



# Approximate Bayesian computation for Lorenz curves from grouped data

Genya Kobayashi<sup>1</sup> · Kazuhiko Kakamu<sup>2</sup>

Received: 10 August 2017 / Accepted: 18 August 2018 / Published online: 22 August 2018  
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

## Abstract

This paper proposes a new Bayesian approach to estimate the Gini coefficient from the grouped data on the Lorenz curve. The proposed approach assumes a hypothetical income distribution and estimates the parameter by directly working on the likelihood function implied by the Lorenz curve of the income distribution from the grouped data. It inherits the advantages of two existing approaches through which the Gini coefficient can be estimated more accurately and a straightforward interpretation about the underlying income distribution is provided. Since the likelihood function is implicitly defined, the approximate Bayesian computational approach based on the sequential Monte Carlo method is adopted. The usefulness of the proposed approach is illustrated through the simulation study and the Japanese income data.

**Keywords** Generalised beta distribution · Gini coefficient · Income distribution · Sequential Monte Carlo

## 1 Introduction

The Gini coefficient plays a fundamental role in measuring inequality and can be a basis of political decision-making. Although it is ideal to utilise individual household

---

This work is partially supported by JSPS KAKENHI (#25245035, #15K17036, #16K03592, #16KK0081, and #18K12754).

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00180-018-0831-x>) contains supplementary material, which is available to authorized users.

---

✉ Genya Kobayashi  
gkobayashi@chiba-u.jp  
Kazuhiko Kakamu  
kakamu@person.kobe-u.ac.jp

<sup>1</sup> Graduate School of Social Sciences, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba 263-8522, Japan

<sup>2</sup> Graduate School of Business Administration, Kobe University, 2-1, Rokkodai, Nada, Kobe 657-8501, Japan

data to estimate the Gini coefficient accurately (see, e.g., Hasegawa and Kozumi 2003), availability of individual data is usually severely limited because of the difficulty in data collection and management and confidentiality of, say, individual income data, the former being particularly true in developing countries. Instead, grouped data, which provide the summary of income and number of individuals or households for several income classes, are widely available. Estimating the Gini coefficient based on grouped data has drawn substantial attention from both the theoretical and empirical perspectives. See, for example, Chotikapanich (2008) for an overview.

There are mainly two approaches to estimate the Gini coefficient from grouped data in the parametric framework (Ryu and Slottje 1999). One is to assume a hypothetical statistical distribution for income and estimate the parameters of the income distribution from the grouped level income data. The Gini coefficient is then calculated from the parameter estimates. An income distribution that provides good fit to the data can be chosen from the wide range of statistical distributions (McDonald and Xu 1995; Kleiber and Kotz 2003). The advantage of this approach is that it provides a straightforward interpretation about the underlying income distribution because the moments can be computed and the shape of the distribution can be visualised through the parameter estimates. In this approach, the likelihood function can be constructed based on the multinomial distribution, since the data can be regarded as an realisation from the multinomial trials with the probabilities derived from the hypothetical income distribution and the number of households in the income groups (McDonald 1984). The likelihood function can also be constructed by regarding the thresholds for the income groups as the selected order statistics (Nishino and Kakamu 2011).

The other approach is to fit a specific functional form directly to the data on the Lorenz curve that include the cumulative population shares and cumulative income shares, instead of using the level income data, and estimate the parameters of the function. Since such a functional form for the Lorenz curve is designed so that the inequality measures can be easily derived, the Gini coefficient is immediately calculated once the parameter estimates are obtained. A notable advantage of this approach is that a large list of functional forms is available (see, e.g., Kakwani and Podder 1973; Basman et al. 1990; Ortega et al. 1991; Rasche et al. 1980; Villaseñor and Arnold 1989; Chotikapanich 1993; Sarabia et al. 1999) in addition to the ones that are derived from the well-known income models such as the lognormal, Singh–Maddala, and Dagum distributions. Also, this approach is known to provide a more accurate estimate of the Gini coefficient than the first approach, because the Lorenz curve is estimated on the bounded interval  $[0, 1]$  (Ryu and Slottje 1996). However, while the implied probability density function exists provided some conditions are satisfied (Iritani and Kuga 1983; Sarabia 2008), the interpretation as an income distribution is less intuitive as the support of the probability density function of the implied distribution is limited to some interval. Moreover, the current practice for parametric Lorenz curve estimation lacks a solid statistical foundation compared with the hypothetical statistical distribution estimation despite the fact that the discussion on the Lorenz curve has a long history since the seminal work by Lorenz (1905).

Given this context, the aim of the present paper is to estimate the Gini coefficient assuming a hypothetical statistical distribution from the grouped data on the Lorenz curve in a general framework. For some flexible hypothetical distributions, such as

the generalised beta distribution and mixture of parametric distributions, an analytical form of the Lorenz curve is not available and the evaluation of the Lorenz curve can be computationally expensive and unstable. Therefore, a new estimation procedure for the case where the Lorenz curve is not explicitly available is required. Moreover, it is worth noting that the present study is motivated by the approach proposed by Chotikapanich and Griffiths (2002, 2005). More recently, Hajargst and Griffiths (2015) proposed a generalised method-of-moment approach for the Lorenz curves in a similar setting. In Chotikapanich and Griffiths (2002, 2005), the expectations of the income shares for the groups are assumed to be equal to the differences in the values of the Lorenz curve for the two consecutive groups. Then, they adopted the likelihood function based on the Dirichlet distribution and proposed the maximum likelihood estimator and Bayes estimator by using the Markov chain Monte Carlo (MCMC) method. Although their pseudo Dirichlet likelihood approach may be convenient, the parameter estimates and the resulting Gini estimate can be highly sensitive with respect to the additional tuning parameter, which is required to construct the Dirichlet likelihood and the data do not contain information about. In the Bayesian framework, the posterior distributions of the parameters and Gini coefficient are sensitive with respect to the prior distribution of this parameter. Furthermore, the evaluation of the likelihood function requires the evaluation of the Lorenz curve for each group. The applicability of this approach is limited to the cases where evaluation of the Lorenz curve derived from the hypothetical distribution is feasible and substantial prior information for the tuning parameter is available.

Motivated by the above issues, in this paper, we attempt to work on the likelihood function implicitly defined from the grouped data on the Lorenz curve of the hypothetical income distribution, instead of utilising the Dirichlet likelihood. We employ the approximate Bayesian computation (ABC) method that avoids the direct evaluation of the likelihood function and simulates data from the model given a candidate parameter value. If the simulated and observed data are similar, the candidate parameter value is a good candidate to have generated the observed data. Then, it can be regarded as a sample from the posterior distribution. The ABC method has a wide variety of fields of application including population genetics, population biology, signal processing, epidemiology, and economics. See, for example, Csilléry et al. (2010), Sisson and Fan (2011), and Marin et al. (2012) for an overview of the ABC methods. The application of the ABC method requires an ability to simulate datasets from the probability model and is well-suited to the present context, because datasets can be easily generated from the frequently used hypothetical income distributions. Since it is difficult to devise an efficient proposal distribution for an MCMC algorithm in the ABC setting, we adopt the sequential Monte Carlo (SMC) algorithm with adaptive weights proposed by Bonassi and West (2015), which is computationally efficient and easy to implement.

The advantage of the proposed approach is threefold. Firstly, the proposed approach inherits the advantage of the two existing approaches. A straightforward interpretation about the hypothetical income distribution can be obtained from the parameter estimates. Furthermore, since the parameters are estimated from the data on the Lorenz curve, the Gini coefficient can be estimated more accurately than estimating it from the grouped level income data. Secondly, we work on the likelihood function implied

from the Lorenz curve, the proposed approach does not require the tuning parameter on which the data do not contain information unlike the Dirichlet approach in Chotikapanich and Griffiths (2002, 2005). Finally, since the ABC method only requires an ability to simulate data from the probabilistic model, which is the case for the frequently used hypothetical income distributions, the proposed approach does not require an analytical form of the Lorenz curve unlike the Dirichlet approach.

The rest of this paper is organised as follows. Section 2 briefly reviews the estimation methods for the Lorenz curve from grouped data and proposes our estimation method based on ABC. We focus on the parametric framework and adopt the five-parameter generalised beta distribution as a flexible hypothetical income distribution. Section 3 illustrates the proposed method by using the simulated data and compares the performance with the existing methods. The application of the proposed method to the real data from the Family Income and Expenditure Survey in Japan is also presented. Finally, Sect. 4 concludes and some remaining issues are discussed.

## 2 Method

### 2.1 Estimating Gini coefficient from Lorenz curve based on grouped data

Suppose that the population is divided into  $k$  groups. Let us denote the observed cumulative shares of households and income by  $\mathbf{p} = (p_0 = 0, p_1, \dots, p_{k-1}, p_k = 1)$  and  $\mathbf{y} = (y_0 = 0, y_1, \dots, y_{k-1}, y_k = 1)$ , respectively, which are usually constructed from a survey on  $n$  individual households. Even if the cumulative shares of households and cumulative shares of income are not directly available,  $\mathbf{p}$  and  $\mathbf{y}$  can be calculated from the income classes and class income means reported in the grouped data. Let us denote the cumulative distribution function and probability density function of the hypothetical income distribution with the parameter  $\theta$  by  $H(\cdot|\theta)$  and  $h(\cdot|\theta)$ , respectively. Then, the Lorenz curve denoted by  $L(y|\theta)$  is defined by

$$L(y|\theta) = \frac{1}{\mu} \int_0^y H^{-1}(z|\theta) dz, \quad y \in [0, 1],$$

where  $\mu$  is the mean of the distribution and  $H^{-1}(z|\theta) = \inf \{x : H(x|\theta) \geq z\}$ . Once the parameter estimate for  $\theta$  is obtained, the Gini coefficient can be estimated by using

$$G = -1 + \frac{2}{\mu} \int_0^\infty x H(x|\theta) h(x|\theta) dx, \quad (1)$$

$$= 1 - 2 \int_0^1 L(z|\theta) dz, \quad (2)$$

There are several methods to estimate the parameters of the Lorenz curves, for example, the least squares (Kakwani and Podder 1973) or generalised least squares (Kakwani and Podder 1976). More recently, Chotikapanich and Griffiths (2002) proposed a maximum likelihood estimator based on the likelihood from the Dirichlet distribution given by

$$f_{\text{DIR}}(\mathbf{q}|\boldsymbol{\theta}, \lambda) = \Gamma(\lambda) \prod_{j=1}^k \frac{q_j^{\lambda(L(p_j|\boldsymbol{\theta}) - L(p_{j-1}|\boldsymbol{\theta})) - 1}}{\Gamma(\lambda(L(p_j|\boldsymbol{\theta}) - L(p_{j-1}|\boldsymbol{\theta})))}, \tag{3}$$

where  $\mathbf{q} = (q_1, \dots, q_k)$ ,  $q_j = y_j - y_{j-1}$  is the income share for the  $j$ th group,  $\Gamma(\cdot)$  is the gamma function, and  $\lambda$  is the additional parameter of the Dirichlet likelihood. This likelihood function is motivated by the assumption given by  $E[q_j] = L(p_j|\boldsymbol{\theta}) - L(p_{j-1}|\boldsymbol{\theta})$ . The variance and covariance of the income shares implied from this likelihood are given by

$$\text{Var}(q_j) = \frac{E[q_j](1 - E[q_j])}{\lambda + 1}, \quad \text{Cov}(q_i, q_j) = -\frac{E[q_i]E[q_j]}{\lambda + 1},$$

where  $\lambda$  acts as a precision parameter. A Larger value of  $\lambda$  suggests that the variation of the income share around the Lorenz curve is small. Based on this likelihood function, Chotikapanich and Griffiths (2005) considered an MCMC method in the Bayesian framework by specifying the prior distributions of  $\boldsymbol{\theta}$  and  $\lambda$ .

