Parametric estimation of income distributions using grouped data: an Approximate Bayesian Computation approach

Mathias Silva

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Parametric estimation of income distributions using grouped data: an Approximate Bayesian Computation approach

Mathias Silva

WP 2023 Nr 10

Parametric estimation of income distributions using grouped data: an Approximate Bayesian Computation approach*

Mathias Silva†

Aix Marseille Univ, CNRS, AMSE, Marseille, France

This version: April 12, 2023

Abstract

Recent empirical analysis of income distributions are often limited by the exclusive availability of data in a grouped format. This data format is made particularly restrictive by a lack of information on the underlying grouping mechanism and sampling variability of the grouped-data statistics it contains. These restrictions often result in the unavailability of an analytical parametric likelihood function exploiting all information available in the grouped data. Building on recent methods for inference on parametric income distributions for this type of data, this paper explores a new Approximate Bayesian Computation (ABC) approach. ABC overcomes the restrictions posed by grouped data for Bayesian inference through a non-parametric approximation of the likelihood function exploiting simulated data from the income distribution model. Empirical applications of the proposed ABC method in both simulated and World Bank's PovCalNet data illustrate the performance and suitability of the method for the typical formats of grouped data on incomes.

Keywords: Grouped data, Bayesian inference, Generalized Lorenz curve, GB2

JEL Code: C11, C18, C63

*I am grateful to Michel Lubrano, Stephen Bazen, Emmanuel Flachaire, Philippe Van Kerm, and seminar participants at the AMSE PhD seminar for helpful contributions and comments on earlier versions of this paper.

†Corresponding author: mathias.silva-vazquez@univ-amu.fr. Aix Marseille Univ, CNRS, AMSE, Marseille, France. AMSE - Aix-Marseille Université 5-9 Boulevard Bourdet 13001 Marseille, France. The project leading to this publication has received funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University - A*MIDEX. Declaration of interest: None. All `R` replication code for the results shown in this paper are available upon request and will soon be made available on the author's Github site.

1 Introduction

Since its early days, the empirical study of income distributions and income inequality has faced the limitations posed by data often being available only in a grouped or tabulated format. Historically, perhaps the earliest cases for such limitations can be traced to the release of tabulated or bracketed data on incomes from tax authorities and other administrative institutions for the purpose of economic research (e.g., Goldsmith et al. (1954), Kuznets (1955), and Lebergott (1959)). This choice of format at the time responded in part to technological limitations in providing and handling full samples of individual-level data and in part as a method of statistical disclosure control avoiding the identification of individuals from their reported incomes.

Recent decades have seen an important increase in the availability of microdata sources to study income distributions from individualized survey or tax data samples, made feasible with the progress of computational resources and with the development of more refined statistical disclosure control methods such as top-coding or multiple imputations (An and Little, 2007). However, it still is the case today that data for income distribution analysis are in many cases only available in restrictive grouped-data formats.

Grouped data is still encountered for instance in datasets constructed with the interest in international comparisons of income distributions. This is the case for all large compilatory datasets on world household incomes publicly available today such as UNU-WIDER's World Income Inequality Database (WIID) (an extension of the World Bank's Deininger and Squire (1996) dataset), the Lakner-Milanovic World Panel Income Distribution (LM-WPID) (Lakner and Milanovic, 2016), the Global Consumption and Income Project (GCIP) dataset (Lahoti et al., 2016), or the World Inequality Database (WID.world) (Piketty et al., 2019), where only summary statistics for different income groups, such as average decile incomes, are provided for each country and period.

This format of data are the outcome of compiling many micro- and grouped-data sources across several countries and periods using compatible definitions of incomes and populations of interest. In doing so, it allows for splitting cross-country comparisons into simple differences in scale-invariant measures such as group income shares and more delicate differences in scale-dependent measures such as per capita incomes measured in a common Purchasing Power Parity (PPP) currency.

Moreover, this format for data provision is not exclusive to questions of international comparisons but is also still the usual format in which income tax data are provided for research purposes (e.g., Piketty and Saez (2003), Piketty (2003), Blanchet et al. (2017)).

The challenge that grouped data poses for its analysis is that it imposes additional limitations to the usual methods used for inference on income distributions: the income distribution underlying the data needs to be infered along with the specific grouping mechanism yielding the observed grouped-data summary statistics. This poses a difficulty for point-estimation of quantities of the income distribution such as inequality measures, as the distribution or level of inequality within groups cannot be directly observed in the data.

Grouped data also poses more general difficulties of inference on the income distribution as computing measures of statistical uncertainty such as standard errors may require analytical expressions or additional information on the underlying individual-level data such as sample sizes which are unavailable in many situations.

This paper seeks to contribute to the recent literature on grouped data methods for income distribution analysis in several aspects. Firstly, following recent approaches approximating the often unavailable grouped-data likelihood function, Approximate Bayesian Computation (ABC) is explored as a new type of non-parametric approximation to the likelihood. Secondly, parametric Generalized Lorenz curves are explored as suitable representations of the income distribution for exploiting all information commonly available in grouped data. Finally, the performance of the proposed ABC method is studied in both a simulated data and a real data setting, evidencing it to perform competitively with other recent methods. In the interest of comparability with recent research on grouped-data methods, the case of the Peruvian income distribution in 2013 is revisited exploiting the same data as in Eckernkemper and Gribisch (2021) through a parametric Generalized Beta distribution of the second kind (GB2) Generalized Lorenz curve.

The rest of this article is presented as follows. The following section summarizes empirically relevant formats of grouped data on incomes and the recent development of methods to study a population's income distribution through them. The third section proposes a grouped-data Bayesian inference method for this purpose. Section four presents firstly a simulated-data experiment assessing the performance of this method in contrast with other recent grouped-data methods. Section four also provides a real-data application of this Bayesian method taking as benchmark recent results on the income distribution of Peru in 2013 with grouped data from the World Bank's *PovCalNet* database. Finally, the fifth section discusses the main virtues and limitations of the proposed approach along with future lines of work which could overcome them.

2 Parametric income distributions and methods for inference from grouped data

Studying an income distribution from grouped data involves a degrouping task. To exploit this type of data fully requires considering two separate components of the process generating them (i.e., the data generating process (DGP)). Firstly, a grouping component which produces the observed grouped-data summary statistics from an empirical income distribution. This grouping mechanism can be exploited for insight on the possible empirical distribution of incomes underlying grouped data. Secondly, an income distribution component which produces the underlying empirical income distribution. This second component allows for exploiting the information on the underlying sample income distribution to make inference on the corresponding population's income distribution.

2.1 Grouped-data formats

Introducing notation, define individual i's observed income by $y_i \sim f_y$ and let the K income groups $[z_{k-1}, z_k)$, $k = 1, ..., K$ be defined by the income levels $\{z_k : k = 0, ..., K\}$, with $z_0 = 0$ and $z_K = \infty$. This gives $K - 1$ relevant income group bounds $\{z_k\}_{k=1}^{K-1} = \{z_k : k = 1, ..., K - 1\}$.

Traditionally, the most common format of grouped data on incomes provides a summary statistic S_k^{Obs} for each of these K income groups from the underlying sample of incomes $\{y_i\}_{i=1}^n$. These $\{S_k^{Obs}\}_{k=1}^K$ summary statistics are often the groups' mean incomes $\{S_k^{Obs}\}_{k=1}^K = \{\bar{y}_k^{Obs}\}_{k=1}^K$ defined as[1]:

$$\bar{y}_k^{Obs} = \frac{1}{n_k} \sum_{i=1}^{n} y_i \times I(z_{k-1} \leq y_i < z_k) \ , \ k \in \{1, ..., K\}$$

with group sizes $\{n_k\}_{k=1}^K$ and sample proportions $\{u_k\}_{k=1}^K$ following

$$n_k = \sum_{i=1}^{n} I(z_{k-1} \leq y_i < z_k) \ , \ u_k \equiv \frac{n_k}{n} \ , \ k \in \{1, ..., K\}$$

or groups' cumulative income share $\{S_k^{Obs}\}_{k=1}^K = \{s_k^{Obs}\}_{k=1}^K$:

$$s_k^{Obs} = \frac{\sum_{j=1}^{k} \bar{y}_j^{Obs}}{\sum_{j=1}^{K} \bar{y}_j^{Obs}} \ , \ k \in \{1, ..., K\} \tag{1}$$

. In the latter case, the data often also provides a sample average income μ^{Obs}

$$\mu^{Obs} = \frac{1}{n} \sum_{i=1}^{n} y_i = \sum_{j=1}^{k} u_k \bar{y}_k^{Obs} \ , \ k \in \{1, ..., K\} \tag{2}$$

along with $\{s_k^{Obs}\}_{k=1}^K$.

This is still today the more widely spread format for such type of data and is therefore the focus of the remainder of this paper. An important observation is that the case $\{S_k^{Obs}\}_{k=1}^K = \{\bar{y}_k^{Obs}\}_{k=1}^K$ always allows for a re-expression in terms of cumulative income shares $\{s_k^{Obs}\}_{k=1}^K$ following (1). Another additional important observation is that although the observed statistics $\{S_k^{Obs}\}_{k=1}^K$ are sample quantities subject to sampling variability, it is very rare for datasets on grouped data to provide an estimate of this variability for them. Estimating standard errors for any distributional estimate obtained from this data is also difficult given the usual lack of information on the underlying sample sizes n (e.g., see Anand and Segal (2008)).

Consider the first component of the DGP: the grouping mechanism relating $\{S_k^{Obs}\}_{k=1}^K$ to the underlying sample on incomes $\{y_i\}_{i=1}^n$. Earlier studies dealing with grouped data on incomes faced a format in which sampled individuals where split into groups defined by their level of income using pre-defined income group bounds $\{z_k\}_{k=1}^{K-1}$ available in the

[1] In what follows $I(.)$ represents the identity function, taking value 1 whenever the condition it takes as argument holds true and 0 otherwise.

data. This format, still common today in sources such as tabulated tax data, determines a grouping mechanism which is trivial whenever the bounds are reported in the data. In this case, which will be referred to as grouped data of type II, sampling variability affects only the sample proportions $\{u_k\}_{k=1}^K$ and the masses (or shares $\{s_k\}_{k=1}^K$) of income of the groups covered in the data but not the $\{z_k\}_{k=1}^{K-1}$ income groups' bounds. In what follows, type II data is defined to contain K observations $\{u_k, S_k^{Obs}, z_k\}_{k=1}^K$ and a sample average income μ^{Obs}.

With the development of survey data and the spread of microdata sources across developped and developping countries in recent decades it has become usual in grouped-data sources that income groups' bounds are themselves sample quantities and not pre-defined by the data provider. This format of data, which will be referred to as grouped data of type I in what follows, essentially differs from type II in that the groups are built under externally fixed proportions $\{u_k\}_{k=1}^K$. Type I data is defined to contain K observations $\{u_k, S_k^{Obs}\}_{k=1}^K$ and a sample average income μ^{Obs} in what follows. The design of type I data concludes in sampling variability affecting the income groups' bounds $\{z_k\}_{k=1}^K$ and the group income masses (or shares $\{s_k\}_{k=1}^K$) but not the sample proportions $\{u_k\}_{k=1}^K$. In the most common case, this grouping is done by setting $\{z_k\}_{k=1}^K$ to be sample income quantiles, such as deciles, fixing all groups to represent a same proportion u_k. This grouping mechanism poses bigger difficulties for inference on income distributions than data of type II as the income groups' bounds are rarely presented in the data.

As an illustrative example of type I data, table 1 below presents data on a sample of incomes from Peru for the year 2013 as available on the World Bank's *PovCalNet*[2] website. In this case, as usual, no information on the underlying $\{z_k\}_{k=1}^K$ bounds is provided with the data. Each row in the data can be interpreted analogously to the following: Approximately 2810 individuals in the underlying sample have incomes positioning between the 60-th and 70-th sample percentiles, their mean income is of 383.64 PPP$ and corresponds to a cumulative income share of $s_7^{Obs} = 38.89\%$.

