can define a new model for a sub-population based on a mixture of the densities $f_\ell(.)$ given in Equation (5.7) or $\tilde{f}_\ell(.)$ in Equation (5.11). In this model, the component densities f_ℓ of the mixture are fixed and the probabilities p_ℓ are re-fitted at the sub-population level.

The initial GB2 fit of p_ℓ, given by the bracket in Equation (5.7) or (5.11) will serve as starting values $p_\ell^{(0)}$.

The estimation method is by pseudo-maximum likelihood as before for the GB2 fit. We can use the procedure in two ways:

1. Fit the p_ℓ on a sub-population.
   It is assumed that we need a much smaller sample size for a good estimate of the probabilities p_ℓ than it was necessary for the estimation of the GB2 parameters.

2. Model the p_ℓ with auxiliary information.
   Auxiliary variables can be used to model the probabilities p_ℓ, without reference to the density $h(\cdot)$. In this way, heterogeneous population structures can be accounted for.

In both cases, an iterative algorithm is constructed. The initial values $p_\ell^{(0)}$ for p_ℓ are given by the GB2 fit, i.e. by the expression in brackets in Equation (5.7).

5.3.2 Pseudo-likelihood

Let us write for simplicity the component densities as f_ℓ. The estimation method is the same for $\tilde{f}_\ell$.

Let n be the sample size. The pseudo-log-likelihood is written as

$$\log L(p_1, ..., p_L) = \sum_{k=1}^{n} w_k \log \left(\sum_{\ell=1}^{L} p_\ell f_\ell(x_k) \right) \tag{5.12}$$

There are only $L-1$ parameters to estimate, because the probabilities p_ℓ sum to 1. Moreover the p_ℓ must be positive. With these constraints in mind, change the parameters p_ℓ, $\ell = 1, ..., L$, to

$$v_\ell = \log(p_\ell / p_L), \qquad \ell = 1, ..., L-1$$

then

$$p_\ell \;=\; \exp(v_\ell)/(1 + \sum_{1,...,L-1} \exp(v_j)), \qquad \ell = 1, ..., L-1$$

$$p_L = 1/(1 + \sum_{1,...,L-1} \exp(v_j)).$$

The partial derivatives are, respectively:

$$\begin{aligned}
\frac{\partial p_\ell}{\partial v_\ell} &= p_\ell(1 - p_\ell), \qquad \ell = 1, ..., L-1, \\
\frac{\partial p_\ell}{\partial v_j} &= -p_\ell p_j,\ j \neq \ell; \qquad \ell = 1, ..., L;\ j = 1, ..., L-1.
\end{aligned}$$

Thus, for $\ell = 1, ..., L-1$, the likelihood equations are:

$$\begin{aligned}
\frac{\partial \log L}{\partial v_\ell} &= \sum_{k=1}^{n} w_k \frac{p_\ell \left[f_\ell(x_k) - \sum_{j=1}^{L} p_j f_j(x_k) \right]}{\sum_{j=1}^{L} p_j f_j(x_k)} = 0 \\
&\iff \sum_{k=1}^{n} w_k \left(\frac{f_\ell(x_k)}{\sum_{j=1}^{L} p_j f_j(x_k)} - 1 \right) = 0
\end{aligned} \tag{5.13}$$

From the set of equations (5.13), we can estimate p_j.

5.3.3 Introduction of auxiliary variables

One can model the probabilities p_ℓ with auxiliary variables. Let $\mathbf{z}_k$ be the vector of auxiliary information for unit k. This auxiliary information modifies the probabilities p_ℓ at the unit level. Let us denote by $p_{k,\ell}$ the weight of the density f_ℓ for unit k. For $\ell = 1, ..., L-1$, we pose a linear model for $v_{k,\ell}$:

$$\log(p_{k,\ell}/p_{k,L}) = v_{k,\ell} = \sum_{i=1}^{I} \lambda_{\ell i} z_{ki} = \mathbf{z}_k \boldsymbol{\lambda}_{\boldsymbol{\ell}} \tag{5.14}$$

The log-likelihood becomes:

$$\log L(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_{L-1}) = \sum_{k=1}^{n} w_k \log \left(\sum_{\ell=1}^{L} p_{k,\ell} f_\ell(x_k) \right) \tag{5.15}$$

One must solve

$$\frac{\partial \log L}{\partial \boldsymbol{\lambda}_\ell} = \sum_k \frac{\partial \log L}{\partial v_{k,\ell}} \frac{\partial v_{k,\ell}}{\partial \boldsymbol{\lambda}_\ell} = 0, \qquad \ell = 1, ..., L-1,$$

which is equivalent to

$$\begin{aligned}
&\sum_{k=1}^{n} w_k \left(\frac{p_{k,\ell} f_\ell(x_k)}{\sum_{j=1}^{L} p_{k,j} f_j(x_k)} - 1 \right) \mathbf{z}_k = \\
&\sum_{k=1}^{n} w_k \left(\frac{\exp(\mathbf{z}_k \boldsymbol{\lambda}_\ell) f_\ell(x_k)}{\sum_{j=1}^{L-1} \exp(\mathbf{z}_k \boldsymbol{\lambda}_j) f_j(x_k) + f_L(x_k)} - 1 \right) \mathbf{z}_k = 0
\end{aligned} \tag{5.16}$$

For each $\ell = 1, ..., L-1$, the number of equations in (5.16) is equal to the dimension of $\mathbf{z}_k$.

5.3.4 Usage

Estimate the GB2 parameters a, b, p, q by pseudo-ML at the population (national) level and choose a partition of the GB2 as in Equation (5.7).

Algorithm without auxiliary variables

For a given sub-population, adapt the GB2 fit by changing the probabilities p_ℓ.

1. Compute the initial probabilities $p_\ell = \hat{p}_\ell^{(0)}$ and the component densities $f_\ell(x)$ according to Equations (5.8) and (5.9), respectively.
2. Starting with the initial values $\hat{p}_\ell^{(0)}$, maximize the pseudo-likelihood with respect to p_ℓ in Equation (5.12) by solving the system (5.13).

Algorithm with auxiliary variables

For the whole population, use the information given by the vector of auxiliary variables $\mathbf{z}_k$ to adapt the GB2 fit by changing the probabilities $p_{k,\ell}$.
Let I be the dimension of $\mathbf{z}_k$.

1. Compute the initial probabilities $p_{k,\ell} = \hat{p}_\ell^{(0)}$ (not depending on k) and the component densities $f_\ell(x)$ according to equations (5.8) and (5.9), respectively.
2. We must find initial values for $\lambda_{\ell i}, i = 1, ..., I$. Let $\bar{z}_i = \sum_k w_k z_{ki} / \sum_k w_k$ be the average value of the i-th explanatory variable. Writing

$$\log(\hat{p}_\ell^{(0)} / \hat{p}_L^{(0)}) = v_\ell^{(0)} = \sum_{i=1}^{I} \lambda_{\ell i}^{(0)} \bar{z}_i,$$

   we can choose

$$\lambda_{\ell i}^{(0)} = v_\ell^{(0)} / (I \bar{z}_i) \tag{5.17}$$

   as starting values.
3. Starting with the initial values $\lambda_{\ell i}^{(0)}$, maximize the pseudo-likelihood with respect to $\lambda_{\ell i}$ in Equation (5.15) by solving the system (5.16).

Choice of partition

The number L of components f_ℓ can be chosen arbitrarily, but it may be reasonable to keep L small. In the examples, we choose $L = 3$ and the integration bounds in Equation (5.8), so that $p_\ell^{(0)} = 1/3$. In this way, the components f_1, f_2, f_3 represent respectively the income distributions with small, medium and high scale parameters, that is with more mass to the left for f_1, more mass to the center for f_2 and more mass to the right for f_3, each having the same weight in the overall GB2 fit. A better founded way to choose the partition has still to be developed.

Chapter 6

Use of mixture distributions in the context of heterogeneous populations

6.1 Introduction

In this chapter, we investigate the use of parametric mixture distributions in the special case of two components, each following the same type of distribution. Mixture distributions are appropriate when the population consists of heterogeneous subpopulations. For instance, in many species the body weight depends on the gender. The weights of the males and the weights of the females might each be approximatively normally distributed, but with different means and standard deviations. In this context, it can make sense to interpret the overall distribution of weights as a mixture of the weight distributions by gender in this species.
The field of analysis of income distributions is not a typical example for the usage of mixture distributions, but in this context they might be useful as well. If there is an income sample of a population, consisting of different subpopulations with heterogeneous income distributions, which can be fitted well by single component models each, a mixture density can be adequate. In some sense the income data set of Amelia, explained in detail in Alfons et al. (2011), can be interpreted as an income distribution of a *synthetic Europe*, which consists of income distributions of several countries. This might justify the usage of a mixture distribution in this context.

After these rather intuitive explanations it is necessary to quote a formal definition of a mixture distribution. A mixture density or mixture distribution can be defined as the following (see Redner and Walker, 1984): Let f_i, $i = 1, \ldots, m$, be densities, with each of them determined unequivocally by parameter vectors $\mathbf{a}_i$, $i = 1, \ldots, m$, where $\mathbf{a}_i \subseteq \Omega_i \subseteq \mathbb{R}^m$, for all i. Then for $x \in \mathbb{R}^n$, $n \in \mathbb{N}$

$$f(x|\mathbf{A}) = \sum_{i=1}^{m} \alpha_i f_i(\mathbf{x}|\mathbf{a}_i), \tag{6.1}$$

is called *(parametric) mixture distribution (with a finite number of components)*, with $\alpha_i \geq 0$, $i = 1, \ldots, m$, $\sum_{i=1}^{m} \alpha_i = 1$ and $\mathbf{A} = (\alpha_1, \ldots, \alpha_m, \mathbf{a}_1, \ldots, \mathbf{a}_m)$. The $f_i(x|\mathbf{a}_i)$ are called

mixture components (or simply components), the α_i are named mixture proportions. So a mixture density can be interpreted as a convex combination of single densities.

Since there are many different component densities used in the field of income distributions and there are infinitely many combinations to combine those to a mixture density, there are infinitely many possible choices of mixture densities. There is always a conflict of goals between the goodness of fit and the simplicity of a model. In general, a large number of parameters increases the flexibility of a model, since its number of degrees of freedom also raises. In addition, the downside of many model parameters is, that the model tends to get more complex and more difficult to fit. Also the economic interpretation of each parameter of a model with many parameters may become more intricate. It seems reasonable to choose the same parametric distribution for all mixture components, which should be a generally accepted model for the subject. In the context of income distributions this means that the GB2 and related distributions should be considered. Since a mixture density of two GB2s would already contain nine parameters (four for each GB2 component and one mixture parameter), it is sensible to fall back on single distributions with less parameters. In this report we choose a mixture distribution of two Dagum distributions due to the fact that this three-parametric distribution has proven to be the best fitting three-parametric special case of the GB2 in the context of income distributions (see BANDOURIAN et al., 2002). In the following, the selected model is referred to as the TCD (two component Dagum) for the sake of clarity.