Although their Dirichlet likelihood approach may appear convenient, it has the following problems. The parameter estimates and the resulting Gini estimate can be highly sensitive with respect to the choice of the value or prior distribution of  $\lambda$ , since data do not contain information on this parameter. The sensitivity is profound especially when the number of groups is small. Furthermore, the evaluation of the likelihood function requires the evaluation of  $L(y_i|\boldsymbol{\theta})$  for  $i = 1, \dots, k$ . Except for some simple standard distributions, such as the lognormal, Singh–Maddala, and Dagum, some flexible classes of hypothetical income distributions do not admit an analytical form of the Lorenz curve or the evaluation of the Lorenz curve is computationally expensive and unstable. Therefore, the inference based on the Dirichlet approach can be unreliable and its applicability would be limited.

### 2.2 Hypothetical income distribution: generalised beta distribution

In order to estimate the Lorenz curve and the related inequality measures accurately, a flexible class of hypothetical distributions is required. This paper adopts the five-parameter generalised beta (GB) distribution denoted by  $\mathcal{GB}(a, b, c, p, q)$  as an interesting and important income distribution. This distribution was proposed by McDonald and Xu (1995) and is the most flexible distribution of the family of beta-type distributions. The probability density function of the GB distribution is given by

$$h_{GB}(x|\boldsymbol{\theta}) = \frac{|a|x^{ap-1} \left[1 - (1 - c) \left(\frac{x}{b}\right)^a\right]^{q-1}}{b^{ap} B(p, q) \left[1 + c \left(\frac{x}{b}\right)^a\right]^{p+q}}, \quad 0 < x^a < \frac{b^a}{1 - c}, \tag{4}$$

where  $\boldsymbol{\theta} = (a, b, c, p, q)'$ ,  $a \in \mathbb{R}$ ,  $b > 0$ ,  $c \in [0, 1]$ ,  $p > 0$ ,  $q > 0$ , and  $B(p, q)$  is the beta function. Using the incomplete beta function  $B_x(p, q)$ , the cumulative distribution

function is given by  $H_{GB}(x|\theta) = B_z(p, q)/B(p, q)$  with  $z = (x/b)^a/(1 + c(x/b)^a)$ . The GB distribution includes a number of special cases. For example, when  $c = 0$  and  $c = 1$ , the GB distribution reduces to the generalised beta distribution of the first and second kind (GB1 and GB2) (McDonald 1984), respectively. Moreover, when  $(c, p) = (1, 1)$  and  $(c, q) = (1, 1)$ , the distribution reduces to the Singh–Maddala (SM) distribution (Singh and Maddala 1976) and Dagum (DA) distribution (Dagum 1977), which are known to perform well in many empirical applications. Detailed relationships among the class of distributions are summarised in McDonald and Xu (1995).

The hypothetical GB distribution can also be directly estimated from the grouped level income data by using the MCMC or maximum likelihood method (Kakamu and Nishino 2018). However, an explicit formula of the Lorenz curve for the GB distribution is not available (McDonald and Ransom 2008). This is also the case for the GB2 distribution, but the result on the Lorenz ordering for GB2 is known (Sarabia et al. 2002). Hence, the likelihood based on the Dirichlet distribution (3) is not explicitly available and the evaluation of the likelihood can be computationally expensive and unstable. The Gini coefficients for the GB and GB2 distributions are also not analytically available. In this paper, given the values for the parameters, the Gini coefficient is computed based on the equality given in McDonald and Ransom (2008):

$$G = 1 - \frac{\int_0^{b/(1-c)^{1/a}} (1 - H_{GB}(t|\theta))^2 dt}{\int_0^{b/(1-c)^{1/a}} 1 - H_{GB}(t|\theta) dt}, \quad (5)$$

where the integrals are evaluated numerically. Figure 1 compares the Gini coefficient for GB under various parameter values with its special cases. Since  $c$  is a distinctive parameter for GB, it is allowed to vary in all panels. The special cases of GB correspond to the coloured vertical dashed lines or symbols shown in each panel. The figure shows that GB can span a wide range of values of the Gini coefficient.

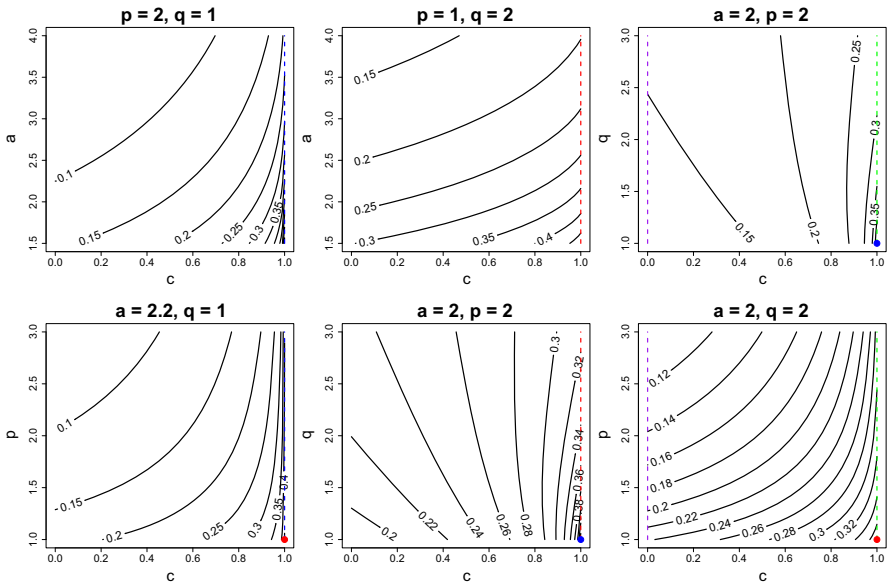
Note that the random variable  $X \sim \mathcal{GB}(a, b, c, p, q)$  can be easily generated by using

$$X = b \left( \frac{Z}{1 - cZ} \right)^{\frac{1}{a}}, \quad Z \sim Be(p, q), \quad (6)$$

where  $Be(p, q)$  is the beta distribution with parameters  $p$  and  $q$  (Kakamu and Nishino 2018). Therefore, the proposed ABC method described in the following would be a convenient approach to estimating the hypothetical GB distribution from the Lorenz curve based on the grouped data.

### 2.3 Approximate Bayesian computation for Lorenz curve

We work on the likelihood function implied from the Lorenz curve of the hypothetical income distribution. This likelihood function is constructed through the statistics of the individual household incomes and is not explicitly available. Thus, the standard MCMC methods cannot be directly applied, because these methods require evaluating the likelihood function and prior density. The approximate Bayesian computation



**Fig. 1** Contour plot of the Gini coefficient of GB under various parameter values. The blue, red, purple and green dashed lines or symbols indicate the Gini coefficient corresponding to DA, SM, GB1 and GB2, respectively (colour figure online)

(ABC) methods avoid direct evaluation of the likelihood function and simulate data from the model, given a candidate parameter value. If the simulated and observed data are similar, the candidate parameter value is a good candidate to generate the observed data. Then, it can be regarded as a sample from the posterior distribution (Sisson and Fan 2011). The posterior distribution can be approximated by weighting the intractable likelihood function. Therefore, ABC is a convenient approach when the likelihood function is not explicitly available or computationally prohibitive to evaluate, because it requires only the ability to simulate data from the probability model.

Let  $\pi(\theta)$  denote the prior density of the parameter  $\theta$ ,  $f(\mathbf{y}|\theta)$  be the likelihood function of the observed data  $\mathbf{y}$ , and  $\pi(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$  be the posterior distribution of  $\theta$ . Here,  $f(\mathbf{y}|\theta)$  corresponds to the likelihood function that is implicitly defined from the Lorenz curve. ABC methods augment the posterior from  $\pi(\theta|\mathbf{y})$  to

$$\pi_\epsilon(\theta, \mathbf{x}|\mathbf{y}) \propto \pi(\theta)f(\mathbf{x}|\theta)I_{A_{\epsilon,\mathbf{y}}}(\mathbf{x}),$$

where  $\epsilon > 0$  is a tolerance level,  $I_B(\cdot)$  is the indicator function of the set  $B$ , and  $\mathbf{x}$  is the simulated data. The set  $A_{\epsilon,\mathbf{y}}$  is defined as  $A_{\epsilon,\mathbf{y}} = \{\mathbf{x} : \rho(\mathbf{x}, \mathbf{y}) < \epsilon\}$ , where  $\rho(\cdot, \cdot)$  is a distance function. The value of  $\epsilon$  and form of  $\rho$  are chosen by the user and can affect the performance of ABC. The marginal distribution

$$\pi_{\epsilon}(\boldsymbol{\theta}|\mathbf{y}) \propto \int \pi(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})I_{A_{\epsilon,\mathbf{y}}}(\mathbf{x})d\mathbf{x}$$

provides an approximation to  $\pi(\boldsymbol{\theta}|\mathbf{y})$  for sufficiently small  $\epsilon$ .

Various ABC algorithms to sample from the approximate posterior distribution based on, for example, rejection sampling (Beaumont et al. 2002), MCMC (Marjoram et al. 2003; Fearnhead and Prangle 2012), and the sequential Monte Carlo (SMC) method (Sisson et al. 2007, 2009; Beaumont et al. 2009; Toni et al. 2009) have been proposed. Furthermore, a number of extensions of the SMC algorithm has been considered by, for example, Del Moral et al. (2012), Lenormand et al. (2013), Filippi et al. (2013), Silk et al. (2013), and Bonassi and West (2015). We employ the SMC approach because it is difficult to construct an efficient proposal distribution for MCMC in the present context and the SMC with adaptive weights proposed by Bonassi and West (2015), among others, is adopted because of its computational efficiency and ease of implementation.

