Jordá, Vanesa; Sarabia Alzaga, José Maria; Jäntti, Markus

Working Paper

Estimation of income inequality from grouped data

LIS Working Paper Series, No. 804

Provided in Cooperation with:
Luxembourg Income Study (LIS)

Suggested Citation: Jordá, Vanesa; Sarabia Alzaga, José Maria; Jäntti, Markus (2020) : Estimation of income inequality from grouped data, LIS Working Paper Series, No. 804, Luxembourg Income Study (LIS), Luxembourg

This Version is available at:
https://hdl.handle.net/10419/247239

LIS
Working Paper Series

No. 804

Estimation of Income Inequality from Grouped Data

Vanesa Jordá, José María Sarabia, Markus Jäntti

December 2020

Luxembourg Income Study (LIS), asbl

Estimation of income inequality from grouped data*

Vanesa Jordá† José María Sarabia Markus Jäntti

University of Cantabria University of Cantabria Stockholm University

Abstract

Grouped data in the form of income shares have conventionally been used to estimate income inequality due to the lack of individual records. We provide guidance on the choice between parametric and nonparametric methods and its estimation, for which we develop the `GB2group` R package. We present a systematic evaluation of the performance of parametric distributions to estimate economic inequality. The accuracy of these estimates is compared with those obtained by nonparametric techniques in more than 5000 datasets. Our results indicate that even the simplest parametric models provide reliable estimates of inequality measures. The nonparametric approach, however, fails to represent income distributions accurately.

JEL Classification: D31, C13, C18

1 Introduction

The analysis of income distribution has a venerable history in economics. Its evolution has been considered essential in explaining not only the causes but also the potential consequences of inequality and poverty. The role of changes in income distribution on different socio-economic aspects, such as growth, consumption and human capital formation, is widely documented in the literature (see e.g Barro, 2000; Krueger et al. 2006). Much empirical research has also been directed at examining geographical differences in inequality and their

*We are grateful to Stephen Jenkins, Prasada Rao, Nora Lustig, Duangkamon Chotikapanich and participants at the UM Sustainability and Development Conference, the 33th Annual Congress of the European Economic Association and the IARIW 35th General Conference for helpful comments on earlier versions of this paper. Vanesa Jordá and Jose María Sarabia acknowledge financial support from the Ministerio de Economía y Competitividad (Project ECO2016-76203-C2-1-P).

†Corresponding author: Tel. +34 942202275. Fax. +34 942201603 . E-mail: jordav@unican.es. Department of Economics, University of Cantabria. Avda. de los Castros, s/n, 39005, Santander (Spain)

evolution over time, considering family structure, medical progress and technological change, to mention a few, as potential determinants of these two phenomena (Deaton, 2013; McLanahan and Percheski, 2008).

Income inequality would be relatively simple to estimate if individual records on personal or household income data were available. Unfortunately, much of the existing scholarship on economic inequality has been plagued by a lack of individual data. Nevertheless, the periodic release of certain summary statistics on the distribution of income has become relatively common. The World Bank's PovcalNet, the World Income Inequality Database (WIID) and the World Wealth and Income Database (WID) are the largest cross-country databases that provide grouped income/consumption data, typically including information on income and population shares. This type of grouped data depicts sparse points of the Lorenz curve, which makes defining a method to link those points an essential requisite for estimating inequality measures.

Much of the academic literature on the estimation of income inequality from grouped data deploys nonparametric techniques to approximate the shape of the Lorenz curve. Linear interpolation of income shares is the most common approach for constructing the so-called empirical Lorenz curve, from which inequality measures are obtained. With very few exceptions, the extant scholarship on the global distribution of income presents inequality trends based on this method (Bourguignon and Morrison 2002; Lakner and Milanovic, 2016; Niño-Zarazua et al., 2017). The popularity of this methodology is explained not only by its simplicity but also because it is argued that there is no need to impose any particular model to fit the empirical data. However this approach rests, albeit not explicitly, on a predefined distributional model. Indeed, it assumes that all individuals within a particular quantile have the same level of income, which is obviously not an accurate representation of the income distribution. As a result, relative inequality measures estimated with this method are lower bound approximations and the actual level of inequality is therefore underestimated (see, e.g., Kakwani, 1980).

Hence, to obtain reliable estimates of inequality measures, it is necessary to deploy a model which defines more plausible assumptions on the income distribution within income shares. Due to its flexibility, some authors have opted for kernel estimation, which avoids imposing a particular functional form on the distribution of income (Sala-i-Martin, 2006, Hong et al., 2019). However, the performance of this approach seems to be extremely sensitive to the bandwidth parameter, which might lead to significant biases in the estimates of poverty and inequality measures (Minoiu and Reddy, 2014).

Parametric models seem to be a suitable alternative to nonparametric techniques for estimating income distributions (Dhongde and Minoiu, 2013). Yet, this approach has hardly ever been used to estimate income inequality. The reason seems to be the need to make ex-ante assumptions on the shape of the distribution. If the choice is not a valid candidate

for representing the distribution of income, the estimates on inequality measures might be severely affected by misspecification bias. Despite this potential limitation, prior research suggests that the parametric approach outperforms other nonparametric techniques for estimating poverty indicators from grouped data (Dhongde and Minoiu, 2013; Bresson, 2009). However, systematic empirical research on the effectiveness of parametric models in estimating inequality measures is surprisingly scarce. Previous studies point towards an excellent performance of parametric models (Cowell and Metha, 1982; Shorrocks and Wan 2008), but these evaluations rely on single case studies and a limited range of distributions, so their findings should be treated with great caution.

Therefore, robust empirical evidence on the reliability of parametric and nonparametric estimates would cast valuable light on the relative merits of these approaches for estimating income inequality from grouped data. This paper explores the implications of using different econometric strategies for 5570 datasets, which cover more than 180 countries over the period 1867-2015. Out of the whole range of parametric distributions, we direct our attention at the generalised beta distribution of the second kind (GB2) and its particular and limit cases. Several distributions from this family have been used to estimate income distribution from grouped data (Chotikapanich et al., 2007; Jorda et al., 2014; Pinkovskiy and Sala-i-Martin, 2014) because it is acknowledged to provide an excellent fit to income data across different periods and countries (Feng et al., 2006, Hajargasht et al., 2012).

Our results show that the nonparametric approach performs very poorly in estimating income inequality. The GB2 distribution is confirmed as the best candidate for estimating income distributions, although the special cases in this family also lead to accurate estimates, which are more reliable than nonparametric estimates in virtually all cases. Even for bimodal income distributions, which are clearly misrepresented by the GB2 distribution, we find no evidence to support the preference for the lower bound approximation of inequality measures and kernel density estimates over parametric estimates.

This analysis therefore confirms that a common failing in much of the research on global inequality is a tendency to avoid using parametric functional forms. Most of those studies that do consider parametric models rely on simple two- or three-parameter distributions. Jorda and Nino-Zarazúa (2016) is the only study that uses the GB2 distribution for estimating the global distribution of income. Country-specific applications are more common, but still scarce (see Burkhauser et al., 2012; Jenkins et al., 2011; Feng et al. 2006). The lack of interest in this distribution may, we believe, be largely attributed to the fact that the efficient estimation of this model is far from straightforward. Seeking to incentivise the use of the GB2 distribution, our estimation procedure implemented in R is conveniently available in the `GB2group` package.

In the next section, we introduce the notation and describe how the grouped data have been generated. We then outline the GB2 distribution and its related models. The following

section discusses the estimation strategy based on minimum distance estimators in a context of limited information. Thereafter, we compare the survey Gini index with both the so-called lower bound of inequality and the estimates based on parametric functional forms. We also present some results for model competition between different functional forms of the GB2 family to assess their performance to estimate the Gini coefficient. We make use of individual records to examine the robustness of the results to inequality measures which are more sensitive to the lower part of the distribution. Monte Carlo simulation is used to compare the performance of the parametric approach and the lower bound approximations in estimating inequality levels of bimodal income distributions. The paper concludes by considering the practical implications of the study.

2 Estimating income inequality from grouped data

To define the estimation strategy, it is crucial to understand how the grouped data are generated. Let $\mathbf{x}$ be an *i.i.d.* random sample of size N from a continuous income distribution $f(x;\boldsymbol{\theta})$ defined over the support $H=[0,\infty)$, where $\boldsymbol{\theta}\in\boldsymbol{\Theta}\subseteq\mathbb{R}^k$, with $\boldsymbol{\Theta}$ being the parameter space. Assume that H is divided into J mutually exclusive intervals $H_j=[h_{j-1},h_j), j=1,\ldots,J$. Denote $c_j=\sum_{i=1}^{N}\mathbf{1}_{[h_{j-1},h_j)}(x_i)x_i/\sum_{i=1}^{N}x_i, j=1,\ldots,J$ as the proportion of total income held by individuals in the j^{th} interval and the cumulative proportion by $s_j=\sum_{k=1}^{j}c_k$. Let $p_j=\sum_{i=1}^{N}\mathbf{1}_{[h_{j-1},h_j)}(x_i)/N, j=1,\ldots,J$ denote the frequency of the sample $\mathbf{x}$ in the j^{th} interval and $u_j=\sum_{k=1}^{j}p_k$ the cumulative frequency. According to this scheme, income shares $(s_j, j=1,\ldots,J)$ are ordinates of the Lorenz curve corresponding to the abscissae $u_j, j=1,\ldots,J$.

Five or ten points of the Lorenz curve are publicly available for a large sample of countries. The Lorenz curve reports the proportion of income accruing to each cumulative share of the population, once incomes are arranged in increasing order. This curve is scale independent, so changes in the unit of measurement of the income variable, for instance, from dollars to thousand of dollars, have no impact on the shape of the curve. Minimum inequality is observed when $s_j=u_j, j=1,\ldots,J$, so the Lorenz curve corresponds to the diagonal from the origin to the point (1, 1), which is known as the egalitarian line. The Lorenz curve is a powerful tool for comparing and ordering distributions according to their inequality levels. If the Lorenz curve of one distribution lies nowhere below and somewhere above the curve of another distribution, the first distribution can be declared to be less unequal than the second (Marshall and Olkin, 1979).

To construct the Lorenz curve with the available information on income shares, a method must be defined for linking the pairs of points $(u_j,s_j), j=1,\ldots,J$. An intuitive approximation would be to interpolate the observed income shares linearly. A major drawback

of using linear interpolation is that these comparisons would be somewhat crude in that all individuals classified in a given population group are assumed to have the same income. Moreover, the Lorenz ordering is partial in the sense that not all distributions can be ranked. To provide a complete ordering of distributions, we need to rely on inequality measures. The Gini index is the main indicator used to measure income inequality mainly due to its intuitive interpretation in terms of the area between the Lorenz curve and the egalitarian line. The nonparametric estimation of the Gini index is defined as twice the area between the egalitarian line and the Lorenz curve obtained by linear interpolation:

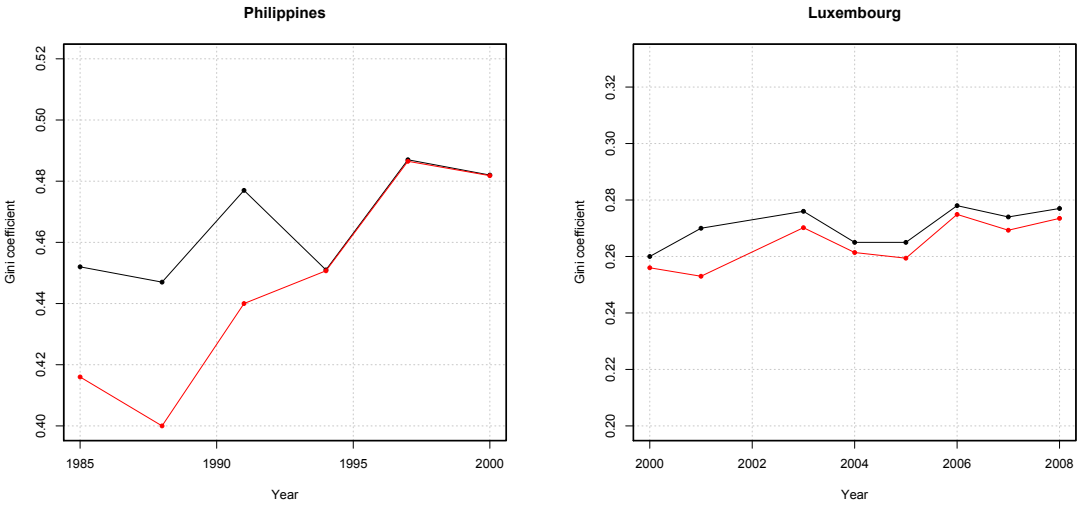
$$G(s_j, u_j) \approx 1 - \sum_{j=1}^{J} (s_j + s_{j-1})(u_j - u_{j-1}). \tag{1}$$

The main limitation of computing the Gini index with the above formula is that it yields biased estimates of inequality because its construction is based on the assumption that all individuals in a given population group get the same income. Hence, this formula is interpreted as the lower bound of the Gini coefficient, which neglects the variation within income shares (Cowell, 2011).