2.2 Parametric income distribution models

Consider the second component of the DGP, the income distribution f_y that yields the sample of incomes $\{y_i\}_{i=1}^n$ underlying the data. Given the generally small samples determined by grouped data, non-parametric approaches such as kernel density estimation have been evidenced to largely underperform parametric approaches (e.g., Minoiu and Reddy (2014), Jorda et al. (2020)). This is, assuming the income distribution f_y to belong to a class of distributions parametrized by the parameter vector $\boldsymbol{\theta}$, $f_y \equiv f_y(.;\boldsymbol{\theta})$, has been evidenced empirically to be a fruitful approach when facing grouped data.

One class of parametric approaches, largely credited to Kakwani and Podder (1973), focuses on exploiting the cumulative aspect of the data in $\{S_k^{Obs}\}_{k=1}^K$ through parametric Lorenz curves. The Lorenz curve of a population's income distribution LC (Lorenz, 1905, Gastwirth, 1971) can be defined as a function relating income-ordered population cumulative shares $u \in [0,1]$ and the cumulative share of the total population's income

[2]PovcalNet: the on-line tool for poverty measurement developed by the Development Research Group of the World Bank. http://iresearch.worldbank.org/PovcalNet/

Table 1: *PovCalNet data for the income distribution of Peru in 2013*

k	u_k	$S_k^{Obs} = \bar{y}_k^{Obs}$
1	0.1	60.67
2	0.2	114.95
3	0.3	162.62
4	0.4	209.53
5	0.5	259.53
6	0.6	317.46
7	0.7	383.64
8	0.8	472.59
9	0.9	629.35
10	1	1267.85
μ^{Obs}		387.82
n		28099

Source: PovCalNet, in same scale as in Eckernkemper and Gribisch (2021) for ease of comparability in what follows.

$LC(u) \in [0,1]$ represented by these population shares. The Lorenz curve is a powerful tool for studying the concentration of a population's incomes along different points of the income distribution and has been used as such for countless studies of income distribution and income inequality.

Following the presentation in Sarabia (2008), the Lorenz curve for the parametric income distribution $f_y(.;\boldsymbol{\theta})$ can be expressed using the definition of its first-order moment cumulative distribution function $F_{(1)}(.;\boldsymbol{\theta})$, defined as in (3) below, where $F_{(1)}(y_i;\boldsymbol{\theta})$ can be interpreted as the proportion of total incomes which correspond to the subpopulation of individuals with incomes below y_i. For each income-ordered population proportion u, the corresponding income level is given by the quantile function associated to f_y:

$$Q(u;\boldsymbol{\theta}) = F^{-1}(u;\boldsymbol{\theta}) = \inf\{y \ : \ F(y;\boldsymbol{\theta}) \geq u\}$$

, such that the Lorenz curve $LC \equiv LC(.;\boldsymbol{\theta})$ corresponding to this parametric distribution can be expressed as

$$\begin{aligned} LC(u;\boldsymbol{\theta}) &= F_{(1)}(Q(u;\boldsymbol{\theta});\boldsymbol{\theta}) \\ &= \frac{\int_0^u Q(x;\boldsymbol{\theta})dx}{\mu} \ , \ u \in [0,1] \end{aligned} \tag{3}$$

with μ denoting the expected value of y_i

$$\mu \equiv E[y_i;\boldsymbol{\theta}] \equiv \int_0^1 Q(x;\boldsymbol{\theta})dx$$

and from which properties of continuity, convexity, non-decreasing, and differentiability almost everywhere in $u \in [0,1]$ can be proved, as well as the properties of $L(0;\boldsymbol{\theta}) = 0$ and $L(1;\boldsymbol{\theta}) = 1$.

Several income concentration and inequality measures are derived from the LC such as the Gini coefficient $G(\boldsymbol{\theta})$:

$$G(\boldsymbol{\theta}) = 1 - 2\int_0^1 LC(u;\boldsymbol{\theta})du$$

Closely related to the LC is the Generalized Lorenz curve GLC (Shorrocks (1983), Kakwani (1984)):

$$GLC(u;\boldsymbol{\theta}) \equiv LC(u;\boldsymbol{\theta}) \times \mu \equiv \int_0^u Q(x;\boldsymbol{\theta})dx$$

This curve, in addition to allowing for comparisons between distributions in terms of relative inequality as any Lorenz curve, allows for comparisons in terms of mean incomes.

This first type of parametric approaches then exploits a chosen parametric form for the Lorenz curve to make inference on its parameter vector $\boldsymbol{\theta}$ by matching points of $LC(u;\boldsymbol{\theta})$ to their empirical counterpart $\{(u_k, s_k^{Obs})\}_{k=1}^K$. Sarabia (2008) presents an overview of conventional parametric forms for the Lorenz curve usually exploited in the literature on income distributions.

One important limitation of this first approach is that directly assuming a parametric form for $LC \equiv LC(.;\boldsymbol{\theta})$ often lacks a known form for its associated distribution $f_y(.;\boldsymbol{\theta})$, limiting the quantities of this distribution on which inference can be made. In particular, the Lorenz curve is a scale-invariant function of the income distribution $f_y(.;\boldsymbol{\theta})$ determined exclusively by proportions and therefore only allows for inference on all but the scale parameters in $\boldsymbol{\theta}$, denoted from here on by $\boldsymbol{\theta}^R$. The GLC is then the closest feasible scale-sensitive alternative to the Lorenz curve allowing for estimation of all parameters in $\boldsymbol{\theta}$.

A second class of parametric approaches directly assumes a parametric form for $f_y(.;\boldsymbol{\theta})$ and derives analytical expressions for the quantities represented in the data $\{S_k(\boldsymbol{\theta})\}_{k=1}^K$. Inference on $\boldsymbol{\theta}$ is then made by fitting the theoretical $\{S_k(\boldsymbol{\theta})\}_{k=1}^K$ to the observed data $\{S_k^{Obs}\}_{k=1}^K$. This approach allows in particular for using the GLC to exploit the cumulative aspect of grouped data while making the underlying distribution model $f_y(.;\boldsymbol{\theta})$ available.

2.3 The GB2 distribution

One parametric form of particular relevance is the GB2 (McDonald (1984)), which has become increasingly popular in the recent literature (e.g., Jenkins (2009), Hajargasht et al. (2012), Graf and Nedyalkova (2014)). The recent popularity of this four-parameter distribution arises firstly from it generalizing many popular income distribution models such as the log-Normal, the Singh-Maddala (Burr XII), and the Weibull distributions. Secondly, the GB2 distribution can be derived as the income distribution resulting from a neoclassical microeconomic model of optimizing firm behaviour as presented in Parker (1999). Finally, this distribution has been empirically evidenced to outperform other candidate distributions in fitting income data and performing not significantly worse than the more complex generalized Beta distribution of which it is a special case (e.g., McDonald and Ransom (2008), Hajargasht and Griffiths (2013), Kobayashi and Kakamu (2019)).

Under the GB2 distribution, the income level $y_i \sim GB2(\alpha, \beta, p, q)$ with parameters $\boldsymbol{\theta}^R = (\alpha, p, q)$ controlling the shape of the distribution and β controlling the scale. The Lorenz curve of this distribution follows (Chotikapanich et al. (2018)):

$$LC^{GB2}(u|\alpha, p, q) = IB\left(IB^{-1}(u|p, q)|p + \frac{1}{\alpha}, q - \frac{1}{\alpha}\right) \ , \ 0 < u < 1$$

where IB denotes the regularized incomplete Beta function[3], the cumulative distribution function (cdf) of the GB2 distribution (Kleiber and Kotz, 2003, chapter 6):

$$F_y^{GB2}(y_i|\alpha, \beta, p, q) = \frac{1}{B(p, q)} \int_0^x t^{p-1}(t - x)^{q-1} dt = IB(x; p, q)$$

where $x \equiv (y_i/\beta)^\alpha/(1 + (y_i/\beta)^\alpha)$ and where $B(p, q)$ denotes the Beta function[4]. Additionally, the GB2 has the following probability density function (pdf)

$$f_y^{GB2}(y_i|\alpha, \beta, p, q) = \frac{\alpha y_i^{\alpha p - 1}}{\beta^{\alpha p} B(p, q) \left(1 + \left(\frac{y_i}{\beta}\right)^\alpha\right)^{p+q}} \ , \ (y_i, \beta, \alpha, p, q) \in \mathbb{R}_+^5$$

and the expected value follows:

$$\mu \equiv \mu^{GB2}(\alpha, \beta, p, q) \equiv E(y_i|\alpha, \beta, p, q) = \beta \frac{B(p + \frac{1}{\alpha}, q - \frac{1}{\alpha})}{B(p, q)} \ , \ -\alpha p < 1 < \alpha q$$

2.4 Parameter inference under grouped data

Regardless of the choice made to model $\{S_k(\boldsymbol{\theta})\}_{k=1}^K$, the main challenge for inference on $\boldsymbol{\theta}$ from type I data stems from the unavailability of the income groups' bounds $\{z_k\}_{k=1}^K$.

In the earlier context of type II data with known bounds, maximum likelihood estimation (MLE) methods exploiting the inherently multinomial likelihood of the distribution of group sizes n_k across groups under this grouping mechanism were developped for inference on $\boldsymbol{\theta}$ (e.g., McDonald (1984)). However, as recent work has illustrated, the multinomial likelihood approach only exploits information on the group bounds and sizes, entirely ignoring all information contained in the income group means or shares and as such is an inefficient approach (e.g., see Eckernkemper and Gribisch (2021)).

Recent methods have placed the focus on overcoming the issue of unknown group bounds in data of type I, and on constructing measures of statistical uncertainty for their estimations without requiring the unavailable data on the standard errors (or underlying sample sizes n) of the statistics $\{S_k^{Obs}\}_{k=1}^K$ presented in the data.

Amongst recent methods for frequentist inference on $f_y(.; \boldsymbol{\theta})$ in the context of type I grouping is the Dirichlet distribution approach of Chotikapanich and Griffiths (2002). This approach avoids the issue of uncertain group bounds by restricting itself to analysis on the LC and measures of relative inequality. Sample non-cumulative income shares $q_k^{Obs} = s_k^{Obs} - s_{k-1}^{Obs}$ are computed and modelled through a parametric Lorenz curve

[3] Also known as the incomplete Beta function ratio.

[4] The Beta function can be defined as $B(p, q) = \int_0^1 t^{p-1}(1 - t)^{q-1} dt$

$q_k(\boldsymbol{\theta}^R) = LC(u_k;\boldsymbol{\theta}^R) - LC(u_{k-1};\boldsymbol{\theta}^R)$ with the assumption of a Dirichlet distribution for $\{q_k(\boldsymbol{\theta})\}_{k=1}^K$. This corresponds to a multivariate distribution assuming each q_k^{Obs} to follow a Beta distribution and the restriction $\sum_{i=1}^K q_k(\boldsymbol{\theta}) = 1$. An additional tuning parameter ruling the inverse of the variance (i.e., the precision) of the estimates must be calibrated or estimated along with $\boldsymbol{\theta}^R$ under this approach. This tuning parameter has been found to be hardly identifiable from grouped data without information on standard errors while having a high impact on the resulting $\boldsymbol{\theta}^R$ parameter estimates (Kobayashi and Kakamu, 2019).

An alternative recent frequentist method is the grouped-data generalized method-of-moments (GMM) approach of Hajargasht et al. (2012) and Griffiths and Hajargasht (2015) . This approach matches group mean incomes $\{\bar{y}_k^{Obs}\}_{k=1}^K$ to their theoretical counterpart $\{S_k(\boldsymbol{\theta})\}_{k=1}^K$ under a GB2 distribution through GMM estimation. The weights for this GMM estimation are given by an estimated asymptotically optimal covariance matrix. Deriving expressions for the group mean incomes and this GMM weights matrix for a given parametric distribution requires expressions for the first- and second-order moments and moment cdf of this distribution. The main virtue of the GMM approach is that it always exploits all information in the data on group bounds, sizes, and shares while offering an estimate of asymptotic standard errors. This allows for frequentist inference on all shape and scale parameters in $\boldsymbol{\theta}$. However, under the GMM method all $\{z_k\}_{k=1}^{K-1}$ income groups' bounds must be estimated as additional parameters along with $\boldsymbol{\theta}$ under type I data, increasing the number of parameters to estimate with the number of groups in the data.