The SMC algorithm proceeds by sampling from a series of intermediate distributions with the user-specified decreasing tolerance levels,  $\pi_{\epsilon_t}(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$  with  $\epsilon_t < \epsilon_{t-1}$  for  $t = 0, \dots, T$  and  $\epsilon_T$ . A large number of particles, denoted by  $(\boldsymbol{\theta}_i, \mathbf{x}_i)$ ,  $i = 1, \dots, N$ , is propagated by using importance sampling and resampling until the target tolerance level  $\epsilon_T$  is reached. Bonassi and West (2015) proposed to approximate the intermediate distribution at each step by kernel smoothing with the joint kernel  $K_t(\boldsymbol{\theta}, \mathbf{x}|\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{x}})$ . They employed the product kernel such that  $K_t(\boldsymbol{\theta}, \mathbf{x}|\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{x}}) = K_{\boldsymbol{\theta},t}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})K_{\mathbf{x},t}(\mathbf{x}|\tilde{\mathbf{x}})$ . Note that  $K_{\mathbf{x},t}(\mathbf{x}|\tilde{\mathbf{x}})$  is uniform over  $A_{\epsilon_t,\mathbf{y}}$  in the standard SMC of Sisson et al. (2007, 2009), Beaumont et al. (2009), and Toni et al. (2009). Algorithm 1 describes the method of Bonassi and West (2015). Introducing a kernel function for  $\mathbf{x}$  makes the perturbation step of the algorithm such that particles for which the simulated  $\mathbf{x}$  is close to  $\mathbf{y}$  are chosen more likely. Bonassi and West (2015) showed that the proposal distribution of their algorithm has higher prior predictive density over the acceptance region for the next step—and hence, higher acceptance probability—than that of the standard SMC algorithm. Finally, the posterior distribution of  $\boldsymbol{\theta}$  at step  $t$  of the algorithm is approximated by

$$\pi_{\epsilon_t}(\boldsymbol{\theta}|\mathbf{y}) \propto \int \int \pi(\tilde{\boldsymbol{\theta}})f(\tilde{\mathbf{x}}|\tilde{\boldsymbol{\theta}})K_{\boldsymbol{\theta},t}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})K_{\mathbf{x},t}(\mathbf{y}|\tilde{\mathbf{x}})I_{A_{\epsilon_t,\mathbf{y}}}(\mathbf{x})d\tilde{\mathbf{x}}d\tilde{\boldsymbol{\theta}}.$$

To estimate the hypothetical income distribution from the Lorenz curve based on the group data by using ABC, we use the cumulative income shares in percentage. In Algorithm 1, we set  $\rho(\mathbf{x}, \mathbf{y}) = 100 \max_j |x_j - y_j|$ , which was also employed in McVinish (2012), because this choice of the tolerance schedule and level of approximation is intuitive. To simulate  $\mathbf{x}$ ,  $n$  observations,  $(z_1, \dots, z_n)$ , that independently and identically follow the hypothetical income distribution with the density function  $h(\cdot|\boldsymbol{\theta})$  are generated. Then, they are sorted in the ascending order, denoted by  $(z_{(1)}, \dots, z_{(n)})$ , and the cumulative income shares are computed using  $x_j = \sum_{i=1}^{n_j} z_{(i)} / \sum_{i=1}^n z_{(i)}$ , where  $n_j = \lfloor np_j \rfloor$  is the number of households in the  $j$ th income class for  $j = 1, \dots, k - 1$ . For the GB distribution and its special cases, the simulated data are generated from  $\mathcal{GB}(a, 1, c, p, q)$  by using (6), as the Lorenz curve is location-free.



---

**Algorithm 1** *SMC with adaptive weights*

---

```

1: Initialise tolerance levels  $\epsilon_0 > \epsilon_1 > \dots > \epsilon_T$  and set  $t = 0$ .
2: for  $i = 1$  to  $N$  do
3:   repeat
4:     Simulate  $\theta_i^{(0)}$  from  $\pi(\theta)$  and  $\mathbf{x}_i^*$  from  $f(\mathbf{x}|\theta_i^{(0)})$ .
5:   until  $\rho(\mathbf{x}_i^*, \mathbf{y}) < \epsilon_0$ 
6:   Set  $\mathbf{x}_i^{(0)} = \mathbf{x}_i^*$  and  $w_i = 1/N$  for  $i = 1, \dots, N$ .
7: end for
8: for  $t = 1$  to  $T$  do
9:   Compute the weights  $v_i^{(t-1)} \propto w_i^{(t-1)} K_{\mathbf{y},t}(\mathbf{y}|\mathbf{x}_i^{(t-1)})$  for  $i = 1, \dots, N$ .
10:  for  $i = 1$  to  $N$  do
11:    repeat
12:      Choose  $\theta_i^*$  from  $\theta_i^{(t-1)}$ 's based on the weights  $v_j^{(t-1)}$ .
13:      Draw  $\theta_i^{(t)}$  from  $K_{\theta,t}(\theta_i^{(t)}|\theta_i^*)$  and simulate  $\mathbf{x}_i^{(t)}$  from  $f(\mathbf{x}|\theta_i^{(t)})$ .
14:    until  $\rho(\mathbf{x}_i^{(t)}, \mathbf{y}) < \epsilon_t$ 
15:    Compute the new weights as  $w_i^{(t)} \propto \frac{\pi(\theta_i^{(t)})}{\sum_{j=1}^N v_j^{(t-1)} K_{\theta,t}(\theta_i^{(t)}|\theta_j^{(t-1)})}$ .
16:  end for
17: end for

```

---

As in Bonassi and West (2015), the product of normal kernels is used for  $\theta$ . Following the rule of thumb for the product of normal kernels, the bandwidth is determined based on  $u_s = \hat{\sigma}_s N^{-1/(d+4)}$ , where  $N$  is the number of particles,  $d$  is the total dimension of the parameter and data, and  $\hat{\sigma}_s$  is the standard deviation for  $s \in \{\theta, \mathbf{x}\}$  (Scott and Sain 2005; Bonassi and West 2015). When the number of groups is large, such as  $k = 10$  in decile data, the performance and computing time of ABC may be affected (see, e.g., Prangle 2015), as Algorithm 1 compares two nine-dimensional vectors. To reduce the dimensionality, we can also use summary statistics that consist of a subset of the elements of the cumulative incomes. For example, when  $k = 10$ , we can replace  $\mathbf{y}$  and  $\mathbf{x}$  in Algorithm 1 with  $S(\mathbf{y}) = (y_1, y_3, y_5, y_7, y_9)$  and  $S(\mathbf{x}) = (x_1, x_3, x_5, x_7, x_9)$ , respectively. Note that if we take  $S(\mathbf{x}) = (x_2, x_4, x_6, x_8)$ , it is identical to the simulated data in the case of  $k = 5$ . The use of the summary statistics in the case of  $k = 10$  is also examined in Sect. 3.

### 3 Numerical examples

#### 3.1 Simulated data 1

A series of simulation studies is conducted to illustrate the proposed approach, which is denoted by ABC hereafter. First, the individual household income follows the Dagum (DA) and Singh–Maddala (SM) distributions, denoted by  $\mathcal{DA}(a, b, p) = \mathcal{GB}(a, b, 1, p, 1)$  and  $\mathcal{SM}(a, b, q) = \mathcal{GB}(a, b, 1, 1, q)$ , respectively. For ABC, the parameters to be estimated are given by  $\theta_{\text{ABC}} = (a, p)$  and  $(a, q)$  for DA and SM, respectively. The performance of ABC is compared with that of the two existing meth-

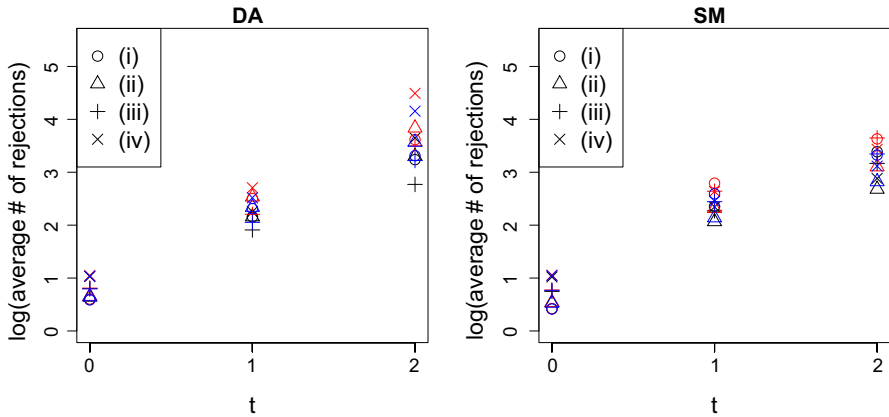
ods. The first method is based on the Dirichlet likelihood given by (3), denoted by DIR hereafter. Note that both distributions allow the explicit forms of the Lorenz curve. The parameters  $\theta_{\text{DIR}} = (a, p, \lambda)$  and  $(a, q, \lambda)$  are estimated by using the Metropolis–Hastings (MH) algorithm. The other method, proposed by Kakamu and Nishino (2018), estimates the hypothetical income distribution from the grouped level income. In this approach, the thresholds for the income classes,  $\mathbf{v} = (v_1, \dots, v_{k-1})$ , are regarded as the selected order statistics (SOS) from the order statistics of the  $n$  observations of the level income,  $(w_{(1)}, \dots, w_{(n)})$ , where  $v_j = w_{(n_j)}$  for  $j = 1, \dots, k - 1$  and  $n_j$  is defined in Sect. 2.3. The likelihood function is given by

$$f_{\text{SOS}}(\mathbf{v}|\theta_{\text{SOS}}) = n! \frac{[H(v_1|\theta_{\text{SOS}})]^{n_1-1}}{(n_1 - 1)!} \times \left[ \prod_{j=2}^{k-1} \frac{[H(v_j|\theta_{\text{SOS}}) - H(v_{j-1}|\theta_{\text{SOS}})]^{n_j-n_{j-1}-1}}{(n_j - n_{j-1} - 1)!} \right] \times \frac{[1 - H(v_{k-1}|\theta_{\text{SOS}})]^{n-n_{k-1}}}{(n - n_{k-1})!} \left[ \prod_{j=1}^k h(v_j|\theta_{\text{SOS}}) \right],$$

where  $\theta_{\text{SOS}} = (a, b, p)$  and  $(a, b, q)$  are estimated from the data in this approach. While this likelihood is similar to the multinomial likelihood, it provides a more accurate representation for the grouped data (David and Nagaraja 2003). The posterior inference is based on the MH algorithm. This approach is denoted by SOS hereafter.

To create the data for this simulation study,  $n = 10,000$  observations are generated from  $\mathcal{DA}(a, 1, p)$  and  $\mathcal{SM}(a, 1, q)$ . Then, the data are sorted in ascending order and are grouped into  $k$  groups of equal size to calculate the cumulative income and household shares. The data are replicated 100 times. For the Dagum distribution, the following four settings for the true parameter values and corresponding Gini coefficients are considered: (i)  $(a, p, G) = (3.8, 1.3, 0.2482)$ , (ii)  $(a, p, G) = (3.0, 1.5, 0.3087)$ , (iii)  $(a, p, G) = (2.5, 2.5, 0.3518)$ , (iv)  $(a, p, G) = (2.3, 1.5, 0.4077)$ . For the Singh–Maddala distribution, the following four settings are considered: (i)  $(a, q, G) = (3.5, 1.5, 0.2429)$ , (ii)  $(a, q, G) = (2.3, 3.0, 0.3041)$ , (iii)  $(a, q, G) = (2.0, 2.5, 0.3567)$ , (iv)  $(a, q, G) = (1.6, 3.5, 0.4052)$ . For the number of groups, we consider  $k = 5$  and 10. These choices respectively correspond to quintile and decile data, which are the most commonly available in practice. In the case of  $k = 10$ , we also implement Algorithm 1 with  $S(\mathbf{y}) = (y_1, y_3, y_5, y_7, y_9)$  and  $S(\mathbf{x}) = (x_1, x_3, x_5, x_7, x_9)$  suggested in Sect. 2.3.