This kind of analysis might yield biased estimates on inequality, but it is still expected to provide valuable information. Indeed, the lower-bound approximation is deemed to be useful because if an upward trend is observed, it could be assured that the actual level of inequality would also rise. Moreover, with optimal grouping, the bias is expected to be relatively small for observations with more than five data points (Davies and Shorrocks, 1989).[1] Nevertheless, the empirical evidence based on this methodology is problematic in several ways. Firstly, the groups are often not optimally selected and, more importantly, the result above is obtained for the particular distribution of Canada. This finding might not necessarily match the distributional dynamics of other countries, so the bias may be considerably higher than expected. Secondly, since the size of the bias might vary over time, it is not possible to obtain conclusions about the overall evolution of income inequality, even in those cases that exhibit an ascending trend. This is illustrated in Figure 1, which shows the trend in the survey Gini coefficient in Luxembourg and the Philippines along with the lower bound of this measure, computed using Eq. (1). In the Philippines, the lower bound of the Gini index points to an increase in income inequality from 1991 to 1994, but the survey Gini index shows a downward trend during the same period. In Luxembourg, the survey Gini index rises one point from 2000 to 2001. The lower bound, however, falls from 0.256 to 0.253.

[1] Davies and Shorrocks (1989) develop an algorithm that maximises the value of the inequality index of interest to arrange the groups, which is equivalent to minimising the loss of distributional information due to grouping.

Figure 1: Estimates of the Gini coefficient using different estimation techniques in Luxembourg and the Philippines



Note: Black lines depict the evolution of the survey Gini index, red lines correspond to the evolution of lower bound of the Gini index .

Parametric models are a sound statistical method for estimating inequality from grouped data. The use of a parametric model seeks to define a more reliable approximation of the shape of the Lorenz curve between the observed income shares than a rough linear interpolation. However, it is key to chose a functional form that models the income distribution accurately. Out of the whole range of alternatives, the GB2 family of distributions seems to be the most appealing option.[2]

2.1 The generalised functions for the size distribution of income

The generalised functions for the size distribution of income proposed by McDonald (1984) include three well-known parametric models: the generalised beta of the first and the second kind (GB1 and GB2 respectively) and the generalised gamma (GG). Among them, the GB2 distribution seems to be particularly suitable for modelling income distributions. It is a general class of distributions that is acknowledged to provide an accurate fit to income data (Jenkins, 2009; McDonald and Xu, 1995; McDonald and Mantrala,1995). The GB2 can be defined in terms of the probability density function (pdf) as follows:

$$f(x;a,b,p,q) = \frac{ax^{ap-1}}{b^{ap}B(p,q)[1+(x/b)^a]^{p+q}}, \; x>0,$$

where $a,b,p,q>0$ and $B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}\,\mathrm{d}t$ is the beta function.

[2] For a comprehensive review on this topic, readers are referred to Kleiber and Kotz (2003).

Let $\mathcal{Z}$ be the class of all non-negative random variables with positive and finite expectation. For a random variable $X \in \mathcal{Z}$ with cumulative distribution function (cdf) $F(x;\boldsymbol{\theta})$, define $F_X^{-1}(y) = \inf\{x; F_X(x) \geq y\}$. The Lorenz curve associated with X is defined as (Gastwirth, 1971)

$$L_X(u) = \frac{\int_0^u F_X^{-1}(y)\,\mathrm{d}y}{\int_0^1 F_x^{-1}(y)\,\mathrm{d}y}, 0 \leq u \leq 1. \tag{2}$$

Following Sarabia and Jordá (2014), the Lorenz curve in Eq. (2) can also be expressed as,

$$L(u) = F_{X_{(1)}}(F_X^{-1}(u)), 0 \leq u \leq 1, \tag{3}$$

where $F_Y^{-1}(u)$ denotes the quantile function and $F_{X_{(1)}}(x) = (1/E(X))\int_0^x tf(t)\,\mathrm{d}t$ is the distribution of the first incomplete moment. To obtain the Lorenz, there is thus a need for closed expressions for the cumulative distribution function and the distribution of the first incomplete moment. These functions along with the kth moment and the Gini index are presented in Table 1.

Following Chotikapanich et al. (2018) and Arnold and Sarabia (2018), the Lorenz curve of the GB2 distribution is given by,

$$L_{GB2}(u;a,p,q) = B\left(B^{-1}(u;p,q); p+\frac{1}{a}, q-\frac{1}{a}\right), \;\; 0 \leq u \leq 1,$$

where $q > 1/a$ and $B^{-1}(x;p,q)$ is the inverse of the incomplete beta function ratio.

This model nests most of the functional forms used to model income distributions including the beta of the second kind (beta 2) when $a = 1$, used by Chotikapanich et al. (2012) to estimate the global distribution of income; the Singh-Maddala (2008) ($p = 1$) and the Dagum (1977) ($q = 1$) distributions, used by Hajargasht et al., (2012) and Bresson (2009).

The Lorenz curves of these distributions can be obtained using Eq. (3). The Lorenz curve of the second kind beta distribution can be expressed as follows:

$$L_{B2}(u;p,q) = B\left(B^{-1}(u;p,q); p+1, q-1\right), \;\; 0 \leq u \leq 1, q > 1. \tag{4}$$

Table 1: Cumulative distribution function, kth moment distribution, kth moment and Gini index for a selection of distributions of the GB2 family

Distribution	CDF	kth moment distribution	$E(X^k)$	Gini Index
GB2	$B\left(\frac{(x/b)^a}{1+(x/b)^a};p,q\right)$	$GB2\left(a,p+\frac{k}{a},q-\frac{k}{a}\right)$	$\frac{b^k B(p+\frac{k}{a},q-\frac{k}{a})}{B(p,q)},\ q>k/a$	see Eq. (7)
Beta 2	$B\left(\frac{x/b}{1+x/b};p,q\right)$	$B2(p+k,q-k)$	$\frac{b^k B(p+k,q-k)}{B(p,q)},\ q>k$	$\frac{2B(2p,2q-1)}{pB^2(p,q)},\ \ q>1.$
Singh-Maddala	$1-\left(1+\left(\frac{x}{b}\right)^a\right)^{-q}$	$GB2\left(a,1+\frac{k}{a},q-\frac{k}{a}\right)$	$\frac{b^k\Gamma(1+\frac{k}{a})\Gamma(q-\frac{k}{a})}{\Gamma(q)},\ q>k/a$	$1-\frac{\Gamma(q)\Gamma(2q-\frac{1}{a})}{\Gamma(q-\frac{1}{a})\Gamma(2q)},\ \ q>1/a.$
Dagum	$\left(1+\left(\frac{x}{b}\right)^{-a}\right)^{-p}$	$GB2\left(a,p+\frac{k}{a},1-\frac{k}{a}\right)$	$\frac{b^k\Gamma(p+\frac{k}{a})\Gamma(1-\frac{k}{a})}{\Gamma(p)},\ k/a<1$	$\frac{\Gamma(p)\Gamma(2p+\frac{1}{a})}{\Gamma(2p)\Gamma(p+\frac{1}{a})}-1,\ \ a>1.$
Lognormal	$\Phi\left(\frac{\log x-\mu}{\sigma}\right)$	$LN(\mu+k\sigma^2,\sigma)$	$\exp\left(k\mu+k^2\sigma^2/2\right)$	$2\Phi\left(\frac{\sigma}{\sqrt{2}}\right)-1.$
Fisk	$1-\left(1+\left(\frac{x}{b}\right)^a\right)^{-1}$	$GB2\left(a,1+\frac{k}{a},1-\frac{k}{a}\right),\ k/a<1$	$b^k\Gamma(1+k)\Gamma(1-k),\ k<1$	$\frac{1}{a},\ \ a>1.$

Source: Arnold and Sarabia (2018), Kleiber and Kotz (2003) and McDonald (1984).
Note: $B(v;p,q)=\int_0^v t^{p-1}(1-t)^{q-1}\,\mathrm{d}t/B(p,q)$ denotes the incomplete beta function ratio. The existence of kth moment distribution, defined as $F_{(k)}(x)=(\int_o^x t^k\,\mathrm{d}F(t))/(\int_0^\infty t^k\,\mathrm{d}F(t)), x>0$, requires the same constraints on the parameters as the kth moment and $E(X^k)<\infty$.

For the Singh-Maddala distribution, the Lorenz curve is given by the following equation:

$$L_{SM}(u;a,q) = B\left(1-(1-u)^{1/q}; 1+\frac{1}{a}, q-\frac{1}{a}\right), \;\; 0 \leq u \leq 1, q > 1/a, \tag{5}$$

and for the Dagum distribution it can be written as:

$$L_D(u;a,p) = B\left(u^{1/p}; p+\frac{1}{a}, 1-\frac{1}{a}\right), \;\; 0 \leq u \leq 1, a > 1. \tag{6}$$

We also consider in this study two-parameter distributions, including the Fisk (1961), which is a particular case of the GB2 making $p = q = 1$; and the lognormal distribution as a limit case of the GB2 distribution, which is one of the most popular candidates for modelling income variables (see e.g. Bresson 2009; Jorda et al., 2014).

Eq. (3) is used to obtain the Lorenz curve of the lognormal distribution as:

$$L_{LN}(u;\sigma) = \Phi(\Phi^{-1}(u) - \sigma), \;\; 0 \leq u \leq 1,$$

where $\Phi(\cdot)$ represents the cdf of the standard normal distribution and $\sigma > 0$.

For the Fisk distribution, the Lorenz curve is given by

$$L_F(u;a) = B\left(u; 1+\frac{1}{a}, 1-\frac{1}{a}\right), \;\; 0 \leq u \leq 1, \;\; a > 1.$$

Closed expressions of the Gini index for some special cases of the GB2 family are summarised in Table 1. For the GB2 distribution, the Gini coefficient is given by (McDonald, 1984),

$$G_{GB2} = \frac{B(2q-1/a, 2p+1/a)}{B(p,q)B(p+1/a, q-1/a)}\left(\frac{1}{p}J^{(1)} - \frac{1}{p+1/a}J^{(2)}\right), \tag{7}$$

where

$$\begin{aligned} J^{(1)} &= {}_3F_2\left(1, p+q, 2p+\frac{1}{a}; p+1, 2(p+q); 1\right), \\ J^{(2)} &= {}_3F_2\left(1, p+q, 2p+\frac{1}{a}; p+\frac{1}{a}+1, 2(p+q); 1\right), \end{aligned}$$

if $q > 1/a$, where ${}_3F_2(a_1, a_2, a_2; b_1, b_2; x)$ is a special case of the generalised hypergeometric function defined by

$${}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; x) = \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_q)_k} \frac{x^k}{k!},$$

where $(a)_k$ represents the Pochhammer symbol defined by $(a)_k = a(a+1)\cdots(a+k-1)$.