Figure 6.1 shows the density of the positive part of the Amelia income data set (variable: EDIS). For a better illustration the highest 8,000 incomes were excluded. It is easy to see that this synthetic Europe income distribution does not have the typical shape of the income distribution of a single country. Hence, in these cases we would expect a good fit with a mixture density.

Figure 6.2 shows a sample of size 15,528 fitted with a TCD. Again, for a better illustration, the highest 28 incomes were excluded. Although there are some fitting problems close to 0 and for high incomes, in general the distribution provides a decent fit.

The next section provides a few definitions and facts about the Dagum distribution and the TCD as well as fitting methods for the named distributions. The third section deals with the numerical calculation of inequality and poverty measures of the TCD. Finally, this chapter concludes with a description of a simulation study. One way for estimating inequality and poverty measures of a population in practice is to draw samples and calculate estimators of the indicators directly from the sample. Those estimators (in the following referred to as *direct estimates*) are in general unbiased, but may have a very high variance for some non-robust indicators like the quintile share ratio (QSR). Another approach is to fit a parametric distribution to the sample and then calculate the indicators out of the fitted distribution. Those estimators will be called *indirect estimates*. In this chapter, the indirect estimation is always associated with a TCD fit. The results of the described simulation study can be found in HULLIGER et al. (2011).

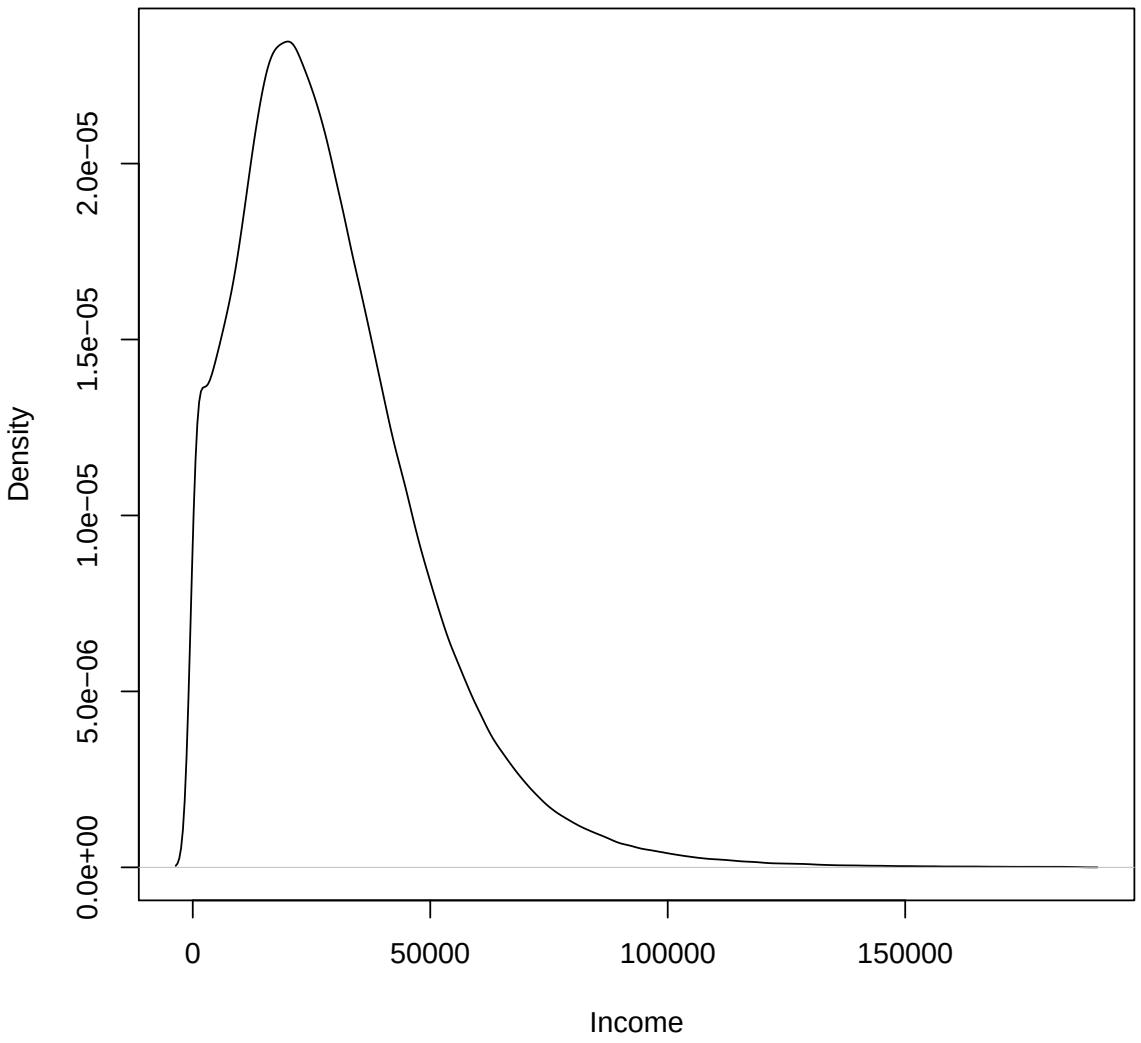

Figure 6.1: Kernel density of the equivalized household income of the Amelia data set

6.2 Fitting of mixtures of Dagum distributions

This section explains how the TCD can be fitted to data. Therefore, some preparatory work has to be done, which leads to the formation of this section. Firstly, the single Dagum distribution and the TCD are introduced very shortly. Since the theoretical distribution is well documented in the literature, only a few key facts are pointed out. More details about the Dagum distribution can be found in (Dagum, 1977 and Kleiber and Kotz, 2003).
Afterwards, the fitting procedure for a single Dagum distribution with the maximum likelihood method is presented. This forms a central component in the EM algorithm, used for fitting a TCD. A subsection about the EM algorithm concludes this section.

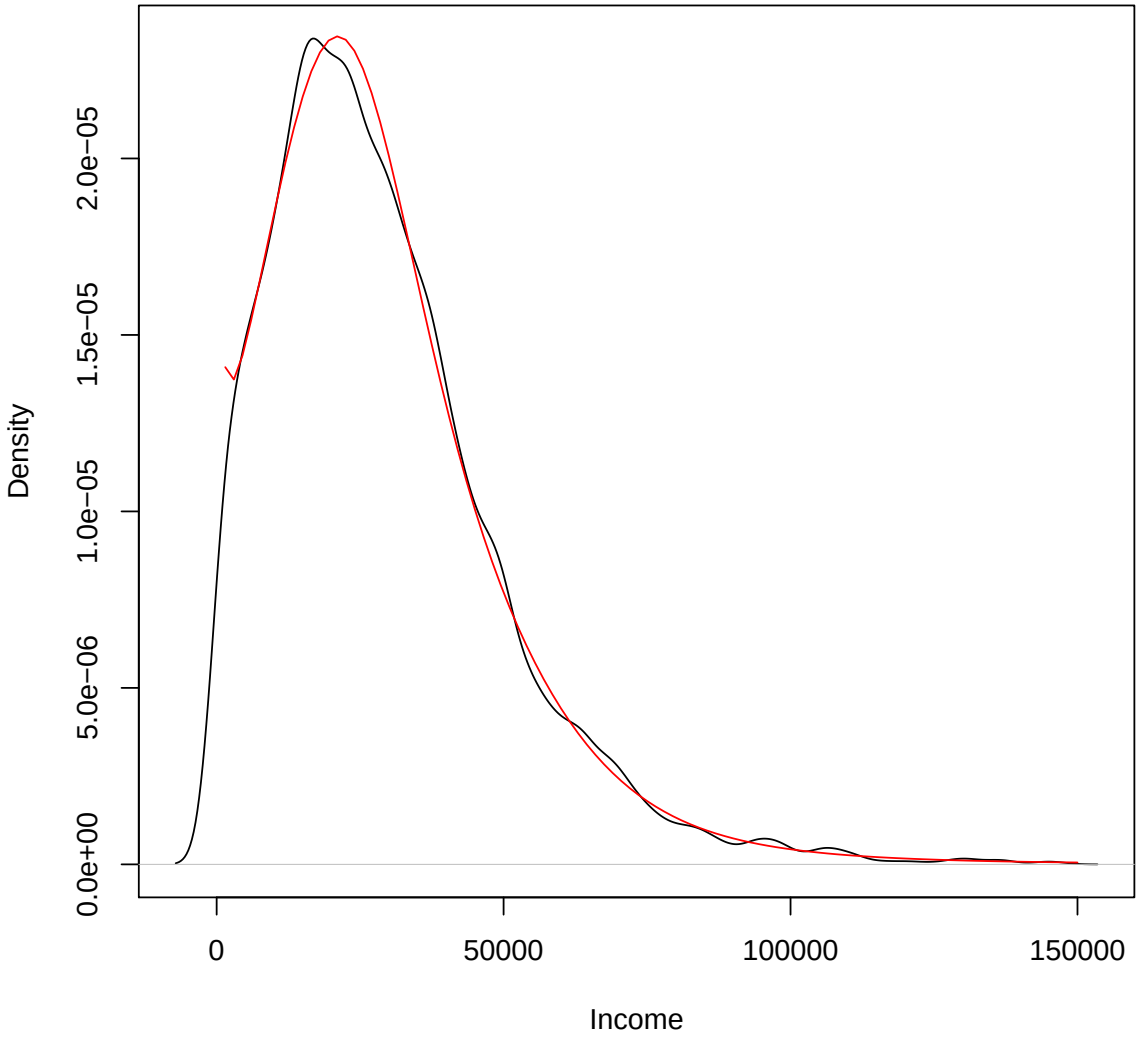

Figure 6.2: Kernel density of an income sample fitted with a TCD

6.2.1 The Dagum distribution and the TCD

The Dagum distribution (D) is a three-parametric model developed by and named after Camilo Dagum in 1977. Other common names for the Dagum distribution are Burr-III-distribution, inverse Burr distribution, (three-parametric) Kappa distribution and Beta-K-distribution (KLEIBER and KOTZ, 2003). Its density is

$$f_D(x;a,b,p) = \frac{apx^{ap-1}}{b^{ap}[1+(x/b)^a]^{p+1}}, \quad x > 0, \tag{6.2}$$

where $a, b, p > 0$.

b is a scale parameter, whereas a and p are shape parameters. It can be shown that the Dagum distribution is a special case of the more general GB2 ($GB2(x; a, b, p, q = 1) = D(x; a, b, p)$). The cdf of the Dagum distribution is

$$F_D(x; a, b, p) = \left(1 + \left(\frac{x}{b}\right)^{-a}\right)^{-p}. \tag{6.3}$$