Focusing on the middle and upper parts of the income distribution, the use of Generalized Pareto curves (GPC) has recently been explored for analysis from grouped data (Blanchet et al. (2017)). Generalized Pareto interpolation is a semi-parametric approach which estimates inverted Pareto coefficients[5] points of the GPC, from the $\{S_k^{Obs}\}_{k=1}^K$ statistics available in the grouped data. It then employs smoothing methods to interpolate a smooth curve passing through these point-estimates. Such interpolation can yield a good approximation to the true GPC for the distribution of incomes only above a pre-defined and sufficiently high threshold. Furthermore, to extrapolate the GPC for the top of the income distribution, this method exploits the Generalized Pareto distribution as a parametric form assumed for the upper tail of the distribution[6].

The GPC method may prove fruitful in studying the upper tail of an income distribution from grouped data, estimating group bounds if these are unknown, yet it is not well suited for studying the lower part of the distribution. Additionaly, this method lacks

[5]The inverted Pareto coefficient $b(p)$ at percentile p of an income distribution following $y_i \sim f_y$ can be expressed as the average income above this percentile $E[y_i|y_i > Q(p)]$ divided by the p-th quantile of the distribution $Q(p)$:

$$b(p) = \frac{E[y_i|y_i > Q(p)]}{Q(p)}$$

[6]The Generalized Pareto is a three-parameter distribution with cdf following:

$$y_i \sim F_y(y_i;\zeta,\mu,\sigma) = \begin{cases} 1 - \left(1 + \frac{\zeta(y_i-\mu)}{\sigma}\right)^{-\frac{1}{\zeta}} & , \text{ if } \zeta \neq 0 \\ 1 - e^{-\left(\frac{y_i-\mu}{\sigma}\right)} & , \text{ if } \zeta = 0 \end{cases}$$

development for estimating standard errors for its estimates, such that overall this degrouping method does not allow for inference on f_y.

Bayesian inference offers an alternative approach when no information on the sampling variability of the grouped-data statistics is given. Expliciting prior uncertainty on the values of $\boldsymbol{\theta}$ within parameter space $\boldsymbol{\Theta}$ in a prior probability distribution $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the goal would be to make inference on the posterior probability distribution $\pi(\boldsymbol{\theta}|\{(u_k, S_k^{Obs})\}_{k=1}^K)$ following Bayes' rule:

$$\pi(\boldsymbol{\theta}|\{(u_k, S_k^{Obs})\}_{k=1}^K) \propto \mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \tag{4}$$

This posterior distribution is proportional to weighting prior probabilities $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ through the likelihood function of the assumed parametric model $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$. For any given parameter value $\tilde{\boldsymbol{\theta}}$, the likelihood $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}})$ reflects the probability of the model having generated the observed grouped data $\{(u_k, S_k^{Obs})\}_{k=1}^K$. In this sense, if the data is informative on $\boldsymbol{\theta}$ then the uncertainty contained in the posterior distribution is lower than that contained in the prior and reflects an 'evidence-weighted' measure of statistical uncertainty.

The main constraint that grouped data poses for studying the target posterior distribution is that, while likelihood functions can be analytically obtained for a parametric distribution over individual-level data[7] $\mathcal{L}(\{y_i\}_{i=1}^n|\boldsymbol{\theta})$, it is often not feasible to do so for grouped data and this is particularly so under grouping of type I with unknown bounds. Recent methods seek to overcome such constraint through approximating the unavailable grouped-data likelihood function $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$.

Eckernkemper and Gribisch (2021) propose a Bayesian approach approximating the unavailable joint likelihood for group mean incomes $\mathcal{L}(\{(u_k, \bar{y}_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$ through a product of Normal distributions. The main virtue of this Bayesian quasi-likelihood (BQL) approach is that it is asymptotically equivalent to the frequentist GMM estimator described above, allowing for inference on all parameters in $\boldsymbol{\theta}$ from grouped data of type I or II. As with the GMM approach, however, all income groups' bounds $\{z_k\}_{k=1}^{K-1}$ must be estimated as additional model parameters under type I data and the asymptotic GMM covariance matrix must be computable.

An alternative approximation to $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$ is available from 'likelihood-free' Bayesian computation methods[8]. These methods are specifically devised for situations in which an analytical likelihood function is unavailable or costly to compute for the model being exploited but in which we can simulate data from such model.

Theoretically, in dealing with an intractable likelihood one could estimate $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}})$ as the probability of observing $\{(u_k, S_k^{Obs})\}_{k=1}^K$ under the model $f_y(.;\tilde{\boldsymbol{\theta}})$ at $\tilde{\theta} \in \boldsymbol{\Theta}$. This probability could be estimated from simulating a sufficiently large M amount of datasets $\tilde{S}^{(m)} = \{S_k(\tilde{\boldsymbol{\theta}})\}_{k=1}^K$, $m = 1, ..., M$ and counting the proportion of them exactly matching the observed $\{S_k^{Obs}\}_{k=1}^K$. However, this is unfeasible in practice as $\{S_k^{Obs}\}_{k=1}^K$ is a real-

[7] For example, in the case of $\boldsymbol{y}^{Obs} = \{y_i^{Obs}\}_{i=1}^n$ being n independent observations following $y_i \sim f_y(.;\boldsymbol{\theta})$ their joint likelihood follows $\mathcal{L}(\boldsymbol{y}^{Obs}|\boldsymbol{\theta}) = \prod_{i=1}^n f_y(y_i^{Obs};\boldsymbol{\theta})$

[8] A likelihood is always exploited in Bayesian inference. The 'likelihood-free' term here refers to not requiring a computable expression for the likelihood function.

isation of a continuous random vector and therefore the probability of simulating $\tilde{S}^{(m)}$ exactly reproducing $\{S_k^{Obs}\}_{k=1}^K$ is theoretically zero. Two strategies to overcome this limitation to 'likelihood-free' Bayesian computation have recently been at the center of the Bayesian literature, with some rare applications in economics.

The first of such approaches is that of Bayesian synthetic likelihood (BSL, Price et al. (2018)). BSL closely resembles BQL in that $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$ is approximated through a multivariate Normal distribution $\{S_k^{Obs}\}_{k=1}^K \sim N_K(\boldsymbol{\mu}_K(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ with means vector $\boldsymbol{\mu}_K(\boldsymbol{\theta})$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. For any given point $\tilde{\boldsymbol{\theta}}$, $p(\tilde{\boldsymbol{\theta}}) > 0$, M independent grouped-data $\tilde{S}^{(m)} = \{S_k(\tilde{\boldsymbol{\theta}})\}_{k=1}^K$, $m = 1, ..., M$ are simulated from the model $f_y(.; \tilde{\boldsymbol{\theta}})$ and the approximate likelihood is computed as $f_{N_K}(\{S_k^{Obs}\}_{k=1}^K; \boldsymbol{\mu}_K(\tilde{\boldsymbol{\theta}}), \boldsymbol{\Sigma}(\tilde{\boldsymbol{\theta}}))$ with

$$\begin{cases} \boldsymbol{\mu}_K(\tilde{\boldsymbol{\theta}}) = \frac{1}{M}\sum_{m=1}^M \tilde{S}^{(m)} \\ \boldsymbol{\Sigma}(\tilde{\boldsymbol{\theta}}) = \frac{1}{M}\sum_{m=1}^M (\tilde{S}^{(m)} - \boldsymbol{\mu}_K(\tilde{\boldsymbol{\theta}}))(\tilde{S}^{(m)} - \boldsymbol{\mu}_K(\tilde{\boldsymbol{\theta}}))' \end{cases}$$

and f_{N_K} denoting the pdf of the multivariate Normal distribution.

One important difference between BSL computational approach and BQL is that no restrictions are imposed on the type and dimension of grouped data statistics to exploit from the data. This allows for exploiting potentially more information on a population's incomes which might eventually be available in the data such as Gini coefficient estimates. The main drawback of BSL is that it still imposes multivariate normality as an appropriate representation of $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$, which might not be adecquate in all cases, particularly so when statistics of different nature are being considered simultaneously.

ABC is another type of 'likelihood-free' Bayesian computation methods[9]. ABC differs from BSL primarily in that it imposes no particular form to approximate $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$ and instead approximates it in a non-parametric fashion. In a nutshell, ABC approximates $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}})$ through a weighting kernel $K_\tau(d(\{S_k(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{S_k^{Obs}\}_{k=1}^K))$ giving higher importance to points $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\theta}$ yielding simulated data $\{S_k(\tilde{\boldsymbol{\theta}})\}_{k=1}^K$ closely resembling the observed $\{S_k^{Obs}\}_{k=1}^K$ under a pre-specified distance $d(.,.)$. The bandwidth parameter τ rules the strictness of the approximation, with $\tau \to 0$ enforcing an exact approximation of $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}})$ and $\tau \to \infty$ enforcing a same likelihood for any point $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ (i.e., considering all points $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ equally likely to have generated $\{S_k^{Obs}\}_{k=1}^K$ from the model).

To illustrate the properties of this approach, the ABC target posterior distribution can be stated as (Drovandi and Frazier (2021)):

$$\pi_\tau(\boldsymbol{\theta}|\{(u_k, S_k^{Obs})\}_{k=1}^K) \propto \underbrace{\int_{\mathbb{R}^K} K_\tau(d(\tilde{S}, \{S_k^{Obs}\}_{k=1}^K)) \times \mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})\, d\tilde{S}}_{\mathcal{L}_\tau(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})} \times p(\boldsymbol{\theta}) \ , \ \tilde{S} = \{S_k(\boldsymbol{\theta})\}_{k=1}^K \tag{5}$$

with the intractable integral defining $\mathcal{L}_\tau(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$ being unbiasedly estimated in practice for any given point $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ using M simulated samples from the model $f_y(.; \tilde{\boldsymbol{\theta}})$

[9]See Sunnåker et al. (2013), Sisson et al. (2018), and Beaumont (2019) for a wide coverage of the history and current state of ABC and its applications within different scientific fields

following:

$$\hat{\mathcal{L}}_\tau(\{(u_k, S_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}}) = \frac{1}{M}\sum_{m=1}^{M} K_\tau(d(\tilde{S}^{(m)}, \{S_k^{Obs}\}_{k=1}^K)) \ , \ \tilde{S}^{(m)} = \{S_k(\tilde{\boldsymbol{\theta}})\}_{k=1}^K$$

The target posterior distribution in (5) can differ from the 'exact' target posterior (4) for several reasons. The degree of similarity between these posterior distributions depends mainly on the quality of the approximation $\hat{\mathcal{L}}_\tau(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$ to the likelihood $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$. The quality of this approximation depends primarily on the choice of τ. The computational cost of the exact approximation $\tau \to 0$ makes it unfeasible, while setting τ too large can severly impact negatively the quality of the approximation. A second parameter ruling the quality of the ABC approximation is the number M of simulated samples generated for each point $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$, despite setting $M = 1$ being common practice. To a lesser extent, the quality of this approximation is also conditioned by the choice for the discrepancy metric $d(.,.)$ and ABC kernel K_τ.

With respect to the target posterior distributions, it is also important to note that group-data posterior distributions in (4) or (5) can differ from the exact microdata posterior distribution $\pi(\boldsymbol{\theta}|\{y_i\}_{i=1}^n)$ as a consequence of the group data statistics $\{S_k(\boldsymbol{\theta})\}_{k=1}^K$ implying too large a loss of information with respect to $\{y_i\}_{i=1}^n$. Such a loss of information can hinder the inference made on $\boldsymbol{\theta}$ from grouped data with respect to microdata, in which case $\pi(\boldsymbol{\theta}|\{y_i\}_{i=1}^n) \not\propto \pi(\boldsymbol{\theta}|\{(u_k, S_k^{Obs})\}_{k=1}^K)$ and the statistics defining $\{S_k^{Obs}\}$ are said to be insufficient.