2.2 Estimation methods

Before going any further, it is key to consider how the groupings are generated. Hajargasht and Griffiths (2016) recognise two different data generating process (DGP) that yield different methods for grouping observations. In the first process, the proportion of observations in each group is specified before sampling, so that the population proportions (p_j) are fixed, whereas income shares (s_j) are random variables. The second type of DGP assumes pre-specified group boundaries (h_j) and, hence, generates random population proportions in each interval. We focus on the first type of DGP because it fits the structure of the largest datasets of grouped data, including the WIID and PovcalNet.

The estimation of parametric distributions from grouped data by maximum-likelihood using a multinomial likelihood function (see McDonald, 1984) would be misspecified under this type of DGP because of the non-stochastic nature of the group frequencies. Moreover, this estimation strategy requires information on the limits of the income groups (h_j), which is often unavailable. Due to this data limitation, non-linear least squares have been conventionally used to estimate the vector of parameters of interest, minimising the distance between income shares and the functional form of the Lorenz curve under the parametric assumptions made on the distribution of income. In this context, non-linear least squares can be referred the as equally weighted minimum distance (EWMD) estimator. Let X be a random variable in $\mathcal{Z}$, with cdf $F(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}$ and Lorenz curve $L(u;\boldsymbol{\theta})$. The estimation problem can be expressed as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \boldsymbol{M}(\boldsymbol{\theta})'\boldsymbol{M}(\boldsymbol{\theta}), \tag{8}$$

where $\boldsymbol{M}(\boldsymbol{\theta})' = [m_1(\boldsymbol{\theta}), \ldots, m_{J-1}(\boldsymbol{\theta})]$ is the vector of moment conditions, which takes the form

$$\boldsymbol{M}(\boldsymbol{\theta}) = L(\boldsymbol{u};\boldsymbol{\theta}) - \boldsymbol{s}, \tag{9}$$

where $\boldsymbol{s}' = (s_1, \ldots, s_{J-1})$ is a vector of cumulative income shares associated with the population proportions $\boldsymbol{u}' = (u_1, \ldots, u_{J-1})$.

As discussed above, the Lorenz curve is scale independent, so using Eq. (8), it is only possible to estimate the subset of $\boldsymbol{\theta}$ corresponding to the shape parameters.[3] The fact that only

[3] Hajargasht and Griffiths (2016) propose using the generalised Lorenz curve to define the moment conditions. The generalised Lorenz curve is the result of upscaling the ordinates of the Lorenz curve by mean income. With their approach, both scale and shape parameters can be estimated because the mean introduces scale into the model.

estimates on shape parameters can be obtained with this estimation procedure should not be interpreted as a limitation. Scale parameters are not needed to estimate relative inequality measures consistent with the Lorenz ordering, such as the Gini index or the Atkinson index. Therefore, if the interest lies in measuring relative inequality, this estimation strategy avoids the need to collect information on mean income. An additional advantage of this estimation strategy compared to the methods proposed in previous studies (see Hajargasht et al., 2012) is that the income limits of the groups (h_j) are not estimated. Thus the dimensionality of the optimisation function is substantially reduced, which makes numerical optimisation simpler, especially when the number of moments is large (Chen, 2018).

EWMD, however, overlooks the fact that the sum of the income shares is, by definition, equal to one, which introduces dependence between the income shares used in Eq. (8). EWMD thus yields inefficient although still consistent estimates of $\boldsymbol{\theta}$ and hence of the functions that depend on this set of parameters, including relative inequality measures. To gain efficiency, we also deploy the optimal minimum distance (OMD) estimator of the following form:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \boldsymbol{M}(\boldsymbol{\theta})' \boldsymbol{\Omega}^{-1} \boldsymbol{M}(\boldsymbol{\theta}), \tag{10}$$

It should be noted that Eqs. (8) and (10) are equivalent if $\boldsymbol{\Omega} = \boldsymbol{I}_{J-1}$. However, the identity matrix is not the optimal choice for $\boldsymbol{\Omega}$, which is why EWMD yields generally inefficient estimates. The optimal choice of the weighting matrix $\boldsymbol{\Omega}$ is the variance and covariance matrix of the moment conditions. Results from Beach and Davison (1983) and Hajargasht and Griffiths (2016) characterise the asymptotic distribution of $\sqrt{N}(L(\boldsymbol{u};\boldsymbol{\theta}) - \boldsymbol{s})$ as a multivariate normal distribution with zero mean and variance and covariance matrix of the form:

$$\boldsymbol{\Omega} = \boldsymbol{\Psi} \boldsymbol{W} \boldsymbol{\Psi}', \tag{11}$$

where

$$\boldsymbol{\Psi} = \left[\begin{array}{ccc:c} 1/\mu & \dots & 0 & -s_1/\mu \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1/\mu & -s_{J-1}/\mu \end{array}\right],$$

with $\mu = \int_0^\infty x f(x)\,\mathrm{d}x$. $\boldsymbol{W}$ is a symmetric matrix whose elements are

$$\{\boldsymbol{W}\}_{i,j} = \mu_i^{(2)} + (u_i h_i - \mu s_i)(h_j - u_j h_j + \mu s_j) - h_i \mu s_i, \text{ for } i \leq j$$

with

$$\mu_i^{(2)} = \int_0^{h_i} x^2 f(x)\,\mathrm{d}x.$$

Because we only have access to grouped data on income shares, it is not possible to compute

the variance and covariance matrix of the moment conditions. In order to obtain an efficient estimator of $\boldsymbol{\theta}$ from (10), we consider a two-step OMD estimator that uses the consistent estimates from EWMD (Eq. (8)) to compute a first stage estimate of $\boldsymbol{\Omega}$, which is used in the second stage to estimate Eq. (10).

The estimation of Eq. (8) involves the definition of starting values for the optimisation algorithm.[4] For the two-parameter distributions, which only have one shape parameter, we propose to solve the following equation to obtain an initial value of θ:

$$g = G(\theta),$$

where g is the sample Gini index, usually reported in the largest datasets of grouped income data, and $G(\theta)$ is the expression of the Gini index of the two-parameter distribution under consideration (see Table 1).

The distributions Singh-Maddala, beta 2 and Dagum are characterised by two shape parameters, which complicates the definition of non-arbitrary initial values. Conventionally, the estimates of a restricted model are taken as initial values. A potential limitation of this method is that as the dimensionality of the parameter space increases it becomes more difficult to achieve global convergence. Although it seems quite intuitive that the moment estimates of a restricted model might be a good starting point, the optimisation of the non-linear function in (8) could converge to a local minimum, which might lead to inaccurate estimates of the parameters and, hence, of inequality measures. The approach presented above for the two-parameter distributions is not feasible for these models in most cases because no information other than the Gini index and the income shares is reported. To provide several non-arbitrary combinations of starting values, we propose the following procedure:

1. Rewrite Eq. (8) using the Lorenz curve of the model to be estimated $L(u; \theta_1, \theta_2)$, which is given in Eq. (4) for the B2 distribution, in Eq. (5) for the Singh-Maddala distribution and in Eq. (6) for the Dagum distribution.
2. Define a grid of integer numbers for the starting values of θ_1, $\theta_1^{(s)} \in [1, 20]$.
3. Solve $g = G(\theta_1^{(s)}, \theta_2)$ for θ_2, to obtain $\theta_2^{(s)}$.
4. Estimate the Eq. (8) using $(\theta_1^{(s)}, \theta_2^{(s)})$, as initial values.
5. Keep the parameter estimates with the lowest residual sum of squares (RSS).

The routine described above enables us to obtain moment estimates of one of the parameters assuming that the other is equal to the grid value. These 20 combinations of initial values are used to undertake 20 different regressions using Eq. (8). Although we cannot ensure that

[4] We use the `optim` package in R to find the minimum of Eq. (8). The BFGS algorithm is implemented by default and L-BFGS is used when this method reports an error. The gradient is computed numerically.

our estimates belong to the global minima our proposed procedure covers a larger proportion of the parametric space than just using the moment estimates of a particular sub-model.

For the GB2 distribution, which has three shape parameters and one scale parameter, the estimates obtained for the three-parameter distributions are used. Eq. (8) is estimated using as initial values the 20 combinations of parameters from the beta 2 distribution, setting $a = 1$; the 20 initial values of the Singh-Maddala distribution with $p = 1$; and the ones obtained for the Dagum distribution, assuming $q = 1$. The estimation that reports the lowest RSS is saved.

For the estimation of $\boldsymbol{\Omega}$, the mean (μ), the second-order moment ($\mu_j^{(2)}$) and the income limits of each group(h_j) must also be computed. Therefore, a consistent estimate of the scale parameter is required. Let η denote the scale parameter of the distribution so that, $\boldsymbol{\theta} = (\eta, \boldsymbol{\lambda})'$. We propose that η is estimated by solving the following equation:

$$\bar{X} = \mu(\eta, \boldsymbol{\lambda}^*), \tag{12}$$

where $\bar{X}$ is the sample mean, $\boldsymbol{\lambda}^*$ are the EWMD estimates of the shape parameters from Eq. (8) and $\mu(\eta, \boldsymbol{\lambda}) = \int_{\mathbb{R}_+} x f(x; \eta, \boldsymbol{\lambda})\, \mathrm{d}x$. Closed expressions for $\mu(\eta, \boldsymbol{\lambda})$ for the distributions belonging to the GB2 family are presented in Table 1.

Let $\theta^* = (\eta^*, \boldsymbol{\lambda}^*)'$ be the consistent estimate of the parameters of the model obtained from Eqs.(8) and (12) used to obtain a first-stage estimate of the weighting matrix ($\boldsymbol{\Omega}^* = \boldsymbol{\Omega}(\theta^*)$). Because the first stage estimate (θ^*) is consistent, so is the weighting matrix $\boldsymbol{\Omega}^*$. Substituting $\boldsymbol{\Omega}^*$ in (10), gives the second-stage estimator of $\boldsymbol{\theta}$ as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \boldsymbol{M}(\boldsymbol{\theta})' \boldsymbol{\Omega}^{*-1} \boldsymbol{M}(\boldsymbol{\theta}). \tag{13}$$

Replacing $\boldsymbol{\Omega}$ by a consistent estimate of this matrix does not affect the asymptotic properties of the OMD estimator. It does however affect to the small-sample behavior of this estimator, which is generally biased (Altonji and Segal, 1996). Prior research based on Monte Carlo simulation suggests that the size of the bias depends on the underlying distribution of the data, being particularly large for heavy-tailed distributions (Altonji and Segal, 1996). The size of the bias also seems to increase with the number of overidentifying restrictions (Clark, 1996). This limitation is overcome as the sample size gets large (Hansen, 1982), but information on the sample size used to construct the grouped data is often unavailable. It is therefore recommended to deploy both estimators and opt for EWMD results if the parameter estimates differ substantially.

3 Results

3.1 Estimation of the Gini coefficient using grouped data

In this section, we explore some practical issues regarding the estimation of economic inequality from grouped data. To consider a diverse set of observations, we use the most comprehensive source of grouped income data: the WIID v3.4, released in January 2017. This database contains information from 8817 datasets for 182 countries over the period 1867-2015. Each dataset may report different types of information: the Gini index is generally provided (99.6% of the observations); less frequently, information is given on five to ten income shares (63.2%); finally, it is less common to report data on mean income (50.1%).

The WIID brings together a heterogeneous collection of datasets in terms of welfare concept, unit of analysis, equivalence scale, quality of data and population and area coverage. For this reason, it also includes metadata about these concepts along with information on the sources from which data are taken. Therefore, even though the WIID is notable in terms of geographical and time coverage, the lack of data comparability is often seen as a potential limitation. In this study, however, we take advantage of its heterogeneity to examine whether these features affect the performance of the different estimation methods in order to bring them to the attention of potential users.