Its quantile function exists analytically and is given as

$$Q_D = F^{-1}(u; a, b, p) = b[u^{-1/p} - 1]^{-1/a}. \tag{6.4}$$

The moments of the Dagum distribution exist for $k < a$ and can be calculated as

$$E_D(x^k) = \frac{b^k B(p + k/a, 1 - k/a)}{B(p, 1)} = \frac{b^k \Gamma(p + k/a)\Gamma(1 - k/a)}{\Gamma(p)}. \tag{6.5}$$

In particular its mean exists for $a > 1$ and can be calculated as

$$\mu_D = \frac{b^k B(p + 1/a, 1 - 1/a)}{B(p, 1)} = \frac{b^k \Gamma(p + 1/a)\Gamma(1 - 1/a)}{\Gamma(p)}. \tag{6.6}$$

It is rather trivial to extend some of these properties to the TCD, since it is a convex combination of two Dagum distributions. The density of the TCD is

$$f_{TCD}(x) = \frac{\alpha a p x^{ap-1}}{b^{ap}[1 + (x/b)^a]^{p+1}} + \frac{(1 - \alpha) a_2 p_2 x^{a_2 p_2 - 1}}{b_2^{a_2 p_2}[1 + (x/b_2)_2^a]^{p_2+1}}, \; x > 0, \tag{6.7}$$

$a, a_2, b, b_2, p, p_2 > 0$ and $\alpha \in [0, 1]$. This leads to the cdf

$$F_{TCD}(x; a, b, p) = \alpha \left(1 + \left(\frac{x}{b}\right)^{-a}\right)^{-p} + (1 - \alpha)\left(1 + \left(\frac{x}{b_2}\right)^{-a_2}\right)^{-p_2}. \tag{6.8}$$

In contrast to the Dagum distribution, the cdf of the TCD is not invertible, so there is no closed form expression of its quantile function. Calculating quantiles of the TCD is one issue in section 6.3.
The mean of the TCD is

$$\mu_{TCD} = \frac{\alpha b \Gamma(p + 1/a)\Gamma(1 - 1/a)}{\Gamma(p)} + \frac{(1 - \alpha) b_2 \Gamma(p_2 + 1/a_2)\Gamma(1 - 1/a_2)}{\Gamma(p_2)}, \tag{6.9}$$

exists at least if $a, a_2 > 1$. After this complementing list of definitions and facts about the TCD, the following subsection deals with the fitting of a single Dagum distribution.

6.2.2 Fitting a Dagum distribution with the maximum likelihood method

Before fitting a TCD distribution, it seems reasonable to have a look at the fitting of a single Dagum distribution. In the original paper Dagum (1977) Dagum presented five

different methods to fit the Dagum distribution. However, the maximum likelihood approach (in the following abbreviated with ML) tends to lead to the best results.
Since the Dagum distribution is a special case of the more general GB2, its fitting procedure can be derived directly from the ML-fit of the GB2. Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ denote a complete random sample of size n, then its log-likelihood function is given by

$$\begin{aligned}\log L_D = \quad & n \log a + n \log p + (ap - 1) \sum_{i=1}^{n} \log x_i - nap \log n \\ & -(p+1) \sum_{i=1}^{n} \log \left[1 + \left(\tfrac{x_i}{b}\right)^a\right]. \end{aligned} \tag{6.10}$$

To maximize the value of $\log L_D$, we need to solve a system of equations which are given by the roots of its partial derivatives. This leads to the following equations (see KLEIBER and KOTZ, 2003):

$$\begin{aligned} \tfrac{n}{a} + p \sum_{i=1}^{n} \log\left(\tfrac{x_i}{b}\right) - (p+1) \sum_{i=1}^{n} \log\left(\tfrac{x_i}{b}\right) \left[\left(\tfrac{b}{x_i}\right)^a + 1\right]^{-1} &= 0 \\ np - (p+1) \sum_{i=1}^{n} \left[1 + \left(\tfrac{b}{x_i}\right)^a\right]^{-1} &= 0 \\ \tfrac{n}{p} + a \sum_{i=1}^{n} \log\left(\tfrac{x_i}{b}\right) - \sum_{i=1}^{n} \log\left[1 + \left(\tfrac{x_i}{b}\right)^a\right] &= 0. \end{aligned} \tag{6.11}$$

For solving this system of equations, methods of non-linear optimization like the BFGS method are required. It is possible to implement weights, for example survey weights, into (6.11), in analogy to Equation (4.5). Indeed for the fitting of a mixture of two Dagum distributions performed by the EM algorithm, the version with weights is used.
Let w_i denote the weight of x_i and let the weights already be standardised, i.e. $\sum_{i=1}^{n} w_i = 1$, then (6.11) turns into the following system of equations:

$$\begin{aligned} \tfrac{1}{a} + p \sum_{i=1}^{n} w_i \log\left(\tfrac{x_i}{b}\right) - (p+1) \sum_{i=1}^{n} w_i \log\left(\tfrac{x_i}{b}\right) \left[\left(\tfrac{b}{x_i}\right)^a + 1\right]^{-1} &= 0 \\ p - (p+1) \sum_{i=1}^{n} w_i \left[1 + \left(\tfrac{b}{x_i}\right)^a\right]^{-1} &= 0 \\ \tfrac{1}{p} + a \sum_{i=1}^{n} w_i \log\left(\tfrac{x_i}{b}\right) - \sum_{i=1}^{n} w_i \log\left[1 + \left(\tfrac{x_i}{b}\right)^a\right] &= 0. \end{aligned} \tag{6.12}$$

It can be solved in analogy to (6.11).

6.2.3 Fitting of a TCD: The EM algorithm

The maximum likelihood method leads to good fitting results for a single Dagum distribution. For a mixture of Dagum distributions, like the TCD, the log-likelihood function tends to have multiple local maxima. Also the additional constraint on the mixture parameter α increases the complexity of the general optimization problem. Therefore it is

advisable to avoid the usage of a ML-fit in this context. A good alternative is the EM algorithm, invented by Dempster/Laird/Rubin in 1977 (Dempster et al., 1977). The following explanations refer to McLachlan and Krishnan (2008).
Let $\mathbf{x}$ denote an income vector of length n, whose density is to approximate with a TCD. We assume that each element of $\mathbf{x}$, x_i can be allocated to one of the mixture components. There exist label vectors $\mathbf{z}_i = (z_{i1}, z_{i2})$ for each x_i which indicate from which component x_i is taken. For all entries of $\mathbf{z}_i$:

$$\begin{aligned} z_{ij} &= 1, \quad \text{if } x_i \text{ belongs to the } j\text{-th component,} \\ z_{ij} &= 0 \quad \text{else.} \end{aligned} \tag{6.13}$$

The z_{ij} will be denoted as *component labels* in the following text. In general, the $\mathbf{z}_i$ are not known. Therefore, the whole issue can be interpreted as a missing data problem. With the same notation as in 6.1 and survey weights w_i $(i = 1, \ldots, n)$, the log-likelihood function $L(f)$, which is to be maximized, can be expressed as

$$\log L(f) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} w_i \left(\log \alpha_j + \log f_j(x_i|\mathbf{a}_j) \right). \tag{6.14}$$

One main idea of the EM algorithm is to assign the data $\mathbf{x}$ to the mixture components f_j. Therefore, the component labels z_{ij} $\forall i$, $\forall j$ have to be estimated, which is performed by one step of the EM algorithm. The EM algorithm requires starting values for all distribution parameters. The algorithm consists of two steps, which justify its name: The **E**xpectation step and the **M**aximization step.

For the kth run the steps can be described as the following:

E-step:

To estimate the $z_{ij}^{(k)}$, calculate the estimated probability that x_i originates from distribution f_j under the condition that the distribution parameters $\hat{\mathbf{a}}_j^{(k)}$ coincide with the *true* parameter values $\mathbf{a}_j$. For the estimators of $z_{ij}^{(k)}$ one gets

$$\hat{z}_{ij}^{(k)} = \frac{\alpha_j^{(k-1)} f_j\left(x_i|\hat{\mathbf{a}}_j^{(k-1)}\right)}{\sum_{l=1}^{m} \alpha_l^{(k-1)} f_l\left(x_i|\hat{\mathbf{a}}_l^{(k-1)}\right)}. \tag{6.15}$$

Indeed $\hat{z}_{ij}^{(k)}$ is a real number $\in [0, 1]$ and not necessarily binary. This arises from the fact that, in general, it is not possible to determine the origination of the label components unequivocally.
Summation of the $\hat{z}_{ij}^{(k)}$ leads directly to the mixture parameters. The mixture parameter of the first mixture component is

$$\hat{\alpha}_1^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{i1}^{(k)}. \tag{6.16}$$

Since we analyse a mixture density which consists of only two components, the mixture parameter of the second component can be calculated as $\alpha_2 = 1 - \alpha_1$ and it is possible to reduce the whole notation to a single mixture parameter α.

M-step:

After the estimation of $z_{ij}^{(k)}$, the M-step realizes the estimation of the distribution parameters of both components by weighted pseudo maximum likelihood estimation, where the survey weights multiplied with the associated component labels are the weights of the estimation procedure. For each mixture component

$$\mathbf{a}_j^{(k)} := \underset{\mathbf{a}_j}{\operatorname{argmax}} \sum_{i=1}^{n} z_{ij} w_i \log f_j(x_i|\mathbf{a}_j) \tag{6.17}$$

has to be determined. With given component labels, $\mathbf{a}_1^{(k)}$ and $\mathbf{a}_2^{(k)}$ minimize the equation (6.14).

The EM algorithm has some very desirable properties, but most of them are not of direct importance for the matters of this study and can be found in (DEMPSTER et al., 1977 and REDNER and WALKER, 1984). Its key property is that it improves the parameter estimation with every step and the associated likelihood function L converges to a value L^*. Unfortunately L^* is not necessarily the global maximum of L. The question whether the global maximum is reached, depends highly on the used starting values for all parameters. Because of that, it is essential to find sufficiently *good* starting values, which is a highly non-trivial problem. One approach applicable in our simulation study can be found in 6.4.2.

6.3 Numerical calculation of inequality measures of the TCD

There are three monetary poverty or respectively inequality measures estimated in the simulation study in section 6.4. The Gini coefficient (for short Gini), the quintile share ratio (QSR) and the at-risk-of-poverty rate (ARPR). All of them are in the set of indicators of povery and social cohesion, formerly known as Laeken indicators, used by the European Commission. For rather complex continuous distributions like the TCD, there are in general no closed formulae for these indicators. That is why they have to be computed numerically. The detailed calculation methodology, also needed in 6.4, is explained in this section.