Opting for its overall flexibility and its non-parametric approximation to the grouped-data likelihod $\mathcal{L}(\{(u_k, S_k^{Obs})\}_{k=1}^K|\boldsymbol{\theta})$, the following section develops an ABC approach for inference on $\boldsymbol{\theta}$ compatible with type I data exploiting the GLC.

3 ABC inference from grouped data through the Generalized Lorenz curve

Kobayashi and Kakamu (2019) present a first application of ABC to study income distributions from grouped data. Their approach exploits type I data with unknown income groups' bounds $\{z_k\}_{k=1}^{K-1}$ through a parametric LC under the generalized Beta distribution, where no expression is available for the corresponding likelihood function. Their analysis under the specific case of the GB2 distribution provides evidence of the suitability of ABC for estimating the posterior distribution of all GB2 shape parameters $\boldsymbol{\theta}^R = (\alpha, p, q)$ and on the income Gini coefficient $G(\boldsymbol{\theta}^R)$ as jointly determined by them. By focusing strictly on the LC, however, their approach only allows for exploiting data in the form of shares and so disregards any information on the scale of the distribution contained in the income groups' means $\{\bar{y}_k^{Obs}\}_{k=1}^K$.

Exploiting the available information on average income μ^{Obs} and the groups' data expressed in cumulative shares $\{(u_k, s_k^{Obs})\}_{k=1}^K$ a sample GLC can be computed as a scale-sensitive representation following:

$$GLC_k^{Obs} = s_k^{Obs} \times \mu^{Obs} \ , \ k = 1, ..., K$$

. This representation offers an approach similar to Kobayashi and Kakamu (2019) while exploiting all the available information to allow for inference on all parameters in $\boldsymbol{\theta}$.

Assuming a GB2 distribution as a model of the population's income distribution, ABC inference on $\boldsymbol{\theta} = (\alpha, \beta, p, q)$ requires firstly being able to simulate data in the same form as $\{(u_k, GLC_k^{Obs})\}_{k=1}^K$. For any given value $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\beta}, \tilde{p}, \tilde{q}) \in \boldsymbol{\Theta}$ this can be computed analytically as:

$$GLC_k^{GB2}(\tilde{\boldsymbol{\theta}}) \equiv GLC^{GB2}(u_k; \tilde{\boldsymbol{\theta}}) = LC^{GB2}(u_k|\tilde{\alpha}, \tilde{p}, \tilde{q}) \times \mu^{GB2}(\tilde{\alpha}, \tilde{\beta}, \tilde{p}, \tilde{q}) \ , \ -\alpha p < 1 < \alpha q$$

or can be approximated from computing sample shares $\{(u_k, \tilde{s_k})\}_{k=1}^K$ using (1) and sample mean $\tilde{\mu}$ using (2) over simulated microdata $\{\tilde{y}_i\}_{i=1}^n$ from $\tilde{y}_i \sim f_y^{GB2}(.; \tilde{\alpha}, \tilde{\beta}, \tilde{p}, \tilde{q})$.

A further step in defining a framework for ABC inference concerns a choice on how to make comparisons between observed $\{GLC_k^{Obs}\}_{k=1}^K$ and simulated $\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K$ data that are most informative on the values of $\boldsymbol{\theta}$. This is, a choice for the discrepancy metric $d(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$. One simple metric that can be informative on greater or smaller relative discrepancies along different segments of the income distributions underlying the data is the following vector-valued distance[10]:

$$\begin{aligned} d(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K) = & \quad (\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K - \{GLC_k^{Obs}\}_{k=1}^K) \\ \circ & \quad (\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K - \{GLC_k^{Obs}\}_{k=1}^K) \\ \oslash & \quad (\{GLC_k^{Obs}\}_{k=1}^K \circ \{GLC_k^{Obs}\}_{k=1}^K) \end{aligned}$$

. Each k-th element in $d(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$, denoted $d_k(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$ is then the squared difference between the observed and simulated GLCs for the k-th group, scaled by the square of the observed value GLC_k^{Obs}:

$$d_k(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K) = \left(\frac{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}}) - GLC_k^{Obs}}{GLC_k^{Obs}} \right)^2$$

Choosing an appropriate discrepancy metric is not only relevant in exploiting the data in a way informative about $\boldsymbol{\theta}$ but also for insight on the goodness-of-fit of the estimated model to the data. Following Ratmann et al. (2009), the distribution of $d(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$ under the estimated ABC posterior distribution $\tilde{\boldsymbol{\theta}} \sim \pi_\tau(\boldsymbol{\theta}|\{(u_k, GLC_k^{Obs})\}_{k=1}^K)$ can be used for the purpose of model criticism. Vector-valued metrics, in particular, can be informative about the degree of fit to the data at each of the observed GLC_k^{Obs} , $k = 1, ..., K$ groups along the income distribution. This information is contained in the corresponding posterior distribution $d_k(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$ under $\tilde{\boldsymbol{\theta}} \sim \pi_\tau(\boldsymbol{\theta}|\{(u_k, GLC_k^{Obs})\}_{k=1}^K)$.

Finally, a form for the ABC kernel K_τ in (5) needs to be specified. Most ABC applications exploit a simple uniform kernel for each $\varepsilon_k \equiv d_k(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$, $k = 1, ..., K$, following:

$$K_\tau^{Uni}(\{\varepsilon_k\}_{k=1}^K) = \prod_{k=1}^K \frac{1}{\tau_k} I\left(|\varepsilon_k| \leq \frac{\tau_k}{2}\right) \ , \ \varepsilon_k \equiv d_k(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$$

[10] $\circ$ and $\oslash$ denote the Hadamard or elementwise product and division operators respectively.

. This amounts to giving a same common weight to points $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\theta}$ yielding $\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K$ that are each within an interval of width τ_k , $k = 1, ..., K$ around the corresponding observed GLC_k^{Obs} , $k = 1, ..., K$. All other values $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\theta}$ yielding simulated data 'too far away' from the observed data are given a weight of $K_\tau^{Uni}(\{\varepsilon_k\}_{k=1}^K) = 0$ as at least one of the $I\left(|\varepsilon_k| \leq \frac{\tau_k}{2}\right)$, $k = 1, ..., K$ conditions do not hold.

For computational purposes in studying $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$, 'smooth' ABC kernels might be prefered, giving increasingly higher weights to values $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\theta}$ with associated simulations reproducing the observed data more closely than others. A multivariate Gaussian kernel may be used together with a vector-valued metric in a spirit similar to the multivariate Uniform kernel (Ratmann (2010)):

$$K_\tau^{gauss}(\{\varepsilon_k\}_{k=1}^K) = \prod_{k=1}^K \frac{1}{\tau_k} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\varepsilon_k}{\tau_k}\right)^2\right\} \ , \ \varepsilon_k \equiv d_k(\{GLC_k^{GB2}(\tilde{\boldsymbol{\theta}})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$$

. This implies a weighting scheme for each of the ABC discrepancies $\{\varepsilon_k\}_{k=1}^K$ following a Normal distribution, centered at zero (i.e., highest weight is given to values $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\theta}$ exactly reproducing $\{GLC_k^{Obs}\}_{k=1}^K$), and with a standard deviation of τ_k.

Despite the approximations involved in ABC, the way in which the posterior distribution $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ is estimated closely follows the approach of most modern Bayesian methods. It is in general feasible to sample values from or converging to this distribution exploiting the Monte Carlo principle: any quantity of $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ which can be expressed as an expectation can be studied through a sufficiently large sample of independent draws $\{\tilde{\boldsymbol{\theta}}^{(j)}\}_{j=1}^J$ from this distribution $\tilde{\boldsymbol{\theta}}^{(j)} \sim \pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$. The simplest of algorithms seeking to obtain samples from the ABC posterior distribution $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ is that of rejection sampling. A standard rejection sampling algorithm follows:

Algorithm 1: An ABC rejection (***ABC-R***) algorithm.

1: Set $j \leftarrow 1$
2: **while** $j \leq J$ **do**
3: Sample $(\tilde{\alpha}^{(j)}, \tilde{\beta}^{(j)}, \tilde{p}^{(j)}, \tilde{q}^{(j)}) \equiv \tilde{\boldsymbol{\theta}}^{(j)} \sim p(\theta)$
4: Generate $\{GLC_k^{(j);GB2}(\tilde{\boldsymbol{\theta}}^{(j)})\}_{k=1}^K$ by simulating from $f_y^{GB2}(.; \tilde{\boldsymbol{\theta}}^{(j)})$
5: Generate $\{\varepsilon_k^{(j)}\}_{k=1}^K = d(\{GLC_k^{(j);GB2}(\tilde{\boldsymbol{\theta}}^{(j)})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$
6: Accept $(\tilde{\boldsymbol{\theta}}^{(j)}, \{\varepsilon_k^{(j)}\}_{k=1}^K)$ with probability:

$$K_\tau(\{\varepsilon_k^{(j)}\}_{k=1}^K) = \prod_{k=1}^K \frac{1}{\tau_k} I\left(|\varepsilon_k^{(j)}| \leq \frac{\tau_k}{2}\right)$$

▷ e.g., if $u^{(j)} \leq K_\tau(\{\varepsilon_k^{(j)}\}_{k=1}^K)$ where $u^{(j)}$ is a draw from a $Uniform(0,1)$ distribution
7: If accepted, store $(\tilde{\boldsymbol{\theta}}^{(j)}, \{\varepsilon_k^{(j)}\}_{k=1}^K)$ and set $j \leftarrow j + 1$, if rejected, re-draw.
end

This rejection sampling algorithm explores, for a sufficiently large J, points $\tilde{\boldsymbol{\theta}}^{(j)} \in \Theta$, $j = 1, ..., J$ sampled independently from the prior $p(\boldsymbol{\theta})$ and returns a sample of accepted parameter values and associated ABC discrepancies $\{(\tilde{\boldsymbol{\theta}}^{(j)}, \{\varepsilon_k^{(j)}\}_{k=1}^K)\}_{j=1}^J$. The output contains only parameter values which have been accepted based on them closely reproducing the observed $\{GLC^{Obs}\}_{k=1}^K$ through the model. These sampled values qualify as a sample proportional to $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$.

Independent sampling algorithms like (***ABC-R***) are in general very computationally inefficient as all samples are drawn from the prior $p(\theta)$. This can force the algorithm to spend many simulations exploring regions of high prior density with low acceptance rates. This can also make it such that an unfeasibly large number of draws $J \to \infty$ must be done before the sample of accepted points $\{(\tilde{\boldsymbol{\theta}}^{(j)}, \{\varepsilon_k^{(j)}\}_{k=1}^K)\}_{j=1}^J$ is representative of the target posterior distribution $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$.

Dependent sampling algorithms, where the acceptance rule depends on previous draws, can be designed instead under Sequential Monte Carlo (SMC) (e.g., see Sisson et al. (2007)) or under the Markov Chain Monte Carlo (MCMC) principle[11] (e.g., see Marjoram et al. (2003), Clarté et al. (2021)). Allowing for dependance in the resulting samples is the cost at which computational efficiency is gained.