The first question that arises when the parametric approach is deployed is: given that grouped data in the WIID is expected to come from surveys with reasonable sample sizes, is the asymptotically efficient OMD a superior method than the unbiased EWMD to estimate income inequality? To answer this question we estimate different parametric distributions belonging to the GB2 family using both econometric strategies. Parameter estimates are then used to compute the Gini index, which is compared with the observed Gini index reported in the survey. We opt for the method that yields more accurate estimates of the Gini coefficient because our interest lies in measuring income inequality.[5]

The optimisation function in Eq. (8) depends only on shape parameters, so EWMD can be deployed in the 5570 country/year datasets which present information on at least five income shares. An estimation of the shape parameters suffices to estimate the Gini index because this inequality measure is scale-independent. To obtain OMD estimates from Eq. (13), we also need information about the mean of the income distribution to estimate the scale parameter, which is used to construct a consistent estimate of $\mathbf{\Omega}$ (Eq.11). As a result,

[5] The fact that one method provides more accurate estimates of the Gini index does not mean that it is the most suitable econometric strategy for modelling relative inequality. To provide strong evidence in this regard, the robustness of this result to the consideration of different inequality measures must be examined. Unfortunately, other measures besides the Gini coefficient are rarely reported. However, we explore this issue further in Section 3.2 with individual records.

Table 2: Comparison of the performance of the OMD and the EWMD estimators of the Gini coefficient

	GB2	Beta 2	SM	Dagum	Lognormal	Fisk
Total	25.17%	24.31%	14.84%	8.99%	14.68%	10.91%
5 income shares	37.66%	52.50%	34.87%	32.47%	35.58%	25.66%
10 income shares	24.55%	22.92%	13.81%	7.73%	13.67%	10.10%

Note: Results based on the 3286 datasets from the WIID, 154 of which provide data on five income shares and 3132 on ten income shares. For all parametric distributions except the GB2, the Gini coefficient is estimated using the formulas in Table 1. The Gini index of the GB2 distribution is estimated by Monte Carlo simulation using samples of size $N = 10^6$.

this estimation method can be implemented only in 3286 country/year datasets from the WIID.

Table 2 presents the proportion of observations for which OMD yields more accurate estimates of the Gini index than EWMD. Our estimates for the GB2 distribution suggest that OMD reports more accurate estimates than EWMD for only 25% of the datasets. This proportion tends to decrease with the number of parameters of the distributions. By contrast, we observe that the OMD estimator outperforms EWMD by a proportionally greater extent when fewer income shares are considered. Therefore, in line with the findings presented in prior simulation studies, these two results confirm that the size of the bias increases with the number of overidentifying restrictions (Clark, 1996).

The second question of interest when examining inequality from grouped data is whether parametric functional forms provide better approximations of the Lorenz curve than the nonparametric methods used in prior studies, i.e. kernel density estimation and linear interpolation. The relevance of this question lies in the overwhelming number of studies that have opted for nonparametric techniques to estimate inequality measures (see Anand and Segal (2008) for a review). The popularity of this approach seems to be supported by the extended argument that parametric functional forms might lead to misspecification bias because ex-ante assumptions must be made about the shape of the income distribution and/or the Lorenz curve. To compare the performance of the GB2 distribution and the related sub-models for estimating income inequality with the nonparametric approach, conventional goodness-of-fit measures, such as the residual sum of squares, are not informative because linear interpolation is designed to perfectly match the income shares. Hence, as a goodness-of-fit measure, we consider the gap between the survey Gini coefficient and the estimated Gini indices using kernel density estimation, the lower bound approximation (Eq. (1)) and different parametric functional forms (Table 1).

To estimate parametric models, we focus on EWMD estimates because this method seems

to yield more accurate estimates of the Gini index.[6] Another advantage of analysing EWMD estimates is that a larger number of datasets can be examined because the estimation of Eq. (8) does not require information on the mean income. We reformulate Eq. (8) to estimate six parametric distributions that belong to the GB2 family for the 5570 country/year datasets with at least five income shares available. Once these models are estimated, we compute the estimated Gini indices and the lower bound of this inequality measure derived from linear interpolation of the Lorenz curve. This approximation of the Gini index assumes equality of incomes within shares. Hence, its value must be lower than the survey Gini index computed with individual records because it does not consider variation within income shares. We find, however, that this relationship is violated in 355 datasets. This incongruent result might have two potential explanations. Because Eq. (1) is an approximation of the lower bound, it has an inherent error that may lead to such inconsistencies. It could also be explained by the use of different data to estimate the Gini index and the income shares included in the WIID database. Hence, we opt for removing these datasets to facilitate the discussion of the results.[7]

Table 3 presents the difference between the survey Gini index and the parametric and nonparametric estimates. To facilitate the comparison of these two methodologies, we report the results in absolute value. Our estimates reveal that the lower bound yields a very poor approximation of the Gini index. The gap with the observed Gini index is more than 0.01 in 56% of the cases. Estimates based on kernel density estimation present larger estimation errors, which are above 0.01 in 75% of the datasets analysed. The parametric approach, however, provides much more accurate results, with substantially smaller differences between the estimated and the observed Gini indices. On average, lower bound estimates report an error three to four times larger than most parametric models. Among the parametric models, the GB2 seems to outperform the other sub-models, with estimation errors of less than 0.01 for 92% of the datasets. For the particular cases of this family, even the two-parameter distributions report fairly accurate Gini indices, which differ by less than 0.01 in 91% of the cases. Estimates of the Gini index with errors greater than 0.1 are more frequent for the nonparametric approach. All parametric specifications report the same proportion of estimates with differences larger than 0.1, the 0.6%, corresponding to three datasets: Mauritius in 1980 and Zambia in 2004 (rural and urban). In these three cases, the parametric and the nonparametric approaches report very similar estimates of the Gini index. For instance, in Mauritius the lower bound is 0.321 and the estimate for the GB2 distribution is 0.341, but the WIID reports a survey Gini coefficient of 0.457. Hence, we believe that the survey data of these datasets might be affected by some kind of measurement error.

[6] We present the results for the OMD estimates in Appendix A (Table A1), which seem to confirm the better performance of the parametric models in general, and the GB2 distribution in particular.

[7] Overall, the estimates for the whole sample with 5570 country/year datasets point to the same conclusions as for the restricted sample (results available upon request).

Table 3: Absolute error in the estimation of the Gini index using linear interpolation and different parametric distributions of the GB2 family.

Distribution	Mean	[0, 0.01)	[0.01, 0.02)	[0.02, 0.05)	[0.05, 0.1)	[0.1,)
Lower bound	0.0140	44.28%	39.50%	14.57%	1.42%	0.23%
KDE	0.0214	24.52%	33.33%	35.88%	5.92%	0.36%
GB2	0.0033	91.95%	4.89%	2.40%	0.71%	0.06%
B2	0.0040	91.68%	4.99%	2.53%	0.75%	0.06%
SM	0.0040	91.47%	5.25%	2.49%	0.73%	0.06%
Dagum	0.0041	91.31%	5.41%	2.51%	0.71%	0.06%
Lognormal	0.0043	91.26%	5.64%	2.24%	0.81%	0.06%
Fisk	0.0043	91.18%	5.45%	2.61%	0.71%	0.06%

Note: Results based on 5215 datasets from the WIID. Parametric models are estimated by EWMD. The lower bound of the Gini index is obtained using Eq. (1). Kernel density estimates (KDE) computed using a Gaussian kernel with optimal bandwidth (Silverman, 2018). For the parametric distributions the Gini index is estimated by Monte Carlo simulation using samples of size $N = 10^6$.

Although the results above point to a better performance of the parametric models than the nonparametric approach, only 16% of the observations show large deviations (higher than 0.02) between the lower bound and the observed Gini index. Therefore, it could be argued that the nonparametric approach provides researchers with an intuitive and fairly accurate tool for assessing inequality in most cases. However, since our sample includes Gini coefficients of very different magnitudes, the error should be evaluated in relative terms. Table 4 shows the difference between the observed and the estimated Gini coefficients relative to the value reported in the survey. These results strongly suggest that the lower bound yields highly inaccurate estimates of the Gini index, which is underestimated by more than 2% in 87% of the datasets.

These estimates reflect not only that linear interpolation provides a poor approximation of the Lorenz curve but also that parametric distributions lead to highly reliable estimates of the Gini index. The GB2 distribution seems to offer the best estimates, with 84% of estimations providing Gini coefficients that deviate by less than 1% from the survey Gini index. For the three-parameter functional forms, the figure drops to 80%. The two-parameter functional forms also present fairly accurate results for slightly more than 70% of the datasets.

Due to the inherent heterogeneity of the WIID in terms of welfare definition and data quality, another relevant question is whether these data characteristics affect the accuracy of the previous estimates. More importantly, does the estimation error decrease with the number of income shares? The answer to the last question is quite obvious for the nonparametric approach: the more income shares there are, the better the approximation of the Lorenz

Table 4: Relative error in the estimation of the Gini index using linear interpolation and different parametric distributions of the GB2 family

Distribution	[0%, 1%)	[1%, 2%)	[2%, 5%)	[5%, 10%)	[10%,)
Lower bound	3.16%	10.11%	72.66%	11.01%	3.07%
KDE	8.28%	9.73%	29.77%	45.07%	7.31%
GB2	84.39%	5.85%	6.40%	2.13%	1.23%
B2	77.01%	12.54%	6.98%	2.24%	1.23%
SM	80.33%	9.15%	7.11%	2.19%	1.23%
Dagum	80.25%	9.03%	7.40%	2.15%	1.17%
Lognormal	72.58%	16.03%	7.96%	2.21%	1.23%
Fisk	74.96%	14.06%	7.54%	2.26%	1.19%

Note: Results based on 5215 datasets from the WIID. Parametric models are estimated by EWMD. The lower bound of the Gini index is obtained using Eq. (1). Kernel density estimates (KDE) computed using a Gaussian kernel with optimal bandwidth (Silverman, 2018). For the parametric distributions except the GB2, the Gini coefficient is estimated using the formulas in Table 1. The Gini index of the GB2 distribution has been estimated by Monte Carlo simulation using samples of size $N = 10^6$.

curve becomes, hence the more reliable the estimate of the Gini coefficient is. For parametric models, however, five income shares might suffice to represent the shape of the Lorenz curve as rigorously as with ten data points.

Table 5 presents a summary of the absolute error in the estimation of the Gini index using the GB2 distribution, as the best parametric approximation of the Lorenz curve, linear interpolation, and kernel density estimation. We present the mean and the standard deviation (in parenthesis) of the difference in absolute terms between the survey and the estimated Gini coefficient for the four new welfare categories introduced in the WIID v3.4: consumption, disposable income, gross income and others;[8] and for different data quality levels: high, average, low and not known. To examine the effect of using a larger number of moments, we present these results broken down into five and ten income shares. In this regard, our results suggest that, when linear interpolation is used, the error in the estimation of the Gini index with only five income shares is two to three times higher than in datasets with ten data points. In the parametric framework, this pattern is not so obvious.

Overall, estimates performed with a larger number of income shares are found to be more accurate. The difference in the estimation error might be considerable in some categories,

[8] This new classification is a simplified version of the previous classification by welfare definition, which combines categories that are close to each other. See https://www.wider.unu.edu/sites/default/files/Data/WIID3.4 for a detailed description of the new labels.