6.3.1 The Gini coefficient

The calculation of the Gini coefficient for continuous distributions can be a rather complex task because obtaining the Lorenz curve of an extensive function is a highly non-trivial challenge. Because of that, it seems reasonable to avoid the calculation of the Lorenz curve of the TCD if possible. As a matter of fact, there exists an old formula, invented by Gumbel in 1929, for the calculation of the Gini of a continuous distribution, which does not require an explicit specification of its Lorenz curve (see GUMBEL, 1929):

Let $F(x)$ denote the cdf of a continuous density function $f(x)$ with mean μ and domain (a, b). Then its Gini coefficient can be calculated as

$$\mathrm{G} = 1 - \frac{a}{\mu} - \frac{1}{\mu}\int_a^b [1 - F(x)]^2\, dx. \tag{6.18}$$

Since the cdf, the mean and the domain (i.e. $(0, \infty)$) of the TCD are known, its Gini coefficient can be calculated with the following formula:

$$\mathrm{G}_{TCD} = 1 - \mu_{TCD}\int_0^\infty \left(1 - \alpha\left(1 + \left(\frac{x}{b}\right)^{-a}\right)^{-p} + (\alpha - 1)\left(1 + \left(\frac{x}{b_2}\right)^{-a_2}\right)^{-p_2}\right)^2 dx. \tag{6.19}$$

The integral in (6.19) has to be calculated numerically and the infinity in the upper bound has to be substituted with an adequate finite value. For the calculation of this integral we use the R function **integrate**.

6.3.2 The quantile function of the TCD

The calculation of some of the inequality and poverty measures, by name, the quintile share ratio and the at-risk-of-poverty rate, requires the calculation of quantiles. As mentioned in 6.2.1, the quantile function of the TCD does not exist analytically. That is why its quantiles have to be computed numerically. Since the distribution function

$$F_{TCD}(x; a, b, p) = \alpha\left(1 + \left(\frac{x}{b}\right)^{-a}\right)^{-p} + (1 - \alpha)\left(1 + \left(\frac{x}{b_2}\right)^{-a_2}\right)^{-p_2} \tag{6.20}$$

is monotonically increasing and its domain is limited to the interval $[0, 1]$, it is an easy task to solve the equation

$$\alpha\left(1 + \left(\frac{u}{b}\right)^{-a}\right)^{-p} + (1 - \alpha)\left(1 + \left(\frac{u}{b_2}\right)^{-a_2}\right)^{-p_2} - Q = 0 \tag{6.21}$$

for a given $Q \in (0, 1)$ with respect to u. The resulting u then is the Q-quantile, e.g. in the case of $Q = 0.5$, u would be the median of the distribution.

6.3.3 The quintile share ratio

The general definition of the QSR of a continuous function $f(x)$ with $u_1 = F^{-1}(0.2)$ and $u_2 = F^{-1}(0.8)$ is

$$\mathrm{QSR} = \frac{0.2\int_{u_2}^{\infty} x f(x)dx}{0.2\int_{-\infty}^{u_1} x f(x)dx} = \frac{\int_{u_2}^{\infty} x f(x)dx}{\int_{-\infty}^{u_1} x f(x)dx}. \tag{6.22}$$

For functions without an invertible cdf it is impossible to calculate u_1 and u_2 analytically. The inverse of a cdf is the quantile function, so the inverse of the TCD has no closed form expression. We tackle this problem by choosing $Q_1 = 0.2$ and $Q_2 = 0.8$ in (6.21) which leads to numerical results for u_1 and u_2. After doing this, u_1 and u_2 can be used in the equation (6.22). Since in the indefinite case

$$\int x f_{TCD}(x)dx = \frac{\alpha a p b^{-ap} x^{ap+1} \; {}_2F_1\left(p+1, p+\frac{1}{a}; p+\frac{1}{a}+1; -\left(\frac{x}{b}\right)^a\right)}{ap+1} + \frac{(1-\alpha)a_2 p_2 b_2^{-a_2 p_2} x^{a_2 p_2+1} \; {}_2F_1\left(p_2+1, p_2+\frac{1}{a_2}; p_2+\frac{1}{a_2}+1; -\left(\frac{x}{b_2}\right)\right)^{a_2}}{a_2 p_2 + 1} \tag{6.23}$$

it is possible to solve (6.22) either analytically or numerically, with both methods leading to the correct result.

6.3.4 The At-risk-of-poverty rate

The At-risk-of-poverty rate is defined as the share of a population with an income lower than 60% of its median income. In analogy to the calculation of the quintile limits in 6.3.3 we get the median x_{TCD}^{med} numerically as described in 6.3.2. Since the TCD's cdf is given by (6.20), the ARPR_{TCD} can be obtained directly as

$$\text{ARPR}_{TCD} = \alpha\left(1+\left(\frac{0.6x_{TCD}^{med}}{b}\right)^{-a}\right)^{-p} + (1-\alpha)\left(1+\left(\frac{0.6x_{TCD}^{med}}{b_2}\right)^{-a_2}\right)^{-p_2}. \tag{6.24}$$

6.4 The TCD in practice: A simulation study on the Amelia data set

6.4.1 General setup of the simulation study

After these rather theoretical remarks, this subsection deals with the TCD in practice. In the simulation study the Amelia equivalized disposable personal income data set restricted to positive incomes was used. Details about the Amelia data set can be found in ALFONS et al. (2011). The simulation results are presented in HULLIGER et al. (2011). The simulation study bases on a repeated drawing of samples according to the different designs and the estimation of the Gini, the QSR and the ARPR with the indirect and the direct approach, explained in section 6.1. The compared sampling designs, explained in MÜNNICH and ZINS (2011) are: Simple random sampling (design 1.2) and stratified random sampling (design 1.4a), with a regional indicator as stratification variable. For each design 1,000 samples of 6,000 households (approximatively 15,888 persons each) are drawn. Afterwards all non-positive incomes are eliminated. Finally, the indicators are estimated in the direct and the indirect way.

With regard to further analysis, it is necessary to provide also some variance estimators for all methods. For the direct approach linearisation methods are available. They are

explained in detail in Münnich and Zins (2011). For the Dagum mixture case of the indirect approach there are no linearisation methods developed yet. That is why we estimated the variances with a bootstrap routine with 50 replications per sample. The variance of the point estimators for each design were used as benchmarks for the variance estimators.

6.4.2 The generation of starting values for the EM algorithm

As already stated in 6.2.3, the result of a fit with the EM algorithm depends highly on the starting values for all parameters. For the simulation study we utilized a peculiarity of the Amelia data set: The income sample of the whole Amelia *continent* can be divided into subsamples coming from Amelia's four subregions. Afterwards, we recombined the subsamples optimally to two samples consisting of two subsamples each, which can be fitted by a single Dagum distribution each. In extenso: There are three possibilities to combine the four subregions to two doublesubregions (dsr) in the explained way:

1. dsr1: region 1 and region 2; and dsr2: region 3 and region 4
2. dsr1: region 1 and region 3; and dsr2: region 2 and region 4
3. dsr1: region 1 and region 4; and dsr2: region 2 and region 3

For each of these combinations we fitted single Dagum distributions to the doublesubregions and summed up the log-likelihood values. The parameters of the combination with the highest sum of log-likelihood values were taken as the starting values for the EM algorithm. The whole concept is based on the fact that it is plausible that the Dagum distribution provides a decently *good* fit to components of the whole sample.

Chapter 7

Summary and Discussion

The reason why parametric estimation may be useful, when empirical data and estimators are available is threefold: 1. to stabilize estimation; 2. to get insight into the relationships between the characteristics of the theoretical distribution and a set of indicators, e.g. by sensitivity plots; 3. to deduce the whole distribution from known empirical indicators, when the raw data are not available. Deliverable 2.1 addresses these points and conveys the experiences done within the AMELI project on the parametric estimation of the EU-SILC monetary indicators.

In Chapter 2, we give a general overview of the state-of-the-art in parametric estimation of income distributions. The literature points out that a specially useful distribution in this context is the Generalized Beta distribution of the second kind (GB2), derived by McDonald (1984). The focus of our study is thus on the GB2 which is a highly flexible four-parameter income distribution. Apart from the scale parameter, this distribution has three shape parameters: the first governing the overall shape, the second the lower tail and the third the upper tail of the distribution. These characteristics give to the GB2 a large flexibility for fitting a wide range of empirical distributions and it has been established that it outperforms other four-parameter distributions for income data (Kleiber and Kotz, 2003).

In Chapter 3, we present the basic properties of the GB2 distribution and give formulas for the indicators of poverty and inequality under the GB2. Our main developments are presented in Chapter 4. We have studied different types of estimation methods, taking into account the design features of the EU-SILC surveys. Pseudo maximum likelihood estimation, using either the full or the profile likelihood, is compared with a *nonlinear fit from the indicators*. We have seen that both methods of ML estimation give similar results, but that the optimization with the profile log-likelihood is much faster. The third estimation method, the *method of nonlinear fit from indicators* uses the GB2 assumption and direct estimates of the main indicators of poverty and inequality (ARPR, RMPG, QSR, Gini and median income) to reproduce the whole income distribution. It is shown that the empirical (direct) distribution is guessed to a good precision.

ML estimation tends to produce a bias in the estimates of ARPR and RMPG (see Tables 4.2 and 4.3). We have developed an ad hoc procedure for robustification of the sampling weights which markedly improves the bias in point estimates.

Variance estimation is done by linearization and different types of simplified formulas for the variance proposed in the literature are evaluated by simulation in Deliverable 7.1.

Chapter 5 focuses on the compounding property of the GB2 distribution. This property implies that the GB2 density can be seen as a mixture of component densities arising from the breaking down of the scale range into intervals. The intervals breakdown can be chosen arbitrarily. For each breakdown, there exist probabilities of the mixture components that reproduce the original GB2 density. It can be highly useful, when we wish to use the overall GB2 fit and adapt for subpopulations by adjusting the mixture probabilities. The advantage of this approach is that we can derive the component densities from the global (population) level using the global GB2 fit and then only readjust the probabilities of the components at the subpopulation level, without changing the components themselves. Because the the components are fixed, the iterative algorithm for searching the optimal probabilities is fast. Of course the way the components are chosen is crucial for the quality of the result. Further development could be to estimate the optimal breakdown and the probabilities by an EM algorithm in the spirit of Chapter 6.

The parametric methods described in Chapters 3 to 5 are programmed in R (R Development Core Team, 2011) and are accessible to the wide public through the GB2 package (Graf and Nedyalkova, 2010), which is part of the output of the AMELI project.

For the methods developed for the GB2, simulation results based on the AMELIA dataset will be presented in the simulation report in Deliverable D7.1 (WP7).

Chapter 6 presents a different approach, useful in the context of heterogeneous populations. The case considered here is the mixture of two Dagum distributions (i.e. GB2 with parameter $q = 1$). However, the difference with the method described in Chapter 5 is that at each step of estimation, the distribution parameters and the mixture parameters are re estimated by the EM algorithm.

This study shows that parametric estimation is perfectly feasible in the context of complex survey designs. It is provides insight into the data. One byproduct is 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.

