Efficient Maximum Likelihood Estimation for Income Distributions using Grouped Data

Tobias Eckernkemper [*] and Bastian Gribisch [†]

(May 30, 2018)

Abstract

We develop a framework for maximum likelihood (ML) estimation of income distributions based on grouped data. We explicitly account for unknown group boundaries and allow for two data generating processes (DGPs) corresponding to two different methods of grouping observations. Dependent on the type of DGP, the likelihood exploits different data information including group means and group boundaries, which have up to now not been included in ML inference for grouped data. A comprehensive simulation experiment shows that the proposed ML framework improves the statistical precision of parameter estimates relative to the classical multinomial likelihood. The results furthermore indicate that the precision is not significantly affected if the group boundaries are not available. We finally provide an empirical application to a set of countries included in the World Bank database *PovcalNet*.

JEL classification: C21, C51, D31

Keywords: Maximum Likelihood; Efficient inference; Simulation based inference; Generalized beta distribution; Inequality and poverty.

[*] Institute of Econometrics and Statistics, University of Cologne, Universitaetsstr. 22a, D-50937 Cologne, Germany. Tel.: +49(0)2214702283; Fax: +49(0)2214705074. *E-mail address:* eckernkemper@statistik.uni-koeln.de

[†] Corresponding author. Institute of Econometrics and Statistics, University of Cologne, Universitaetsstr. 22a, D-50937 Cologne, Germany. Tel.: +49(0)2214707711; Fax: +49(0)2214705074. *E-mail address:* bastian.gribisch@statistik.uni-koeln.de

1 Introduction

The empirical analysis of welfare, income inequality and poverty requires precise estimates of the distribution of income.[1] If the data is fully released, the distribution can be estimated by standard parametric or non-parametric methods like Maximum Likelihood (ML) or kernel density estimation. Especially for developing countries it is however common that researchers can only access grouped income data which is e.g. provided by the World Bank and the World Institute for Development Economics Research (WIDER). The data typically consists of population shares and group-specific mean-incomes for ten to twenty income groups, where the group boundaries are not provided. The apparent problem of limited data then makes the parametric approach to income distributions more popular and also more natural than non-parametric techniques.

The literature provides a variety of parametric income distributions including, but not limited to Pareto's distribution, the lognormal distribution, Champernowne's distribution, Fisk's distribution, the gamma-, generalized gamma-, Weibull-, Singh-Maddala- and Dagum distribution (see e.g. Kleiber and Kotz, 2003). McDonald (1984) proposed the generalized beta distribution of the second kind (GB2 distribution), which nests the lognormal, generalized gamma, Singh-Maddala, Beta-2 and Dagum distributions. Parker (1999) showed that the GB2 distribution can be derived from microeconomic principles and the distribution has therefore become very popular in applied economic research. An alternative, flexible way of income modeling is based on mixture distributions, which are e.g. analyzed by Griffiths and Hajargasht (2012).

Contributions on statistical inference for grouped income data are rare. The traditional and most frequently applied method is ML based on sample proportions using a multinomial likelihood function (see e.g. McDonald, 1984, and Bandourian et al, 2003). This approach is inefficient in the majority of practical applications since it neglects the information content of observed group means and cannot account for unknown group boundaries. Recent work then focused on nonlinear least squares and GMM estimation, where relative population- and income-shares are effectively matched to their theoretical counterparts (see e.g. Wu and Perloff, 2005, Wu, 2006, Chotikapanich et al., 2007 and 2012). Hajargasht et al. (2012) propose an efficient GMM framework which accounts for unknown group boundaries but lacks a solid statistical foundation with respect to the underlying data generating process (DGP). Hajargasht and Griffiths (2016) shift the focus from income distributions to parametric Lorenz curves and provide

[1] An overview on the vast and growing literature on statistical inference for income distributions is e.g. provided by Kleiber and Kotz (2003), Chotikapanich (2008) and Bandourian et al. (2003).

a GMM framework covering two DGPs of empirical relevance. Finally, Chen (2016) generalizes the GMM framework to incorporate varying data information.

The present paper contributes to the literature by offering the first comprehensive discussion of efficient ML estimation of parametric income distributions for grouped income data with potentially unknown boundaries. ML is well known for its asymptotic efficiency and derived estimates of poverty or inequality measures directly inherit this property. The likelihood approach also opens the door to Bayesian inference for grouped income data. We explicitly account for two DGPs corresponding to two methods of grouping observations. The first method (DGP1) builds on proportions of observations in each income group, which have been fixed prior to sampling. As a result the group income means and group boundaries are random. In the second method of grouping (DGP2) the group boundaries are predetermined prior to sampling. Hence both the number of observations and the income means in each group are random. Income data from the World Bank or WIDER typically correspond to DGP1 with unknown group boundaries. Dependent on the type of DGP the likelihood comprises varying data information including group means and group boundaries. The multinomial ML method of McDonald (1984) fits DGP2 with known boundaries. The according likelihood is misspecified in case of DGP1 and the method cannot be applied if boundaries are unknown.

Efficient ML requires the derivation of the joint (conditional) density of group mean-incomes. This distribution is unknown for all relevant income distributions, but converges to the Gaussian by standard central limit arguments. We approximate the density by a product of Normals with moments given by their asymptotic counterparts. Under DGP1 the group boundaries constitute random order statistics and can easily be included in the likelihood (known boundaries) or integrated out from the joint density of group means and boundaries via MC integration (unknown boundaries: e.g. World Bank or WIDER data). Under DGP2 both group means and relative population shares are random and the likelihood results from the product of the joint conditional density of group means and the standard multinomial likelihood. If group boundaries are unknown, we can simply estimate them along with the remaining model parameters.

We focus on the GB2 distribution and provide an extensive simulation experiment which shows the efficiency gains of our new ML method. Our results indicate significant improvements over the conventional multinomial approach and we obtain very accurate parameter estimates which come close to those obtained for individual income data. We also note that the estimation efficiency does not suffer if group boundaries are unknown. This finding is of practical relevance, since group boundaries are usually not provided in the World Bank or WIDER data sets.

We finally apply our method to income data for four countries obtained from the World

Bank *PovcalNet* data base. An evaluation of the goodness of fit using likelihood ratio tests and an income share prediction exercise strongly favors the GB2 distribution relative to its nested competitors.

The remainder of this paper is organized as follows. Section 2 gives general definitions and discusses the relevant data generating processes. Section 3 introduces our ML approach and Section 4 provides a simulation experiment in order to assess the finite sample performance of the estimators. Section 5 provides the empirical application and Section 6 concludes.

2 Definitions and Data Generating Processes

Let $y_1, ..., y_n$ be a random sample from a parametric distribution with density function $f_y(y;\theta)$, distribution function $F_y(y;\theta)$ and moment distribution function

$$F_\ell(y;\theta) = \frac{1}{E[y^\ell]} \int_0^y t^\ell f_y(t;\theta)\ dt, \quad \ell = 1, 2, \ldots \tag{1}$$

where y denotes income and θ comprises the model parameters. In the following we assume that the first and second moments of y exist. For the GB2 distribution we e.g. obtain $\theta = (a, b, p, q)'$ with $a, b, p, q > 0$ and

$$\begin{aligned} f_y(y;\theta) &= \frac{ay^{ap-1}}{b^{ap}\mathsf{B}(p,q)(1+(y/b)^a)^{p+q}}, \\ F_y(y;\theta) &= B_u(p,q), \\ F_\ell(y\ ;\ \theta) &= B_u(p+\ell/a,\ q-\ell/a), \\ E[y^\ell] &= b^\ell \frac{\mathsf{B}(p+\ell/a,\ q-\ell/a)}{\mathsf{B}(p,q)}, \end{aligned}$$

where $u = (y/b)^a/[1+(y/b)^a]$, $\mathsf{B}(\cdot)$ denotes the beta function and $B_u(\cdot)$ denotes the Beta distribution function evaluated at u. An overview of the GB2 and its nested distributions is provided in Table 1, which has been taken from Hajargasht et al. (2012).

Place Table 1 here.

The sample is grouped into K income groups where the boundaries are denoted by $\{z_{i-1}, z_i\}_{i=1}^K$ with $z_0 = 0$ and $z_K = \infty$. Let n_i denote the number of observations in income group i such that the sample size obtains as $n = \sum_{i=1}^K n_i$. Typical income data (e.g. World Bank or WIDER) contains information on relative population shares $c_i = n_i/n$ and group-specific mean incomes

$\bar{y}_i = (1/n_i)\sum_{j=1}^{n} y_j g_i(y_j)$, for $i = 1, \ldots, K$, where

$$g_i(y) = \begin{cases} 1 & \text{if } z_{i-1} < y \leq z_i \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In some cases we do not have data on mean incomes directly but observe the overall mean income $\bar{y}$ together with income shares $\{s_i\}_{i=1}^{K}$ instead, where $s_i = (n\bar{y})^{-1}\sum_{j=1}^{n} y_j g_i(y_j)$. Group-specific mean incomes are then obtained via $\bar{y}_i = s_i\bar{y}/c_i$. Group boundaries $\{z_i\}_{i=1}^{K-1}$ are usually not provided.

The method of grouping individuals into income classes is not unique and likelihood functions for ML estimation of θ must be tailored to the respective DGP in order to enable solid statistical inference. The upcoming subsections therefore define two distinct DGPs which are of particular relevance in practice.

2.1 DGP1: Fixed n_i and random z_i, $\bar{y}_i$

Under DGP1 the relative proportions of observations in each income group, $c_i = n_i/n$, are prespecified. This is the case for the majority of the data sets in the World Bank and the WIDER data base. Respective data consist of constant relative population shares corresponding e.g. to deciles or quintiles together with the respective mean incomes.

Denote the cumulative number of group observations by $n_i^c = \sum_{\ell=1}^{i} n_\ell$. Under DGP1 the group boundary z_i $(i = 1, \ldots, K-1)$ corresponds to the n_i^c'th order statistic $y_{[n_i^c]}$ from f_y, which represents a random variable.[2] The upper panel of Figure 1 depicts a schematic illustration of DGP1 for $n = 20$.