Table 5: Absolute error in the estimation of the Gini index for different welfare definitions and quality standards

		Lower bound		KDE		GB2 distribution	
		10 shares	5 shares	10 shares	5 shares	10 shares	5 shares
Welfare definition	Consumption	0.0111	0.0292	0.0178	0.0165	0.0047	0.0158
	(857, 124)	(0.0097)	(0.0264)	(0.014)	(0.0216)	(0.0137)	(0.0329)
	Income, disposable	0.0124	0.0312	0.0228	0.0205	0.0031	0.0118
	(2686, 112)	(0.0085)	(0.0206)	(0.0155)	(0.021)	(0.0081)	(0.0184)
	Income, gross	0.0116	0.0272	0.0247	0.0239	0.0108	0.0121
	(384, 325)	(0.012)	(0.0162)	(0.0216)	(0.0211)	(0.015)	(0.0147)
	Other	0.0132	0.0329	0.0224	0.0279	0.0031	0.0081
	(1031, 51)	(0.0157)	(0.0158)	(0.0197)	(0.0286)	(0.0148)	(0.0151)
Data quality	Average	0.0108	0.0268	0.0197	0.0177	0.003	0.0075
	(1645, 139)	(0.0076)	(0.0159)	(0.0141)	(0.0193)	(0.0089)	(0.0135)
	High	0.0122	0.0324	0.0223	0.0246	0.0025	0.0135
	(2559, 167)	(0.0086)	(0.0243)	(0.0156)	(0.0224)	(0.0079)	(0.0298)
	Low	0.0159	0.0265	0.0272	0.0213	0.0125	0.0126
	(642, 261)	(0.021)	(0.0167)	(0.0255)	(0.0219)	(0.0225)	(0.0155)
	Not known	0.0131	0.0352	0.0185	0.031	0.0025	0.0228
	(112, 45)	(0.0076)	(0.0224)	(0.0128)	(0.0269)	(0.0029)	(0.0164)

Note: The number of datasets used to compute the mean and the standard deviation of the error in the estimation of the Gini index are presented in parenthesis below the label of the corresponding category, for ten and five income shares respectively. The GB2 distribution is estimated by EWMD and the Gini index is computed by Monte Carlo simulation using samples of size $N = 10^6$. The lower bound of the Gini index is computed using Eq. (1). Kernel density estimates (KDE) computed using a Gaussian kernel with optimal bandwidth (Silverman, 2018).

such as *disposable income* or *high data quality*, with estimation errors four to five times larger when five income shares are used. In other categories, however, the accuracy of the estimates does not seem to be affected by the number of moments. Low-quality datasets show, on average, estimation errors of the same magnitude, but with greater variation, for the estimation with ten data points. Hence, we might find larger estimation errors in datasets with ten income points than with five for this particular category. This result should not be interpreted as a recommendation to use five income shares for the estimation of parametric models with datasets of poor quality. Instead, this should be seen as an argument in favour of the parametric approach, which even with very few points of the Lorenz curve might yield reliable estimates.

Despite the fact that we are primarily interested in assessing income inequality, by comparing Gini indices alone we are not able to assert the supremacy of any parametric model. To provide a complete picture of the goodness-of-fit of the different parametric models, we now turn our attention to measures that assess the performance of nested models considering not only the accuracy but also the parsimony of the model by penalising for the number of parameters. Table 6 presents the proportion of observations for which the models in rows outperform the distributions in columns according to the Akaike information criterion (AIC), which, for the EWMD estimator in (8), takes the form (see Bresson, 2009)

$$aic = \frac{e^{2k/J}}{J}\boldsymbol{M}(\boldsymbol{\theta})'\boldsymbol{M}(\boldsymbol{\theta}),$$

where k is the number of parameters and $e^{2k/J}$ is a penalty term that increases with the numbers of parameters of the model.[9]

Our results suggest again that the GB2 distribution is the most suitable model for income and consumption variables, although the three-parameter models seem to be preferred in about 15 percent of the cases. As regards the three-parameter distributions, the beta 2 and the Dagum distributions seem to perform equally well, but the Singh-Maddala distribution seems to yield more accurate estimates in most cases. As expected, the two-parameter models rarely improve the goodness-of-fit of the GB2 and the three-parameter functional forms.

3.2 Estimation of distributionally-sensitive inequality measures from grouped data

So far, our analysis suggests that the GB2 family includes several models for obtaining reliable estimates of the Gini index. However, the analysis of income inequality rarely relies on

[9] The results for the Schwarz Bayesian information criterion can be found in Appendix A, Table A2.

Table 6: Goodness-of-fit matrix based on the AIC

	GB2	Beta 2	SM	Dagum	Lognormal	Fisk
GB2	..	87%	85%	88%	98%	98%
Beta 2	13%	..	42%	48%	96%	83%
Singh-Maddala	15%	58%	..	62%	93%	89%
Dagum	12%	52%	38%	..	87%	89%
Lognormal	2%	4%	7%	13%	..	44%
Fisk	2%	17%	11%	11%	56%	..

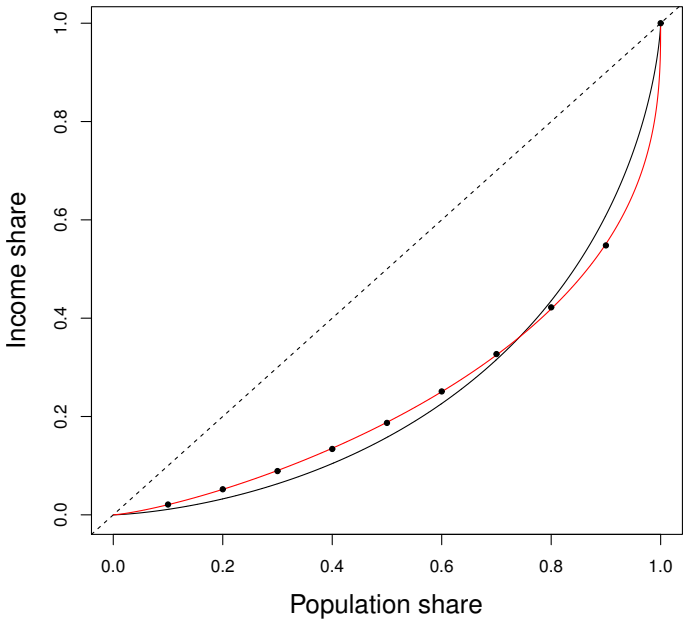
Note: Proportion of observations for which distributions in rows present better fit in terms of the AIC statistic than distributions in columns. Results based on 5215 datasets of the WIID. Parametric models are estimated by EWMD.

just one measure. Depending on the properties of the inequality measure and its sensitivity to the top or the bottom part of the distribution, different inequality indices may reflect diverging trends. Hence, assessments of the performance of different models in estimating income inequality should not be based solely on the Gini coefficient. Because this measure is proportional to the area between the Lorenz curve and the egalitarian line, differences below the observed income shares can be offset by overestimated income shares. This case is illustrated in Figure 2, which presents the survey income shares (black points) for Argentina in 1961 and the fits of the Singh-Maddala (red line) and the lognormal (black line) distributions. This graph reveals that the Singh-Maddala distribution provides a highly accurate fit and clearly outperforms the lognormal distribution. However, a comparison of the survey and the estimated Gini coefficients suggests a better performance of the lognormal distribution for estimating income inequality: the survey Gini index is 0.531 and the estimated Gini indices of the Singh-Maddala and the lognormal distributions are 0.516 and 0.522 respectively.

The apparently better performance of the lognormal distribution is, therefore, a statistical artifact caused by the manner in which the Gini coefficient is defined. If we compared the observed and the estimated values of inequality measures more sensitive to the left tail, the Singh-Maddala distribution would be declared as a superior model. For this kind of measures, the lognormal distribution would overestimate inequality levels because its Lorenz curve lies far below the sample income shares at the bottom of the distribution. Unfortunately, no measures other than the Gini coefficient are reported in the WIID.

To extend the insights about the estimation of income inequality from grouped data, we rely on data from the Luxembourg Income Study (LIS, 2020). The LIS database contains harmonised microdata on disposable income collected from nearly 50 countries for the period from 1980 to 2016. Using the 278 datasets of individual records available in the ten waves of the LIS database, we reconstruct grouped data with the same structure as the WIID: five and

Figure 2: Lorenz curve for Argentina (1961): Singh-Maddala (red) and lognormal (black) distributions



ten income shares, the mean and the Gini coefficient. These statistics are obtained following the methodological guidelines of the LIS.[10] We consider equivalised disposable income which is equal to household income divided by the square root of household size. We exclude all missing observations and records with zero disposable income. For the remaining sample, LIS proposes applying top and bottom coding. Equivalised income is bottom-coded at 1% of equivalised mean and top-coded at 10 times the median household income.[11] Finally, household weights are multiplied by household size to obtain person-adjusted weights.

The advantage of working now with individual data is that the analysis does not have to be restricted to the Gini coefficient. To examine the reliability of the parametric models in estimating different inequality measures, we also calculate the Atkinson index of the surveys

[10] For a detailed description of these guidelines see http://www.lisdatacenter.org/data-access/key-figures/methods/ and the R code used to compute of inequality measures can be downloaded from http://www.lisdatacenter.org/wp-content/uploads/files/access-key-programs-r-ineq.txt

[11] The aim of this section is to expand the results presented in Section 3.1. To that end, it is essential to replicate as accurately as possible the context of limited information under which those results were obtained. Hence, although bottom and top coding applied the income variable might be debatable, we apply this procedure not only because of LIS recommendations but because WIID data from LIS is reported with censoring. For this reason, even though we are deliberately introducing a potential source of measurement error, we do not consider the double censoring in the estimation of the parametric models because we do not have such information when using grouped data.

using the following expressions:

$$A_\epsilon = 1 - \left(\frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i}{\mu} \right)^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}, \epsilon \neq 1,$$

$$A_\epsilon = 1 - \frac{1}{\mu} \prod_{i=1}^{N} x_i^{1/N}, \epsilon = 1,$$

where ϵ is an inequality aversion parameter, which makes this measure more sensitive to the left tail of the income distribution as it increases.

The income shares from the LIS database are used to replicate the analysis in Section 3.1, comparing estimated and survey inequality measures. Tables 7 and 8 show the absolute error in the estimation of the Gini coefficient and the Atkinson indices from grouped data in the form of ten and five income shares respectively. Parametric models are estimated by EWMD following the procedure presented in Section 2.2.[12] Since the two-parameter distributions seem to lead to less reliable estimates of inequality measures, we only include the results of the lognormal distribution because this model has conventionally been employed to estimate the size distribution of income.[13]

Our results suggest that all functional forms seem to lead to very accurate estimates of the Gini coefficient, with estimates that differ from the observed value by less than 0.02. For the Atkinson index, the accuracy of the estimates seems to depend on the inequality aversion parameter. Our results suggest that the estimates of this inequality measure become less reliable as the value of this parameter increases. The GB2, the Singh-Maddala and the Dagum distributions show accurate estimates of the Atkinson measure for parameter values lower than one. When the sensitivity parameter is greater than 1, the measure is highly sensitive to the lower end of the distribution, meaning that the value of this inequality measure is largely influenced by the left tail of the distribution. Hence, even if the bulk of the distribution is adequately modeled by the parametric model, relatively small errors in the representation of the left tail might bias the estimates of the Atkinson index. Although the reliability of the estimates is inversely associated with the inequality aversion parameter, the GB2 and the Singh-Maddala distributions report relatively accurate estimates of the

[12] Results based on the estimation of parametric distributions by OMD are presented in Tables A3 and A4 in Appendix A. Our estimates suggest that the accuracy does not seem to be strongly affected by the estimation method in the case of the GB2 distribution, although, on average, EWMD estimates are generally more reliable than OMD. For the three-parameter distributions, however, OMD estimates of inequality measures highly sensitive to the bottom tail are more accurate than EWMD estimates. For the lognormal distribution, EWMD estimates seem to present much larger estimation errors than OMD for all inequality measures except for the Gini index.

[13] We have also computed the absolute error in the estimation of different inequality measures for the Fisk distribution. These results are available upon request.