Metropolis-Hastings (MH) samplers are a general-purpose class of MCMC algorithms feasible for ABC inference. In general terms, the MH algorithm performs a global exploration of the support of the target posterior distribution $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ through local accept-reject steps. Any j-th, $j = 1, ..., J$, local accept-reject step is defined by the following acceptance probability:

$$\rho^{(j)} = \min\left\{1, \left(\frac{\pi_\tau(\tilde{\boldsymbol{\theta}}^{(j)}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)}{\pi_\tau(\tilde{\boldsymbol{\theta}}^{(j-1)}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)}\right) \times \left(\frac{g_{\Sigma^{(j-1)}}(\tilde{\boldsymbol{\theta}}^{(j-1)}, \tilde{\boldsymbol{\theta}}^{(j)})}{g_{\Sigma^{(j-1)}}(\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\theta}}^{(j-1)})}\right)\right\}$$

with $g_{\Sigma^{(j-1)}}(\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\theta}}^{(j-1)})$ denoting a candidate function from which the j-th candidate value $\tilde{\boldsymbol{\theta}}^{(j)}$ is sampled, given the previously retained value $\tilde{\boldsymbol{\theta}}^{(j-1)}$. Given a same probability of sampling $\tilde{\boldsymbol{\theta}}^{(j)}$ after $\tilde{\boldsymbol{\theta}}^{(j-1)}$ and of sampling $\tilde{\boldsymbol{\theta}}^{(j-1)}$ after $\tilde{\boldsymbol{\theta}}^{(j)}$, the j-th draw $\tilde{\boldsymbol{\theta}}^{(j)}$ has a higher probability $\rho^{(j)}$ of being accepted as a draw from $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ the higher its posterior density $\pi_\tau(\tilde{\boldsymbol{\theta}}^{(j)}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ is with respect to the previous draw's posterior density $\pi_\tau(\tilde{\boldsymbol{\theta}}^{(j-1)}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$.

One popular choice for a MH proposal g is the Adaptive Random-Walk Metropolis (AM) algorithm (Haario et al., 2001). g_Σ is in this case defined by the following adaptive random walk process:

[11] The MCMC principle states that any quantity of the posterior distribution of θ which can be expressed as an expectation can be studied through a sufficiently large sample of dependent draws from a Markov Chain having this posterior distribution as stationary distribution.

$$(\alpha^{(j)}, \beta^{(j)}, p^{(j)}, q^{(j)}) \equiv \boldsymbol{\theta}^{(j)} \sim g_{\Sigma^{(j-1)}}(\boldsymbol{\theta}^{(j)}, \tilde{\boldsymbol{\theta}}^{(j-1)}) \Rightarrow \tilde{\boldsymbol{\theta}}^{(j)} = \tilde{\boldsymbol{\theta}}^{(j-1)} + \tilde{\boldsymbol{\eta}}^{(j)}$$
$$\boldsymbol{\eta}^{(j)} \sim N_4(0, \Sigma^{(j-1)})$$
$$\Sigma^{(j-1)} = \begin{cases} \Sigma^{(0)} \ , \text{ if } j \leq J_0 \\ s_4 \times \frac{1}{(j-1)} \left(\sum_{d=1}^{(j-1)} \tilde{\boldsymbol{\theta}}^{(d)} \tilde{\boldsymbol{\theta}}^{(d)'} - d \times \bar{\boldsymbol{\theta}} \bar{\boldsymbol{\theta}}' \right) + s_4 \times \chi \times I_4 \ , \text{ if } j > J_0 \ , \ 0 < \chi \ll 1 \end{cases}$$

with $\bar{\boldsymbol{\theta}}$ denoting the mean value of all draws up to and including the $(j-1)$-th and with s_4 suggested, following Gelman et al. (1996), to be set to $s_4 = \frac{2.4^2}{4}$ given the 4 parameters in $\boldsymbol{\theta}$ under the GB2 distribution[12]. This proposal distribution draws a candidate value $\tilde{\boldsymbol{\theta}}^{(j)}$ by drawing from a multivariate Gaussian distribution centered at the previously retained draw $\tilde{\boldsymbol{\theta}}^{(j-1)}$ and with covariance matrix $\Sigma^{(j-1)}$. Being initially set to a given matrix $\Sigma^{(0)}$, this covariance matrix starts adapting exploiting all past draws after a sufficiently large initial period J_0 following the sample covariance matrix. An AM algorithm can thus focus on sampling more densely in regions near values $\tilde{\boldsymbol{\theta}}$ with high posterior density and more sparsely in regions of low posterior density.

An AM algorithm for ABC inference on $\boldsymbol{\theta}$ through a parametric model $f_y^{GB2}(.;\theta)$ under type I grouped data follows:

[12]The addition of the diagonal matrix $\chi \times I_4$ is needed with an insignificantly small but non-zero χ to assure the non-singularity of $\Sigma^{(j-1)}$ and assure the ergodicity properties of the Markov chain allowing for the MCMC principle.

Algorithm 2: An AM ABC (***ABC-AM***) algorithm.

Initialization:

Set $\Sigma^{(0)}$

Until $K_\tau(\{\varepsilon_k^{(0)}\}_{k=1}^K) > 0$:

  Sample $(\tilde{\alpha}^{(0)}, \tilde{\beta}^{(0)}, \tilde{p}^{(0)}, \tilde{q}^{(0)}) \equiv \tilde{\boldsymbol{\theta}}^{(0)}$ from $p(\boldsymbol{\theta})$

  Generate $\{GLC_k^{(0);GB2}(\tilde{\boldsymbol{\theta}}^{(0)})\}_{k=1}^K$ by simulating from $f_y^{GB2}(.;\tilde{\boldsymbol{\theta}}^{(0)})$

  Generate $\{\varepsilon_k^{(0)}\}_{k=1}^K = d(\{GLC_k^{(0);GB2}(\tilde{\boldsymbol{\theta}}^{(0)})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$

Sampling: **for** $j = 1, ..., J$ **do**

Sample $\tilde{\boldsymbol{\theta}}^{(j)} \sim g_{\Sigma^{(j-1)}}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{(j-1)})$ from the candidate $g_{\Sigma^{(j-1)}}$

Generate $\{GLC_k^{(j);GB2}(\tilde{\boldsymbol{\theta}}^{(j)})\}_{k=1}^K$ by simulating from $f_y^{GB2}(.;\tilde{\boldsymbol{\theta}}^{(j)})$

Generate $\{\varepsilon_k^{(j)}\}_{k=1}^K = d(\{GLC_k^{(j);GB2}(\tilde{\boldsymbol{\theta}}^{(j)})\}_{k=1}^K, \{GLC_k^{Obs}\}_{k=1}^K)$

Accept and store $(\tilde{\boldsymbol{\theta}}^{(j)}, \{\varepsilon_k^{(j)}\}_{k=1}^K)$ with probability:

$$\rho^{(j)} = \min\left\{1, \frac{\overbrace{K_\tau(\{\varepsilon_k^{(j)}\}_{k=1}^K)}^{\hat{\mathcal{L}}_\tau(\{(u_k, GLC_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}}^{(j)})} \times p(\tilde{\boldsymbol{\theta}}^{(j)}) \times g_{\Sigma^{(j-1)}}(\tilde{\boldsymbol{\theta}}^{(j-1)}, \tilde{\boldsymbol{\theta}}^{(j)})}{K_\tau(\{\varepsilon_k^{(j-1)}\}_{k=1}^K) \times p(\tilde{\boldsymbol{\theta}}^{(j-1)}) \times g_{\Sigma^{(j-1)}}(\tilde{\boldsymbol{\theta}}^{(j)}, \tilde{\boldsymbol{\theta}}^{(j-1)})}\right\}$$

  ▷ e.g., if $u^{(j)} \leq \rho^{(j)}$ where $u^{(j)}$ is a draw from a $Uniform(0,1)$ distribution

  otherwise store $(\tilde{\boldsymbol{\theta}}^{(j)}, \{\varepsilon_k^{(j)}\}_{k=1}^K) = (\tilde{\boldsymbol{\theta}}^{(j-1)}, \{\varepsilon_k^{(j-1)}\}_{k=1}^K)$

  if $j > J_0$ **then**

Update $\Sigma^{(j)}$

  end

end

As usual in MCMC algorithms, the unavailability of $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ makes it that the acceptance probability $\rho^{(j)}$ is instead computed from the product of a prior probability $p(\tilde{\boldsymbol{\theta}})$ and the approximated likelihood $\hat{\mathcal{L}}_\tau(\{(u_k, GLC_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}})$, proportional to $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ following (5). For computing $\hat{\mathcal{L}}_\tau(\{(u_k, GLC_k^{Obs})\}_{k=1}^K|\tilde{\boldsymbol{\theta}})$ with $M > 1$ it suffices to sample M times in step 9 of (***ABC-AM***).

One first important property to assess on output from an MCMC sampler is the convergence of the sampled values to a stationary distribution. Without such convergence, it is uncertain whether the achieved sample is representative of $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$. Typically, these class of samplers contain an initial 'burn-in' period in which the sample relative frequencies are rarely representative of their respective posterior density thus it is usual in practice to discard an initial set of draws from the chain before analysis.

Three parameters must be pre-defined in (***ABC-AM***). Firstly, $\Sigma^{(0)}$, the initial covariance matrix for the proposal distribution. Secondly, J_0, defining the initial 'burn-in' number of draws before this covariance matrix starts incorporating information from all previous draws of the chain. Finally, the $\tau = \{\tau_k\}_{k=1}^K$ bandwidths ruling the strictness of the ABC approximation in terms of discrepancies with each of the observed $\{GLC_k^{Obs}\}_{k=1}^K$. Without previous informative guesses for these parameters, these may be set on the basis of the degree and speed of convergence and acceptance rates across several pilot runs of the algorithm.

Once a sample $\{(\tilde{\boldsymbol{\theta}}^{(j)}, \{\varepsilon_k^{(j)}\}_{k=1}^K)\}_{j=1}^J$ has been obtained as a valid sample from $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$, sample quantities such as average income $\mu^{GB2}(\alpha^{(j)}, \beta^{(j)}, p^{(j)}, q^{(j)})$ or Gini coefficient $G(\alpha^{(j)}, p^{(j)}, q^{(j)})$ can be computed over each set of parameter values yielding a sample representative of their posterior distributions.

The posterior distribution of the ABC discrepancies, as represented by the sample $\{\{\varepsilon_k^{(j)}\}_{k=1}^K\}_{j=1}^J$, can be informative of segments of the income distribution in the data $\{GLC_k^{Obs}\}_{k=1}^K$ which the fitted model systematically fails to reproduce correctly. In particular, if the sample distribution for a given k-th group discrepancy $\{\varepsilon_k^{(j)}\}_{j=1}^J$ has little density at 0 this can be interpreted as the estimated posterior distribution $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ poorly fitting the income distribution at the k-th income group's GLC_k^{Obs}.

In dealing with the unknown income groups' bounds $\{z_k\}_{k=1}^{K-1}$ this ABC approach treats them as nuisance parameters. In fact, with known $\{u_k\}_{k=1}^K$ sample proportions, these bounds are not necessary to make inference on $\boldsymbol{\theta}$ under this framework. Furthermore, if information on $\{z_k\}_{k=1}^{K-1}$ were available then a goodness-of-fit check can be devised by computing the posterior predictive distribution[13] $\{\tilde{z}_k(\tilde{\boldsymbol{\theta}}^{(j)})\}_{j=1}^J = \{Q_y^{GB2}(u_k; \tilde{\boldsymbol{\theta}}^{(j)})\}_{j=1}^J$ for each of the $k = 1, ..., K-1$ group bounds and observing for significant discrepancies with the $\{z_k\}_{k=1}^{K-1}$.

The following section assesses the relative performance of the ABC approach proposed in this section through a Monte Carlo experiment, firstly, and on *PovCalNet* data, secondly.

[13]Here $Q_y^{GB2}(.; \boldsymbol{\theta})$ denotes the quantile function of the GB2 distribution.