We summarize that DGP1 generates random group boundaries and group means, while relative proportions c_i and $n_i = n \cdot c_i$ are preset and therefore deterministic. The non-stochastic nature of the group frequencies renders the classical multinomial ML method of McDonald (1984) misspecified and ML estimation for DGP1 can only be based on the information contained in the group boundaries (if available) and the group means.

2.2 DGP2: Fixed z_i and random n_i, $\bar{y}_i$

DGP2 assumes pre-specified fixed group boundaries resulting in a random number of observations in each income group. Respective data-sets contain group means and relative population

[2] Note that, strictly speaking, the group boundary can take any value in $[y_{[n_i^c]}, y_{[n_i^c+1]})$. Corresponding data-generating processes are however observationally equivalent and $z_i \widehat{=} y_{[n_i^c]}$ therefore constitutes an identifying restriction.

shares which vary over income groups. Such data is rather infrequently met in practice - a few examples are found in the *PovcalNet* data base of the World Bank for selected countries and years. A schematic illustration of DGP2 is provided in the lower panel of Figure 1. The multinomial ML method of McDonald (1984) is designed under DGP2 with known group boundaries.

We summarize that DGP2 generates random population shares and group means, while group boundaries are preset and therefore deterministic. ML estimation for DGP2 can therefore be based on the information contained in both, the group means and the population shares. Note that the multinomial ML method of McDonald (1984) fits DGP2 with known boundaries but remains inefficient since the informational content of the group-specific mean incomes is not exploited.

Place Figure 1 here.

3 Maximum Likelihood Inference

3.1 DGP1: Fixed n_i and random $z_i,\ \bar{y}_i$

Under DGP1 and known group boundaries (KB) the likelihood for the complete set of observable data obtains as

$$L_{\text{DGP1, KB}}(\theta;\underline{\bar{y}},\underline{z}) \quad = \quad f(\underline{\bar{y}}|\underline{z};\theta)\cdot f(\underline{z};\theta), \tag{3}$$

where $\underline{\bar{y}} = \{\bar{y}_i\}_{i=1}^K$ and $\underline{z} = \{z_i\}_{i=1}^{K-1}$. Dependence on $\{n_i\}_{i=1}^K$ is suppressed for notational convenience.

The i'th group boundary z_i corresponds to the n_i^c'th order statistic of iid random variables from f_y. Exploiting the Markov property of order statistics (see e.g. David and Nagaraja, 2003, Theorem 2.5) we obtain the joint density of group boundaries in (3) as

$$f(\underline{z};\theta) = f(z_1;\theta)\cdot f(z_2|z_1;\theta)\cdot\ \ldots\ \cdot f(z_{K-1}|z_{K-2};\theta), \tag{4}$$

where standard calculus for order statistics gives

$$\begin{aligned} f(z_1;\theta) &= \frac{n!}{(n_1^c-1)!(n-n_1^c)!}\ F_y(z_1;\theta)^{n_1^c-1}\ [1-F_y(z_1;\theta)]^{n-n_1^c}\ f_y(z_1;\theta), & (5)\\ f(z_i|z_{i-1};\theta) &= \frac{(n-n_{i-1}^c)!}{(n_i^c-n_{i-1}^c-1)!(n-n_i^c)!} \\ &\quad \cdot\frac{[1-F_y(z_i;\theta)]^{n-n_i^c}}{[1-F_y(z_{i-1};\theta)]^{n-n_{i-1}^c}}\ [F_y(z_i;\theta)-F_y(z_{i-1};\theta)]^{n_i^c-n_{i-1}^c-1}\ f_y(z_i;\theta). & (6)\end{aligned}$$

By exploiting conditional independence the joint density of group means in Eq. (3) can be decomposed into

$$f(\underline{\bar{y}}|\underline{z};\theta) = f(\bar{y}_1\ |z_1;\theta) \cdot f(\bar{y}_2\ |z_1,z_2;\theta) \cdot ... \cdot f(\bar{y}_{K-1}\ |z_{K-2},z_{K-1};\theta) \cdot f(\bar{y}_K\ |z_{K-1};\theta), \tag{7}$$

where $z_0 = 0$ and $z_K = \infty$. The distribution of the arithmetic mean is unknown for any income distribution of practical relevance.[3] We therefore replace the individual constituents of $f(\underline{\bar{y}}|\underline{z};\theta)$ in Eq. (7) by approximations, which are consistent in the sense that the resulting approximation error diminishes to zero as $n \to \infty$. Employing the standard Lindberg Levy Central Limit Theorem (CLT) for iid random variables we obtain

$$f(\underline{\bar{y}}|\underline{z};\theta) \approx f_N(\bar{y}_1\ |z_1;\theta) \cdot f_N(\bar{y}_2\ |z_1,z_2;\theta) \cdot ... \cdot f_N(\bar{y}_{K-1}\ |z_{K-2},z_{K-1};\theta) \cdot f_N(\bar{y}_K\ |z_{K-1};\theta)\ , \tag{8}$$

where $f_N(\bar{y}_i|\ \cdot)$ denotes the density function of a Gaussian distribution. Since conditional on the group boundaries (z_{i-1}, z_i), the $n_i - 1$ individual stochastic incomes[4] in group i are independent and identically distributed with density function $f(y|z_{i-1}, z_i;\theta) = f(y|z_{i-1} < y < z_i\ ;\ \theta)$ (see David and Nagaraja, 2003, Theorem 2.5) we obtain the moments

$$\begin{aligned} \mu_i &= E(\bar{y}_i|z_{i-1}, z_i\ ;\ \theta) \\ &= \frac{n_i - 1}{n_i} E(y|z_{i-1} < y < z_i\ ;\ \theta) + \frac{z_i}{n_i} \\ &= \frac{n_i - 1}{n_i} \frac{[F_1(z_i;\theta) - F_1(z_{i-1};\theta)] \cdot E(y)}{F_y(z_i;\theta) - F_y(z_{i-1};\theta)} + \frac{z_i}{n_i}, \end{aligned} \tag{9}$$

and

$$\begin{aligned} \sigma_i^2 &= Var(\bar{y}_i|z_{i-1}, z_i\ ;\ \theta) \\ &= \frac{n_i - 1}{n_i^2} Var(y|z_{i-1} < y < z_i\ ;\ \theta) \\ &= \frac{n_i - 1}{n_i^2} \left[\frac{(F_2(z_i;\theta) - F_2(z_{i-1};\theta)) \cdot E(y^2)}{F_y(z_i;\theta) - F_y(z_{i-1};\theta)} - \mu_i^2 \right], \end{aligned} \tag{10}$$

where $i = 2, \dots, K-1$. Mean and variance for the first and the last income group are obtained by analogy while recognizing that $z_0 = 0$ and $z_K = \infty$.

Inserting the previously derived expressions into Eq. (3) the log-likelihood under DGP1 and known group boundaries obtains as

$$\begin{aligned} \mathcal{L}_{\text{DGP1, KB}}(\theta;\underline{\bar{y}},\underline{z}) \ = \ & \Omega - \frac{1}{2}\Big[\ln(\sigma_K^2) + \frac{(\bar{y}_K - \mu_K)^2}{\sigma_K^2}\Big] + \sum_{i=1}^{K-1} \Bigg\{ -\frac{1}{2}\Big[\ln(\sigma_i^2) + \frac{(\bar{y}_i - \mu_i)^2}{\sigma_i^2}\Big] \\ & +(n - n_i^c)\ln[1 - F_y(z_i;\theta)] - (n - n_{i-1}^c)\ln[1 - F_y(z_{i-1};\theta)] \\ & +(n_i^c - n_{i-1}^c - 1)\ln[F_y(z_i;\theta) - F_y(z_{i-1};\theta)] + \ln f_y(z_i;\theta) \Bigg\}, \end{aligned} \tag{11}$$

[3] See e.g. Nadarajah (2005) for the complex derivation of the distribution of the sum of only two GB2 distributed random variables.

[4] Note that conditional on z_i the last summand in $\bar{y}_i$ is deterministic and given by z_i.

where $n_0^c = F_y(z_0;\theta) = 0$, μ_i and σ_i^2 are given in Eqs. (9) and (10), and

$$\Omega = -\frac{K}{2}\ln(2\pi) + \sum_{i=1}^{K-1} \ln\left[(n-n_{i-1}^c)!\right] - \ln\left[(n-n_i^c)!\right] - \ln\left[(n_i^c - n_{i-1}^c - 1)!\right]. \tag{12}$$

ML estimation of θ is carried out by maximizing the log-likelihood function in Eq. (11) using numerical techniques routinely available in standard software packages. Asymptotic standard errors are obtained by inverting a numerical approximation to the hessian at the ML estimates.

Note that the approximation in (8) is consistent for $n \to \infty$ by standard CLT arguments for iid random variables, leaving the asymptotic properties of the ML estimator unaffected. The quality of the approximation is further analyzed in Sections 4 and 5. The results imply overall accurate approximations even for relatively low sample-sizes with $n_i = 1,000 \; \forall \; i$.

The majority of the WIDER and World Bank data sets do not report group boundaries. We then marginalize the joint density in Eq. (3) w.r.t. the mean incomes. The resulting likelihood under DGP1 and unknown boundaries (UB) boils down to a $(K-1)$-dimensional interdependent integral given by

$$L_{\text{DGP1, UB}}(\theta;\underline{\bar{y}}) = \int f(\underline{\bar{y}}|\underline{z};\theta)\cdot f(\underline{z};\theta)\; d\underline{z}, \tag{13}$$

which cannot be solved analytically. Typical income data consists of ten to twenty income groups. Such low-dimensional integration problems are easily solved by standard simulation based integration routines up to any desired degree of accuracy.