Appendix A

Partial derivatives of the log density of the GB2 distribution

Knowing that $\partial y/\partial a = (1/a)y\log(y)$ and $\partial y/\partial b = (-a/b)y$, and denoting as ψ the digamma function (the derivative of the natural logarithm of the gamma function), the partial derivatives of the log density with respect to a, b, p and q are:

$$\begin{aligned}
\frac{\partial \log(f)}{\partial a} &= \frac{1}{a} + p\log(x/b) - (p+q)\log(x/b)\frac{y}{1+y},\\
\frac{\partial \log(f)}{\partial b} &= -\frac{a}{b}p + \frac{a}{b}(p+q)\frac{y}{y+1},\\
\frac{\partial \log(f)}{\partial p} &= \psi(p+q) - \psi(p) + \log(y) - log(1+y),\\
\frac{\partial \log(f)}{\partial q} &= \psi(p+q) - \psi(q) - log(1+y).
\end{aligned}$$

Let denote $g(y) = y/(y+1)$ and ψ' the derivative of the digamma function. Knowing that

$$\frac{\partial g(y)}{\partial a} = \frac{\partial y}{\partial a}\frac{1}{(y+1)^2} = \frac{y\log(y)}{a(y+1)^2},$$

and that

$$\frac{\partial g(y)}{\partial b} = \frac{\partial y}{\partial b}\frac{1}{(y+1)^2} = \frac{-ay}{b(y+1)^2},$$

we have:

$$\begin{aligned}
\frac{\partial^2 \log(f)}{\partial a^2} &= -\frac{1}{a^2} - \frac{(p+q)}{a^2}\frac{y\log^2(y)}{(y+1)^2},\\
\frac{\partial^2 \log(f)}{\partial a \partial b} &= -\frac{p}{b} + \frac{(p+q)}{b}\left[\frac{y}{y+1} + \frac{ylog(y)}{(y+1)^2}\right],\\
\frac{\partial^2 \log(f)}{\partial a \partial p} &= \frac{log(y)}{a}\frac{1}{(y+1)},
\end{aligned}$$

$$\frac{\partial^2 \log(f)}{\partial a \partial q} = -\frac{log(y)}{a}\frac{y}{(y+1)},$$
$$\frac{\partial^2 \log(f)}{\partial b \partial a} = \frac{\partial^2 \log(f)}{\partial a \partial b},$$
$$\frac{\partial^2 \log(f)}{\partial b^2} = \frac{ap}{b^2} - \frac{a(p+q)}{b^2}\left[\frac{y}{y+1} + \frac{ay}{(y+1)^2}\right],$$
$$\frac{\partial^2 \log(f)}{\partial b \partial p} = -\frac{a}{b}\frac{1}{(y+1)},$$
$$\frac{\partial^2 \log(f)}{\partial b \partial q} = \frac{a}{b}\frac{y}{(y+1)},$$

$$\frac{\partial^2 \log(f)}{\partial p \partial a} = \frac{\partial^2 \log(f)}{\partial a \partial p},$$
$$\frac{\partial^2 \log(f)}{\partial p \partial b} = \frac{\partial^2 \log(f)}{\partial b \partial p},$$
$$\frac{\partial^2 \log(f)}{\partial p^2} = \psi'(p+q) - \psi'(p),$$
$$\frac{\partial^2 \log(f)}{\partial p \partial q} = \psi'(p+q)$$
$$\frac{\partial^2 \log(f)}{\partial q \partial a} = \frac{\partial^2 \log(f)}{\partial a \partial q},$$
$$\frac{\partial^2 \log(f)}{\partial q \partial b} = \frac{\partial^2 \log(f)}{\partial b \partial q},$$
$$\frac{\partial^2 \log(f)}{\partial q \partial p} = \frac{\partial^2 \log(f)}{\partial p \partial q},$$
$$\frac{\partial^2 \log(f)}{\partial q^2} = \psi'(p+q) - \psi'(q).$$

Appendix B

Proofs to Chapter 5

B.1 Derivation of the GB2 as a compound distribution

It is instructive to derive the result in Equation (5.4). We have

$$f(x;a,b,p,q) = \frac{a^2}{b\,\Gamma(p)\,\Gamma(q)} J(x;a,b,p,q) \tag{B.1}$$

where $J(x;a,b,p,q) =$

$$\int_0^\infty x^{ap-1}\theta^{-ap+1}\theta^{-1}\exp\left(-(x/\theta)^a\right)(\theta/b)^{-aq-1}\exp\left(-(\theta/b)^{-a}\right)d\theta =$$

$$(x/b)^{ap-1}\int_0^\infty (\theta/b)^{-a(p+q)}\theta^{-1}\exp[-((x/b)^a+1)(\theta/b)^{-a}]\,d\theta$$

Letting $t = (x/b)^a + 1$ and changing the variable to $u = (\theta/b)^{-a}$, $du = -(a/b)(\theta/b)^{-a-1}d\theta = -a(\theta/b)^{-a}\theta^{-1}d\theta$, we obtain

$$\begin{aligned} J(x;a,b,p,q) &= \frac{(x/b)^{ap-1}}{a}\int_0^\infty u^{p+q-1}\exp(-tu)\,du \\ &= \frac{(x/b)^{ap-1}}{a}\frac{\Gamma(p+q)}{t^{p+q}} \\ &= \frac{(x/b)^{ap-1}}{a}\frac{\Gamma(p+q)}{((x/b)^a+1)^{p+q}} \end{aligned}$$

Introducing this expression into Equation (B.1), we obtain the GB2 density in Equation (5.1):

$$f(x;a,b,p,q) = \frac{a^2}{b\,\Gamma(p)\,\Gamma(q)}\frac{(x/b)^{ap-1}}{a}\frac{\Gamma(p+q)}{((x/b)^a+1)^{p+q}}$$

$$= \frac{a}{b\,B(p,q)} \frac{(x/b)^{ap-1}}{((x/b)^a+1)^{p+q}}$$

B.2 Derivation of the right tail discretization

The computation of $f_\ell(x)$ in Equation (5.9) and of the initial value of p_ℓ in Equation (5.8) is analogous to the computation of $J(x;a,b,p,q)$ in Equation (B.1).

$$\begin{aligned}
p_\ell &= \frac{1}{\Gamma(q)} \int_{u_{\ell-1}}^{u_\ell} u^{q-1}\exp(-u)\,du = P(u_\ell,q) - P(u_{\ell-1},q) \\
f_\ell(x) &= \frac{1}{p_\ell}\frac{a}{b\Gamma(p)\Gamma(q)}\left[(x/b)^{ap-1}\right] \int_{u_{\ell-1}}^{u_\ell} u^{p+q-1}\exp(-tu)\,du \\
&= \frac{1}{p_\ell}\frac{a}{b}\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\frac{(x/b)^{ap-1}}{t^{p+q}}\left[P(tu_\ell,p+q) - P(tu_{\ell-1},p+q)\right] \\
&= f(x)\frac{P(tu_\ell,p+q) - P(tu_{\ell-1},p+q)}{P(u_\ell,q) - P(u_{\ell-1},q)}
\end{aligned}$$

B.3 Derivation of the left tail discretization

Starting now from Equation (5.5), parameters a', b', p', q' and

$$t' = t'(y) = (y/b')^{a'} + 1 = (x/b)^{-a} + 1 = t'(x),$$

we can write new component densities in function of the inverse income y as:

$$f_\ell(y;a',b',p',q') = f(y;a',b',p',q')\frac{P(t'u'_\ell,p'+q') - P(t'u'_{\ell-1},p'+q')}{P(u'_\ell,q') - P(u'_{\ell-1},q')},$$

where t' is viewed as a function of y.

Changing to the variable $x = 1/y$, we obtain the new component densities $\tilde{f}_\ell(x)$:

$$\begin{aligned}
\tilde{f}_\ell(x) &= \frac{1}{x^2} f(\frac{1}{x};a',b',p',q')\frac{P(t'u'_\ell,p'+q') - P(t'u'_{\ell-1},p'+q')}{P(u'_\ell,q') - P(u'_{\ell-1},q')} \\
&= f(x;a,b,p,q)\frac{P(t'u'_\ell,p+q) - P(t'u'_{\ell-1},p+q)}{P(u'_\ell,p) - P(u'_{\ell-1},p)}
\end{aligned}$$

where t' is viewed as a function of x. This proves Equation (5.10).

Appendix C

An efficient algorithm for the computation of the Gini coefficient of the Generalized Beta Distribution of the Second Kind

Author: Monique Graf
(Published in JSM Proceedings, Business and Economic Statistics Section, Alexandria, VA: American Statistical Association, pages 4835-4843.)

Abstract

The analytical expression for the Gini coefficient of the Generalized Beta Distribution of the Second Kind (GB2) has been derived by McDonald (1984). This formula involves the computation of two generalized hypergeometric functions at $z = 1$, for which a direct evaluation can lead to a very slow convergence. The proposed algorithm selects among the ten Thomae (1879) equivalent representations the one with the fastest convergence. The gain can be extremely large. The implementation has been done in the open source language R.

Keywords: Income distribution; Gini coefficient; GB2 distribution; algorithm; convergence; R language; hypergeo package.

C.1 Introduction

Theoretical income distributions have attracted a lot of interest. A huge literature emerged and many equivalent distributions have appeared under different names. A encyclopedic overview of income and size distributions can be found in Kleiber and Kotz (2003). One of the main contributions of Kleiber and Kotz's book is the unification of the terminology. In this paper, their terminology will be followed. A panorama of the modeling of income distributions and inequality measures, from seminal papers to current research, has been published by Chotikapanich (2008). The Generalized Beta Distribution of the Second

Kind (GB2) is a four parameter distribution that has been introduced by McDonald (1984) as a flexible and widely applicable income distribution. It encompasses many distributions used in the context of incomes as special cases: Singh-Maddala, Dagum, Fisk, and the Generalized Gamma as a limiting case. Empirical findings, summarised in Kleiber and Kotz (2003), show that family income distributions are generally best fitted by the GB2 or one of its particular cases. McDonald and Xu (1995) have embedded the GB1 and the GB2 into a five parameter distribution, called Generalized Beta (GB) (see also McDonalds and Ransom, 2008, Chap.8), but did not derive the Gini index in this general case. Their empirical findings show that the GB2 fit is competitive with regard to the GB. Thus the GB2 (or its subdistributions) still seems to remain the generally best fitting parametric distribution to family income data.
Inequality can be assessed by several different indices. A widely used inequality index is the Gini index, defined by a ratio of expectations:

$$G = \frac{\mathrm{E}(|X - Y|)}{2\,\mathrm{E}(X)}$$

where X and Y are two independent identically distributed random variables. In the GB2 case, it takes the form of a linear combination of two generalized hypergeometric functions ${}_3F_2$ at $z = 1$, for which a direct evaluation can lead to a very slow convergence. The proposed algorithm selects, among the ten Thomae (1879) equivalent representations of the ${}_3F_2$, the one with the fastest convergence. The algorithm thus provides a more efficient evaluation.