For ABC, we used 3000 particles with the schedule of tolerance levels given by  $\{\epsilon_t\} = \{0.1, 0.01, 0.002\}$ . Algorithm 1 is implemented by using Ox Professional version 7.10 (Doornik 2013) with six parallel threads for the lines between 10 and 16. We assume that  $a, b, p$ , and  $q$  independently follow  $\mathcal{G}(3, 1)$  to reflect the results in the existing literature on the GB distributions (e.g., McDonald and Ransom 2008). For DIR, the following prior distributions of  $\lambda$  with the different prior means and same prior variances are considered:  $\mathcal{G}(10, 1)$ ,  $\mathcal{G}(1, 0.5)$ , and  $\mathcal{G}(1, 1)$ . For SOS and DIR, the MCMC algorithms are run for 40,000 iterations

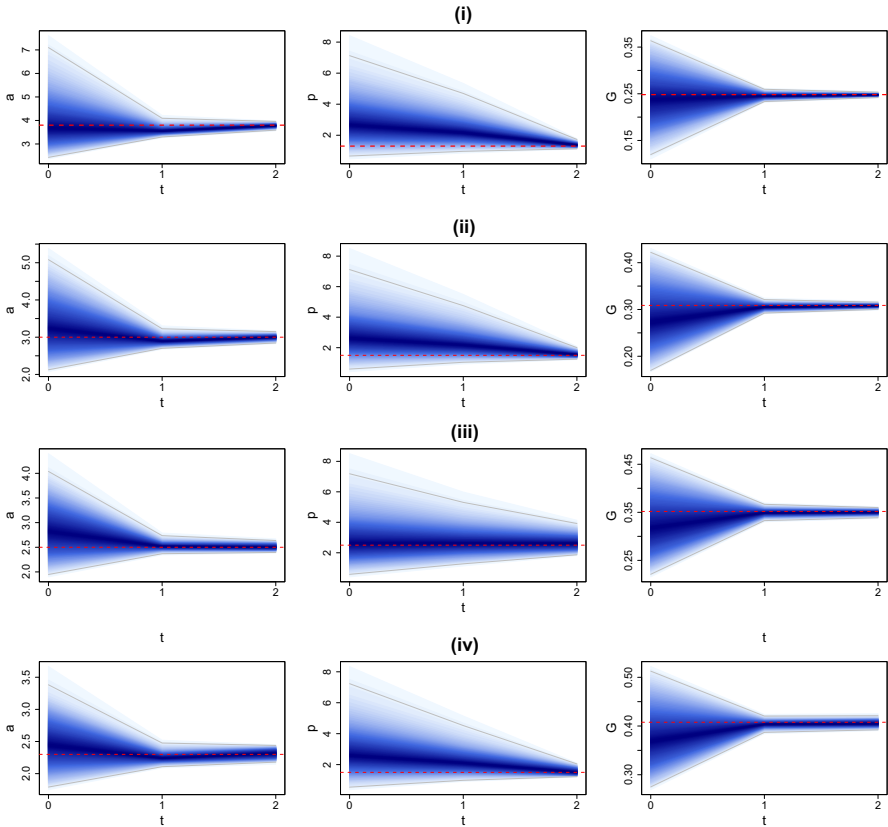


**Fig. 2** Log average numbers of rejections per particle for DA and SM:  $k = 5$  (black),  $k = 10$  (red),  $k = 10$  with summary (blue) (colour figure online)

including the 10,000 initial burn-in period. To reduce the undesired autocorrelation among the MCMC samples, every 10th draw is retained for posterior inference.

Figure 2 presents the log average numbers of rejections per particle for each step of Algorithm 1 for DA and SM. A large number of rejections implies longer computing time because the lines between 11 and 14 of Algorithm 1 are repeated for an increased number of times. The figure shows that the computing time of the algorithm increases as the number of groups increases in all cases. Two nine-dimensional vectors are compared when  $k = 10$ , leading to large numbers of rejections, but the computing time can be decreased by using the summary statistics. In addition, the figure shows that the computing time may depend on the true Gini coefficient. The average number of rejections tend to increase as the Gini coefficient increases in the case of DA while this tendency is less clear in the case of SM. Figures 3 and 4 present the typical trajectories of Algorithm 1 for  $k = 5$  for DA and SM, respectively. In the figures, the red horizontal dashed lines represent the true parameter values and the grey curves represent the 2.5% and 97.5% quantiles at each step. The figures show that the learning about the parameters and corresponding Gini coefficients occurs as the algorithm proceeds and the posterior distributions are concentrated around the true values under the target tolerance level.

Now, the performance of the three methods is compared. Table 1 presents the averages of the posterior means of the parameters and Gini coefficient and root mean squared errors (RMSE) for DA and SM over the 100 replicates. Overall, ABC appears to work well. In the case of  $k = 5$ , ABC resulted in the smallest RMSE for the Gini coefficients for both DA and SM. In the case of  $k = 10$ , ABC and SOS produced the comparable result for DA and ABC and Dirichlet produced almost identical performance for SM in terms of the RMSE for the Gini coefficient. As the available information increases from  $k = 5$  to  $k = 10$ , the performance of ABC seems to improve slightly, but the degree of improvement is small. On the other hand, we observe a clear improvement in the performance of SOS and Dirichlet as the number of groups increases.

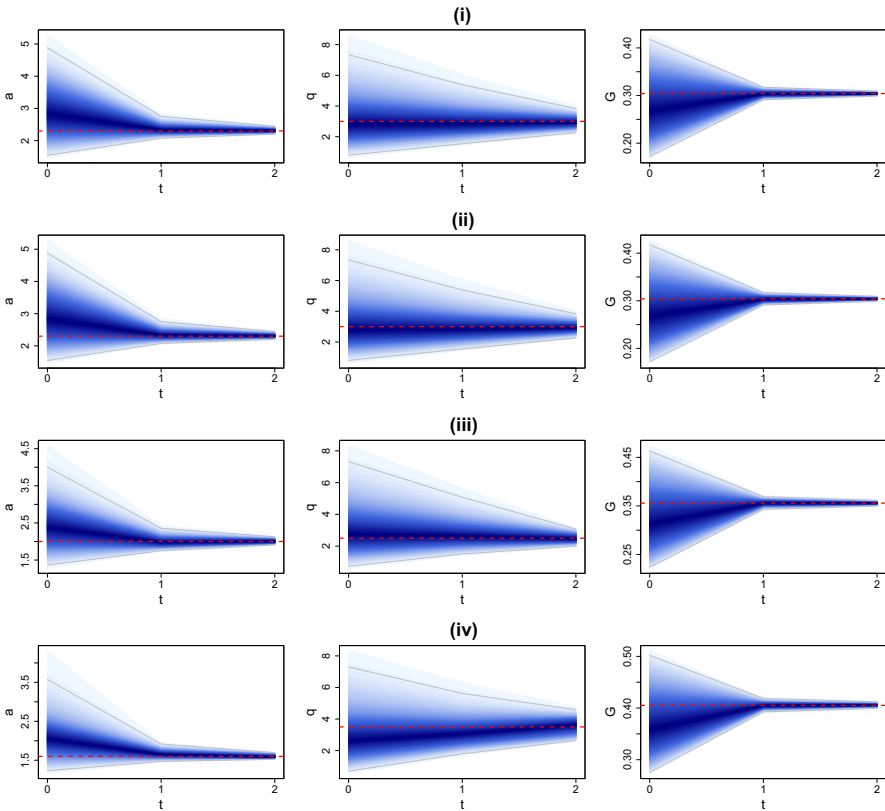


**Fig. 3** Typical trajectories of Algorithm 1 for DA ( $k = 5$ ) with the 2.5% and 97.5% quantiles (grey solid lines) and true parameter values (red dashed lines) (colour figure online)

The table also shows that the parameter and Gini estimates for DIR in the case of  $k = 5$  can be sensitive with respect to the prior specification for  $\lambda$ . More specifically, the smaller the prior means, the larger RMSE for the parameters and Gini coefficient are resulted. The influence of the prior distribution for  $\lambda$  appears to vanish in the case of  $k = 10$ . We also considered the two alternative prior specifications for the parameters, namely,  $\mathcal{G}(1.5, 0.5)$  and  $\mathcal{G}(6, 2)$ . They have the same prior means as the default prior, but the prior variance is two times inflated in the first alternative prior and it is two times deflated in the second alternative prior. The results are presented in the supplementary material, where it is reported that ABC and SOS are fairly robust with respect to the choice of prior distribution for the parameters other than  $\lambda$ , but the results for DIR can be influenced by the prior specification especially when  $\mathcal{G}(0.1, 0.1)$  is used for  $\lambda$  in the case of  $k = 5$ .

### 3.2 Simulated data 2

To study the potential of the proposed approach, the more flexible alternatives to the Dagum and Singh–Maddala distributions, namely the generalised beta distribution



**Fig. 4** Typical trajectories of Algorithm 1 for SM ( $k = 5$ ) with the 2.5% and 97.5% quantiles (grey solid lines) and true parameter values (red dashed lines) (colour figure online)

of the second kind (GB2) denoted by  $\mathcal{GB2}(a, b, p, q) = \mathcal{GB}(a, b, 1, p, q)$ , and the five-parameter GB distribution, are additionally considered. For GB2, the data are generated from  $\mathcal{GB2}(a, 1, p, q)$  based on the following four settings: (i)  $(a, p, q, G) = (2.5, 2.3, 1.7, 0.2572)$ , (ii)  $(a, p, q, G) = (2.1, 1.8, 2.0, 0.3037)$ , (iii)  $(a, p, q, G) = (1.8, 3.0, 1.5, 0.3536)$ , (iv)  $(a, p, q, G) = (1.5, 2.5, 1.8, 0.4064)$ . For GB, the following five settings covering various values of the parameters and Gini coefficient are considered: (i)  $(a, c, p, q, G) = (2.0, 0.95, 3.0, 2.0, 0.2456)$ , (ii)  $(a, c, p, q, G) = (1.2, 0.4, 1.7, 2.5, 0.3062)$ , (iii)  $(a, c, p, q, G) = (1.5, 0.9, 1.7, 1.7, 0.3589)$ , (iv)  $(a, c, p, q, G) = (1.2, 0.1, 1.3, 3.5, 0.3397)$ , (v)  $(a, c, p, q, G) = (1.5, 0.99, 1.2, 3.0, 0.4105)$ . The data are replicated 50 times. We implement only ABC and SOS, because analytical forms of the Lorenz curves for GB and GB2 are not known (McDonald and Ransom 2008).