We approximate the likelihood function in (13) using importance sampling (see e.g. Robert and Casella, 2004) and rewrite the integral as

$$\begin{aligned} L_{\text{DGP1, UB}}(\theta;\underline{\bar{y}}) &= \int \frac{f(\underline{\bar{y}}|\underline{z};\theta)\cdot f(\underline{z};\theta)}{m(\underline{z};\theta)}\cdot m(\underline{z};\theta)\; d\underline{z} \\ &= E_{m(\underline{z};\theta)}\left[\frac{f(\underline{\bar{y}}|\underline{z};\theta)\cdot f(\underline{z};\theta)}{m(\underline{z};\theta)}\right], \end{aligned} \tag{14}$$

where $m(\underline{z};\theta)$ denotes an appropriately chosen importance density. A consistent estimate of the log-likelihood function then obtains as

$$\mathcal{L}_{\text{DGP1, UB}}(\theta;\underline{\bar{y}}) = \ln\left[\frac{1}{S}\sum_{i=1}^{S}\frac{f(\underline{\bar{y}}|\underline{z}^{(i)};\theta)\cdot f(\underline{z}^{(i)};\theta)}{m(\underline{z}^{(i)};\theta)}\right], \tag{15}$$

where $\underline{z}^{(i)} = \{z_\ell^{(i)}\}_{\ell=1}^{K-1}$ denotes a trajectory simulated from the importance density $m(\underline{z};\theta)$ and S is the simulation sample size.

A standard approach to constructing importance densities employs a local approximation of the logarithm of the integrand in Eq. (13) as a function in $\underline{z}$. Using a 2'nd order Taylor series expansion around the mode we obtain a multivariate Normal importance density with mean vector $\underline{z}_*$ and covariance matrix H_*^{-1}, where $\underline{z}_*$ is the maximizer of $\ln(f(\underline{\bar{y}}|\underline{z};\theta)\cdot f(\underline{z};\theta))$ and

H_* denotes the corresponding hessian evaluated at $\underline{z}_*$. We obtain both $\underline{z}_*$ and H_* via numerical approximations using standard routines implemented in the MATLAB software package. Existence of the expectation in (14) requires the importance density to have fatter tails than the target density $f(\underline{\bar{y}}, \underline{z})$. We therefore replace the Gaussian by a corresponding Student's-t importance density, where we set the d.o.f to 8.[5] Simulated ML estimates of θ are then obtained by maximizing the likelihood approximation in (15) using a standard numerical optimizer. The convergence of such an optimizer requires the likelihood to be continuous in θ. This is achieved by computing $\mathcal{L}_{\text{DGP1, UB}}(\theta; \underline{\bar{y}})$ for different values of θ under a set of Common Random Numbers (CRNs) (see e.g. Liesenfeld and Richard, 2006). This means that all $\{\underline{z}^{(i)}\}_{i=1}^{S}$ draws for different values of θ are obtained by transformation of a common set of canonical random numbers, here standardized Student-t's. Our simulation experiment in Section 4 shows that $S \mathrel{\widehat{=}} 10,000$ results in fast and accurate importance-approximations of the likelihood function.

3.2 DGP2: Fixed z_i and random n_i, $\bar{y}_i$

DGP2 generates random numbers of observations n_i and random mean incomes $\bar{y}_i$ for each group. The (efficient) likelihood comprising all available data information then obtains as

$$\mathcal{L}_{\text{DGP2}}(\theta; \underline{\bar{y}}, \underline{n}) \quad = \quad f(\underline{\bar{y}} | \underline{n}; \theta) \cdot f(\underline{n}; \theta), \tag{16}$$

where $\underline{n} = \{n_i\}_{i=1}^{K}$ and dependence on $\underline{z}$ is suppressed for notational convenience.

The distribution of $\underline{n}$ is multinomial with density function

$$f(\underline{n}; \theta) = \frac{n!}{n_1! \cdot \ldots \cdot n_K!} \cdot \pi_1^{n_1} \cdot \ldots \cdot \pi_K^{n_K}, \tag{17}$$

where

$$\pi_i = \Pr(z_{i-1} < y \leq z_i) = F_y(z_i; \theta) - F_y(z_{i-1}; \theta), \quad i = 1, \ldots, K.$$

We then obtain the log-likelihood under DGP2 via inserting (17) and the Gaussian approximation (8) in Eq. (16):

$$\mathcal{L}_{\text{DGP2}}(\theta; \underline{\bar{y}}, \underline{n}) \quad = \quad \Omega + \sum_{i=1}^{K} \left\{ -\frac{1}{2} \left[\ln(\sigma_i^2) + \frac{(\bar{y}_i - \mu_i)^2}{\sigma_i^2} \right] + n_i \ln(\pi_i) \right\}, \tag{18}$$

where μ_i and σ_i^2 are given in Eqs. (9) and (10), and

$$\Omega = -\frac{K}{2} \ln(2\pi) + \ln(n!) - \sum_{i=1}^{K} \ln(n_i!). \tag{19}$$

[5] Our estimation results presented in Sections 4 and 5 are found to be robust to variations of the d.o.f of the Student's-t importance density.

ML estimation of θ is carried out by maximizing the log-likelihood function in Eq. (18) over θ (known boundaries) or jointly over $\underline{z}$ and θ (unknown boundaries). Finally note that the maximization of (17) for known group boundaries corresponds to the multinomial ML method of McDonald (1984).

4 Simulation Experiment

We now perform a simulation experiment in order to investigate (i) the quality of the likelihood approximation through the normality assumption in (8), (ii) the numerical properties of the simulated ML estimates under DGP1 and unknown boundaries (see Eq. 15) and (iii) the finite sample performance of the ML approach under DGP1 and DGP2 and both, known and unknown boundaries. We consider a GB2 distribution with parameters $a = 1.5$, $b = 100$, $p = 1$ and $q = 1.5$. This parameter setup implies a very heavy-tailed income distribution with a Gini coefficient of 0.53, which renders estimation via our Gaussian likelihood approximation of Eq. (8) comparatively challenging. The corresponding density function is depicted in Figure 2.

Place Figure 2 here.

We start with analyzing the quality of the Gaussian approximation to the joint density of group means in Eq. (8). For this purpose we simulate $N = 100,000$ independent data sets, each of sample size $n = 10,000$, from the GB2 distribution with parametrization as given above. We then construct $K = 10$ income groups, where the group boundaries are set to the theoretical deciles of the GB2 distribution (DGP2), and compute the K group mean incomes for each of the N data sets. The sample size n is chosen to be empirically realistic for the World Bank and WIDER data. Figure 3 depicts kernel density approximations to the true density of the group means (based on the N simulations) together with the Gaussian approximations with moments given by (9) and (10). We obtain very accurate approximations for groups 1 to 9. The heavy right tail of the income distribution however causes significant deviations from normality for the last income group. This finding differs from our empirical results of Section 5 (see Figure 7), where the Gaussian fit appears very accurate, but represents an interesting challenge for our normality approximation. In order to analyze the effect of the approximation error, we introduce a more accurate but very time-consuming approximation of the conditional density of the last group mean in Eq. (8): In each single likelihood evaluation we simulate the distribution of the last group mean using 1,000 CRNs and evaluate the respective likelihood contribution at a kernel density estimate of the true distribution, which takes the observed skewness into account. Note that this estimation approach is very time-consuming in general and by far too time-consuming to be applied for our simulation based estimator under DGP1 and unknown boundaries, since the whole density simulation would have to be performed $S = 10,000$ times for each likelihood evaluation. We therefore do not recommend to employ this estimation approach in practice. In this simulation experiment we restrict the application of the kernel density approach to known boundaries, resulting in the additional likelihood functions $\mathcal{L}^{\text{kernel}}_{\text{DGP1, KB}}$ and $\mathcal{L}^{\text{kernel}}_{\text{DGP2}}$. We include these two new estimators in our analysis of the statistical finite sample performance further below.

Place Figure 3 here.

We now turn to the analysis of the numerical properties of the simulated ML estimates under DGP1 and unknown group boundaries (see Eq. 15). We simulate a single data-set of size $n = 10,000$ and estimate the parameter vector θ for $N = 10,000$ different sets of CRNs, each consisting of $S = 10,000$ standard Student-t draws with 8 d.o.f. for each of the $K - 1 = 9$ latent group boundaries. We use numerical standard errors in order to assess the numerical uncertainty arising through the simulation-based approximation of the likelihood. These standard errors are computed as the sample standard deviations over the N different estimates. We obtain 0.000011 for a, 0.000584 for b, 0.000013 for p and 0.000025 for q. These values indicate a very high level of numerical precision and amount to less than 0.02% of the corresponding small sample standard errors given in Table 2, which will be discussed below.

Place Table 2 here.

We now analyze the statistical finite sample performance of our ML frameworks under both, DGP1 and DGP2. We consider sample sizes of $n = 10,000$ and $n = 100,000$ individuals and construct $K = 10$ income classes. These settings correspond to typical income data e.g. provided by the World Bank. Under DGP1 we set $c_i = 1/K = 0.1$ and group boundaries are implicitly defined by the according order statistics. Under DGP2 we set the group boundaries to the theoretical GB2 quantiles corresponding to group-probabilities of 0.1. We consider both, known and unknown group boundaries, and compare the performance of the proposed ML techniques to the empirically infeasible ML for individual observations (denoted by *ML Raw Data*) and the alternative kernel-approximation based likelihood estimators using $\mathcal{L}^{\text{kernel}}_{\text{DGP1, KB}}$ and $\mathcal{L}^{\text{kernel}}_{\text{DGP2}}$, as introduced above. For DGP2 and known group boundaries we also consider the classical multinomial ML method (denoted by *ML Multinomial*).