In Section C.2 the principal characteristics of the Generalized Beta Distribution of the Second Kind and the formula for the Gini coefficient are recalled. Section C.3 states Thomae's theorem; the algorithm is described in Section C.4. Section C.5 concludes with evaluations and comparisons.

C.2 Generalized Beta Distribution of the Second Kind (GB2)

The Generalized Beta Distribution of the Second Kind is a four-parameter distribution and is denoted $GB2(a, b, p, q)$. Its density takes the form:

$$f_{GB2}(y; a, b, p, q) = \frac{|a|}{bB(p, q)} \frac{(y/b)^{ap-1}}{(1 + (y/b)^a)^{p+q}} \tag{C.1}$$

where $B(p, q)$ is the beta function, $b > 0$ is a scale parameter, $p > 0, q > 0$ and a real are shape parameters. The extension to a negative a parameter is unessential, because $GB2(-|a|, b, p, q) = GB2(|a|, b, q, p)$, as can be readily seen on multiplying the numerator and denominator of the density in Equation (C.1) by $(y/b)^{-a(p+q)}$. Moreover, the GB2 has been shown to be closed under inversion, i.e. if Y follows a $GB2(a, b, p, q)$, then $1/Y$ follows $GB2(a, 1/b, q, p)$ (see Kleiber and Kotz, 2003, Equation (6.14)). We see from this formula that $1/Y$ has the same shape parameter a as Y. Thus from now on, we suppose that $a > 0$.
The moment of order k exists, when

$$-ap < k < aq$$

The Gini coefficient is only defined when the expectation exists, that is when

$$q - 1/a > 0 \tag{C.2}$$

The formula for the Gini index involves the generalized hypergeometric function ${}_3F_2$, defined by

$${}_3F_2(U, L; z) = {}_3F_2\left[\begin{array}{lll} u_1, & u_2, & u_3 \\ & l_1, & l_2 \end{array}; z\right] = 1 + \sum_{n=1}^{\infty} \frac{(u_1)_n\,(u_2)_n\,(u_3)_n}{(l_1)_n\,(l_2)_n\,n!}\,z^n \tag{C.3}$$

where $(x)_n = \prod_{k=0}^{n-1}(x+k)$ is the Pochhammer's symbol, and $U = (u_1, u_2, u_3)$ and $L = (l_1, l_2)$ are the vectors defining the coefficients of the infinite series.
For $|z| = 1$, the series in Equation (C.3) converges absolutely, if

$$s = l_1 + l_2 - u_1 - u_2 - u_3 > 0 \tag{C.4}$$

(see e.g. Henrici, 1977). The parameter s is called the excess. Representing the Pochhammer's symbols as ratios of gamma functions, $(x)_n = \Gamma(x+n)/\Gamma(x)$, and using Stirling's formula, we can see that the series is (up to a constant not depending on n) asymptotic to n^{-s-1}. Thus the speed of convergence is directly related to the excess.

The Gini index of the GB2 distribution is given by (McDonald, 1984):

$$G_{GB2} = \frac{B(2p + 1/a, 2q - 1/a)}{B(p, q)B(p + 1/a, q - 1/a)} \left\{\frac{1}{p}\,G_1 - \frac{1}{p + 1/a}\,G_2\right\} \tag{C.5}$$

where

$$G_1 = {}_3F_2\left[\begin{array}{lll} 1, & p+q, & 2p+1/a \\ & p+1, & 2(p+q) \end{array}; 1\right] \tag{C.6}$$

and

$$G_2 = {}_3F_2\left[\begin{array}{lll} 1, & p+q, & 2p+1/a \\ & p+1+1/a, & 2(p+q) \end{array}; 1\right] \tag{C.7}$$

The parameters a, p, q in Equations (C.6) and (C.7) are all positive. Thus the convergence condition (C.4) translates to $s = q - 1/a > 0$ for G_1, and to $s = q > 0$ for G_2. The second condition is always fullfilled by hypothesis on the parameter space. The first condition is exactly the condition for the existence of the expectation given in Equation (C.4).

C.3 Thomae's Theorem

Thomae (1879) derived equivalent representations for ${}_3F_2(U, L; 1)$. His result is nicely expressed by Krattenthaler and Rao (2004) as the following theorem:

Thomae's theorem *The expression*

$$\frac{1}{\Gamma(l_1)\Gamma(l_2)\Gamma(l_1+l_2-u_1-u_2-u_3)}\,{}_3F_2\left[\begin{array}{ccc;1} u_1, & u_2, & u_3 & \\ & l_1, & l_2 & \end{array}\right] \tag{C.8}$$

is a symmetric function of the five arguments

$$\begin{aligned} g_1 &= l_1+l_2-u_2-u_3 \\ g_2 &= l_1+l_2-u_1-u_3 \\ g_3 &= l_1+l_2-u_1-u_2 \\ g_4 &= l_1 \\ g_5 &= l_2. \end{aligned}$$

When all parameters in ${}_3F_2$ are real, the argument in the third gamma factor in Equation (C.8) is the excess s. The condition $s>0$ implies that all $g_i>0, i=1,...,5$.
The function ${}_3F_2$ is invariant by permutations of (u_1,u_2,u_3) and of (l_1,l_2) , so it is easy to see that there are only 10 equivalent expressions for ${}_3F_2(U,L;1)$: the combinations of two among the five arguments above as possible candidates for the two components of the vector L. These 10 expressions are listed e.g. in MILGRAM (2006). Let $s_g=(1/2)\sum g_i$. The excess corresponding to a specific choice is given by

$$s_{ij}=s_g-g_i-g_j \tag{C.9}$$

C.4 Computation of the Gini coefficient in the GB2 case

The five arguments $g_1,...,g_5$, computed from the parameters in Equations (C.6) and (C.7), for G_1 and G_2 respectively, are:

	G_1 parameters	G_2 parameters
$g_1=$	$q+1-1/a$	$q+1$
$g_2=$	$p+2q-1/a$	$p+2q$
$g_3=$	$2p+q$	$2p+q+1/a$
$g_4=$	$p+1$	$p+1+1/a$
$g_5=$	$2(p+q)$	$2(p+q)$
$s_g=$	$3(p+q)+1-1/a$	$3(p+q)+1+1/a$

To each pair of lower arguments $L^{ij}=(g_i,g_j)$ corresponds another excess parameter s_{ij} given by Equation (C.9), see Table C.1 for G_1. The last combination $(4,5)$ is the original one in Equation (C.6) and it is clear that the excess s_{45} is never the largest possible (e.g. s_{34} is always larger). Let $1\le k_1<k_2<k_3\le 5$ be the 3 distinct integers different from $\{i,j\}$. Then the vector of upper parameters U^{ij} is given by $u_n=g_{k_n}-s_{ij}, n=1,2,3$.

Table C.1: The 10 possible lower arguments $L = (g_i, g_j)$ of ${}_3F_2$, corresponding excess and upper arguments for the equivalent representations of $G1$

	Lower arguments L^{ij}		Excess	Upper arguments U^{ij}		
(i,j)	$l_1 = g_i$	$l_2 = g_j$	s_{ij}	u_1	u_2	u_3
$(1,2)$	$q - 1/a + 1$	$p + 2q - 1/a$	$2p + 1/a$	$q - 1/a$	$1 - p - 1/a$	$2q - 1/a$
$(1,3)$	$q - 1/a + 1$	$2p + q$	$p + q$	$q - 1/a$	$1 - q$	$p + q$
$(1,4)$	$q - 1/a + 1$	$p + 1$	$2(p+q) - 1$	$1 - p - 1/a$	$1 - q$	1
$(1,5)$	$q - 1/a + 1$	$2(p+q)$	p	$2q - 1/a$	$p + q$	1
$(2,3)$	$p + 2q - 1/a$	$2p + q$	1	$q - 1/a$	p	$2(p+q) - 1$
$(2,4)$	$p + 2q - 1/a$	$p + 1$	$p + q$	$1 - p - 1/a$	p	$p + q$
$(2,5)$	$p + 2q - 1/a$	$2(p+q)$	$1 - q$	$2q - 1/a$	$2(p+q) - 1$	$p + q$
$(3,4)$	$2p + q$	$p + 1$	$2q - 1/a$	$1 - q$	p	$2p + 1/a$
$(3,5)$	$2p + q$	$2(p+q)$	$1 - p - 1/a$	$p + q$	$2(p+q) - 1$	$2p + 1/a$
$(4,5)$	$p + 1$	$2(p+q)$	$q - 1/a$	1	$p + q$	$2p + 1/a$

We suppose that $s_{45} = q - 1/a > 0$. The Thomae's representations with negative excess are discarded. Negative excess can occur if either (i) $1 - q < 0$ or (ii) $1 - p - 1/a < 0$ or (iii) $2(p+q) - 1 < 0$. It is easy to see that (i) and (iii) cannot occur simutaneously; the same for (ii) and (iii). Thus there will always be more than one feasible combination. Moreover, there will always be at least one combination (i,j) with $s_{ij} > s_{45}$ (s_{34} fullfills the condition). In conclusion, we can always improve the convergence by exchanging the original combination $(4,5)$ by the one with the maximum excess. Moreover, it is shown in the Appendix C.6.1 that only 4 combinations out of 10 need to be tested. More details on the optimal combination can be found in Appendix C.6.1. Once the optimal combination (i,j) is found, the correction factor $C = C_{ij}$ from Thomae's theorem (Equation C.10) is determined and multiplied by the ratio of beta functions in Equation (C.5).

$$C_{ij} = \frac{\Gamma(g_4)\Gamma(g_5)\Gamma(s_{45})}{\Gamma(g_i)\Gamma(g_j)\Gamma(s_{ij})} \tag{C.10}$$

The function hypergeo_series from R package *hypergeo*, Hankin (2008) has been used for ${}_3F_2$ evaluations. Extensive use of mathematical functions provided in the R language R Development Core Team (2011) is acknowleged. The description of the algorithm will be done for G_1 and is analogous for G_2.

Algorithm

1. Input a, p, q.

   G_1 case:

2. Compute U and L from Equation (C.6).

3. Choose the combination with maximum excess in Table C.1.

4. Compute ${}_3F_2$ for the chosen combination.

5. Compute the sum of the logarithm of the correction factor in Equation (C.10) and of the logratio of the beta functions appearing in Equation (C.5).