In addition to the prior distributions for  $a, b, p, q$  specified in Sect. 3.1,  $c \sim \mathcal{U}(0, 1)$  is assumed for GB. The MCMC algorithm for SOS is run for 70,000 iterations including the 10,000 initial burn-in period and every 20th draw is retained for posterior inference. For ABC, the same setting for Algorithm 1 as in Sect. 3.1 is used.

**Table 1** Result for DA and SM under the default prior

Model	$k$	Setting	Parameter	True	ABC		ABC(sum)		SOS		Dirichlet 1		Dirichlet 2		Dirichlet 3	
					Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
DA	5	(i)	$a$	3.8	3.7898	0.0662			3.7689	0.1089	3.7997	0.0645	3.7804	0.0706	3.7015	0.2913
			$p$	1.3	1.3383	0.0861			1.3699	0.1584	1.3180	0.0719	1.3875	0.1304	2.1106	1.0338
			$G$	0.2482	0.2483	0.0023			0.2491	0.0035	0.2480	0.0024	0.2478	0.0025	0.2464	0.0160
	(ii)	$a$	3.0	2.9973	0.0568			2.9743	0.0907	3.0032	0.0562	2.9955	0.0555	2.9640	0.1783	
		$p$	1.5	1.5424	0.1192			1.6089	0.2771	1.5158	0.1016	1.5662	0.1262	2.1379	0.9944	
		$G$	0.3087	0.3088	0.0034			0.3102	0.0050	0.3084	0.0035	0.3083	0.0035	0.3063	0.0143	
10	(iii)	$a$	2.5	2.4939	0.0506			2.4880	0.0696	2.5052	0.0525	2.5019	0.0508	2.5185	0.1028	
		$p$	2.5	2.6983	0.3989			2.7611	0.6833	2.5501	0.3108	2.6594	0.3472	3.0494	0.9423	
		$G$	0.3518	0.3523	0.0046			0.3532	0.0063	0.3514	0.0050	0.3512	0.0050	0.3485	0.0159	
	(iv)	$a$	2.3	2.2997	0.0506			2.2840	0.0645	2.3059	0.0524	2.3022	0.0530	2.2694	0.0962	
		$p$	1.5	1.5478	0.1428			1.5891	0.2044	1.5072	0.1239	1.5509	0.1657	2.1047	0.8959	
		$G$	0.4077	0.4081	0.0056			0.4098	0.0073	0.4071	0.0060	0.4070	0.0060	0.4057	0.0149	
(i)	(i)	$a$	3.8	3.7893	0.0704	3.7900	0.0726	3.7916	0.0640	3.8042	0.0702	3.8020	0.0700	3.8038	0.0701	
		$p$	1.3	1.3374	0.0945	1.3396	0.0994	1.3185	0.0748	1.3073	0.0739	1.3141	0.0760	1.3077	0.0743	
		$G$	0.2482	0.2483	0.0023	0.2483	0.0023	0.2485	0.0025	0.2480	0.0025	0.2480	0.0025	0.2480	0.0025	
	(ii)	$a$	3.0	2.9934	0.0584	3.0007	0.0637	2.9937	0.0498	3.0055	0.0624	3.0043	0.0626	3.0051	0.0630	
		$p$	1.5	1.5494	0.1325	1.5361	0.1460	1.5247	0.0959	1.5067	0.1114	1.5148	0.1155	1.5110	0.1218	
		$G$	0.3087	0.3090	0.0033	0.3088	0.0033	0.3092	0.0033	0.3084	0.0037	0.3084	0.0037	0.3084	0.0037	
(iii)	$a$	2.5	2.4945	0.0488	2.4939	0.0510	2.4960	0.0395	2.5065	0.0591	2.5060	0.0588	2.5077	0.0603		
	$p$	2.5	2.6868	0.4077	2.7127	0.4477	2.5705	0.2433	2.5378	0.3741	2.5562	0.3800	2.5361	0.3942		
	$G$	0.3518	0.3523	0.0043	0.3524	0.0044	0.3524	0.0040	0.3514	0.0054	0.3513	0.0054	0.3514	0.0055		
(iv)	$a$		2.3	2.2962	0.0477	2.2978	0.0521	2.2960	0.0382	2.3074	0.0591	2.3069	0.0590	2.3068	0.0591	

Table 1 continued

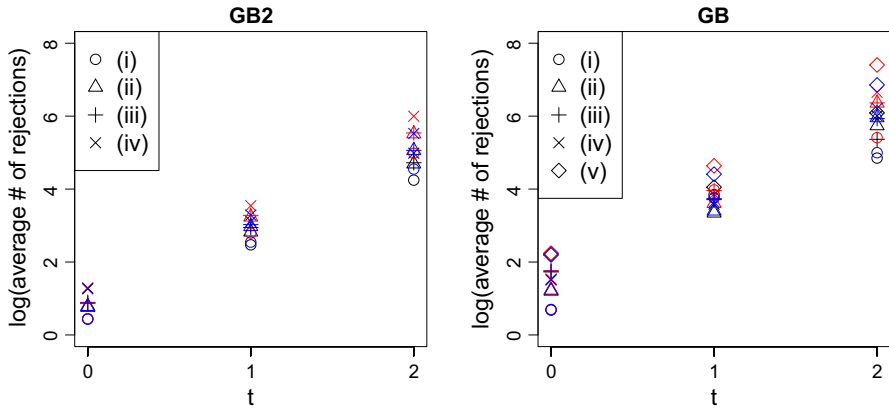
Model	$k$	Setting	Parameter	True	ABC		ABC(sum)		SOS		Dirichlet 1		Dirichlet 2		Dirichlet 3			
					Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
SM	5	(i)	$p$	1.5	1.5560	0.1471	1.5643	0.1660	1.5229	0.0954	1.5034	0.1486	1.5098	0.1522	1.5093	0.1573		
			$G$	0.4077	0.4084	0.0051	0.4083	0.0054	0.4084	0.0047	0.4071	0.0065	0.4071	0.0065	0.4071	0.0065		
			$a$	3.5	3.4741	0.0587			3.4684	0.0993	3.4833	0.0529	3.4646	0.0623	3.4646	0.0623	3.3102	0.3329
		(ii)	$q$	1.5	1.5526	0.0906			1.6053	0.2148	1.5358	0.0774	1.5827	0.1135	1.5827	0.1135	2.1009	0.8839
			$G$	0.2429	0.2430	0.0018			0.2424	0.0038	0.2428	0.0019	0.2428	0.0019	0.2428	0.0019	0.2448	0.0173
			$a$	2.3	2.2913	0.0351			2.2939	0.0537	2.2948	0.0328	2.2934	0.0321	2.2934	0.0321	2.3477	0.1791
	(iii)	$q$	3.0	3.1518	0.2884			3.3001	0.6636	3.0965	0.2476	3.1668	0.2839	3.1668	0.2839	3.2595	0.6596	
		$G$	0.3041	0.3042	0.0021			0.3042	0.0041	0.3040	0.0021	0.3040	0.0021	0.3039	0.0021	0.3023	0.0122	
		$a$	2.0	1.9916	0.0337			1.9893	0.0505	1.9951	0.0299	1.9920	0.0296	1.9920	0.0296	1.9962	0.1447	
	10	(iv)	$q$	2.5	2.6054	0.2182			2.7626	0.5330	2.5591	0.1767	2.6135	0.2021	2.8304	0.6453		
			$G$	0.3567	0.3557	0.0026			0.3552	0.0056	0.3555	0.0026	0.3554	0.0025	0.3557	0.0156		
			$a$	1.6	1.5934	0.0271			1.6019	0.0350	1.5980	0.0236	1.5990	0.0228	1.7070	0.2819		
(i)		$q$	3.5	3.7155	0.4137			3.7560	0.7196	3.5928	0.3192	3.6296	0.3303	3.4775	0.4149			
		$G$	0.4052	0.4051	0.0028			0.4060	0.0059	0.4049	0.0028	0.4049	0.0028	0.3977	0.0240			
		$a$	3.5	3.4806	0.0592	3.4838	0.0603	3.4860	0.0659	3.4897	0.0514	3.4869	0.0522	3.4901	0.0507			
(ii)	$q$	1.5	1.5397	0.0856	1.5371	0.0863	1.5389	0.1076	1.5224	0.0721	1.5283	0.0749	1.5217	0.0714				
	$G$	0.2429	0.2430	0.0019	0.2430	0.0019	0.2427	0.0024	0.2429	0.0019	0.2429	0.0019	0.2429	0.0019				
	$a$	2.3	2.2941	0.0356	2.2941	0.0369	2.2947	0.0408	2.2963	0.0318	2.2957	0.0319	2.2962	0.0319				
		$q$	3.0	3.1075	0.2617	3.1133	0.2742	3.1434	0.3879	3.0652	0.2254	3.0792	0.2312	3.0663	0.2270			
		$G$	0.3041	0.3042	0.0021	0.3042	0.0021	0.3040	0.0028	0.3040	0.0021	0.3040	0.0021	0.3040	0.0021			

Table 1 continued

Model	k	Setting	Parameter	True	ABC		ABC(sum)		SOS		Dirichlet 1		Dirichlet 2		Dirichlet 3	
					Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
(iii)			a	2.0	1.9951	0.0344	1.9958	0.0365	1.9941	0.0362	1.9968	0.0291	1.9963	0.0291	1.9965	0.0294
				2.5	2.5700	0.1996	2.5733	0.2127	2.6047	0.2784	2.5409	0.1662	2.5469	0.1678	2.5431	0.1690
				G	0.3567	0.3557	0.0025	0.3557	0.0025	0.3554	0.0036	0.3556	0.0025	0.3556	0.0025	0.3556
(iv)			a	1.6	1.5968	0.0281	1.5950	0.0306	1.5985	0.0275	1.5985	0.0234	1.5986	0.0233	1.5984	0.0234
				q	3.6390	0.3775	3.6758	0.4182	3.6640	0.4959	3.5698	0.3031	3.5762	0.3044	3.5716	0.3041
				G	0.4052	0.0027	0.4052	0.0027	0.4050	0.0039	0.4050	0.0027	0.4050	0.0027	0.4050	0.0027

ABC(sum) denotes the ABC method using the summary statistics when  $k = 10$ . Dirichlet 1, 2, and 3 denote the results under the Dirichlet likelihood with  $\mathcal{G}(10, 1)$ ,  $\mathcal{G}(2.5, 0.5)$ , and  $\mathcal{G}(0.1, 0.1)$  priors for  $\lambda$ , respectively