Table 2 reports biases, standard errors and MSEs for 500 Monte-Carlo replications under DGP1. We start with discussing the results for known group boundaries and $n = 10,000$. To begin with, we observe that the moderate fit of the Gaussian approximation in the last income group has only minor effects on the statistical performance: The MSEs obtained under $\mathcal{L}^{\text{kernel}}_{\text{DGP1, KB}}$ are in fact lower than the ones obtained under the Gaussian approximation in $\mathcal{L}_{\text{DGP1, KB}}$, but relative deviations are rather small, amounting to a maximum of 15%, which does not meet the computational complexity and time effort induced by the simulation based kernel density approach. We therefore safely recommend to apply the Gaussian approximation technique. We now turn to the results for our proposed ML approach using $\mathcal{L}_{\text{DGP1, KB}}$: As expected, we observe an overall increase of MSEs for ML under grouped data relative to ML for raw data. The MSEs of b, p and q increase by about 30% while the MSE of a increases by 11%. These efficiency losses are statistically significant at the 5% level. The right panel of Table 2 shows the results for unknown group boundaries. Interestingly, the MSEs for the proposed ML method are close to those obtained for ML under known boundaries. This finding is of considerable relevance in practice since group boundaries are usually not provided for international income data. We now analyze the effect of the observed parameter uncertainty on estimates of the income distribution itself. Figure 4 depicts mean estimated income distributions for both grouped data with unknown boundaries and raw data along with corresponding 95% pointwise confidence intervals, which are computed using the 500 estimated income distributions from the simulation experiment. The figure also reports mean estimates of the Gini coefficient and according standard errors. Deviations of estimates under grouping and raw data appear minor. We conclude that the grouping itself generates very moderate losses in estimation efficiency regarding the income distribution itself and derived measures like the Gini coefficient, even for unknown group boundaries. This is an important finding, since international income data is usually provided in grouped form and one might reasonably expect severe statistical limitations by this data format compared to raw data. Our results imply that this is actually not the case. The closeness of the density estimates under ML for raw data and our ML approach with Gaussian approximations also indicates that the attainable MSE reductions by using kernel approximations do not result in economically significant improvements in density estimation. Increasing the sample size to $n = 100,000$ induces considerable reductions in MSEs under both, known and unknown boundaries. This finding illustrates the consistency of the estimation approaches.

Place Figure 4 here.

The results for DGP2 are reported in Table 3 and appear very similar to those obtained under DGP1. We now also observe efficiency gains compared to the multinomial approach under known boundaries: For $\mathcal{L}_{\mathrm{DGP2}}$ we find MSE reductions of 74% to 85% for $n = 10,000$ and 72% to 77% for $n = 100,000$. All MSE differences are significant at any conventional significance level. Figure 5 depicts mean income distributions along with 95% pointwise confidence intervals for known boundaries and both, the proposed ML approach of Eq. (16) and the classical multinomial ML method. Deviations in the estimated distributions appear minor. We therefore conclude that the inclusion of the group mean incomes in the likelihood of Eq. (16) contributes to the statistical efficiency regarding estimation of the model parameters, but the effect on the income distribution and derived measures of income inequality appears moderate.

Place Table 3 and Figure 5 here.

5 Empirical Application

We now apply our ML estimation framework to grouped household income data from the World Bank website *PovcalNet* provided for the year 2013.[6] We consider a selection of four countries: Malaysia, Thailand, Bangladesh and Poland. The data consists of group-specific mean incomes $\bar{y}_i$ and population shares c_i for 10 income groups, where the grouping mechanism corresponds to DGP1 with unknown boundaries (constant population shares $c_i = 0.1\ \forall i$). The complete data set is given in Table 4.

Place Table 4 here.

We start with an assessment of the adequacy of the GB2 for modeling the observed group mean incomes and compare the goodness of fit relative to three nested distributions: the B2 ($a = 1$), Singh-Maddala ($p = 1$) and Dagum distribution ($q = 1$). Parameter estimates are obtained by numerical maximization of the simulated log-likelihood $\mathcal{L}_{\mathrm{DGP1,\ UB}}$ provided in Eq.

[6] The income is measured in purchasing power parity Dollar rates. See *PovcalNet* for details.

(15) with $S = 10,000$. The likelihood-ratio test serves as natural testing-device against the GB2. We also consider the ability to forecast observed income shares s_i for $i = 1, \ldots, 10$ as an additional criterion for the goodness of fit (see also Hajargasht et al., 2012). Predicted cumulative income shares η_i are obtained by the first-moment distribution function, $\hat{\eta}_i = F_1(z_i; \hat{\theta})$. The group boundaries z_i are unknown and therefore replaced by the inverse distribution function evaluated at the cumulative income shares (compare Hajargasht et al., 2012). Predicted income shares are then obtained as $\hat{s}_i = \hat{\eta}_i - \hat{\eta}_{i-1}$.

Table 5 reports the log-likelihood values for the four income distributions. At the 1% level the GB2 turns out as the best fitting distribution for all considered data-sets. The Root Mean Squared Errors (RMSEs) for the forecasted income shares are reported in Table 6. Note that we do not assess the significance of RMSE differences, since each RMSE is based on only ten observations. The GB2 performs best in all cases. Our findings therefore support the adequacy of the GB2 for modeling international income data. However note that the obtained RMSEs are very low for all considered income distributions.

Place Tables 5-6 here.

Table 7 reports the parameter estimates under the GB2 distribution along with estimates of the Gini coefficient and the headcount ratio (HC) while Figure 6 depicts the estimated income distributions along with asymptotic 95% pointwise confidence intervals.[7] For a given poverty line x, the headcount ratio is the proportion of population with income less than x. Hence

$$HC = F(x; \hat{\theta}),$$

where we set $x = 57.79$ as provided by the World Bank. The Gini coefficient is obtained by

$$\text{Gini} = -1 + \frac{2}{E[y]} \int_0^\infty y \; F(y;\theta) \; f(y;\theta) \; dy,$$

where the integral is evaluated numerically. Table 7 also includes predictions of the income shares for the first and the last group. Accurate predictions for the first group are of special importance for poverty measurement, while predictions for the last group suffer from the thick right tail of typical income data. Estimated standard deviations are computed by inverting a numerical approximation to the hessian of the log-likelihood at the estimates. Standard deviations for the Gini and the headcount ratio are obtained by the delta method.

[7] Results for the nested distributions are available upon request.

The asymptotic standard errors reported in Table 7 are low and indicate a high level of estimation precision, in particular for the Gini coefficient and the headcount ratio, which are of special interest in applied economic research. The income share predictions for the first and the last income group are very accurate with a RMSE of 0.0003 for s_1 and 0.0059 for s_{10}. The highest absolute prediction error is obtained for Thailand (3.43%) for the last income group, where the heavy right tail of the income distribution makes accurate predictions rather hard to obtain. We conclude that the flexible GB2 distribution delivers very precise income share predictions and outperforms its nested competitors w.r.t the in-sample fit.

Place Table 7 and Figure 6 here.

We finally check the quality of our Gaussian approximation to the joint density of group means at the estimated parameters of the four income distributions. We again simulate $N = 100,000$ independent data sets, each of sample size corresponding to the actual data, from the estimated GB2 distributions for Malaysia, Thailand, Bangladesh and Poland. We then construct $K = 10$ income groups for each country according to DGP1 and compute the K group mean incomes for each of the N data sets. Figure 7 depicts kernel density approximations to the true densities of the group means (based on the N simulations) together with the Gaussian approximations with moments given by (9) and (10). We observe a very accurate fit which confirms the quality of our approximation for actual income data.

Place Figure 7 here.

6 Conclusion

In this paper we develop a general framework for maximum likelihood estimation of parametric income distributions for grouped data with potentially unknown group boundaries. Our ML approach accounts for two data generating processes and incorporates the information of group mean incomes into the likelihood. The method is therefore more efficient than the traditional multinomial ML approach of McDonald (1984) which neglects the informational content of the group mean incomes and is furthermore misspecified under DGP1 and/or unknown boundaries. This is a considerable shortcoming, since empirical studies typically employ data from the World Bank and/or WIDER, which correspond to this particular DGP.

A Monte-Carlo simulation experiment shows that the proposed ML framework results in improved statistical efficiency relative to the multinomial likelihood under DGP2. We also find comparable estimation precision under known and unknown group boundaries for both, DGP1 and DGP2. This finding is of considerable relevance in practice since the group boundaries are usually not provided for international income data. Even compared to ML for raw (ungrouped) income data, the proposed ML framework performs very well and resulting reductions in estimation efficiency appear moderate.

We finally apply the ML approach to World Bank data for four countries and find strong evidence for the GB2 distribution relative to its nested competitors such as the Beta2, Singh-Maddala and the Dagum distribution. The obtained estimates of inequality and poverty measures as well as predictions of income shares show a high degree of accuracy. These findings confirm the appropriateness of the GB2 distribution and the proposed ML estimation framework.

References

[1] Bandourian, R., McDonald, J.B., Turley, R.S., (2003). Income distributions: an inter-temporal comparison over countries. Estadistica 55: 135-152.

[2] Chen, Y., (2017). A unified approach to estimating and testing income distributions with grouped data. Journal of Business and Economic Statistics. Accepted Manuscript.

[3] Chotikapanich, D., Griffiths, W.E., Rao, D.S.P., (2007). Estimating and combining national income distributions using limited data. Journal of Business and Economic Statistics 25: 97-109.

[4] Chotikapanich, D. (ed.), (2008). Modeling income distributions and Lorenz curves. Springer, New York.

[5] Chotikapanich, D., Griffiths, W.E., Rao, D.S.P., Valencia, V. (2012). Global income distributions and inequality, 1993 and 2000: incorporating country-level inequality modelled with beta distributions. The Review of Economics and Statistics 94: 52-73.

[6] David, H.A., Nagaraja, H.N., (2003). Order Statistics. Wiley, Hoboken, New Jersey.

[7] Griffiths, W.E., Hajargasht, G., (2012). GMM estimation of mixtures from grouped data: an application to income distributions. Working Paper.

[8] Hajargasht, G., Griffiths, W.E., Brice, J.,Rao, P., Chotikapanich, D., (2012). Inference for Income Distributions Using Grouped Data. Journal of Business & Economic Statistics 30(4): 563-575.

[9] Hajargasht, G., Griffiths, W.E., (2016). Inference for Lorenz curves. Working Paper.