6. Similar steps are performed for G_2.

7. The Gini coefficient is computed by using Equation (C.5).

C.5 Results and Discussion

Table IV in McDONALD (1984) gives estimated distribution functions to the 1975 U.S. family income data and corresponding Gini coefficients. For the GB2, the estimated parameters are $a = 3.4977$, $p = 0.4433$, $q = 1.1372$ and the Gini is estimated at 0.352. The different feasible combinations (positive excess) for the same parameter set are shown in Table C.2, where niter is the number of iterations. Using the third combination $(i, j) = (1, 4)$, that gives the maximum excess in this application, the algorithm converged in 55 iterations for G_1 and 78 for G_2 (not shown). It can be seen (Table C.2) that the gain in efficiency is large, the number of iterations until convergence for the original combination $(i, j) = (4, 5)$ being 9914[1]. The tolerance has been set to 1e-07 in the function evaluating ${}_3F_2$. The resulting Gini is 0.35364 and is nearer to the Census estimate of 0.358.

The algorithm has been tested with the Fisk distribution, which is GB2 with $p = q = 1$. In this case, the Gini takes a very simple form: $G = 1/a$. If $p = q = 1$, the combination with maximum excess is always $(1, 4)$ (see Appendix C.6.2) and for this combination, $u_2 = 0$ (Table C.1). This implies that all coefficients in Equation C.3 are zero, except the first and ${}_3F_2 = 1$. Thus the convergence occurs in one iteration, whereas for the original combination 10000 iterations do not suffice. In the Fisk case, the algorithm automatically finds the closed form.

The lack of convergence is not the only numerical problem that can be encountered. In (McDONALD, 1984, Table III) the fit of a B2 distribution (which is GB2 with $a = 1$) to the 1970 U.S. family income data gave $p = 2.5556$, $q = 22.8234$ and $G = 0.355$. The feasible combinations are shown in Table C.3. It can be observed that for combination (1,3), C_{13} is of the order 10e+11, compensating ${}_3F_2$ converging to a value near zero. This implies rounding errors and the estimated $G_1 = 2.1755$ is far away from the correct value of $G_1 = 4.3504$. In this case, combination (4,5) gives a reasonable value for G_1 and the

[1]The maximum number of iterations has been set higher than it would be in practice to make the gain visible.

same Gini is found as by the optimal combination (1,4) although in a higher number of iterations. It is observed in Appendix C.6.1 that combination $(1, 3)$ never corresponds to the maximum excess. Moreover, one can see (Appendix C.6.3) that the maximum factor C_{ij} would occur when $g_i = g_j = s_{ij}$ and it can be observed in Table C.3 that indeed g_1, g_3 and s_{13} are the closest of all combinations. By contrast, the combination with maximum excess implies a discrepancy between s_{ij}, and g_i and g_j, thus won't give rise to comparatively high coefficients, except in extreme cases. One such case would be when q is large and $q - 1/a$ is near zero, but such occurrence is unlikely to appear in practice.

C.6 Appendix

C.6.1 Combinations in Table C.1 with maximum excess

We have the following relationships between the excesses s_{ij}, that are valid for the G_1 as well as for the G_2 parameters:

$$s_{45} < s_{34} \; ; \; s_{25} < s_{23} \; ; \; s_{35} < s_{23} \; ; \; s_{15} < s_{12} \tag{C.11}$$

thus all combinations involving g_5 can be discarded. Moreover all the excesses $s_{12}, ..., s_{34}$ are identical for G_1 and for G_2 and the following relations hold:

$$s_{12} + s_{34} = 2s_{13} = 2s_{24} \Rightarrow \max(s_{12}, s_{34}) \geq s_{13} \text{ and } s_{24} \tag{C.12}$$

$$s_{14} = s_{12} + s_{34} - 1 \Leftrightarrow s_{14} = \max(s_{12}, s_{34}) + \min(s_{12}, s_{34}) - 1 \tag{C.13}$$

Thus, if

- $\min(s_{12}, s_{34}) > 1 \Rightarrow s_{14} > \max(s_{12}, s_{34}) > s_{23}$, so the maximum excess is s_{14}.
- $\min(s_{12}, s_{34}) < 1 \Rightarrow s_{14} < \max(s_{12}, s_{34})$
  - If $\max(s_{12}, s_{34}) > 1 \Rightarrow$ the maximum excess is $\max(s_{12}, s_{34})$.
  - If $\max(s_{12}, s_{34}) < 1 \Rightarrow$ the maximum excess is $s_{23} = 1$.

When equalities occur, then there is more than one solution for which the maximum excess is attained. In any case, the maximum excess is greater than or equal to 1.

C.6.2 Special GB2 distributions

Special parameter values give rise to a simpler Gini formula (see KLEIBER and KOTZ, 2003, for an exposition of all the special cases). If one of the upper arguments vanishes, ${}_3F_2 = 1$, so that the algorithm converges in one iteration.

- Dagum : $q = 1$
  In this case, $s_{34} = 2 - 1/a > 1$, because by hypothesis $0 < q - 1/a = 1 - 1/a$, thus the maximum excess is s_{14} (see above). For the combination $(1, 4)$, in the G_1 expression, $u_2 = q - 1 = 0$ and the corresponding ${}_3F_2 = 1$.
- Singh-Maddalla : $p = 1$
  In this case, $s_{12} = 2 + 1/a > 1$, thus the maximum excess is also s_{14}. For the combination $(1, 4)$, in the G_2 expression, $u_1 = 1 - p = 0$ and the corresponding ${}_3F_2 = 1$.
- Fisk: $p = q = 1$
  In this case, the algorithm converges in one iteration for G_1 and G_2 and the exact value of Gini which is $1/a$ is returned.

C.6.3 Maximum C factor

From Equation (C.9), we see that $g_i + g_j + s_{ij} = s_g$ is constant. C_{ij} in Equation (C.10) is maximum, when $\Gamma(g_i)\Gamma(g_j)\Gamma(s_{ij})$ is minimum. Writing A, B, D for g_4, g_5, s_{45} respectively, and $A-x, B-y, D+x+y$ for g_i, g_j, s_{ij} (which is possible by Equation C.9), the logarithm of the above product of gamma's is expressed as

$$\log\Gamma(A-x) + \log\Gamma(B-y) + \log\Gamma(D+x+y)$$

Taking the partial derivatives with respect to x and y and denoting the logarithmic derivative of the Gamma function by ψ, we obtain

$$\begin{aligned} -\psi(A-x) + \psi(D+x+y) &= 0 \\ -\psi(B-y) + \psi(D+x+y) &= 0 \end{aligned}$$

On the positive range of the argument, the ψ function is monotonic. Thus the above system has only one solution, which is

$$A - x = B - y = D + x + y = s_g/3$$

It is easy to see that the eigenvalues of the Hessian are $\psi'(s_g/3)$ and $3\psi'(s_g/3)$ and are strictly positive, thus the above solution gives the minimum of $\Gamma(A-x)\Gamma(B-y)\Gamma(D+x+y)$. This implies that

$$\Gamma(g_4)\Gamma(g_5)\Gamma(s_{45})/[\Gamma(s_g/3)]^3$$

is a superior bound for C_{ij}.

Table C.2: Feasible combinations for $a = 3.4977$, $p = 0.4433$, $q = 1.1372$

	i	j	g_i	g_j	excess	u_1	u_2	u_3	C	${}_3F_2$	G_1	niter
1	1	2	1.8513	2.4318	1.1725	0.8513	0.2708	1.9885	2.0657	1.2072	2.4937	995
2	1	3	1.8513	2.0238	1.5805	0.8513	-0.1372	1.5805	2.6940	0.9257	2.4939	227
3	1	4	1.8513	1.4433	2.1610	0.2708	-0.1372	1.0000	2.5385	0.9824	2.4939	55
4	1	5	1.8513	3.1610	0.4433	1.9885	1.5805	1.0000	0.5207	4.6946	2.4446	10000
5	2	3	2.4318	2.0238	1.0000	0.8513	0.4433	2.1610	1.7910	1.3922	2.4935	2312
6	2	4	2.4318	1.4433	1.5805	0.2708	0.4433	1.5805	2.2915	1.0883	2.4938	271
8	3	4	2.0238	1.4433	1.9885	-0.1372	0.4433	1.1725	2.5774	0.9676	2.4939	86
9	3	5	2.0238	3.1610	0.2708	1.5805	2.1610	1.1725	0.2924	7.7233	2.2581	10000
10	4	5	1.4433	3.1610	0.8513	1.0000	1.5805	1.1725	1.0000	2.4927	2.4927	9914

First eight columns, see Table C.1.
C is the correction factor in Equation C.10; $(C)({}_3F_2) = G_1$.
niter is the number of iterations.

Table C.3: Feasible combinations for $a = 1$, $p = 2.5556$, $q = 22.8234$

	i	j	g_i	g_j	excess	u_1	u_2	u_3	C	${}_3F_2$	G_1	niter
1	1	2	22.8234	47.2024	6.1112	21.8234	-2.5556	44.6468	1.0973e+03	0.0040	4.3504	2339
2	1	3	22.8234	27.9346	25.3790	21.8234	-21.8234	25.3790	1.0406e+11	0.0000	2.1755	99
3	1	4	22.8234	3.5556	49.7580	-2.5556	-21.8234	1.0000	2.2800e+00	1.9080	4.3504	20
4	1	5	22.8234	50.7580	2.5556	44.6468	25.3790	1.0000	1.1710e-01	37.1499	4.3504	10000
5	2	3	47.2024	27.9346	1.0000	21.8234	2.5556	49.7580	1.1800e-02	367.6974	4.3377	10000
6	2	4	47.2024	3.5556	25.3790	-2.5556	2.5556	25.3790	1.3987e+01	0.3110	4.3504	57
8	3	4	27.9346	3.5556	44.6468	-21.8234	2.5556	6.1112	5.7357e+01	0.0758	4.3504	27
10	4	5	3.5556	50.7580	21.8234	1.0000	25.3790	6.1112	1.0000e+00	4.3504	4.3504	180

First eight columns, see Table C.1.
C is the correction factor in Equation C.10; $(C)({}_3F_2) = G_1$.
niter is the number of iterations.

Bibliography

Alfons, A., **Filzmoser, P.**, **Hulliger, B.**, **Kolb, J.-P.**, **Kraft, S.**, **Münnich, R.** and **Templ, M.** (**2011**): *Synthetic Data Generation of SILC Data*. Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Atkinson, A. B. and **Bourguignon, F.** (editors) (**2000**): Handbook of Income Distribution. Elsevier.

Bandourian, R., **McDonald, J.** and **Turley, R. S.** (**2002**): *A Comparison of Parametric Models of Income Distribution Across Countries and Over Time*. Luxembourg Income Study Working Paper No. 305.
URL http://www.lisproject.org/publications/liswps/305.pdf

Berberi, Z. and **Silber, J.** (**1985**): *The Gini Coefficient and Negative Income: a comment*. Oxford Economic Papers, 37, pp. 525–526.
URL http://www.jstor.org/pss/2663310

Biewen, M. and **Jenkins, S.** (**2005**): *A Framework for the Decomposition of Poverty Differences with an Application to Poverty Differences between Countries*. Empirical Economics, 30, pp. 331–358.