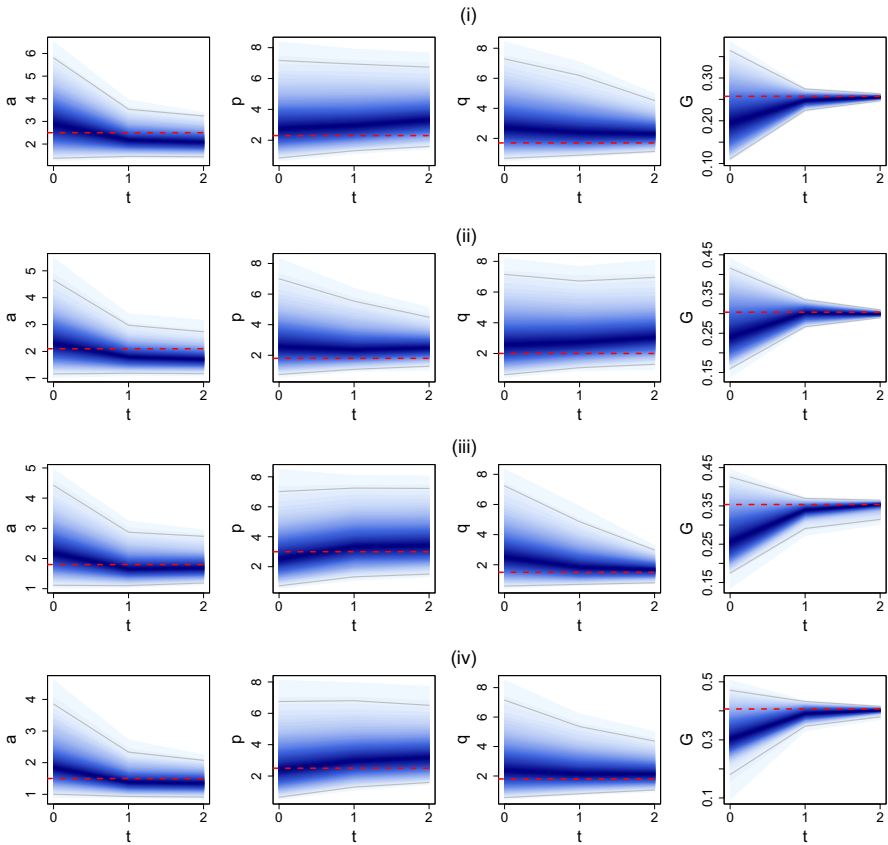


**Fig. 5** Log average number of rejections per particle for GB2 and GB:  $k = 5$  (black),  $k = 10$  (red),  $k = 10$  with summary (blue) (colour figure online)

Figure 5 presents the log average numbers of rejections per particle for each step of Algorithm 1 for GB and GB2. The figure shows that the overall numbers of rejections are larger for GB and GB2 than for DA and SM leading to the increased computing time, since it is required to estimate more parameters for GB and GB2. The average number of rejections tends to increase when the true Gini coefficient increases. In addition, using the summary statistics in the case of  $k = 10$  results in the shorter computing time.

Figures 6 and 7 present the typical trajectories of Algorithm 1 for  $k = 5$  for GB2 and GB, respectively. The red horizontal dashed lines represent the true parameter values and the grey curves represent the 2.5% and 97.5% quantiles. In contrast to the cases of DA and SM, the figures show that not all parameters are simultaneously identified from the data under the present simulation setting. This could be because the information contained in the data is limited. For example, in Setting (i) for GB2, Fig. 6 shows that the learning about  $a$  and  $q$  occurs and the posterior distributions concentrate as the algorithm proceeds, but little learning about  $p$  occurs. Similarly, for GB, Fig. 7 shows we can only learn about  $a$  and  $c$  in Setting (i) and about  $a$  and  $p$  in Setting (iv). Which parameters we can learn seems to depend on the simulation setting. Nonetheless, the figures also show that in all cases the learning about the Gini coefficient does occur as the algorithm proceeds and the posterior distributions under the target tolerance are concentrated around the true values.

Table 2 presents the averages of the posterior means and RMSE for the parameters and the Gini coefficient for GB2 and GB under the two methods. In all cases, ABC estimated the Gini coefficient well and produced the smaller RMSE for the Gini coefficient than SOS. For both methods, the performance improves as the number of groups increases, but the degree of improvement for ABC is small compared to SOS, as in the cases of DA and SM. The large RMSE for the parameters in the table corresponds to the cases where the parameters are not well identified from the data as indicated by Figs. 6 and 7. The large RMSE for SOS could be also due to the poor mixing and convergence failure of the MCMC algorithm, as the convergence

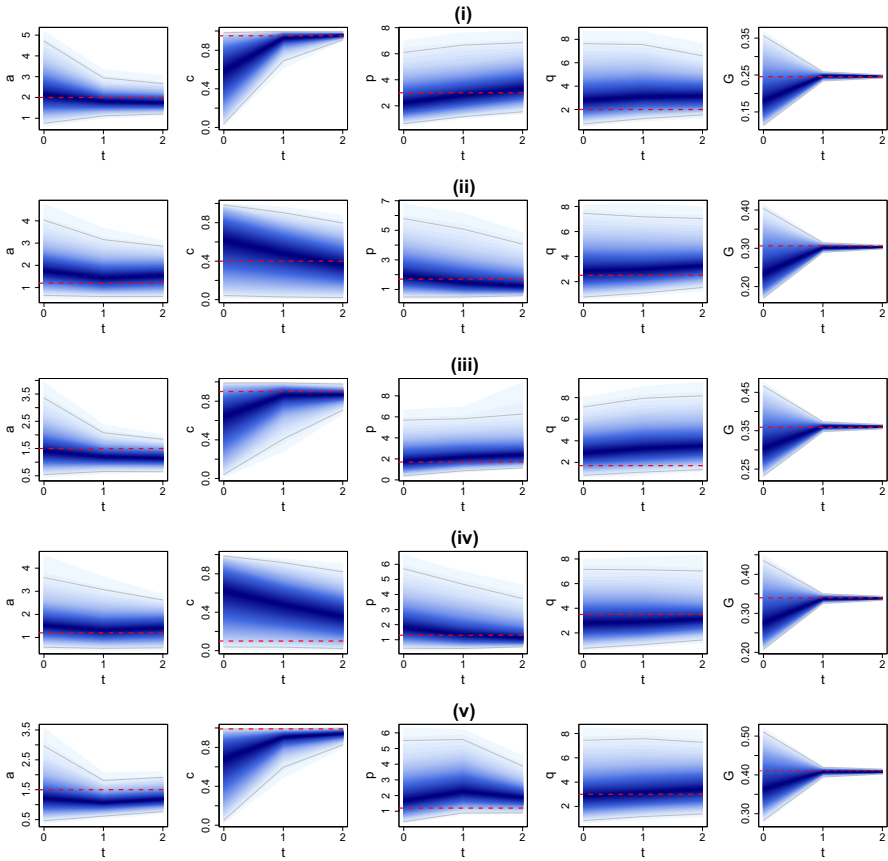


**Fig. 6** Typical trajectories of Algorithm 1 for GB2 ( $k = 5$ ) with the 2.5% and 97.5% quantiles (grey solid lines) and true parameter values (grey dashed lines)

of MCMC in the context of grouped data is typically difficult to ensure (Kakamu 2016). The supplementary material provides the results under the two alternative prior specifications,  $\mathcal{G}(1.5, 0.5)$  and  $\mathcal{G}(6.2)$ , for  $a$ ,  $p$ , and  $q$ . The Gini estimate for the proposed ABC method is robust with respect to the prior specification, while that for SOS exhibits prior sensitivity especially for the quintile data. For both methods, the estimates for the parameters that are difficult to identify from the data are influenced by the prior specification.

### 3.3 Real data: Family Income and Expenditure Survey in Japan

The proposed method is now applied to estimate the Gini coefficient of the data from the Family Income and Expenditure Survey (FIES) in 2012 prepared by Ministry of Internal Affairs and Communications of Japan. The FIES data are based on  $n = 10,000$  households and are available in the forms of quintile and decile data. The datasets are available in the supplementary material. For the hypothetical income distributions,



**Fig. 7** Typical trajectories of Algorithm 1 for GB ( $k = 5$ ) with the 2.5% and 97.5% quantiles (grey solid lines) and true parameter values (red dashed lines) (colour figure online)

DA, SM, GB2, and GB are fitted. The same default prior distributions and algorithm settings as in the simulation studies are used.

Table 3 presents the posterior means and 95% credible intervals under the target tolerances. For GB, we are able to learn about  $a$  and  $c$ , while little learning about  $p$  and  $q$  occurred, similar to Setting (i) of the simulation study. Similarly, for GB2, some learning about  $a$  and  $p$  occurred similar to Setting (ii) of the simulation study. We can still obtain some insights on the shape of the underlying income distribution. Figure 8 presents the implied income distributions which are obtained by generating the random numbers from each distribution with the parameters fixed to their posterior means and scaling them with the theoretical standard deviations under these parameter values. The distribution shapes of GB and GB2 are almost identical. The figure also shows that DA and SM have higher density in the low income region and the right tails decays more quickly compared to GB and GB2.

The goodness of fit of the income models can be quantified through the marginal likelihood, which is calculated following Didelot et al. (2011). The log marginal like-

**Table 2** Result for GB2 and GB distributions under the default prior

Model	Setting	Parameter	True	$k = 5$			$k = 10$			SOS				
				ABC		RMSE	ABC		RMSE	ABC(sum)		RMSE	SOS	
				Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	
GB2	(i)	$a$	2.5	2.1391	0.3630	0.4645	2.1610	0.3473	2.1572	0.3497	2.3658	0.4295		
		$p$	2.3	3.3641	1.0741	0.6206	3.2156	0.9449	3.2277	0.9545	2.0440	0.6048		
		$q$	1.7	2.5028	0.8085	1.8285	2.4675	0.7792	2.4978	0.8067	2.9374	1.6008		
	(ii)	$G$	0.2572	0.2588	0.0029	0.0099	0.2584	0.0026	0.2582	0.0026	0.2493	0.0085		
		$a$	2.1	1.8001	0.3066	0.3392	1.8325	0.2848	1.8301	0.2843	1.9047	0.3653		
		$p$	1.8	2.4669	0.6866	1.0547	2.3497	0.5920	2.3524	0.5904	2.6887	1.1674		
	(iii)	$q$	2.0	3.1481	1.1573	0.6511	3.0366	1.0690	3.0634	1.0853	2.3170	0.6595		
		$G$	0.3037	0.3027	0.0031	0.0084	0.3022	0.0033	0.3021	0.0033	0.3096	0.0077		
		$a$	1.8	1.7017	0.1076	0.2945	1.6945	0.1275	1.6907	0.1301	1.7386	0.2506		
	(iv)	$p$	3.0	3.7787	0.8002	1.3589	3.6963	0.7577	3.7207	0.7754	1.6814	1.3617		
		$q$	1.5	1.8040	0.3212	2.7102	1.8054	0.3337	1.8267	0.3550	3.8035	2.6010		
		$G$	0.3536	0.3721	0.0192	0.0259	0.3715	0.0187	0.3713	0.0185	0.3280	0.0259		
GB	(i)	$a$	1.5	1.3756	0.1311	0.2812	1.3900	0.1374	1.3875	0.1357	1.3864	0.2359		
		$p$	2.5	3.2409	0.7699	0.7222	3.1050	0.6968	3.1176	0.6985	2.2012	0.6111		
		$q$	1.8	2.3654	0.5745	1.8263	2.3084	0.5468	2.3326	0.5635	3.3054	1.8158		
	(ii)	$G$	0.4064	0.4138	0.0085	0.0221	0.4134	0.0081	0.4132	0.0080	0.3862	0.0208		
		$a$	2.0	1.8520	0.1494	0.2602	1.8109	0.1910	1.8212	0.1833	1.9644	0.2668		
		$c$	0.95	0.9538	0.0060	0.0187	0.9490	0.0048	0.9489	0.0051	0.9592	0.0141		
(iii)	$p$	3.0	3.6033	0.6149	0.6437	3.6951	0.7078	3.6701	0.6980	3.2986	0.8269			
	$q$	2.0	3.0596	1.0637	0.4223	3.1261	1.1391	3.1288	1.1399	2.3891	0.6884			
	$G$	0.2456	0.2462	0.0017	0.0056	0.2461	0.0016	0.2461	0.0016	0.2475	0.0036			