Table 7: Absolute difference between estimated and observed inequality measures. Ten income shares

		Lower bound	KDE	GB2	SM	Dagum	Lognormal
	Mean	0.0071	0.0180	0.0007	0.0012	0.0014	0.0022
	[0, 0.01)	85.97%	11.51%	99.64%	98.92%	98.92%	100%
Gini	[0.01, 0.02)	14.03%	52.16%	0.36%	1.08%	1.08%	0%
Index	[0.02, 0.05)	0%	35.25%	0%	0%	0%	0%
	[0.05, 0.1)	0%	1.08%	0%	0%	0%	0%
	[0.1,)	0%	0%	0%	0%	0%	0%
	Mean	0.0066	0.0096	0.0020	0.0031	0.0038	0.0045
Atkinson	[0, 0.01)	98.56%	65.11%	98.92%	94.96%	88.13%	96.4%
index ($\epsilon = 0.5$)	[0.01, 0.02)	1.44%	31.65%	1.08%	3.96%	10.79%	3.6%
	[0.02, 0.05)	0%	2.88%	0%	1.08%	1.08%	0%
	[0.05, 0.1)	0%	0.36%	0%	0%	0%	0%
	[0.1,)	0%	0%	0%	0%	0%	0%
	Mean	0.0131	0.0261	0.0048	0.0065	0.0072	0.0119
	[0, 0.01)	27.7%	8.99%	89.93%	80.94%	75.18%	47.48%
Atkinson	[0.01, 0.02)	67.63%	16.19%	8.99%	17.27%	19.78%	38.13%
index ($\epsilon = 1$)	[0.02, 0.05)	4.68%	74.10%	1.08%	1.08%	3.96%	14.03%
	[0.05, 0.1)	0%	0.72%	0%	0.72%	1.08%	0.36%
	[0.1,)	0%	0%	0%	0%	0%	0%
	Mean	0.0294	0.0588	0.0160	0.0173	0.0183	0.0297
	[0, 0.01)	3.96%	2.16%	44.24%	38.49%	42.45%	20.5%
Atkinson	[0.01, 0.02)	27.7%	3.24%	23.74%	28.78%	25.54%	20.5%
index ($\epsilon = 1.5$)	[0.02, 0.05)	59.71%	29.14%	30.58%	28.42%	25.9%	40.65%
	[0.05, 0.1)	8.63%	60.43%	1.44%	3.6%	5.04%	17.27%
	[0.1,)	0%	5.04%	0%	0.72%	1.08%	1.08%

Note: Results based on 278 datasets of the LIS database. Parametric models are estimated by EWMD. All inequality measures are estimated by Monte Carlo simulation using samples of size $N = 10^6$. The lower bound of the Gini index is computed using Eq. (1). Kernel density estimates (KDE) computed using a Gaussian kernel with optimal bandwidth (Silverman, 2018).

Table 8: Absolute difference between estimated and observed inequality measures. Five income shares

		Lower bound	KDE	GB2	SM	Dagum	Lognormal
	Mean	0.0225	0.0157	0.0013	0.0017	0.0019	0.0027
	[0, 0.01)	0%	30.94%	98.92%	98.92%	98.92%	98.92%
Gini	[0.01, 0.02)	46.76%	53.60%	1.08%	0.72%	0%	1.08%
Index	[0.02, 0.05)	53.24%	12.59%	0%	0.36%	1.08%	0%
	[0.05, 0.1)	0%	2.88%	0%	0%	0%	0%
	[0.1,)	0%	0%	0%	0%	0%	0%
	Mean	0.0133	0.0085	0.0028	0.0036	0.0041	0.0048
Atkinson	[0, 0.01)	26.26%	82.37%	97.84%	92.45%	87.05%	96.76%
index ($\epsilon = 0.5$)	[0.01, 0.02)	58.99%	7.55%	1.8%	6.47%	11.87%	2.88%
	[0.02, 0.05)	14.75%	8.99%	0.36%	1.08%	1.08%	0.36%
	[0.05, 0.1)	0%	1.08%	0%	0%	0%	0%
	[0.1,)	0%	0%	0%	0%	0%	0%
	Mean	0.0247	0.0211	0.0055	0.0063	0.0064	0.0124
	[0, 0.01)	0%	15.11%	86.69%	82.37%	78.42%	43.53%
Atkinson	[0.01, 0.02)	32.37%	34.17%	12.59%	15.83%	17.27%	42.09%
index ($\epsilon = 1$)	[0.02, 0.05)	67.63%	48.92%	0.72%	1.8%	4.32%	14.03%
	[0.05, 0.1)	0%	1.80%	0%	0%	0%	0.36%
	[0.1,)	0%	0%	0%	0%	0%	0%
	Mean	0.0458	0.0513	0.0166	0.0163	0.0159	0.0302
	[0, 0.01)	0%	3.24%	43.53%	40.65%	45.68%	17.99%
Atkinson	[0.01, 0.02)	4.32%	7.91%	21.58%	27.7%	28.06%	21.22%
index ($\epsilon = 1.5$)	[0.02, 0.05)	56.47%	37.05%	33.09%	28.06%	21.22%	42.09%
	[0.05, 0.1)	38.49%	47.84%	1.8%	3.6%	5.04%	17.63%
	[0.1,)	0.72%	3.96%	0%	0%	0%	1.08%

Note: Results based on 278 datasets of the LIS database. Parametric models are estimated by EWMD. All inequality measures are estimated by Monte Carlo simulation using samples of size $N = 10^6$. The lower bound of the Gini index is computed using Eq. (1). Kernel density estimates (KDE) computed using a Gaussian kernel with optimal bandwidth (Silverman, 2018).

Atkinson index ($\epsilon = 1.5$), which differ by less than 0.02 in about 70% of the datasets.

A comparison of the figures in Tables 7 and 8 reveals that the error in the estimation of inequality measures is slightly larger if the estimates are obtained from 5 data points. However, even with five income shares, the GB2 family of income distributions yields reliable estimates of inequality measures, which confirms the claims made in Section 3.1 that parametric models provide an excellent methodological framework for estimating income inequality from grouped data, even when only a few points of the Lorenz curve are available.

3.3 Estimation of income inequality in bimodal distributions

A great deal of the criticism directed at the use of parametric models to estimate the Lorenz curve from grouped data is centered on the misspecification error that may arise as a consequence of imposing a particular functional form. Although the GB2 family is acknowledged to be an outstanding candidate to model income variables, it is only able to represent unimodal and zeromodal distributions. These are the expected shapes of the distribution of income in most countries: unimodal distributions are conventionally observed in developed countries with a well-established middle class, while zeromodal distributions are characteristic of developing countries, which have very high poverty rates. However, the conjunction of these two factors leads to bimodal distributions, which are typically observed in economies in transition.

Grouped data on a few points of the Lorenz curve are not informative enough to ascertain the number of modes of the distribution.[14] The parametric approach requires that the functional form of the distribution be defined ex-ante, so a conservative strategy is to estimate a general model that fits the regular features of the income distribution, typically with zero or one mode. When fitted to bimodal distributions, the GB2 family approximates the bimodality by unimodal/zeromodal functional forms, thus leading to inaccurate estimates. Although bimodal distributions are the exception rather than the rule, prior research has repeatedly emphasised the potential consequences of using parametric models in those cases, thus justifying the use of nonparametric techniques on grounds of reliability and practicability.

This section examines the size of the error in the estimation of inequality measures using the GB2 family when the income distribution has a bimodal pdf. We use Monte Carlo simulation to obtain synthetic samples from a mixture of a Weibull and a truncated normal distribution with pdf of the following form:

$$f(x_i;\beta,\alpha,\omega,\mu,\sigma)=\omega\frac{\beta}{\alpha^{\beta}}x_i^{\beta-1}\exp\left[-\left(\frac{x_i}{\alpha}\right)^{\beta}\right]+(1-\omega)\frac{\phi(x_i;\mu,\sigma^2)}{\Phi(\mu/\sigma)}, \tag{14}$$

where $\omega, 0\leq\omega\leq 1$, represents the mixing proportion of the Weibull distribution with scale parameter α and shape parameter β; $\phi(x,\mu,\sigma^2)$ is the pdf of a normal distribution with mean μ and variance σ^2 and $\Phi()$ represents the cdf of the standard normal distribution.

The pdf in (14) is used by Paap and van Dijk (1998) to estimate the cross-sectional distribution of income in 120 countries in six periods of time from 1960 to 1989. We rely on their estimates because, as observed in Figure A1, they depict bimodal shapes which are particularly representative of income variables. The simulated samples show a variety of shapes of the density function, ranging from a unimodal distribution with a heavy right tail

[14] Krause (2014) develops a method for determining whether a distribution has zero or one mode, but gives no insights on the potential bimodality of the distribution.

to a bimodal distribution where the two components of the mixture are clearly identified.

To illustrate the situation in which only grouped data is available, we compute five and ten income shares from the simulated samples of bimodal distributions to obtain limited information with the same structure as the WIID. We also compute the Gini index and the Atkinson measure setting $\epsilon = 0.5, 1, 1.5$. These values of the inequality measures are taken as a benchmark for evaluating the performance of the GB2 family to estimate income inequality in bimodal distributions. The *simulated* grouped data are used to estimate different models of the GB2 family by EWMD, deploying the estimation techniques described in Section 2.2. The corresponding Gini index and Atkinson measures are estimated by Monte Carlo simulation.

To assess the size of the error in the estimation of relative inequality measures, we calculate the absolute difference between estimated measures and those obtained from the bimodal distributions. A summary of this information is presented in Table 9, which shows the average error in the estimation of the Gini coefficient and the Atkinson index for different models of the GB2 family. We also rely on kernel density estimation and compute the corresponding lower bound of these measures to analyse whether the nonparametric approach leads to more accurate estimates when the underlying income distribution has two modes.

Nonparametric techniques are argued to let the data speak for themselves because they only need to make very weak assumptions on the distribution. Despite its flexibility, our estimates suggest that kernel estimation reports large estimation errors in bimodal distributions that are estimated using five or ten income shares. Our estimates also reveal that the gap between the estimation errors of the parametric and the lower bound is substantially narrower than in the previous results (Tables 3, 7 and 8), which are mostly based on unimodal distributions. As expected, the GB2 seems to provide more accurate estimates of inequality measures than the three-parameter models. Our results also suggest that this model yields, on average, more accurate estimates of the Gini index than the lower bound. This result is also observed for the Atkinson measures which are very sensitive to the left tail of the distribution. By contrast, when the value of the inequality aversion parameter is low, the lower bound yields more reliable estimates than the GB2 distribution. It is worth mentioning, however, that any of the estimation techniques systematically leads to more reliable estimates of the inequality measures.[15]

As mentioned above, the size of the estimation error using the lower bound approximation increases substantially if the estimation is based on five rather than ten income shares. It is therefore surprising that, on average, it still leads to a better approximation of the Atkinson index with $\epsilon = 0.5$ than the GB2 distribution. The estimation of the parametric models is dominated by the first mode, thus providing an accurate fit for the bottom part

[15] The complete results for the six simulated samples are available upon request.

Table 9: Average absolute difference between estimated and observed inequality measures: ten and five income shares

Ten income shares							
	Lower bound	KDE	GB2	Beta 2	SM	Dagum	Lognormal
Gini index	0.0085	0.0360	0.0071	0.0082	0.0092	0.0110	0.0057
Atkinson ($\epsilon = 0.5$)	0.0053	0.0260	0.0108	0.0153	0.018	0.0261	0.0093
Atkinson ($\epsilon = 1$)	0.0132	0.0283	0.0161	0.0268	0.0349	0.0535	0.0291
Atkinson ($\epsilon = 1.5$)	0.0331	0.0302	0.0259	0.0532	0.0739	0.1332	0.0576
Five income shares							
	Lower bound	KDE	GB2	B2	SM	DA	LN
Gini index	0.0308	0.0497	0.0181	0.0177	0.0199	0.0213	0.0149
Atkinson ($\epsilon = 0.5$)	0.0160	0.0324	0.0276	0.0258	0.0327	0.0359	0.0148
Atkinson ($\epsilon = 1$)	0.0315	0.0345	0.0119	0.0117	0.0210	0.0285	0.0293
Atkinson ($\epsilon = 1.5$)	0.0610	0.0300	0.0081	0.0109	0.0148	0.0392	0.0571

Note: Results based on simulated samples of size 10^4 of mixtures of a Weibull and a normal distribution with the following parameter values: (β, μ, α, σ, ω)= (2.02, 5.24, 1.4, 6.27, 0.7), (1.79, 6.68, 1.68, 6.5, 0.73), (1.63, 8.29, 2.03, 7.05, 0.73), (1.38, 10.66, 2.76, 3.13, 0.82), (1.35, 11.77, 2.95, 2.18, 0.82), (1.25, 13.32, 3.15, 3.02, 0.84). All inequality measures are estimated by Monte Carlo simulation using samples of size $N = 10^6$. The lower bound of the Gini index is computed using Eq. (1). Kernel density estimates (KDE) computed using a Gaussian kernel with optimal bandwidth (Silverman, 2018).

of the distribution at the expense of a relatively poor fit in the right tail of the distribution. Hence, the GB2 tends to yield more accurate estimates for the measures that are particularly sensitive to the left tail of the distribution.