Brazauskas, V. (**2002**): *Fisher Information Matrix for the Feller-Pareto Distribution*. Statistics and Probability Letters, 59, pp. 159–167.

Burkhauser, R., **Feng, S.**, **Jenkins, S.** and **Larrimore, J.** (**2008**): *Estimating Trends in US Income Inequality using the Current Population Survey: the Importance of Controlling for Censoring*. Technical report, Institute for social and Economic research.
URL http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2008-25.pdf

Butler, R., **McDonald, J.**, **Nelson, R.** and **White, S.** (**1990**): *Robust and Partially Adaptive Estimation of Regression Models*. The review of economics and statistics, 72, pp. 321–327.

Chen, C.-N., **Tsaur, T.-W.** and **Rhai, T.-S.** (**1982**): *The Gini Coefficient and Negative Income*. Oxford Economic Papers, 34, pp. 473–476.
URL http://132.203.59.36/DAD/technical_notes/note12/Ref/CTR_1982.pdf

Chen, C.-N., **Tsaur, T.-W.** and **Rhai, T.-S.** (**1985**): *The Gini Coefficient and Negative Income: Reply*. Oxford Economic Papers, 37, pp. 527–528.

Chiappero-Martinetti, E. and **Civardi, M.** (**2006**): *Measuring Poverty within and between Population Subgroups*. Technical report, IRISS Working Paper 2006-06, CEPS/INSTEAD, Differdange, Luxembourg.
URL http://ideas.repec.org/p/irs/iriswp/2006-06.html

Chotikapanich, D. (editor) (**2008**): Modeling Income Distributions and Lorenz Curves. Springer: Economic Studies in Equality, Social Exclusion and Well-Being, Vol. 5, doi: 10.1007/978-0-387-72796-7.

Cowell, F. A. and **Victoria-Feser, M.-P.** (**2003**): *Distribution-free Inference for Welfare Indices under Complete and Incomplete Information*. Journal of Economic Inequality, 00, pp. 1–29.

Dagum, C. (**1977**): *A New Model of Personal Income Distribution: Specification and Estimation*. Economie Appliquée, 30, pp. 413–437.

Dagum, C., **Grenier, G.**, **Norris, D.** and **Bédard, M.** (**1984**): *Male-Female Income Distributions in Four Canadian Metropolitan Areas: An Application Using Personal Income Tax Data*. Statistics of Incomes and Relative Administrative Records, Washington: Department of the Treasury, pp. 103–110, W. Alvey et B. Kilss.

Dastrup, S. R., **Hartshorn, R.** and **McDonald, J. B.** (**2007**): *The Impact of Taxes and Transfer Payments on the Distribution of Income: A Parametric Comparison*. J. Econ. Inequal., 5, pp. 353–369.

Davison, A. (**2003**): Statistical Models. Cambridge University Press, 33-35 pp.

Dempster, A. P., **Laird, N. M.** and **Rubin, D. B.** (**1977**): *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society Series B (Methodological), 39, 1, pp. 1–38.

Eurostat (editor) (**2007**): Comparative EU Statistics on Income and Living Conditions: Issues and Challenges. Proceedings of the EU-Silc Conference (Helsinki, 6-8-November 2006), Methodologies and Working Papers, European Communities.

Eurostat (**2009**): *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.

Freedman, D. A. (**2006**): *On the so-called "Huber sandwich estimator" and "robust standard errors"* . The American Statistician, 60, pp. 299–302.

Graf, M. (**2007**): *Use of Distributional Assumptions for the Comparison of four Laeken Indicators on EU-SILC Data*. 56th Session ISI 2007.

Graf, M. (**2009**): *An Efficient Algorithm for the Computation of the Gini Coffcient of the Generalized Beta Distribution of the Second Kind*. JSM Proceedings, Business and Economic Statistics Section, Alexandria, VA: American Statistical Association, pp. 4835–4843.

Graf, M. and **Nedyalkova, D.** (**2010**): *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation*. R package version 1.0.
URL http://cran.r-project.org/web/packages/GB2/index.html

Gumbel, E. J. (**1929**): *Das Konzentrationsmaß*. Allgemeines Statistisches Archiv, 18, pp. 279–300.

Hankin, R. K. S. (**2008**): *The hypergeo Package: The hypergeometric function*.

Henrici, P. (**1977**): Applied and Computational Complex Analysis, vol. two: Special Functions - Integral Transforms - Asymptotics - Continued Fractions. Wiley Interscience.

Huber, P. J. (**1967**): *The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 221–233.

Huber, P. J. (**1981**): Robust Statistics. New York: John Wiley & Sons.

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**): *Report on the Simulation Results*. Research Project Report WP7 – D7.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Jenkins, S. P. (**2007**): *Inequality and the GB2 Income Distribution*. ECINEQ Working Paper Series, Society for the Study of Economic Inequality.
URL http://www.ecineq.org/milano/WP/ECINEQ2007-73.pdf

Jenkins, S. P. (**2008**): *Inequality and the GB2 Income Distribution*. ISER Working Paper 2007-(Revised May 2008), 12. Colchester: University of Essex.

Johnson, N. L., **Kotz, S.** and **Balakrishnan, N.** (**1995**): Continuous Univariate Distributions, vol. 2. New York: John Wiley, 2nd ed. ed.

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

Krattenthaler, C. and **Rao, S.** (**2004**): *Group Theoretical Aspects of Hypergeometric Functions*. **Gruber, B.**, **Marmo, G.** and **Yoshinaga, N.** (editors) Symmetries in Science, vol. XI, Kluwer.

Lilla, M. (**2007**): *Income Inequality and Education Premia*. Working Paper 2007-11, IRISS, Luxembourg.
URL http://ideas.repec.org/p/irs/iriswp/2007-11.html

Lumley, T. (**2010**): *survey: analysis of complex survey samples*. R package version 3.23-3.
URL http://cran.r-project.org/web/packages/survey/index.html

Luzi, O., **Waal, T. D.** and **Hulliger, B.** (**2007**): EDIMBUS. Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys.

Makdissi, P. and **Mussard, S.** (**2006**): *Decomposition of s-Concentration Curves*. Working Paper 2006-09, IRISS, Luxembourg.
URL http://ideas.repec.org/p/irs/iriswp/2006-09.html

McDonald, J. (**1984**): *Some Generalized Functions for the Size Distribution of Income*. Econometrica, 52 (3), pp. 647–663.

McDonald, J. (**1989**): *Alternative Beta Estimation for the Market Model using Partially Adaptive Techniques*. Communications in Statistics Theory and Methods, 16, pp. 4032–4058.

McDonald, J. B. and **Butler, R. J.** (**1987**): *Some Generalized Mixture Distributions with an Application to Unemployment Duration*. The Review of Economics and Statistics, 69, pp. 232–240.

McDonald, J. B. and **Butler, R. J.** (**1990**): *Regression Models for Positive Random Variables*. Journal of Econometrics, 43, pp. 227–251.

McDonald, J. B. and **Xu, Y. J.** (**1995**): *A Generalization of the Beta Distribution with Applications*. Journal of Econometrics, 66, pp. 133–152, erratum: *Journal of Econometrics*,69, 427-428.

McDonalds, J. B. and **Ransom, M.** (**2008**): *The Generalized Beta Distribution as a Model for the Distribution of Income: Estimation and Related Measures of Inequality*. **Chotikapanich, D.** (editor) Modeling Income Distributions and Lorenz Curves, Springer.

McLachlan, G. J. and **Krishnan, T.** (**2008**): The EM Algorithm and Extensions. John Wiley & Sons.

Milgram, M. (**2006**): *On Hypergeometric 3F2(1)*.
URL http://arxiv.org/ftp/math/papers/0603/0603096.pdf

Münnich, R. and **Zins, S.** (**2011**): *Variance Estimation for Indicators of Poverty and Social Exclusion*. Research Project Report WP3 – D3.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Mussard, S. (**2007**): *Between-Group Pigou-Dalton Transfers*. Working Paper 2007-02, IRISS, Luxembourg.
URL http://ideas.repec.org/p/irs/iriswp/2007-02.html

Mussard, S. and **Terraza, M.** (**2007**): *Décompositions des Mesures d'Inégalité : le cas des coefficients de Gini et d'entropie*. Working Paper 2007-03, IRISS, Luxembourg.
URL http://ideas.repec.org/p/irs/iriswp/2007-03.html

Neocleous, T. and **Portnoy, S.** (**2008**): *A Partially Linear Censored Quantile Regression Model for Unemployment Duration*. Working Paper 2008-07, IRISS, Luxembourg.
URL http://ideas.repec.org/p/irs/iriswp/2008-07.html

Pfeffermann, D. and **Sverchkov, M. Y.** (**2003**): *Fitting generalized linear model under informative sampling*. **Skinner, C.** and **Chambers, R.** (editors) Analysis of Survey Data, pp. 175–195, New York, USA: Wiley.

Prentice, R. (**1975**): *Discrimination among some parametric models*. Biometrika, 62, pp. 607–614.

R Development Core Team (**2011**): *R: A language and Environment for Statistical Computing*. ISBN 3-900051-07-0.
URL http://www.R-project.org

Redner, R. A. and **Walker, F. W.** (**1984**): *Mixture Densities, Maximum Likelihood and The EM Algorithm*. SIAM Review, 26 (2), pp. 195–239.

Skinner, C., **Holt, D.** and **Smith, T.** (editors) (**1989**): Analysis of Complex Surveys. New York, USA: Wiley.

Thomae, J. (**1879**): *Ueber die Funktionen, welche durch Reihen von der Form dargestellt werden*. Journal für die Reine und angewandte Mathematik, 87, pp. 26–74.

Van Kerm, P. (**2007**): *Extreme Incomes and the Estimation of Poverty and Inequality Indicators from EU-SILC*. IRISS-C/I Working Paper.
URL http://iriss.ceps.lu/documents/irisswp69.pdf

Victoria-Feser, M.-P. (**2000**): *Robust Methods for the Analysis of Income Distribution, Inequality and Poverty*. International Statistical Review, 68, 3, pp. 277–293.

Victoria-Feser, M.-P. and **Ronchetti, H.** (**1994**): *Robust Methods for Personal-Income Distribution Models*. The Canadian Journal of Statistics, 22,2, pp. 247–258.

Xu, K. (**2004**): *How Has the Literature on Gini Index Evolved in the Past 80 Years?* Technical report, Department of Economics Dalhousie University Halifax, Nova Scotia.
URL http://economics.dal.ca/RePEc/dal/wparch/howgini.pdf

Yu, K., **Van Kerm, P.** and **Zhang, J.** (**2004**): *Bayesian Quantile Regression: An application to the wage distribution in 1990s Britain*. Technical report, CEPS/INSTEAD, G.-D. Luxembourg.
URL http://iriss.ceps.lu/documents/irisswp51.pdf