4 Applications on simulated and real *PovCalNet* data

4.1 Simulated data applications

To study the performance of the proposed ABC method in comparison with other recent methods, the following relevant aspects were explored. Firstly, three samples of $n = 10000$ individual-level incomes were generated from a $GB2(\alpha, \beta, p, q)$ distribution, under three parameter settings representing different empirically relevant scenarios: (i) $(\alpha, \beta, p, q) = (3.85, 10000, .4, 2)$ representing a hypothetical upper middle-income country's distribution with moderate levels of inequality, (ii) $(\alpha, \beta, p, q) = (2.3, 10000, 1.75, 1.25)$ representing a hypothetical high-income country's distribution with moderate levels of inequality, and (iii) $(\alpha, \beta, p, q) = (1.64, 386, 1.22, 1.83)$ representing a hypothetical low-income country's distribution with high levels of inequality. These simulated samples are summarized in table 2 below. To explore the loss in information due to not observing microdata and observing grouped-data statistics instead, MLE estimates were obtained over these microdata samples[14]. Secondly, these individuals samples were grouped into 10 groups with bounds set by the sample deciles to obtain a corresponding grouped dataset of type I. Finally, over these grouped-data samples, GMM (Jorda et al. 2021), GPC[15], and ABC estimates were obtained.

In applying the ABC estimator presented in the previous section, the prior distribution for the vector of GB2 parameters (α, β, p, q) was set as the product of a $Gamma(1, 1)$ distribution for each of the shape parameters α, p, and q, representing a strictly positive distribution heavily right-skewed and with a single mode at 0. A $Gamma(5, 2)$ was assumed for the β parameter under settings (i) and (ii), representing a less right-skewed distribution spreading over a wider range of values, and a $Gamma(5, 1)$ prior was assumed for this parameter on setting (iii) to spread over a lower range of values given the significantly lower observed mean income in this scenario. The mean income in the data μ^{Obs} was scaled by a constant power of 10 to lie within the range of this prior for computational purposes. Additionally, a non-zero mode restriction $\alpha \times p > 1$, and mean existence conditions $\alpha \times q > 1$ and $-\alpha \times p < 1$ were also enforced. The MCMC algorithm presented in the previous section was applied and set with the following parameters. Initially, the proposal distribution was defined with a covariance matrix following $\Sigma^{(0)} = diag(.5, .5, .5, .5)$, and was updated based on retained simulations after a 'burn-in' period of $t_0 = 25000$. Finally, a Gaussian ABC kernel was used with bandwidth parameter set at $\tau = (.001, .001, .001, .001, .0005, .0005, .0005, .0005, .0005, .0005)$, being overall strict in fitting the data across all deciles but particularly so above the median.

Results from all four estimation methods explored over these same datasets are summarized in table 2 below. As a first remark, all methods yield estimes which closely resemble the data in terms of income shares, Gini coefficient, and mean income, with no particular differences between the MLE estimates obtained over the individual-level

[14]The estimation routine used for MLE estimation of GB2 parameters from individual-level samples is provided in the `GB2` (Graf and Nedyalkova. 2015) library in `R`.

[15]GPC esimation was performed following Blanchet et al. (2017) and their publicly-available `gpinter` library for `R`.

samples and the grouped-data estimates. The fact that income shares estimated by the GMM and GPC estimator exactly match the data should come as no surprise, as these methods are built under the condition that this matching holds on resulting estimates. Additionally, comparing the three parametric estimators for the GB2 distribution's parameters evidences no particular gains or losses in terms of bias from estimating using individual-level data or grouped data, a result consistent with those explored in simulated data in Eckernkemper and Gribisch (2021). In terms of standard errors, however, MLE estimation over individual-level samples and ABC result in parameter point-estimates which are generally more precise than those obtained through GMM.

Figure 1 below graphically illustrates the true and fitted GB2 distributions for all point-estimates and scenarios considered. These density curves suggest that while all MLE, GMM, and ABC estimates yield accurate representations of the true density curve the GPC method is suited for representing the upper part of the distribution but behaves very erratically on the middle and bottom segments.

Finally, a remark concerning the point-estimates presented for the ABC estimated posterior distribution is that the shape parameters p and q are in all cases slightly above their true values. These point-estimates are simply obtained as the mean value of the respective estimated posterior distribution (also known as the minimum mean square error (MMSE) estimates), and are therefore sensitive to skewness in the estimated posterior distribution. In this case, this could be the result of the estimated posterior distribution being influenced by the skewed shape of the prior distribution chosen for these parameters. The fact that the estimates for the α shape parameter do not share this behaviour despite having an identical prior can likewise suggest that this format of grouped data is relatively more informative on the value of this parameter and therefore yields a respective estimated posterior distribution less influenced by the skewness of its prior distribution.

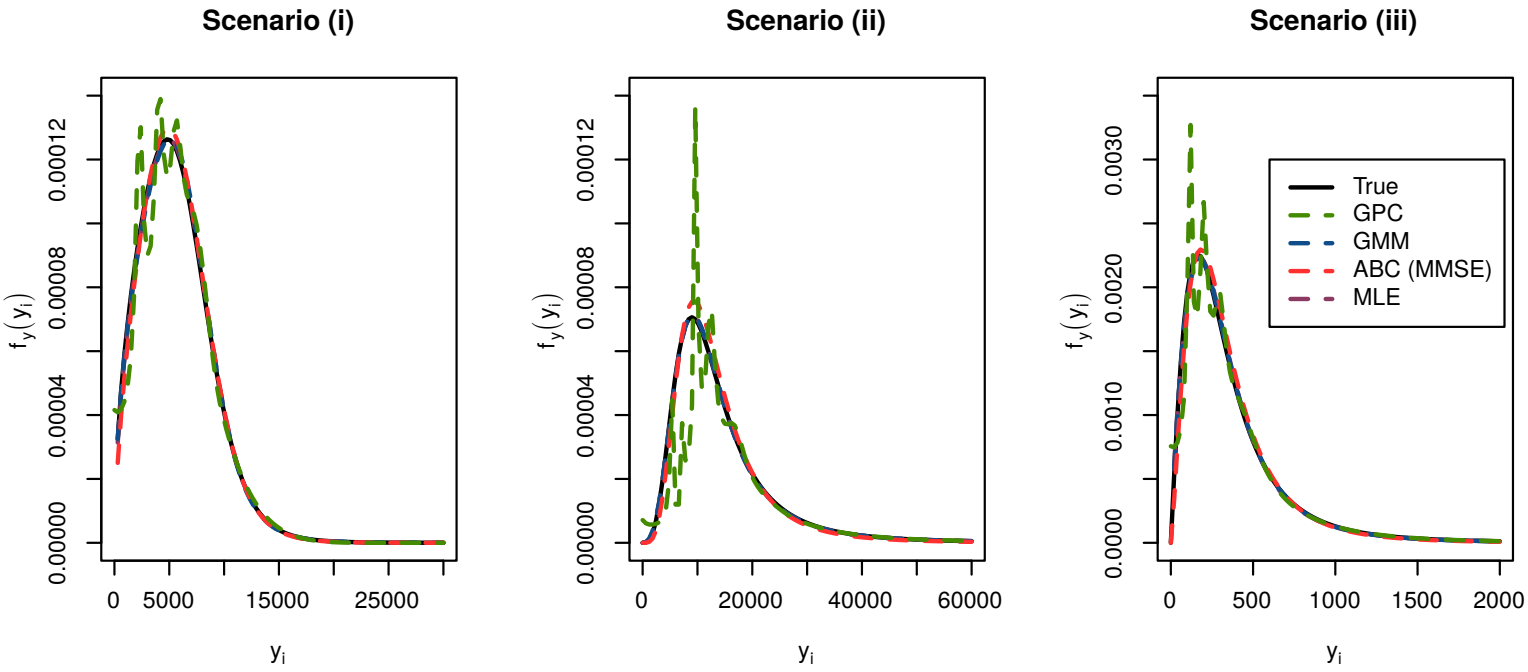

Figure 1: Fitted and true income distribution density curves for each estimation method and the three GB2 scenarios considered.

Table 2: *Simulated-data estimates under all GB2 scenarios and methods considered*

	Data			MLE (individual level)			GMM			ABC (MMSE)			GPC		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)
α	3.85	2.3	1.64	3.986	2.72	2.123	4.402	2.753	1.893	3.568	2.236	1.604	-	-	-
				(0.031)	(0.054)	(0.078)	(0.363)	(0.215)	(0.145)	(0.032)	(0.036)	(0.016)			
β	10000	10000	386	9979.11	10251.948	361.745	9435.045	10388.142	380.266	10513.482	10076.65	406.926	-	-	-
				(0.000)	(0.000)	(0.001)	(443.973)	(293.289)	(16.223)	(67.792)	(45.427)	(2.323)			
p	.4	1.75	1.22	0.388	1.334	0.841	0.346	1.294	0.979	0.464	2.139	1.408	-	-	-
				(0.274)	(0.069)	(0.145)	(0.033)	(0.161)	(0.106)	(0.004)	(0.042)	(0.018)			
q	2	1.25	1.83	1.958	0.972	1.217	1.587	0.966	1.487	2.491	1.551	2.21	-	-	-
				(0.029)	(0.107)	(0.086)	(0.257)	(0.107)	(0.187)	(0.045)	(0.029)	(0.035)			
u_k	s_k^{Obs}			s_k			s_k			s_k			s_k		
.10	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.03	0.02
.20	0.06	0.07	0.05	0.06	0.07	0.05	0.06	0.07	0.05	0.06	0.08	0.05	0.06	0.07	0.05
.30	0.12	0.13	0.09	0.12	0.13	0.09	0.12	0.13	0.09	0.12	0.14	0.10	0.12	0.13	0.09
.40	0.19	0.19	0.14	0.19	0.19	0.14	0.19	0.19	0.14	0.20	0.21	0.16	0.19	0.19	0.14
.50	0.28	0.27	0.21	0.28	0.27	0.21	0.28	0.27	0.21	0.29	0.29	0.23	0.28	0.27	0.21
.60	0.38	0.35	0.29	0.38	0.35	0.29	0.38	0.35	0.29	0.39	0.38	0.31	0.38	0.35	0.29
.70	0.50	0.45	0.39	0.50	0.45	0.39	0.50	0.45	0.39	0.50	0.48	0.41	0.50	0.45	0.39
.80	0.63	0.57	0.51	0.63	0.57	0.51	0.63	0.57	0.51	0.64	0.60	0.54	0.63	0.57	0.51
.90	0.79	0.72	0.68	0.79	0.71	0.67	0.79	0.72	0.68	0.79	0.75	0.70	0.79	0.72	0.68
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Gini	.309	.353	.435	0.31	0.35	0.44	0.31	0.35	0.44	0.30	0.31	0.40	0.31	0.35	0.44
μ	5676.5	15142.9	385.5	5676.98	15180.78	389.25	5678.30	15138.11	385.41	5701.10	14277.30	374.40	5676.51	15142.89	385.47

Note: The Data columns presents true parameter values and sample decile shares, sample Gini coefficient, and sample mean income for each of the three settings explored. MLE estimates are obtained over the simulated individual-level samples, while GMM, ABC, and GPC are the corresponding grouped-data estimates. Estimated standard errors for parameter point-estimates given in parenthesis. Presented ABC results summarize the estimated posterior distribution by its respective mean (i.e., the minimum mean square error (MMSE) estimator) and standard errors were computed using a batch means estimator (Flegal et al. 2020) to account for the autocorrelations inherent to MCMC output.

4.2 Applications to real data from *PovCalNet*

The *PovCalNet* data on the income distribution of Peru in 2013, already presented in table 1, provides an interesting scenario for empirically assesing the proposed ABC inference method on real data on incomes in light of other recent methods. This is so not only because of the popularity and conventional format of the *PovCalNet* data but also because it covers a scenario recently explored in Eckernkemper and Gribisch (2021) and therefore results from their survey on grouped-data methods can serve as a benchmark.

Under the same specifications as in the previous section, the GMM, ABC, and GPC methods were used to estimate an income distribution from these grouped-data statistics. The resultings point-estimates are presented in table 3 below. As a first observation, in common with what is observed in the simulated data setting, all three methods yield estimates which match the data well in terms of income shares, Gini, and mean income. The relationship between GMM and ABC MMSE parameter estimates is also similar to that obtained under simulated data, with GMM yielding a value for the α parameter above the point-estimate obtained under ABC and the latter yielding values for the β, p, and q slightly above those obtained under the former. Additionaly, here also the standard errors of the GMM estimates are above those of ABC. Finally, these parameter estimates obtained under the GMM and ABC methods closely match those obtained for the same data in Eckernkemper and Gribisch (2021, table 7) under their quasi-likelihood approximation. In particular, the GMM estimates yield identical parameter estimates and standard errors as this and their method are asymptotically equivalent.