[10] Kleiber, , C., Kotz, S., (2003). Statistical size distributions in Economics and actuarial sciences. Wiley, New York.

[11] Liesenfeld, R., Richard, J.-F., (2006). Classical and Bayesian analysis of univariate and multivariate stochastic volatility models. Econometric Reviews 25: 335-360.

[12] McDonald, J.B., (1984). Some generalized functions of the size distribution of income. Econometrica 52: 647-663.

[13] Nadarajah, S. (2005). Sums, products and ratios of generalized beta variables. Statistical Papers 47: 69-90.

[14] Parker, S.C., (1999). The generalized beta as a model for the distribution of earnings. Economics Letters 62(2): 197-200.

[15] Robert, C., Casella, G. (2004). Monte Carlo Statistical Methods. Springer, Berlin.

[16] Wu, X., Perloff, J., (2005). China's income distribution: 1985-2001. The Review of Economics and Statistics 87: 763-775.

[17] Wu, X., (2006). Inference and Density Estimation with Interval Statistics. Working Paper.

	Density function	Moments	Distribution function	Moment distribution function	Gini coefficient
GB 2	$f_y(y;\theta)$ $= \frac{ay^{ap-1}}{b^{ap}B(p,q)(1+(y/b)^a)^{p+q}}$	$E[y^\ell] = \frac{b^\ell B(p+\ell/a,\ q-\ell/a)}{B(p,q)}$	$F_y(y;\theta) = B_u(p,q)$, with $u = (y/b)^a/[1+(y/b)^a]$	$F_\ell(y;\theta)$ $= B_u(p+\ell/a,\ q-\ell/a)$, with $u = (y/b)^a/[1+(y/b)^a]$	Integral evaluated numerically
Beta-2 (a=1)	$f_y(y;\theta)$ $= \frac{y^{p-1}}{b^p B(p,q)(1+(y/b))^{p+q}}$	$\mu = bp/(q-1)$ $E[y^2] = bp(p+1)/(q-1)(q-2)$	$F_y(y,\theta) = B_u(p,q)$, with $u = (y/b)/[1+(y/b)]$	$F_\ell(y;\theta)$ $= B_u(p+\ell, q-\ell)$, with $u = (y/b)/[1+(y/b)]$	$G = \frac{2B(2p,2q-1)}{pB^2(p,q)}$
Singh-Maddala $(p=1)$	$f_y(y,\theta)$ $= \frac{aqy^{a-1}}{b^a(1+(y/b)^a)^{1+q}}$	$E[y^\ell] = \frac{b^\ell \Gamma(1+\ell/a,q-\ell/a)}{\Gamma(q)}$	$F_y(y,\theta)$ $= 1 - \left[1+(\frac{y}{b})^a\right]^{-q}$	$F_\ell(y;\theta)$ $= B_u(1+\ell/a, q-\ell/a)$ with $u = (y/b)^a/[1+(y/b)^a]$	$G = 1 - \frac{\Gamma(q)\Gamma(2q-1/a)}{\Gamma(q-1/a)\Gamma(2q)}$
Dagum $(q=1)$	$f_y(y;\theta)$ $= \frac{apy^{ap-1}}{b^{ap}(1+(y/b)^a)^{p+1}}$	$E[y^\ell] = \frac{b^\ell \Gamma(p+\ell/a,1-\ell/a)}{\Gamma(q)}$	$F_y(y;\theta)$ $= \left[1+(\frac{y}{b})^{-a}\right]^{-p}$	$F_\ell(y;\theta)$ $= B_u(p+\ell/a, 1-\ell/a)$, with $u = (y/b)^a/[1+(y/b)^a]$	$G = \frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(p+1/a)\Gamma(2p)} - 1$

Table 1: Overview on the GB2 and the nested Beta-2, Singh-Maddala and Dagum distributions.

		n = 10,000											
					Known Bounds						Unknown Bounds		
		ML Raw Data			ML: $\mathcal{L}_{\text{DGP1, KB}}$			ML: $\mathcal{L}^{\text{kernel}}_{\text{DGP1, KB}}$			ML: $\mathcal{L}_{\text{DGP1, UB}}$		
True Param.		BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE
a	1.5	0.0017	0.1055	0.0111	-0.0227	0.1087	0.0123	-0.0012	0.1103	0.0121	-0.0233	0.1104	0.0127
b	100	0.4725	5.1001	26.1821	1.7425	5.5402	33.6687	0.4583	5.3599	28.8810	1.7929	5.5248	33.6769
p	1	0.0083	0.1042	0.0109	0.0320	0.1149	0.0142	0.0130	0.1119	0.0127	0.0328	0.1169	0.0147
q	1.5	0.0182	0.1813	0.0331	0.0687	0.1972	0.0435	0.0234	0.1909	0.0369	0.0706	0.1998	0.0448
		n = 100,000											
					Known Bounds						Unknown Bounds		
		ML Raw Data			ML: $\mathcal{L}_{\text{DGP1, KB}}$			ML: $\mathcal{L}^{\text{kernel}}_{\text{DGP1, KB}}$			ML: $\mathcal{L}_{\text{DGP1, UB}}$		
True Param.		BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE
a	1.5	0.0005	0.0336	0.0011	0.0003	0.0386	0.0015	0.0026	0.0370	0.0014	0.0001	0.0385	0.0015
b	100	-0.0093	1.4652	2.1427	0.0318	1.7830	3.1739	-0.1136	1.6689	2.7925	0.0285	1.7896	3.1969
p	1	0.0007	0.0318	0.0010	0.0011	0.0363	0.0013	-0.0009	0.0351	0.0012	0.0013	0.0364	0.0013
q	1.5	0.0008	0.0551	0.0030	0.0020	0.0648	0.0042	-0.0030	0.0613	0.0038	0.0023	0.0648	0.0042

Table 2: Monte-Carlo simulation results for the finite-sample performance of the proposed ML estimators under DGP1. The results are obtained for 500 Monte Carlo replications.

		n = 10,000											
		Known Bounds									Unknown Bounds		
		ML Raw Data			ML: $\mathcal{L}_{\text{DGP2}}$			ML: $\mathcal{L}^{\text{kernel}}_{\text{DGP2}}$			ML Multinomial			ML: $\mathcal{L}_{\text{DGP2}}$		
True Param.		BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE
a	1.5	0.0017	0.1055	0.0111	-0.0191	0.1083	0.0121	0.0022	0.1117	0.0125	-0.0029	0.2139	0.0457	-0.0208	0.1103	0.0126
b	100	0.4725	5.1001	26.1821	1.7070	5.5164	33.2831	0.4525	5.3433	28.6979	2.9379	12.6678	168.7824	1.7926	5.5458	33.9074
p	1	0.0083	0.1042	0.0109	0.0277	0.1129	0.0135	0.0093	0.1122	0.0126	0.0480	0.2527	0.0660	0.0298	0.1160	0.0143
q	1.5	0.0182	0.1813	0.0331	0.0627	0.1941	0.0415	0.0188	0.1913	0.0369	0.1183	0.5196	0.2834	0.0670	0.1993	0.0441
		n = 100,000														
		Known Bounds												Unknown Bounds		
		ML Raw Data			ML: $\mathcal{L}_{\text{DGP2}}$			ML: $\mathcal{L}^{\text{kernel}}_{\text{DGP2}}$			ML Multinomial			ML: $\mathcal{L}_{\text{DGP2}}$		
True Param.		BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE	BIAS	Stand. Dev.	MSE
a	1.5	0.0005	0.0336	0.0011	0.0006	0.0386	0.0015	0.0028	0.0369	0.0014	0.0033	0.0740	0.0055	0.0003	0.0384	0.0015
b	100	-0.0093	1.4652	2.1427	0.0226	1.7864	3.1854	-0.1185	1.6770	2.8207	0.1065	3.3797	11.4107	0.0325	1.7864	3.1859
p	1	0.0007	0.0318	0.0010	0.0008	0.0363	0.0013	-0.0012	0.0350	0.0012	0.0017	0.0704	0.0050	0.0011	0.0363	0.0013
q	1.5	0.0008	0.0551	0.0030	0.0015	0.0650	0.0042	-0.0033	0.0613	0.0038	0.0048	0.1359	0.0185	0.0020	0.0646	0.0042

Table 3: Monte-Carlo simulation results for the finite-sample performance of the proposed ML estimators under DGP2. The results are obtained for 500 Monte Carlo replications.

	n	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$	$\bar{y}_6$	$\bar{y}_7$	$\bar{y}_8$	$\bar{y}_9$	$\bar{y}_{10}$
Malaysia	35138	136.5788	219.2135	296.6459	378.4292	474.6554	592.7998	745.3447	953.7475	1311.7348	2702.0503
Thailand	40606	129.4897	181.8031	222.4733	262.1732	310.4708	367.4738	443.6799	547.7945	713.7275	1312.8651
Bangladesh	12224	47.2415	61.3172	71.5175	81.2813	91.8090	103.9869	119.3536	140.8674	178.5301	330.1005
Poland	37139	164.4776	234.4808	285.6283	335.4045	386.9538	445.9620	518.1049	609.7213	755.2327	1286.8993

Table 4: Household income data for the empirical application of Section 5 obtained from the World Bank website *PovcalNet* for the year 2013 measured in purchasing power parity Dollar rates. See *PovcalNet* for details.

	Malaysia		Thailand		Bangladesh		Poland	
Distribution	LL	LL-Diff	LL	LL-Diff	LL	LL-Diff	LL	LL-Diff
GB2	-139094.19	–	-160730.06	–	-48391.87	–	-146963.80	–
B2 ($a = 1$)	-139160.40	66.21	-160743.48	13.42	-48431.61	39.74	-146967.98	4.18
Singh-Maddala ($p = 1$)	-139426.91	332.72	-161059.57	329.51	-48434.13	42.26	-147130.54	166.73
Dagum ($q = 1$)	-139380.39	286.21	-160771.35	41.30	-48395.45	3.58	-147059.00	95.20

Table 5: Log-likelihood values (LL) and log-likelihood differences to the GB2 (LL-Diff) obtained at the ML-estimates for the *PovcalNet* data. Likelihood-Ratio critical values for the B2, Singh-Maddala and Dagum: 3,3174 (1%), 1,9207 (5%).