Table 2 continued

Model	Setting	Parameter	True	$k = 5$				$k = 10$							
				ABC		SOS		ABC		SOS		ABC(sum)		SOS	
				Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
(ii)		$a$	1.2	1.4097	0.2276	1.4815	0.4231	1.3348	0.1646	1.3531	0.1756	1.4260	0.4240		
		$c$	0.4	0.3698	0.0338	0.4168	0.1255	0.3418	0.0616	0.3466	0.0568	0.3824	0.1307		
		$p$	1.7	1.7556	0.1512	1.4145	0.4345	1.8399	0.2225	1.8189	0.1915	1.5187	0.4737		
		$q$	2.5	3.2228	0.7494	3.1716	0.8801	3.0909	0.6281	3.0885	0.6222	3.1904	1.0233		
		$G$	0.3062	0.3058	0.0018	0.3078	0.0039	0.3057	0.0019	0.3057	0.0018	0.3066	0.0028		
(iii)		$a$	1.5	1.1748	0.3267	1.3904	0.2113	1.1168	0.3842	1.1197	0.3813	1.2950	0.3040		
		$c$	0.9	0.8484	0.0530	0.9062	0.0364	0.8250	0.0762	0.8240	0.0772	0.8950	0.0339		
		$p$	1.7	2.6511	0.9607	1.9257	0.4743	2.8834	1.1949	2.8755	1.1871	2.2284	0.7669		
		$q$	1.7	3.7431	2.0464	2.7211	1.1804	3.5490	1.8728	3.5753	1.8998	2.9870	1.5539		
		$G$	0.3589	0.3593	0.0023	0.3652	0.0099	0.3589	0.0022	0.3589	0.0022	0.3631	0.0066		
(iv)		$a$	1.2	1.4435	0.2658	1.3215	0.2574	1.4049	0.2448	1.4117	0.2532	1.3215	0.2574		
		$c$	0.1	0.3384	0.2389	0.2711	0.2122	0.3022	0.2027	0.3092	0.2095	0.2711	0.2122		
		$p$	1.3	1.4112	0.1812	1.2735	0.3131	1.4078	0.2036	1.4194	0.2139	1.2735	0.3131		
		$q$	3.5	3.1348	0.4202	3.3790	0.8128	3.1263	0.4480	3.1159	0.4599	3.3790	0.8128		
		$G$	0.3397	0.3390	0.0022	0.3396	0.0053	0.3390	0.0022	0.3390	0.0022	0.3396	0.0053		
(v)		$a$	1.5	1.1889	0.3176	1.2656	0.2895	1.2986	0.2260	1.2841	0.2367	1.3605	0.2692		
		$c$	0.99	0.9123	0.0792	0.8669	0.1348	0.9474	0.0462	0.9452	0.0483	0.9162	0.0866		
		$p$	1.2	2.0412	0.8583	1.6509	0.5614	1.7250	0.5649	1.7807	0.6166	1.5149	0.5439		
		$q$	3.0	3.7646	0.7763	3.3201	0.7662	3.6695	0.7539	3.6939	0.7594	3.2912	1.1308		
		$G$	0.4105	0.4098	0.0028	0.3968	0.0152	0.4105	0.0028	0.4105	0.0028	0.4035	0.0084		

ABC(sum) denotes the ABC method using the summary statistics

**Table 3** Posterior summary for the FIES data

Data	Parameter	GB		GB2		DA		SM	
		Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Quintile	<i>a</i>	1.9334	(1.2888, 2.9330)	2.0548	(1.3746, 3.1105)	4.2602	(4.0546, 4.4624)	3.6679	(3.4981, 3.8495)
	<i>c</i>	0.9686	(0.9308, 0.9974)						
	<i>p</i>	3.5721	(1.4269, 7.0048)	2.8919	(1.2766, 5.5099)	0.8210	(0.7371, 0.9264)	1.2568	(1.1043, 1.4290)
	<i>q</i>	2.9111	(1.2464, 5.6645)	3.4733	(1.5834, 6.5084)			0.2474	(0.2431, 0.2519)
	<i>G</i>	0.2459	(0.2414, 0.2504)	0.2464	(0.2420, 0.2509)	0.2482	(0.2432, 0.2538)	3.5924	(3.4697, 3.7289)
Decile	<i>a</i>	1.9594	(1.3073, 2.9622)	1.9800	(1.3841, 2.7915)	4.2962	(4.1005, 4.5123)		
	<i>c</i>	0.9762	(0.9470, 0.9982)						
	<i>p</i>	3.4502	(1.4621, 6.7507)	3.0384	(1.4702, 5.6317)	0.7938	(0.7141, 0.8853)	1.3302	(1.2005, 1.4797)
	<i>q</i>	3.0114	(1.3057, 5.6468)	3.5858	(1.9112, 6.3023)			0.2470	(0.2427, 0.2513)
	<i>G</i>	0.2461	(0.2418, 0.2503)	0.2465	(0.2420, 0.2508)	0.2486	(0.2434, 0.2544)		

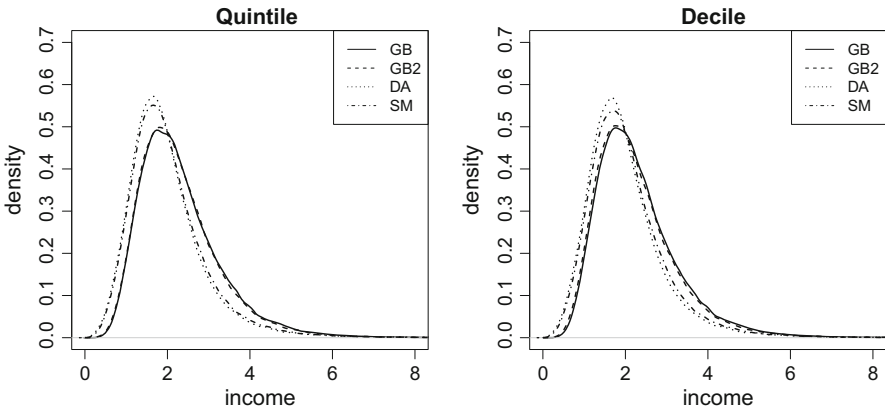


Fig. 8 Implied income distributions for the quintile and decile data

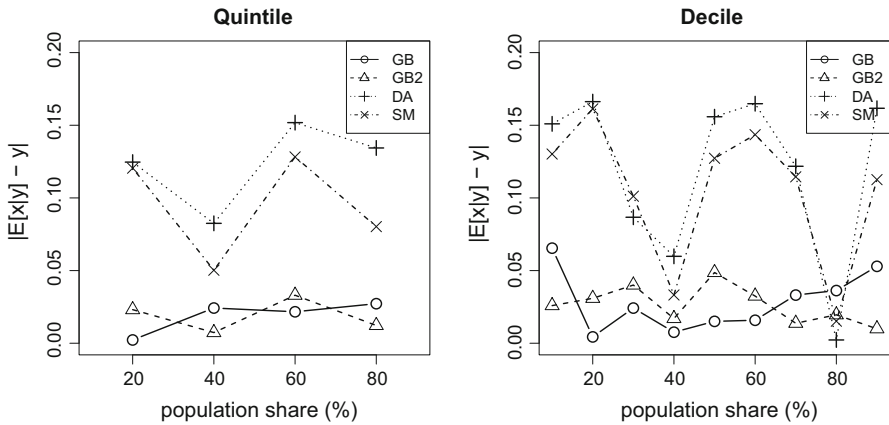
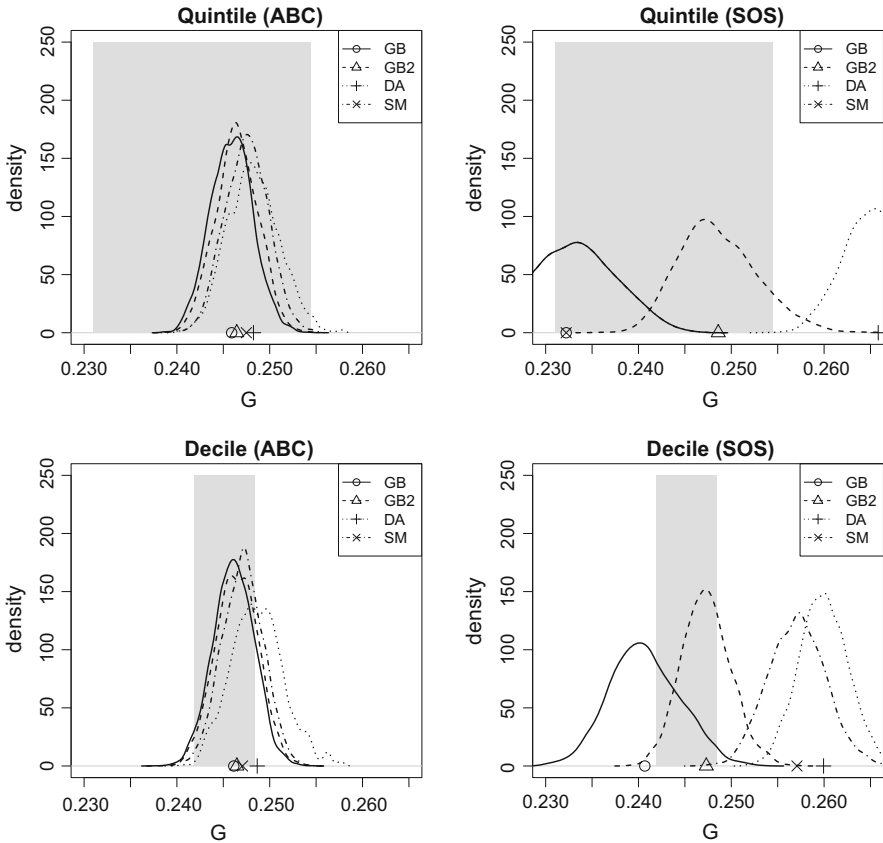


Fig. 9 Plots of  $|E[x_j|y] - y_j|$  for the quintile and decile data

likelihoods for GB, GB2, DA, and SM are, respectively,  $-3.971$ ,  $-2.110$ ,  $-8.037$ , and  $-5.675$  for the quintile data and  $-4.064$ ,  $-2.100$ ,  $-8.736$ , and  $-5.427$  for the decile data. Based on the marginal likelihoods, GB2 is supported the most by the both data followed by GB. This result is consistent with McDonald and Xu (1995) and is also in line with the argument made by Kleiber and Kotz (2003). The goodness of fit can be also checked through the simulating function by plotting the absolute difference between the posterior mean of the simulated income share  $x_j$  and the observed income share  $y_j$ ,  $|E[x_j|y] - y_j|$ , for  $j = 1, \dots, k - 1$  under each model. Figure 9 shows that the absolute differences under GB and GB2 are generally smaller than those under DA and SM for both quintile and decile data, also suggesting the use of a more flexible class of income distributions.