Comparing the density curves implied by these three estimates, figure 2 evidences that both GMM and ABC point-estimates yield an almost identical distribution, while GPC once again gives estimates for the middle and lower part of the distribution strongly in discrepancy with that obtained under the other methods.

Another interesting point of comparison across the three methods concerns the estimation of the unknown bounds $\{z_k\}_{k=1}^{K-1}$ underlying the data. Both GPC and GMM must explicitly employ optimization steps to estimate these bounds. The ABC estimator can be seen as more versatile as no analytical expressions or specific optimizations are required for working around the problem of unknown bounds as long as the assumed parametric model for the income distribution allows for simulating data. Figure 3 below presents estimated posterior marginal distributions for each of the 9 unknown bounds (i.e., deciles) in the data used in this application, each being a function of the estimated joint posterior distribution for the model parameters $\{\tilde{z}_k(\tilde{\boldsymbol{\theta}}^{(j)})\}_{j=1}^{J}$, $\tilde{\boldsymbol{\theta}} \sim \pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^{K})$, along with the ABC MMSE, GMM, and GPC point-estimates for them. This figure evidences very similar estimates for all bounds for the ABC and GMM methods, while the GPC estimates give strongly constrasting results for the lower groups in the data. Interestingly, the 9-th decile, defining the lower bound of the top income group in the data is estimated at an almost identical value by all three methods. The strong similarities between the estimated bounds between GMM and ABC suggest that the latter can be a fruitful approach to parameter inference in cases where the expressions required to estimate group bounds under GMM or BQL for a specific model are unavailable.

Finally, in assesing the fit of the estimated GB2 distribution under ABC to the

data, figure 4 presents the ABC posterior marginal discrepancies using violin plots[16]. These plots describe, for the metric chosen in the ABC algorithm, the distances between the observed data GLC_k^{Obs} and the simulated data $GLC_k(\tilde{\boldsymbol{\theta}}^{(j)})$ for each point in the estimated joint posterior distribution $\tilde{\boldsymbol{\theta}}^{(j)} \sim \pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$ and can therefore help identify specific aspects of the data in which the estimates might systematically err. In this case, the marginal posterior distribution of the discrepancies with respect to each of the ten $\{GLC_k^{Obs}\}_{k=1}^K$ points in the data do not suggest any such problems in terms of goodness-of-fit. Having set the ABC rejection kernel bandwidth parameter at $\tau = (.001, .001, .001, .001, .0005, .0005, .0005, .0005, .0005, .0005)$ allows for a relatively looser fit with respect to the decile groups below the median, a behaviour clearly present in the resulting estimated posterior discrepancies for the first and second decile groups, while all discrepancies have an estimated marginal posterior distribution heavily concentrated below their respective value in τ.

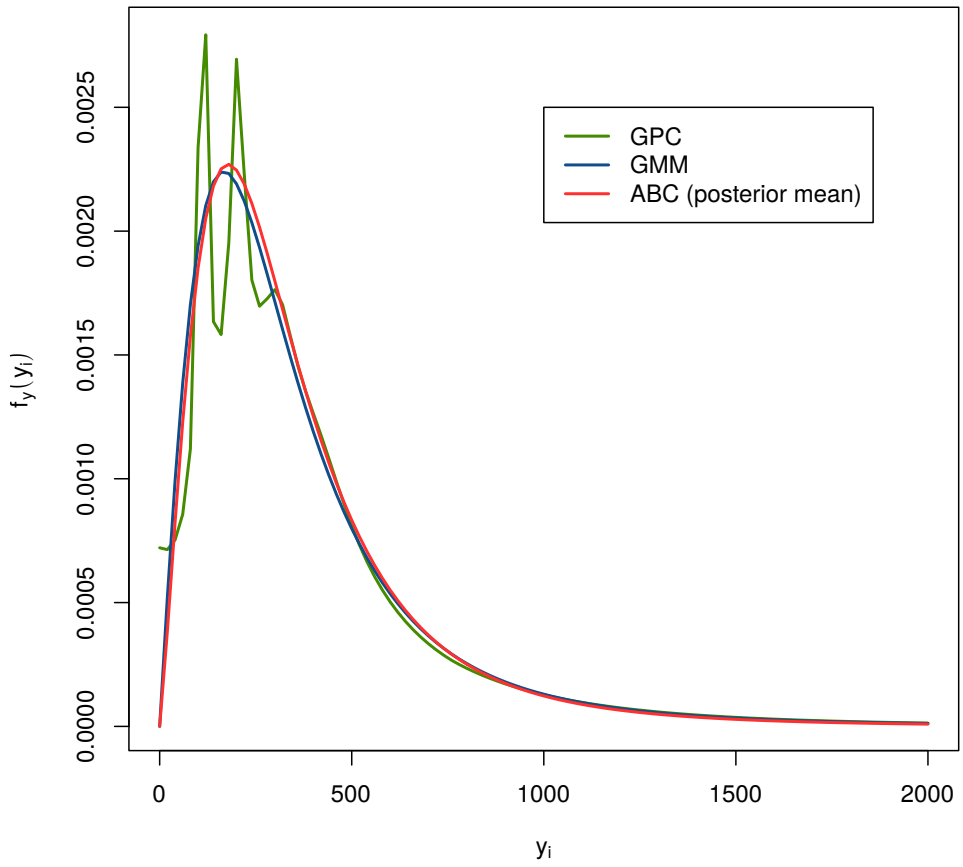

Figure 2: Fitted income distributions for PovCalNet data on Peru's income distribution for 2013.

[16]Violin plots are a graphical extension of the conventional box plot to include kernel density estimates over the same data

Table 3: *Peru 2013 income distribution estimates under all grouped-data methods considered*

	Data	GMM	ABC (MMSE)	GPC
α	-	1.645	1.579	-
		(0.08)	(0.018)	-
β	-	388.737	412.926	-
		(8.337)	(2.727)	-
p	-	1.213	1.385	-
		(0.088)	(0.02)	-
q	-	1.825	2.206	-
		(0.152)	(0.041)	-
u_k	s_k^{Obs}			
.10	0.02	0.02	0.02	0.02
.20	0.05	0.05	0.05	0.05
.30	0.09	0.09	0.10	0.09
.40	0.14	0.14	0.15	0.14
.50	0.21	0.21	0.22	0.21
.60	0.29	0.29	0.31	0.29
.70	0.39	0.39	0.41	0.39
.80	0.51	0.51	0.53	0.51
.90	0.67	0.68	0.70	0.67
1	1.00	1.00	1.00	1.00
Gini	0.43	0.44	0.41	0.44
μ	387.82	387.83	377.65	387.82

Note: The Data columns presents sample decile shares, sample Gini coefficient, and sample mean income for PovCalNet data on Peru's income distribution for 2013. GMM, ABC, and GPC are the corresponding grouped-data estimates. Estimated standard errors for parameter point-estimates given in parenthesis. As in table 2, presented ABC results summarize the estimated posterior distribution by its respective mean and standard errors were computed using a batch means estimator (Flegal et al. 2020).

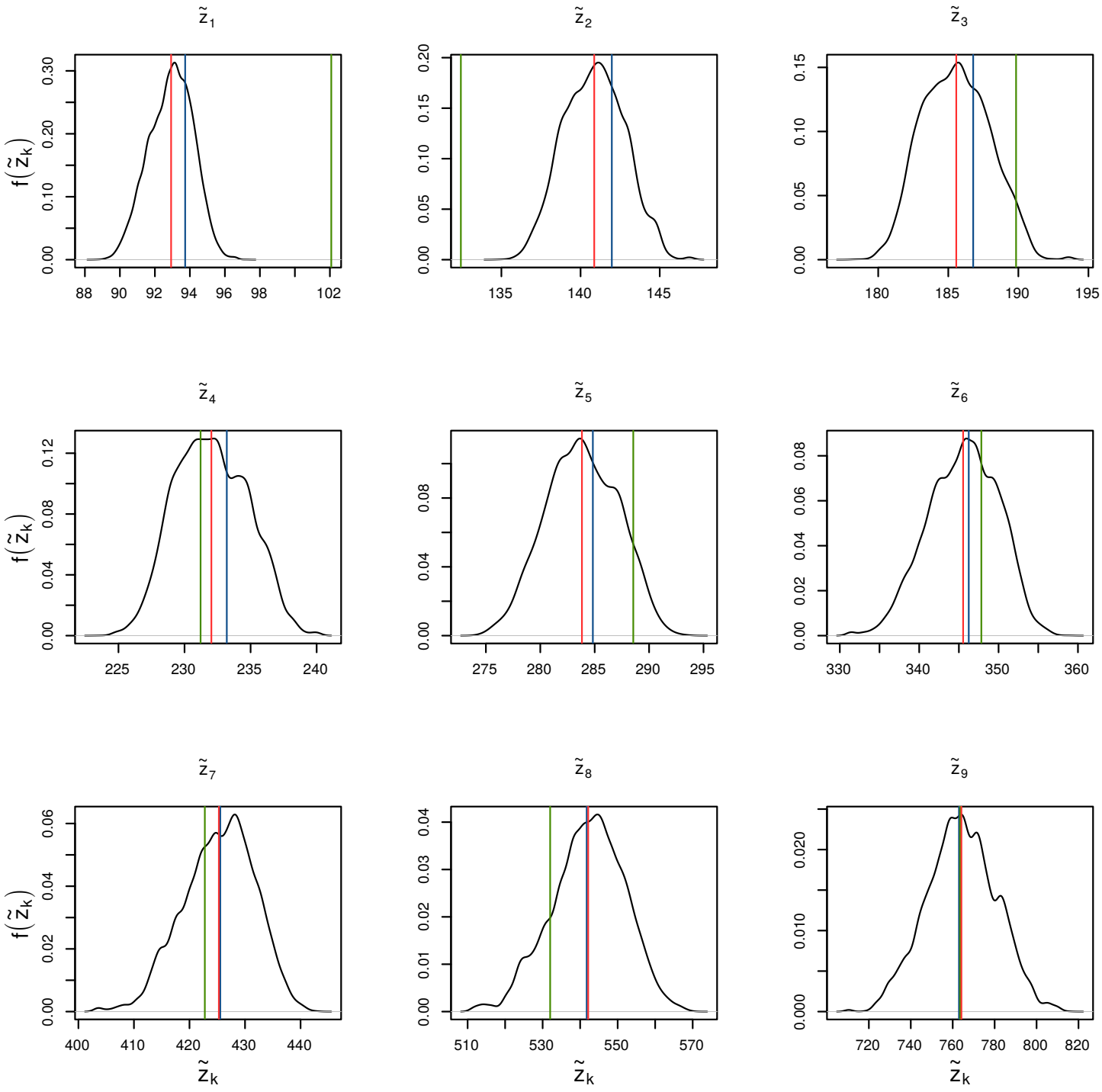

Figure 3: Estimated ABC posterior marginal distributions for the unknown decile group bounds in PovCalNet data on Peru's income distribution for 2013. Posterior mean estimate in red, GMM estimate in blue, and GPC estimate in green.