4 Conclusions

Over the past few decades, there has been growing interest in the distributional patterns of income in both, economic literature and the international policy arena. The introduction of the Sustainable Development Goals has highlighted the relevance of this topic since Goal 10 calls for a decrease in income inequalities, thus positioning disparities as a key concern, not only because wellbeing is a prerogative of all citizens, but also because sustained development itself is impeded by high inequalities. Hence, addressing inequality trends has become essential, but individual data on income or consumption are often unavailable. Instead, grouped data from nationally representative surveys are used in most cases to assess the trends in inequality levels.

In this context of limited information, most prior research on global inequality relies on lower bounds of inequality measures, constructed under the assumption of equality of in-

comes within each income share. While being an intuitive method, it obviously leads to biased estimates of inequality measures. To provide reliable results, we must define more plausible assumptions on income dynamics within shares. In this paper, we have explored the practical implications of using parametric and nonparametric models to estimate income inequality from grouped data. We have focused first on the estimation of parametric models, comparing the performance of EWMD and OMD estimators to estimate the GB2 family of distributions. Our estimates reveal that EWMD yields more accurate estimates of the Gini index than OMD in most cases. Therefore, when the priority is to obtain unbiased estimates of inequality measures, EWMD should be preferred to OMD, even though this means sacrificing asymptotic efficiency.

One potential limitation of using parametric models is the requirement to impose a particular functional form to describe the income distribution, which could lead to biased estimates if the model is unable to represent income dynamics adequately. Indeed, misspecification bias has been the central argument in favour of using lower-bound approximations of inequality from grouped data. To address this issue, we have compared the performance of the GB2 family in estimating different inequality measures to kernel density estimation and the lower-bound approximation. Our results suggest that the parametric approach provides much more accurate results. Even two-parameter distributions yield more reliable estimates of inequality measures than the lower bound, although more complex models are generally preferred to the simplest ones. Only for bimodal distributions, do the lower bound and the parametric approach report estimates with similar precision.

Our estimates therefore suggest that much of the research on economic inequality relies on severely biased estimates. We show that the GB2 distribution provides an excellent approximation of the income distribution, which yields reliable estimates of relative inequality measures in virtually all cases. In the light of these findings, we expect the development of the `GB2group` R package, which deploys the estimation of this model from grouped data in the form of income shares by EWMD and OMD, to help promote the use of this family of distributions to obtain improved estimates on income inequality.

References

Altonji, J. G. & Segal, L. M. (1996), 'Small-sample bias in gmm estimation of covariance structures', *Journal of Business & Economic Statistics* **14**(3), 353–366.

Anand, S. & Segal, P. (2008), 'What do we know about global income inequality?', *Journal of Economic Literature* **46**, 57–94.

Arnold, B. C. & Sarabia, J. M. (2018), *Majorization and the Lorenz order with applications in applied mathematics and economics*, Switzerland: Springer.

Barro, R. J. (2000), 'Inequality and growth in a panel of countries', *Journal of Economic Growth* **5**(1), 5–32.

Beach, C. M. & Davidson, R. (1983), 'Distribution-free statistical inference with Lorenz curves and income shares', *The Review of Economic Studies* **50**(4), 723–735.

Bourguignon, F. & Morrisson, C. (2002), 'Inequality among world citizens: 1820-1992', *American Economic Review* **92**(4), 727–744.

Bresson, F. (2009), 'On the estimation of growth and inequality elasticities of poverty with grouped data', *Review of Income and Wealth* **55**(2), 266–302.

Burkhauser, R. V., Feng, S., Jenkins, S. P. & Larrimore, J. (2012), 'Recent trends in top income shares in the United States: reconciling estimates from March CPS and IRS tax return data', *Review of Economics and Statistics* **94**(2), 371–388.

Chen, Y.-T. (2018), 'A unified approach to estimating and testing income distributions with grouped data', *Journal of Business & Economic Statistics* **36**(3), 438–455.

Chotikapanich, D., Griffiths, W. E., Hajargasht, G., Karunarathne, W. & Rao, D. (2018), 'Using the GB2 income distribution', *Econometrics* **6**(2), 21.

Chotikapanich, D., Griffiths, W. E., Prasada Rao, D. & Valencia, V. (2012), 'Global income distributions and inequality, 1993 and 2000: Incorporating country-level inequality modeled with beta distributions', *Review of Economics and Statistics* **94**(1), 52–73.

Chotikapanich, D., Rao, D. & Tang, K. K. (2007), 'Estimating income inequality in china using grouped data and the generalized beta distribution', *Review of Income and Wealth* **53**(1), 127–147.

Clark, T. E. (1996), 'Small-sample properties of estimators of nonlinear models of covariance structure', *Journal of Business & Economic Statistics* **14**(3), 367–373.

Cowell, F. (2011), *Measuring inequality*, Oxford: Oxford University Press.

Cowell, F. A. & Mehta, F. (1982), 'The estimation and interpolation of inequality measures', *The Review of Economic Studies* **49**(2), 273–290.

Dagum, C. (1977), 'New model of personal income-distribution-specification and estimation', *Economie Appliquée* **30**(3), 413–437.

Davies, J. B. & Shorrocks, A. F. (1989), 'Optimal grouping of income and wealth data', *Journal of Econometrics* **42**(1), 97–108.

Deaton, A. (2013), *The great escape: health, wealth, and the origins of inequality*, Oxford and New Jersey: Princeton University Press.

Dhongde, S. & Minoiu, C. (2013), 'Global poverty estimates: A sensitivity analysis', *World Development* **44**(1), 1–13.

Feng, S., Burkhauser, R. V. & Butler, J. (2006), 'Levels and long-term trends in earnings inequality: overcoming current population survey censoring problems using the GB2 distribution', *Journal of Business & Economic Statistics* **24**(1), 57–62.

Fisk, P. (1961), 'Estimation of location and scale parameters in a truncated grouped sech square distribution', *Journal of the American Statistical Association* **56**(295), 692–702.

Gastwirth, J. L. (1971), 'A general definition of the Lorenz curve', *Econometrica: Journal of the Econometric Society* **39**(6), 1037–1039.

Hajargasht, G. & Griffiths, W. E. (2016), Inference for Lorenz curves, Technical report, The University of Melbourne.

Hajargasht, G., Griffiths, W. E., Brice, J., Rao, D. P. & Chotikapanich, D. (2012), 'Inference for income distributions using grouped data', *Journal of Business & Economic Statistics* **30**(4), 563–575.

Hansen, L. P. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica: Journal of the Econometric Society* **50**(4), 1029–1054.

Hong, S., Han, H. & Kim, C. S. (2019), 'World distribution of income for 1970–2010: dramatic reduction in world income inequality during the 2000s', *Empirical Economics* pp. 1–34.

Jenkins, S. P. (2009), 'Distributionally-sensitive inequality indices and the GB2 income distribution', *Review of Income and Wealth* **55**(2), 392–398.

Jenkins, S. P., Burkhauser, R. V., Feng, S. & Larrimore, J. (2011), 'Measuring inequality using censored data: a multiple-imputation approach to estimation and inference', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**(1), 63–81.

Jorda, V. & Niño-Zarazúa, M. (2016), Global inequality: How large is the effect of top incomes?, Technical report, WIDER Working Paper.

Jordá, V., Sarabia, J. M. & Prieto, F. (2014), 'On the estimation of the global income distribution using a parsimonious approach', *Research on Economic Inequality* **22**(1), 115–145.

Kakwani, N. (1980), 'On a class of poverty measures', *Econometrica: Journal of the Econometric Society* **48**(2), 437–446.

Kleiber, C. & Kotz, S. (2003), *Statistical size distributions in economics and actuarial sciences*, Vol. 470, New Jersey: John Wiley & Sons.

Krause, M. (2014), 'Parametric Lorenz curves and the modality of the income density function', *Review of Income and Wealth* **60**(4), 905–929.

Krueger, D. & Perri, F. (2006), 'Does income inequality lead to consumption inequality? Evidence and theory', *The Review of Economic Studies* **73**(1), 163–193.

Lakner, C. & Milanovic, B. (2016), 'Global income distribution: From the fall of the Berlin Wall to the Great Recession', *World Bank Economic Review* **30**(1), 203–232.

Luxembourg Income Study (LIS) Database (2020), *http://www.lisdatacenter.org (multiple countries; March 2020)*, Luxembourg: LIS.

Marshall, A. W., Olkin, I. & Arnold, B. C. (1979), *Inequalities: theory of majorization and its applications*, Vol. 143, New York: Academic Press.

McDonald, J. B. (1984), 'Some generalized functions for the size distribution of income', *Econometrica* **52**(3), 647–665.

McDonald, J. B. & Mantrala, A. (1995), 'The distribution of personal income: revisited', *Journal of Applied Econometrics* **10**(2), 201–204.

McDonald, J. B. & Xu, Y. J. (1995), 'A generalization of the beta distribution with applications', *Journal of Econometrics* **66**(1), 133–152.

McLanahan, S. & Percheski, C. (2008), 'Family structure and the reproduction of inequalities', *Annual Review of Sociology* **34**(1), 257–276.

Minoiu, C. & Reddy, S. G. (2014), 'Kernel density estimation on grouped data: the case of poverty assessment', *The Journal of Economic Inequality* **12**(2), 163–189.

Niño-Zarazúa, M., Roope, L. & Tarp, F. (2017), 'Global inequality: Relatively lower, absolutely higher', *Review of Income and Wealth* **63**(4), 661–684.

Paap, R. & Van Dijk, H. K. (1998), 'Distribution and mobility of wealth of nations', *European Economic Review* **42**(7), 1269–1293.

Pinkovskiy, M. & Sala-i Martin, X. (2014), 'Africa is on time', *Journal of Economic Growth* **19**(3), 311–338.

Sala-i Martin, X. (2006), 'The world distribution of income: falling poverty and convergence, period', *The Quarterly Journal of Economics* **121**(2), 351–397.

Sarabia, J. M. & Jordá, V. (2014), 'Explicit expressions of the Pietra index for the generalized function for the size distribution of income', *Physica A: Statistical Mechanics and its Applications* **416**(1), 582–595.

Shorrocks, A. & Wan, G. (2008), Ungrouping income distributions: Synthesising samples for inequality and poverty analysis, Technical report, Research Paper, UNU-WIDER, United Nations University (UNU).

Silverman, B. W. (2018), *Density estimation for statistics and data analysis*, London: Routledge.

Singh, S. & Maddala, G. S. (2008), *A function for size distribution of incomes*, New York: Springer Science & Business Media, pp. 27–35.