Distribution	Malaysia	Thailand	Bangladesh	Poland
GB2	**0.002077**	**0.003725**	**0.000284**	**0.000581**
B2	0.013808	0.006482	0.005749	0.001008
Singh-Maddala	0.015882	0.020181	0.018725	0.007512
Dagum	0.015690	0.024754	0.003873	0.007345

Table 6: Root mean squared errors (RMSEs) for income share predictions. The RMSEs are obtained as $RMSE = \sqrt{K^{-1}\sum_{i=1}^{K}(\hat{s}_i - s_i)^2}$.

	Malaysia		Thailand		Bangladesh		Poland	
	Par	SE	Par	SE	Par	SE	Par	SE
a	0.2942	0.0002	0.7722	0.0010	1.8884	0.0075	1.3135	0.0046
b	1.3869	0.0062	3.8401	0.0213	31.1157	0.1345	256.9676	4.0607
p	110.2819	0.1444	132.5090	0.2223	10.7327	0.0808	5.6115	0.1027
q	19.5085	0.0253	4.4873	0.0057	1.5178	0.0100	3.1537	0.0103
Gini	0.4649	0.0002	0.3840	0.0002	0.3219	0.0005	0.3249	0.0004
HC	0.002349	0.000021	0.000269	0.000003	0.117264	0.000536	0.000189	0.000005
s_1 emp	0.0175	–	0.0288	–	0.0385	–	0.0327	–
$\hat{s}_1$	0.0172	–	0.0283	–	0.0385	–	0.0327	–
s_{10} emp	0.3459	–	0.2923	–	0.2693	–	0.2562	–
$\hat{s}_{10}$	0.3515	–	0.3026	–	0.2699	–	0.2554	–

Table 7: Estimates of model parameters, poverty and inequality measures and income shares $\hat{s}_1$ and $\hat{s}_{10}$ obtained under the GB2 distribution for the *PovcalNet* data. HC: Headcount ratio; s_i emp: observed income share for the i'th group.

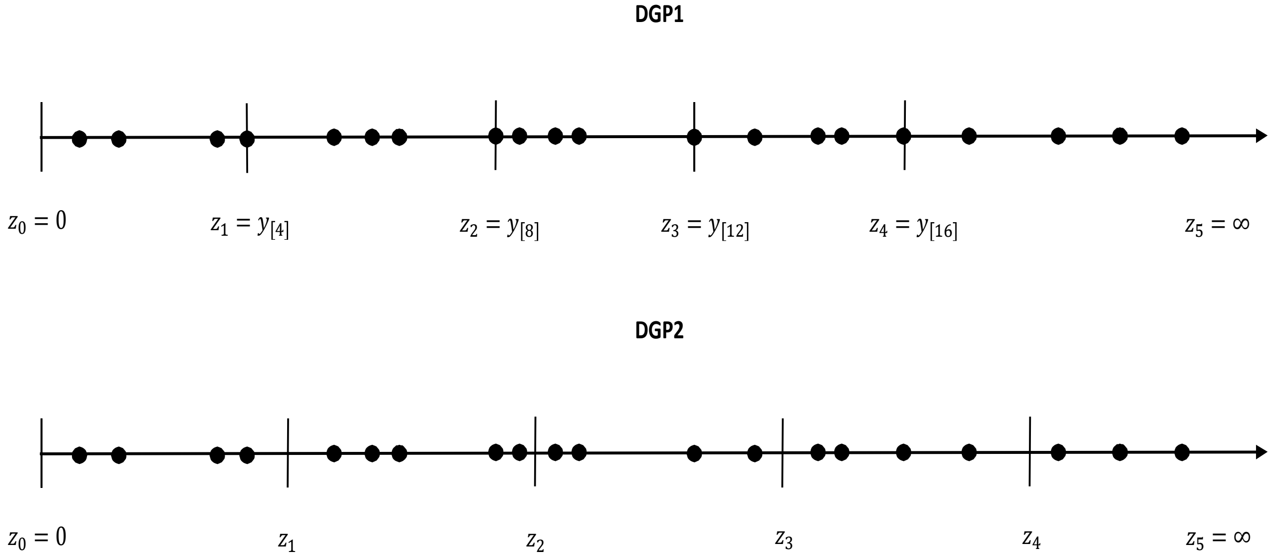

Figure 1: Schematic illustration of the two data generating processes DGP1 and DGP2 for $n = 20$. Black bullets denote individual income y_i on the real line. The example for DGP1 assumes $c_i = 0.2 \; \forall \; i$.

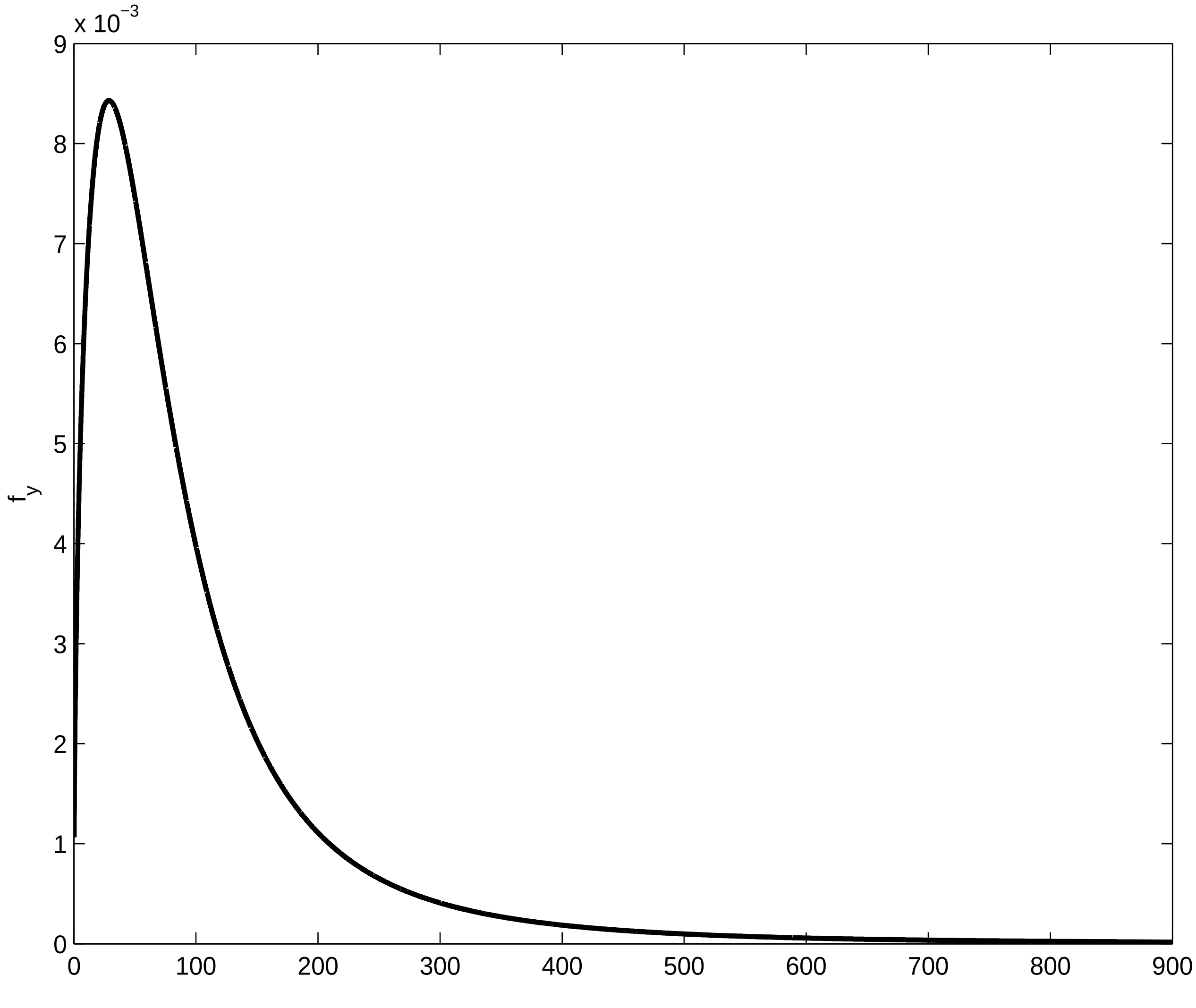

Figure 2: Probability density function of a GB2 distribution with parameters $a = 1.5$, $b = 100$, $p = 1$ and $q = 1.5$.