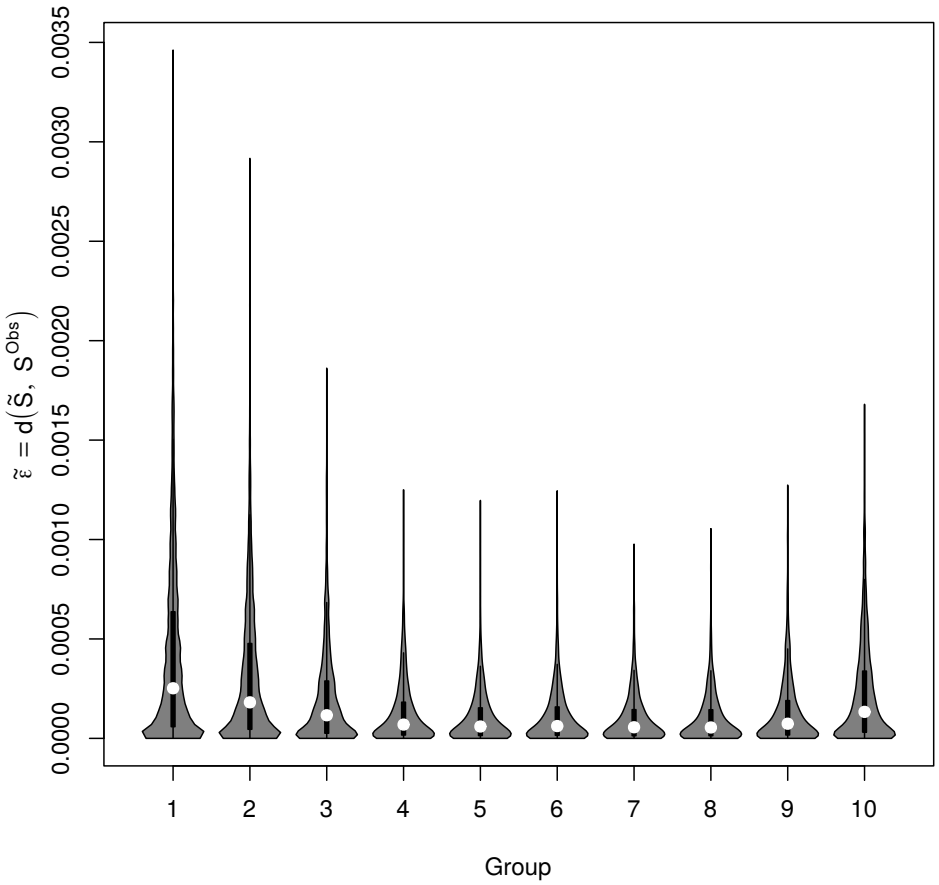

Figure 4: ABC posterior marginal discrepancies between the fitted GLC and empirical GLC over decile groups in PovCalNet data on Peru's income distribution for 2013. ABC rejection kernel bandwidth parameter set at
$\tau = (.001, .001, .001, .001, .0005, .0005, .0005, .0005, .0005, .0005)$.

5 Concluding remarks

A new Bayesian approach to estimating income distributions from grouped data with non-deterministic and unknown bounds is explored in this paper. In following with recent Bayesian approaches bypassing the issue of unavailable likelihood functions for these type I data through simulation-based inference, ABC methods can offer a particularly versatile approach due to the non-parametric nature of its approximation to said likelihood.

The versatility of this method distinguishes it from other recent methods such as the GMM or the BQL which require computable expressions for various aspects of the assumed parametric model for the distribution of incomes which might not be available and/or implemented in statistical software for all such models. As ABC avoids this requirement by only requiring that data can be simulated from the specified model it can potentially be applied to a wider range of models than these methods.

The particular ABC approach explored in this paper exploits the GLC as a representation of the typically available information in type I grouped data. This representation is informative on both the shape and the scale of the underlying income distribution and therefore allows for inference on all parameters of the commonly assumed parametric distributions for modelling incomes. In this sense, the proposed method overcomes the limitations of the previously explored application of ABC to grouped data on incomes of Kobayashi and Kakamu (2019) which focused solely on inference in shape parameters and scale-invariant quantities of the income distribution, and places it amongst the latest approaches seeking to exploit all information available in the data such as the BQL approach.

Additionally, in the interest of exploiting the ABC discrepancies $\{\{\varepsilon_k^{(j)}\}_{k=1}^K\}_{j=1}^J$ for model criticism the proposed algorithm extends standard ABC MCMC algorithms to perform online computing by simply storing the measured ABC discrepancies at each step of the algorithm. This avoids the typical computational cost of producing these measures *ex-post* by re-drawing new simulated datasets from the estimated posterior distribution of the model parameters.

Applications presented on both simulated data and real data suggest this ABC approach to be suited no worse than other recent methods devised for type I grouped data. Both income shares and scale-dependent quantities of the income distribution such as the unknown group bounds are estimated with a high level of precision through ABC in these applications, with estimated standard errors for the model parameter estimates being significantly below those obtained by other methods.

As implemented in this paper, however, the ABC estimator suffers from certain drawbacks which future research could overcome. In similarity with the simulated-data applications of the ABC method in Kobayashi and Kakamu (2019) the applications presented in this paper assuming the GB2 distribution suggest that this format of grouped data is potentially more informative on the α shape parameter than the other p and q shape parameters, with the respective prior distributions being potentially less revised by the ABC approximated likelihood and therefore yielding an estimated posterior distribution closer in shape to said prior than it is the case for α. Two calibration exercises could be

done for potentially improving on these aspects of the method. Firstly, the performance of the estimator on simulated data could be explored under alternative less skewed prior distributions for the p and q parameters. Secondly, aspects of the data, such as specific deciles along the income distribution, could potentially be more informative on these parameters than others. The particular ABC estimator as implemented in these applications employs a multivariate metric and rejection kernel with a bandwidth parameter τ that allows for a more strict fit of the model to specific decile groups in the data. Studying the performance of the estimator under different specifications for this bandwidth parameter can help understand which aspects of the data should be particularly emphasized to obtain estimated posterior distributions for p and q which are less influenced by their prior.

A second drawback of the ABC method in this paper involves the MCMC sampling algorithm used in the interest of obtaining samples from the ABC target posterior distribution $\pi_\tau(\boldsymbol{\theta}|\{u_k, GLC_k^{Obs}\}_{k=1}^K)$. This type of sampling algorithm is an invalid approach if the Markov chain fails to converge, and so convergence must be assesed by the analyst before interpreting MCMC output as a sample from the target posterior distribution. An alternative sampling strategy for ABC inference, often more computationally efficient than MCMC, is that of Sequential Monte Carlo (SMC, Sisson et al. 2007). Future research exploring SMC sampling in the context of the present ABC framework could pose an alternative to the issue of checking convergence posed by MCMC, as SMC samplers explicitly require an *ex-ante* automatic rule determining when the sampler is considered to have converged.

Finally, the versatility of the ABC approach could be exploited in future work by considering more complex models for grouped data of type I than those explored in this paper. In particular, additional components could be included as part of the assumed DGP for the data, such as parametric models of high-income under-reporting or 'missing rich' coverage errors, for which deriving a grouped-data likelihood is yet more challenging. These extensions could allow for Bayesian inference on the income distribution contemplative of these phenomena which are commonly believed to affect virtually any survey-based dataset on incomes.

References

An, D. and Little, R. J. (2007). Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):923–940.

Anand, S. and Segal, P. (2008). What do we know about global income inequality? *Journal of Economic Literature*, 46(1):57–94.

Beaumont, M. A. (2019). Approximate bayesian computation. *Annual review of statistics and its application*, 6:379–403.

Blanchet, T., Fournier, J., and Piketty, T. (2017). Generalized Pareto curves: Theory and applications. *WID.world Working Paper 2017/3*.

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

Chotikapanich, D. and Griffiths, W. E. (2002). Estimating lorenz curves using a dirichlet distribution. *Journal of Business & Economic Statistics*, 20(2):290–295.

Clarté, G., Robert, C. P., Ryder, R. J., and Stoehr, J. (2021). Componentwise approximate bayesian computation via gibbs-like steps. *Biometrika*, 108(3):591–607.

Deininger, K. and Squire, L. (1996). A new data set measuring income inequality. *The World Bank Economic Review*, 10(3):565–591.

Drovandi, C. and Frazier, D. T. (2021). A comparison of likelihood-free methods with and without summary statistics. *arXiv preprint arXiv:2103.02407*.

Eckernkemper, T. and Gribisch, B. (2021). Classical and bayesian inference for income distributions using grouped data. *Oxford Bulletin of Economics and Statistics*, 83(1):32–65.

Flegal, J. M., Hughes, J., Vats, D., and Dai, N. (2020). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN. R package version 1.4-1.

Gastwirth, J. L. (1971). A general definition of the lorenz curve. *Econometrica: Journal of the Econometric Society*, pages 1037–1039.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient metropolis jumping rules. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics V. Proceedings of the Fifth Valencia International Meeting*. Oxford University Press.

Goldsmith, S., Jaszi, G., Kaitz, H., and Liebenberg, M. (1954). Size distribution of income since the mid-thirties. *The Review of Economics and Statistics*, pages 1–32.

Graf, M. and Nedyalkova, D. (2014). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth*, 60(4):821–842.

Graf, M. and Nedyalkova., D. (2015). *GB2: Generalized Beta Distribution of the Second Kind: Properties, Likelihood, Estimation*. R package version 2.1.

Griffiths, W. and Hajargasht, G. (2015). On gmm estimation of distributions from grouped data. *Economics Letters*, 126:122–126.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hajargasht, G. and Griffiths, W. E. (2013). Pareto–lognormal distributions: Inequality, poverty, and estimation from grouped income data. *Economic Modelling*, 33:593–604.

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

Jenkins, S. P. (2009). Distributionally-sensitive inequality indices and the gb2 income distribution. *Review of Income and Wealth*, 55(2):392–398.

Jorda, V., Sarabia, J. M., and Jäntti, M. (2020). Estimation of income inequality from grouped data. *LIS Working Paper*, No. 804.

Jorda, V., Sarabia, J. M., and Jäntti., M. (2021). *GB2group: Estimation of the Generalised Beta Distribution of the Second Kind from Grouped Data*. R package version 0.3.0.

Kakwani, N. (1984). Welfare ranking of income distributions. *Advances in econometrics*, 3:191–213.

Kakwani, N. C. and Podder, N. (1973). On the estimation of lorenz curves from grouped observations. *International Economic Review*, pages 278–292.

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

Kobayashi, G. and Kakamu, K. (2019). Approximate Bayesian computation for Lorenz curves from grouped data. *Computational Statistics*, 34(1):253–279.

Kuznets, S. (1955). Economic growth and income inequality. *The American economic review*, 45(1):1–28.

Lahoti, R., Jayadev, A., and Reddy, S. (2016). The global consumption and income project (gcip): An overview. *Journal of Globalization and Development*, 7(1):61–108.

Lakner, C. and Milanovic, B. (2016). Global income distribution: From the fall of the Berlin wall to the great recession. *The World Bank Economic Review*, 30(2):203–232.

Lebergott, S. (1959). The shape of the income distribution. *The American Economic Review*, 49(3):328–347.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52(3):647–663.

McDonald, J. B. and Ransom, M. (2008). The generalized beta distribution as a model for the distribution of income: estimation of related measures of inequality. In *Modeling Income Distributions and Lorenz Curves*, pages 147–166. Springer.

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

Parker, S. C. (1999). The generalised beta as a model for the distribution of earnings. *Economics Letters*, 62(2):197–200.

Piketty, T. (2003). Income inequality in france, 1901–1998. *Journal of political economy*, 111(5):1004–1042.

Piketty, T., Alvaredo, F., Chancel, L., Saez, E., and Zucman, G. (2019). Distributional national accounts. In Stiglitz, J. E., Fitoussi, J.-P., and Durand, M., editors, *For Good Measure: An Agenda for Moving Beyond GDP*. The New Press.

Piketty, T. and Saez, E. (2003). Income inequality in the united states, 1913–1998. *The Quarterly journal of economics*, 118(1):1–41.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11.

Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106(26):10576–10581.

Ratmann, O. R. (2010). *Approximate Bayesian Computation under model uncertainty*. PhD thesis, Imperial College London (University of London).

Sarabia, J. M. (2008). Parametric Lorenz curves: Models and applications. In *Modeling income distributions and Lorenz curves*, pages 167–190. Springer.

Shorrocks, A. F. (1983). Ranking income distributions. *Economica*, 50(197):3–17.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Comput Biol*, 9(1):e1002803.