Appendix A

Table A1: Absolute error in the estimation of the Gini index using linear interpolation and different parametric distributions of the GB2 family. OMD estimation

Distribution	Mean	[0, 0.01)	[0.01, 0.02)	[0.02, 0.05)	[0.05, 0.1)	[0.1,)
Lower bound	0.0103	70.70%	22.31%	5.71%	0.94%	0.34%
GB2	0.0040	91.70%	5.26%	2.22%	0.71%	0.11%
B2	0.0069	80.95%	12.06%	5.75%	1.13%	0.11%
SM	0.0090	75.13%	12.21%	10.37%	2.10%	0.19%
Dagum	0.0110	65.65%	18.67%	13.67%	1.92%	0.11%
Lognormal	0.0140	51.95%	24.53%	21.37%	1.99%	0.15%
Fisk	0.0186	39.67%	25.69%	29.26%	4.85%	0.53%

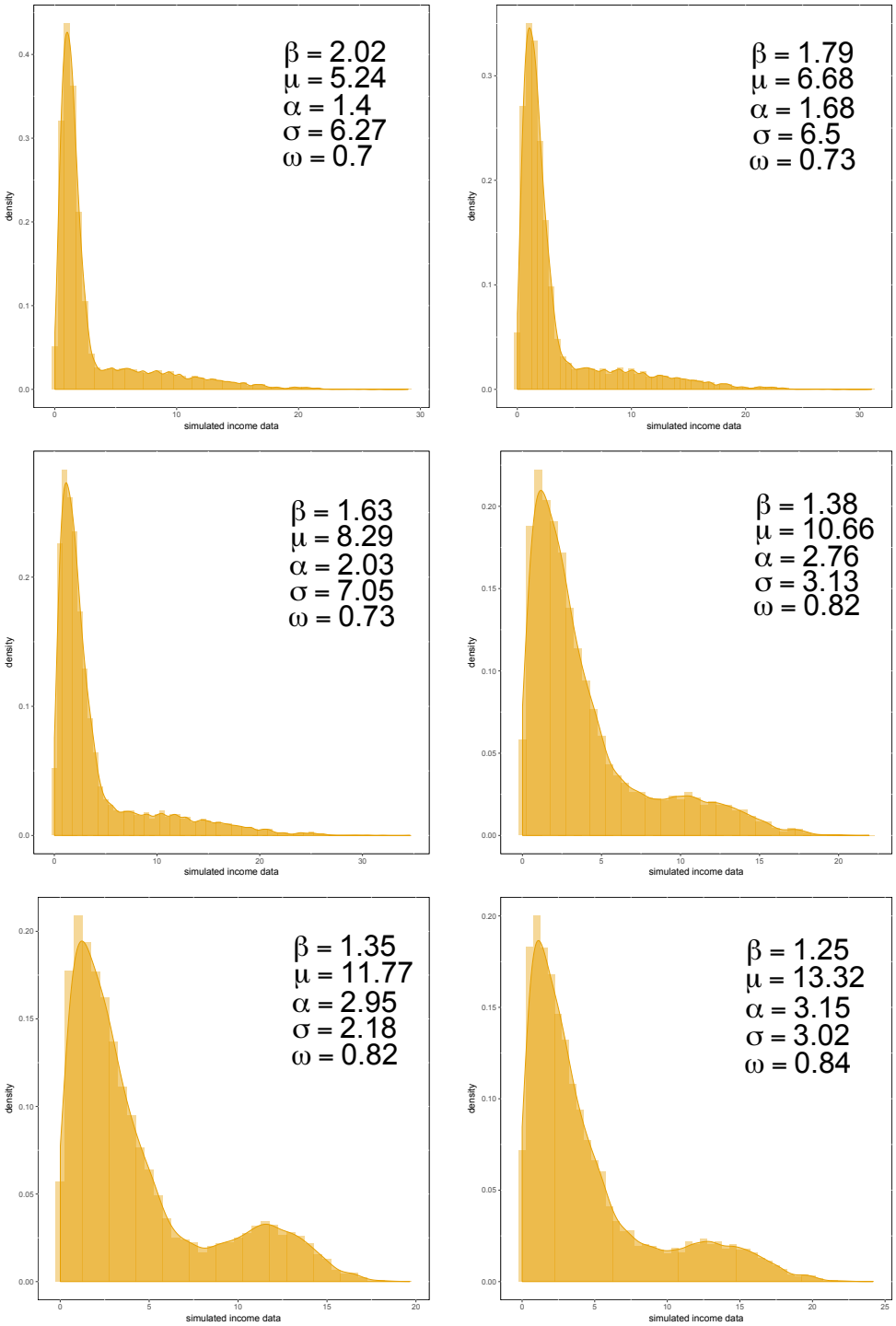
Note: Results based on 2662 datasets from the WIID. Parametric models are estimated by OMD. For the parametric distributions, the Gini coefficient is estimated by Monte Carlo simulation using samples of size $N = 10^6$.

Table A2: Goodness-of-fit matrix based on the Schwarz Bayesian information criterion

	GB2	Beta 2	Singh-Maddala	Dagum	Lognormal	Fisk
GB2	..	87%	85%	88%	98%	98%
Beta 2	13%	..	42%	48%	96%	84%
Singh-Maddala	15%	58%	..	62%	93%	89%
Dagum	12%	52%	38%	..	87%	89%
Lognormal	2%	4%	7%	13%	..	44%
Fisk	2%	16%	11%	11%	56%	..

Note: Proportion of observations for which distributions in rows present better fit than distributions in columns in terms of the Schwarz Bayesian information criterion expressed as $bic = J^{k/J-1}\boldsymbol{M}(\boldsymbol{\theta})'\boldsymbol{M}(\boldsymbol{\theta})$. Results based on 5215 datasets from the WIID. Parametric models are estimated by EWMD.

Figure A1: Histograms and kernel density functions of simulated samples from a mixture of a Weibull and a truncated normal distribution for different parameter values



Note: Simulated samples of size $N = 10^5$. The histograms have been normalised so that the area under the bars is equal to one, making them comparable with the kernel density functions

Table A3: Absolute error in the estimation of inequality measures by OMD using ten income shares

Inequality measure	Absolute difference	GB2	Beta 2	Singh-Maddala	Dagum	Lognormal
	Mean	0.0014	0.0026	0.0035	0.0080	0.0075
	[0, 0.01)	99.64%	96.04%	90.29%	79.5%	71.94%
Gini	[0.01, 0.02)	0.36%	3.60%	5.76%	8.27%	23.02%
Index	[0.02, 0.05)	0%	0.36%	3.96%	11.15%	4.68%
	[0.05, 0.1)	0%	0%	0%	1.08%	0.36%
	[0.1,)	0%	0%	0%	0%	0%
	Mean	0.0024	0.0040	0.0050	0.0088	0.0042
	[0, 0.01)	97.84%	96.4%	86.33%	80.94%	93.17%
Atkinson	[0.01, 0.02)	2.16%	2.16%	6.83%	5.4%	5.04%
index ($\epsilon = 0.5$)	[0.02, 0.05)	0%	1.44%	6.83%	12.23%	1.8%
	[0.05, 0.1)	0%	0%	0%	0.3597%	0%
	[0.1,)	0%	0%	0%	1.0791%	0%
	Mean	0.0049	0.0080	0.0076	0.0100	0.0070
	[0, 0.01)	90.65%	69.42%	76.26%	73.02%	77.7%
Atkinson	[0.01, 0.02)	9.35%	28.78%	18.35%	12.95%	20.86%
index ($\epsilon = 1$)	[0.02, 0.05)	0%	1.80%	5.40%	12.59%	1.44%
	[0.05, 0.1)	0%	0%	0%	0.72%	0%
	[0.1,)	0%	0%	0%	0.72%	0%
	Mean	0.0147	0.0200	0.0165	0.0161	0.0214
	[0, 0.01)	47.12%	27.34%	40.65%	43.17%	21.94%
Atkinson	[0.01, 0.02)	25.18%	30.22%	25.18%	25.90%	31.30%
index ($\epsilon = 1.5$)	[0.02, 0.05)	26.26%	39.57%	31.30%	26.98%	43.17%
	[0.05, 0.1)	1.44%	2.88%	2.88%	3.96%	3.60%
	[0.1,)	0%	0%	0%	0%	0%

Note: Results based on 278 datasets from the LIS database. Parametric models are estimated by OMD. All inequality measures are estimated by Monte Carlo simulation using samples of size $N = 10^6$.

Table A4: Absolute error in the estimation of inequality measures by OMD using five income shares

Inequality measure	Absolute difference	GB2	Beta 2	Singh-Maddala	Dagum	Lognormal
	Mean	0.0016	0.0023	0.0031	0.0061	0.0057
	[0, 0.01)	98.56%	98.56%	92.42%	82.91%	86.33%
Gini	[0.01, 0.02)	1.44%	0.36%	4.33%	9.09%	11.87%
Index	[0.02, 0.05)	0%	1.08%	2.89%	8.00%	1.80%
	[0.05, 0.1)	0%	0%	0.36%	0%	0%
	[0.1,)	0%	0%	0%	0%	0%
	Mean	0.0031	0.0043	0.0048	0.0067	0.0043
	[0, 0.01)	96.40%	97.12%	88.45%	81.82%	94.24%
Atkinson	[0.01, 0.02)	2.88%	1.80%	7.22%	6.91%	5.04%
index ($\epsilon = 0.5$)	[0.02, 0.05)	0.72%	0.72%	3.61%	10.91%	0.72%
	[0.05, 0.1)	0%	0.36%	0.72%	0.36%	0%
	[0.1,)	0%	0%	0%	0%	0%
	Mean	0.0058	0.0086	0.0073	0.0084	0.0089
	[0, 0.01)	84.53%	64.39%	80.14%	74.18%	63.67%
Atkinson	[0.01, 0.02)	15.11%	32.73%	14.08%	13.45%	34.53%
index ($\epsilon = 1$)	[0.02, 0.05)	0.36%	2.88%	5.05%	12.36%	1.80%
	[0.05, 0.1)	0%	0%	0.72%	0%	0%
	[0.1,)	0%	0%	0%	0%	0%
	Mean	0.0168	0.0214	0.0165	0.0158	0.0248
	[0, 0.01)	42.81%	23.02%	40.43%	40.73%	15.47%
Atkinson	[0.01, 0.02)	23.02%	30.58%	26.35%	30.55%	30.22%
index ($\epsilon = 1.5$)	[0.02, 0.05)	32.37%	41.73%	29.60%	24.73%	48.20%
	[0.05, 0.1)	1.8%	4.68%	3.61%	4%	6.12%
	[0.1,)	0%	0%	0%	0%	0%

Note: Results based on 278 datasets from the LIS database. Parametric models are estimated by OMD. All inequality measures are estimated by Monte Carlo simulation using samples of size $N = 10^6$.

Appendix B: The GB2group package

We have implemented the estimation of the GB2 distribution and some of its particular and limit cases, including the Dagum, the beta 2 and the Singh-Maddala distributions, in a user-friendly R package. The package is publicly available from the Comprehensive R Archive Network, so it can be directly dowloaded into R. The `GB2group` package deploys two different econometric strategies to estimate these parametric distributions, equally-weighted minimum distance estimators (EWMD) and optimal minimum distance estimators (OMD). The estimation by EWMD is performed with minimum data requirements: only five income shares and the Gini index are needed to obtain parameter estimates. For the OMD estimation, an estimate of per capita income is also required. The functions for estimating distributions with more than two parameters allow the user to define the grid to be used as initial values in the second step of our estimation procedure. By default, a sequence of integer numbers from 1 to 20 is used (see Section 2.2).

If specified, standard errors of the parameters are also provided. Asymptotic standard errors are reported for the OMD estimates. Using results from Hajargasht and Griffiths (2016), the asymptotic covariance matrix of the estimator of $\boldsymbol{\theta}$ in Eq. (10) is defined as,

$$var(\hat{\theta}) \approx \frac{1}{N}\left(\frac{\partial L(u_j;\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}}\boldsymbol{\Omega}^{-1}\frac{\partial L(u_j;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1},$$

where $\boldsymbol{\Omega}$ is given in Eq.(11) and $L(u_j;\boldsymbol{\theta})$ is the theoretical Lorenz curve evaluated at u_j.

Standard errors for the OMD estimates of the parameters are given by the square root of the elements in the diagonal of the matrix above. Standard errors of EWMD estimates are computed by Monte Carlo simulation, being possible to choose the number of repetitions. The main limitation to obtain estimates of the standard errors is that the available data do not often include information about the size of the sample (N).

As goodness-of-fit measures, the estimation functions in the package report the residual sum of squares for EWMD, defined as:

$$RSS = \boldsymbol{M}(\hat{\boldsymbol{\theta}})'\boldsymbol{M}(\hat{\boldsymbol{\theta}}),$$

where $\boldsymbol{M}(\hat{\boldsymbol{\theta}})$ is the vector with moment conditions defined in Eq. (9).

For OMD estimates, the weighted residual sum of squares is reported:

$$WRSS = \boldsymbol{M}(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{M}(\hat{\boldsymbol{\theta}}).$$

OMD and EWMD estimates of the Gini index are also reported to be compared with its

survey value. The package also includes functions to create goodness-of-fit plots which represent survey income shares and the theoretical Lorenz curve of the fitted model.