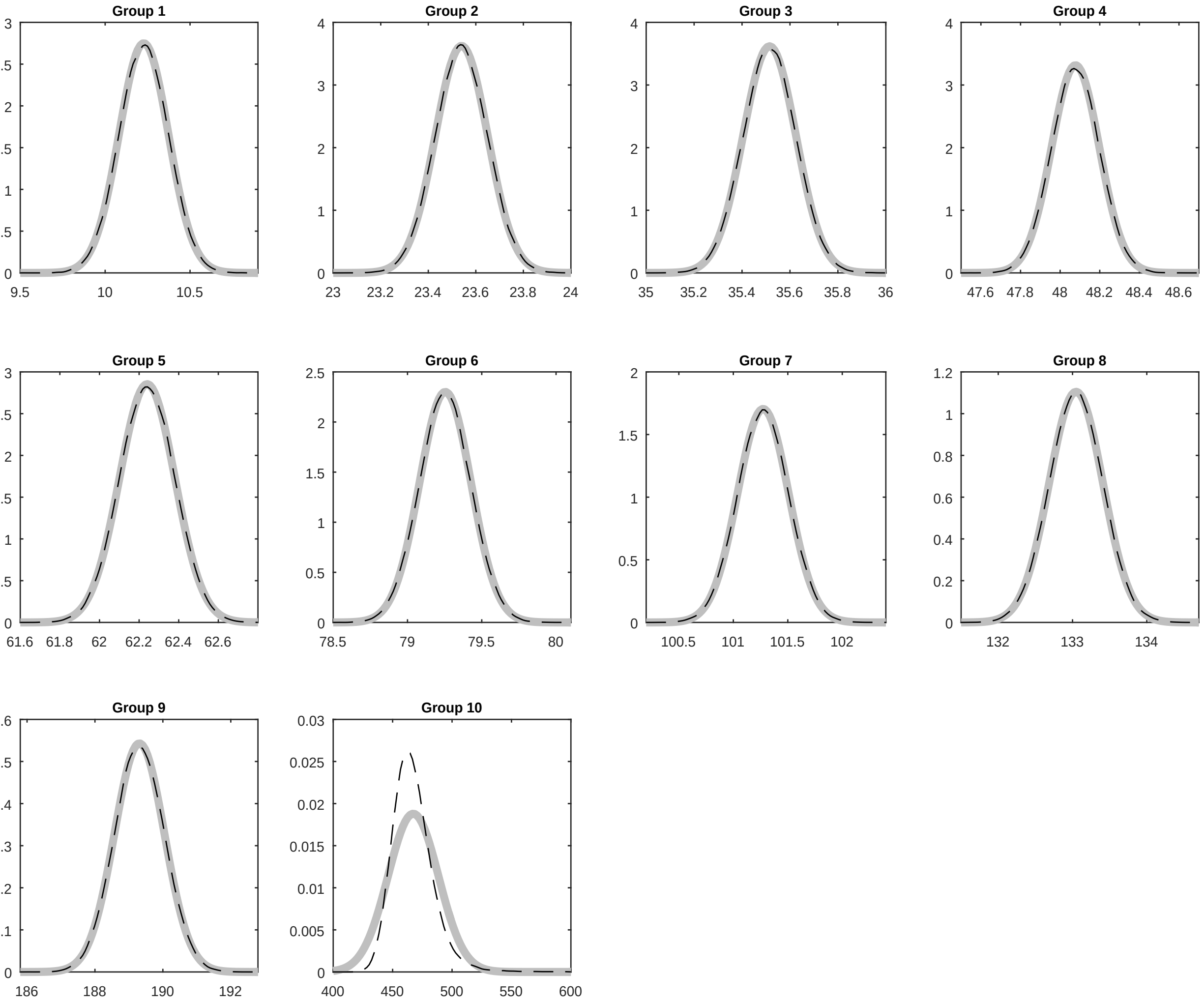

Figure 3: Approximate Gaussian distributions of $K = 10$ group means together with kernel density estimates of their true counterparts. Dashed black line: kernel density estimate; thick gray line: Gaussian approximation using the moments provided in Eqs. (9) and (10). The kernel density estimates are based on 100,000 simulations from a GB2 distribution with parameters $a = 1.5$, $b = 100$, $p = 1$ and $q = 1.5$. The sample size is $n = 10,000$. The group boundaries are set to the theoretical deciles of the GB2 distribution (DGP2).

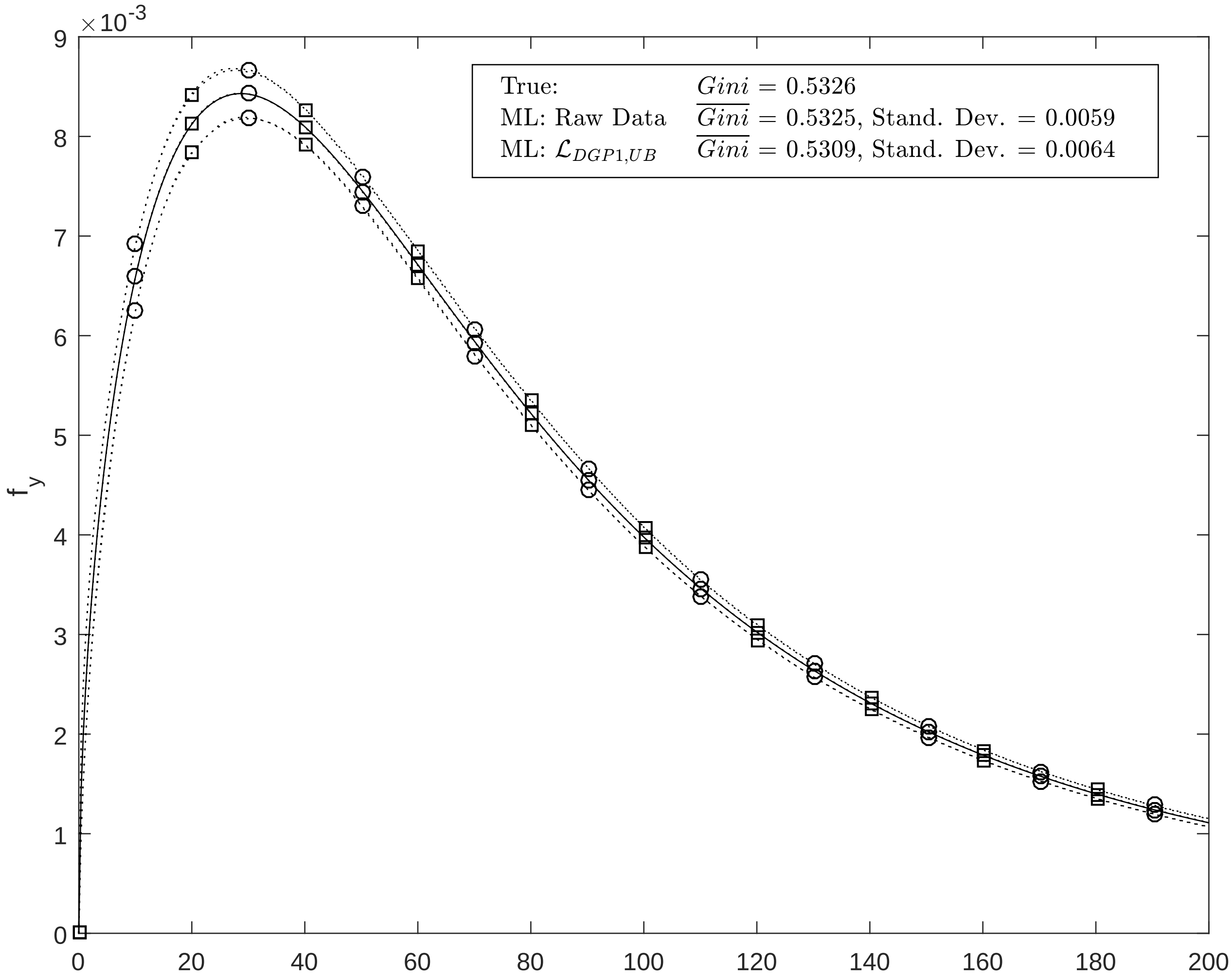

Figure 4: Mean estimated income distributions for both, grouped data with unknown boundaries ($K = 10$ income groups, circles) and raw data (boxes), along with corresponding 95% pointwise confidence intervals under DGP1 with $n = 10,000$. The mean distributions and confidence intervals are computed using the 500 estimated income distributions from the Monte-Carlo experiment of Section 4. The figure also reports mean estimates of the Gini coefficient and according finite sample standard errors which are computed as the sample standard deviation over the 500 Gini estimates.