The posterior distributions of the Gini coefficient are compared with the nonparametric bounds of Gastwirth (1972), in which a Gini estimate should be included. The nonparametric bounds are given by  $(0.2310, 0.2545)$  and  $(0.2419, 0.2484)$  for the



**Fig. 10** Posterior distributions of the Gini coefficient using the proposed ABC approach and the MCMC approach using the level income for the quintile and decile data. The shaded area and symbols on the horizontal axis indicate the nonparametric bounds of Gastwirth (1972) and the posterior means of the Gini coefficient, respectively

quintile and decile data, respectively. This can be seen from Fig. 10, which presents the posterior distributions of the Gini coefficient. The shaded area in the figure represents the region inside the nonparametric bounds. The symbols on the horizontal axis represent the posterior means. For the quintile data, all models resulted the posterior distributions of the Gini coefficient which are fairly concentrated within the nonparametric bounds. The posterior probabilities that the Gini coefficient is included in the bounds are 1.000, 0.999, 0.984, and 0.998 for GB, GB2, DA, and SM, respectively. In the case of the decile data, the figure shows that the bodies of the posterior distributions under GB, GB2, and SM are included in the nonparametric bounds. Under DA, only the left half of the posterior distribution is include in the bounds and the posterior mean is outside the bounds. The posterior probabilities of the Gini coefficient included inside the bounds are 0.818, 0.776, 0.468, and 0.718 for GB, GB2, DA, and SM, respectively. While this result indicates the limitation that the posterior distribution obtained by using the proposed method does not shrink as fast as the nonparametric bounds, it



is consistent with the results of the simulation studies. Nonetheless, GB2 appears to be the most appropriate income model among the four in terms of goodness of fit and the Gini coefficient.

For comparison purpose, Fig. 10 also presents the posterior distributions of the Gini coefficient from SOS. For the quintile data, the posterior distributions appear to be more dispersed and scattered across regions. For the decile data, GB and GB2 produced the posterior distributions concentrated around the bounds with the posterior probabilities given by 0.335 and 0.656, respectively. Contrary, the posterior distributions under DA and SM are located away from the bounds. Therefore, the proposed ABC method also provides more reliable estimates of Gini coefficient in terms of the nonparametric bounds.

## 4 Discussion

We have proposed a new Bayesian approach to estimate the Gini coefficient assuming a hypothetical income distribution based on the grouped data on the Lorenz curve by using the ABC method via the SMC algorithm. From the simulation study, the proposed approach is found to perform comparably with or better than the existing methods. Our approach is found to be particularly valuable in the cases where the number of group is small as in quintile data. In the application to the Japanese data, the usefulness of the proposed approach assuming the class of GB distribution is illustrated by showing that the posterior distributions of the Gini coefficient are included within the nonparametric bounds with relatively high posterior probabilities and by presenting the income distributions implied from the hypothetical distributions. The numerical examples presented in this paper illuminated the limitation of the present study. Some parameters of the hypothetical distribution may not be identified when the number of parameters is large as in the cases of GB and GB2, because the information contained in grouped data is severely limited. Further, the posterior distribution of the Gini coefficient from the proposed approach does not shrink as fast as the nonparametric bounds as the number of income classes increases. Therefore, reconciling the goodness of fit and the accuracy of the Gini estimate when we have more groups in the data would be a direction for the future research.

## References

- Beaumont MA, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035
- Beaumont MA, Cornuet J-M, Marin J-M, Robert CP (2009) Adaptive approximate Bayesian computation. *Biometrika* 96:983–990
- Basmann RL, Hayes KJ, Slottje DJ, Johnson JD (1990) A general functional form for approximating the Lorenz curve. *J Econom* 43:77–90
- Bonassi FV, West M (2015) Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Anal* 10:171–187
- Chotikapanich D (1993) A comparison of alternative functional forms for the Lorenz curve. *Econ Lett* 41:129–138
- Chotikapanich D (ed) (2008) *Modeling income distributions and Lorenz curves*. Springer, New York

- Chotikapanich D, Griffiths WE (2002) Estimating Lorenz curves using a Dirichlet distribution. *J Bus Econ Stat* 20:290–295
- Chotikapanich D, Griffiths WE (2005) Averaging Lorenz curves. *J Income Inequal* 3:1–19
- Csilléry K, Blum M, Gaggiotti O, François O (2010) Approximate Bayesian computation (ABC) in practice. *Trends Ecol Evol* 25:410–418
- Dagum C (1977) A new model of personal income distribution: specification and estimation. *Écon Appl* 30:413–437
- David HA, Nagaraja HN (2003) *Order statistics*, 3rd edn. New York, Wiley
- Del Moral P, Doucet A, Jasra A (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat Comput* 22:1009–1020
- Didelot X, Everitt RG, Johansen AM, Lawson DJ (2011) Likelihood-free estimation of model evidence. *Bayesian Anal* 6:49–76
- Doornik JA (2013) *Ox™ 7: an object-oriented matrix programming language*. Timberlake, London
- Fearnhead P, Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with discussion). *J R Stat Soc Ser B* 74:419–474
- Filippi S, Barnes CP, Cornebise J, Stumpf MP (2013) On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat Appl Genet Mol Biol* 12:87–107
- Gastwirth JL (1972) The estimation of the Lorenz curve and Gini index. *Rev Econ Stat* 54:306–316
- Hajargasht G, Griffiths WE (2015) Inference for Lorenz curves. Technical report, Department of Economics, University of Melbourne. [https://fbe.unimelb.edu.au/\\_\\_data/assets/pdf\\_file/0006/1965867/2022HajargashtGriffiths.pdf](https://fbe.unimelb.edu.au/__data/assets/pdf_file/0006/1965867/2022HajargashtGriffiths.pdf). Accessed 21 Aug 2018
- Hasegawa H, Kozumi H (2003) Estimation of Lorenz curves: a Bayesian nonparametric approach. *J Econom* 115:277–291
- Iritani J, Kuga K (1983) Duality between the Lorenz curves and income distribution functions. *Econ Stud Q* 34:9–21
- Kakamu K (2016) Simulation studies comparing Dagum and Singh–Maddala income distributions. *Comput Econ* 48:593–605
- Kakamu K, Nishino H (2018) Bayesian estimation of beta-type distribution parameters based on grouped data. *Comput Econ*. <https://doi.org/10.1007/s10614-018-9843-4>
- Kakwani NC, Podder N (1973) On the estimation of Lorenz curves from grouped observations. *Int Econ Rev* 14:278–292
- Kakwani NC, Podder N (1976) Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations. *Econometrica* 44:137–148
- Kleiber C, Kotz S (2003) *Statistical size distributions in economics and actuarial science*. Wiley, New York
- Lenormand M, Jabot F, Deffuant G (2013) Adaptive approximate Bayesian computation for complex models. *Comput Stat* 28:2777–2796
- Lorenz MO (1905) Methods of measuring the concentration of wealth. *Publ Am Stat Assoc* 9:209–219
- Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci* 100:15324–15328
- Marin J-M, Pudlo P, Robert CP, Ryder R (2012) Approximate Bayesian computational methods. *Stat Comput* 22:1167–1180
- McDonald JB (1984) Some generalized functions for the size distribution of income. *Econometrica* 52:647–663
- McDonald JB, Ransom M (2008) The generalized beta distribution as a model for the distribution of income: estimation of related measures of inequality. In: Chotikapanich D (ed) *Modeling income distributions and Lorenz curves*. Springer, New York, pp 167–190
- McDonald JB, Xu YJ (1995) A generalization of the beta distribution with applications. *J Econom* 66:133–152
- McVinish R (2012) Improving ABC for quantile distributions. *Stat Comput* 22:1199–1207
- Nishino H, Kakamu K (2011) Grouped data estimation and testing of Gini coefficients using lognormal distributions. *Sankhya Ser B* 73:193–210
- Ortega P, Martín G, Fernández A, Ladoux M, García A (1991) A new functional form for estimating Lorenz curves. *Rev Income Wealth* 14:447–452
- Prangle D (2015) Summary statistics in approximate Bayesian computation. [arXiv:1512.05633](https://arxiv.org/abs/1512.05633)
- Rasche RH, Gaffney J, Koo AYC, Obst N (1980) Functional forms for estimating the Lorenz curve. *Econometrica* 48:1061–1062

- Ryu HK, Slottje DJ (1996) Two flexible functional form approaches for approximating the Lorenz curve. *J Econom* 72:251–274
- Ryu HK, Slottje DJ (1999) Parametric approximations of the Lorenz curve. In: Silber J (ed) *Handbook on income inequality measurement*. Kluwer, Boston, pp 291–312
- Sarabia JM (2008) Parametric Lorenz curves: models and applications. In: Chotikapanich D (ed) *Modeling income distributions and Lorenz curves*. Springer, New York, pp 167–190
- Sarabia JM, Castillo E, Slottje DJ (1999) An ordered family of Lorenz curves. *J Econom* 91:43–60
- Sarabia JM, Castillo E, Slottje DJ (2002) Lorenz ordering between McDonalds generalized functions of the income size distribution. *Econ Lett* 75:265–270
- Scott DW, Sain SR (2005) Multidimensional density estimation. In: Rao CR, Wegman EJ, Solka JL (eds) *Handbook of statistics*, vol 24. North-Holland, Amsterdam, pp 229–261
- Silk D, Filippi S, Stumpf MP (2013) Optimizing threshold-schedules for sequential Monte Carlo approximate Bayesian computation: applications to molecular systems. *Stat Appl Genet Mol Biol* 12:603–618
- Singh SK, Maddala GS (1976) A function for size distribution of income. *Econometrica* 47:1513–1525
- Sisson SA, Fan Y (2011) Likelihood-free Markov chain Monte Carlo. In: Brooks SP, Gelman A, Jones G, Meng XL (eds) *Handbook of Markov chain Monte Carlo*. Chapman and Hall, CRC Press, Boca Raton, pp 313–333
- Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci* 104:1760–1765
- Sisson SA, Fan Y, Tanaka MM (2009) Correction for Sisson et al, Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci* 106:16889
- Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* 6:187–202
- Villaseñor JA, Arnold BC (1989) Elliptical Lorenz curves. *J Econom* 40:327–338