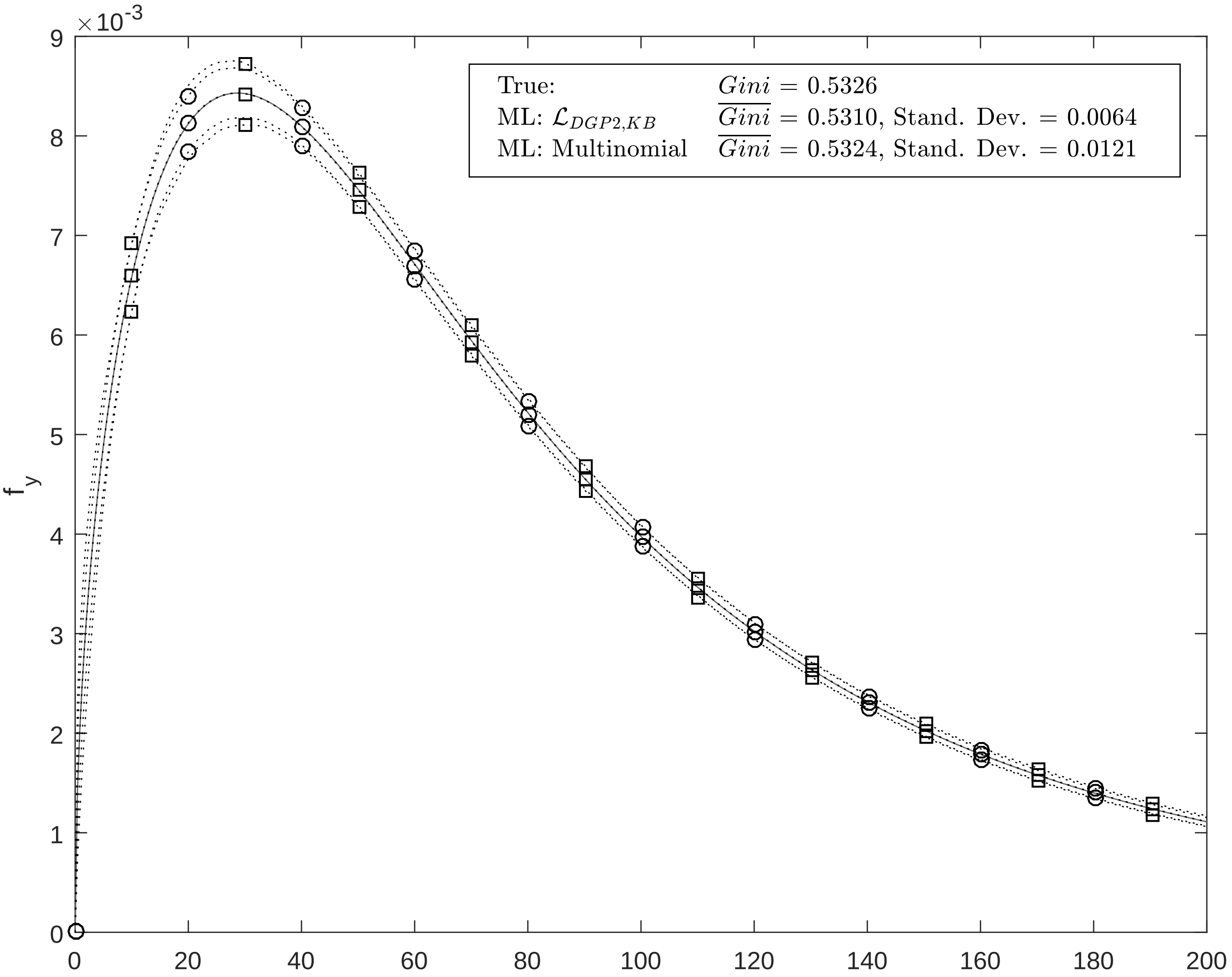

Figure 5: Mean estimated income distributions for known boundaries and both, the proposed ML approach of Eq. (16) (circles) and the multinomial ML method (boxes), along with corresponding 95% pointwise confidence intervals under DGP2 with $n = 10,000$. The mean distributions and confidence intervals are computed using the 500 estimated income distributions from the Monte-Carlo experiment of Section 4. The figure also reports mean estimates of the Gini coefficient and according finite sample standard errors which are computed as the sample standard deviation over the 500 Gini estimates.

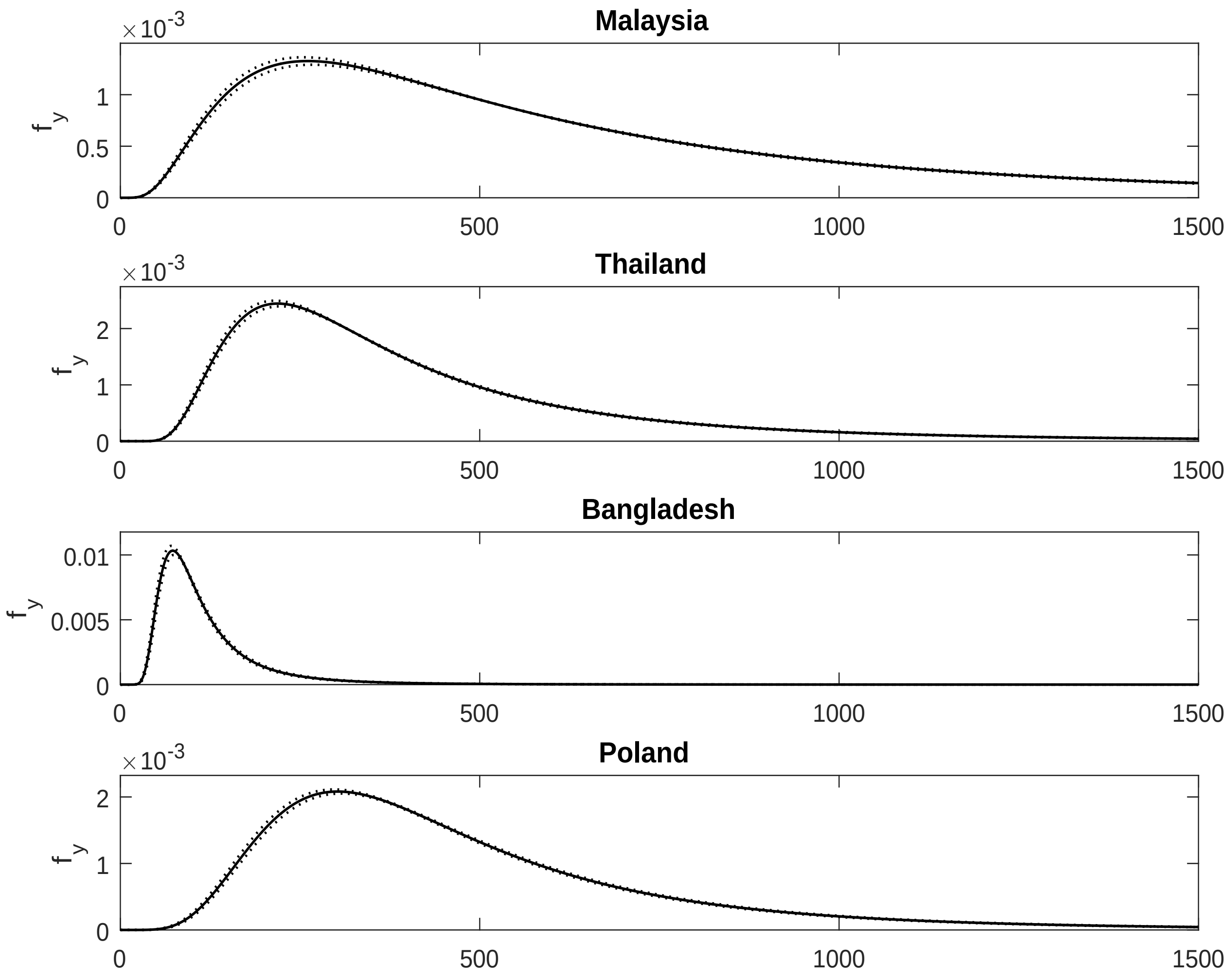

Figure 6: Estimated GB2 income distributions along with asymptotic 95% pointwise confidence bounds for Malaysia, Thailand, Bangladesh and Poland.

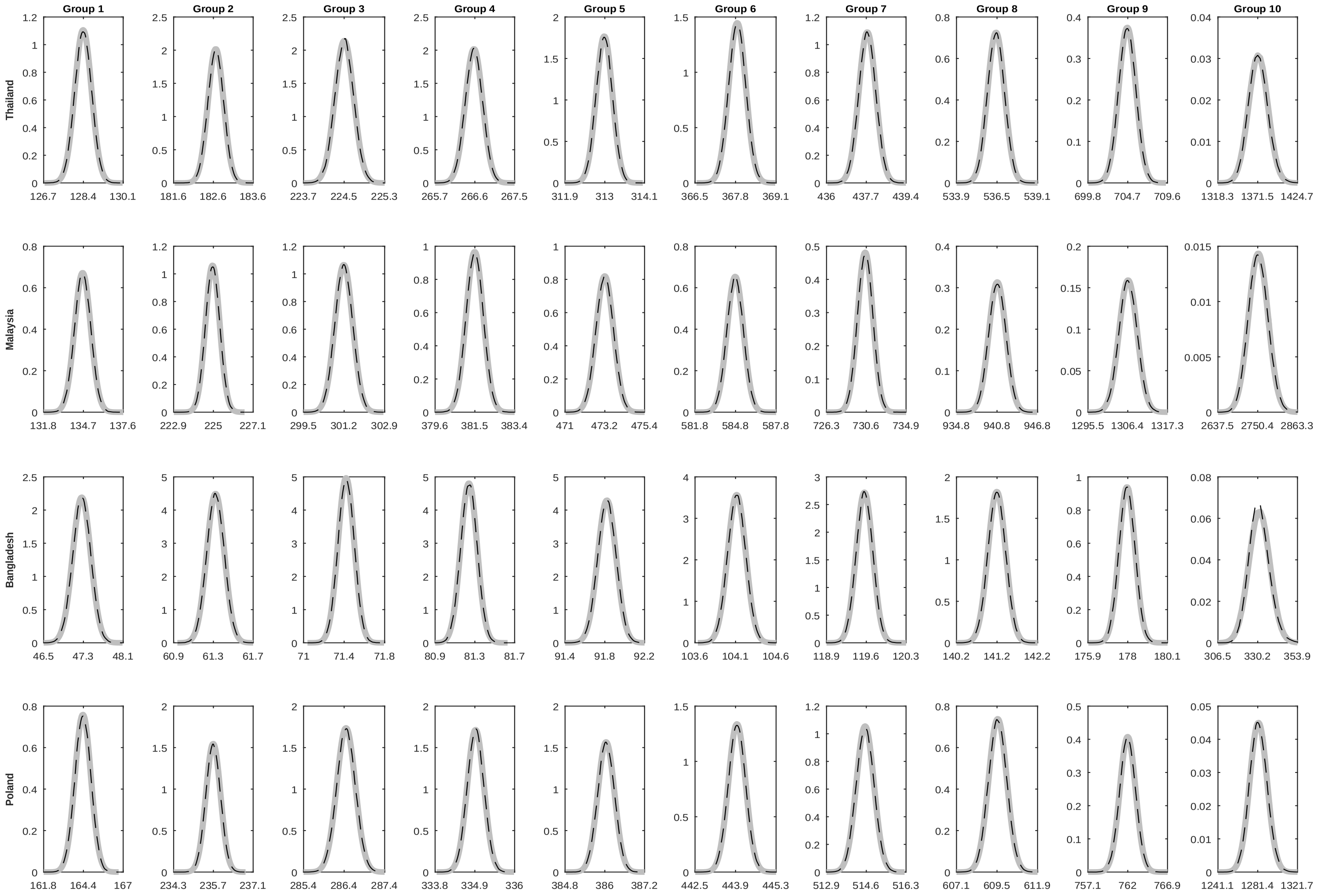

Figure 7: Approximate Gaussian distributions of $K = 10$ group means for Thailand, Malaysia, Bangladesh and Poland together with corresponding kernel density estimates of their true counterparts obtained under the parameter estimates reported in Table 7. Dashed black line: kernel density estimate; thick gray line: Gaussian approximation using the moments provided in Eqs. (9) and (10). The kernel density estimates are based on 100,000 simulations from the estimated GB2 distributions with sample sizes as given in Table 4. The boundaries are set to the according deciles (DGP1).