BAYESIAN ESTIMATION OF BETA TYPE DISTRIBUTION PARAMETERS BASED UPON GROUPED DATA *

Kazuhiko Kakamu † Haruhisa Nishino

July 31, 2012

Abstract

This paper considers the estimation method of generalized beta distribution of the second kind (GB2 distribution) parameters based upon grouped data from a Bayesian point of view. As the GB2 distribution includes several kinds of familiar distributions of income such as Singh-Maddala distribution and so on, it is reasonable to consider the distribution. However, when the number of groups is small, it is sometimes difficult to estimate the parameters of the distribution utilizing the existing estimation methods such as maximum likelihood, minimum chi-square and so on. Thus, we propose a Markov chain Monte Carlo method to estimate the parameters of the distribution. The concept of the selected order statistics is utilized to construct the likelihood function. This method is applied to the Japanese quintile data from 1969 to 2007. Empirical results capture the mobility of income, which is consistent with the history of postwar Japan.

JEL classification: C11; C51; D31.

Key words: Generalized beta distribution of the second kind (GB2 distribution); Grouped data; Markov chain Monte Carlo method; Quintile data; Selected order statistics.

1 Introduction

To estimate the income distribution or income inequality is one of the important themes in considering economic policy for income redistribution. One way is to estimate it from individual or household data and the other way is to estimate it from grouped data. However, it is sometimes difficult to access to individual or household data in developing countries and even in developed countries. Moreover, the number of groups are small in some cases. For example, Family Income and Expenditure Survey in Japan is available data in Japan and the

*This work is partially supported by KAKENHI.

†Corresponding author. Faculty of Law and Economics, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba, 263-8522, Japan. *Email*: kakamu@le.chiba-u.ac.jp

quintile and decile data are announced. However, if we are interested in old data, only quintile data is available. Therefore, we are required to estimate the income distribution or income inequality from small grouped data such as quintile data.

The grouped data contains the information about interval and frequency of the groups as is shown in Table 1. Let x be the observed data and the intervals and the frequency are defined by x_i and n_i^* $(i = 1, \ldots, k)$, respectively. A raw dataset can be organized by constructing a table showing the frequency distribution of the variable (whose values are given in the raw dataset). Such a frequency table is often referred to as a grouped data. However, we can also regard a grouped data as a selected order statistics. Let a sample be $\{X(j); j = 1, \cdots, n\}$ with size n and subsample, $\{X(n_i); i = 1, 2, \ldots, k\}$ with size k, where the subsequence $\{n_i; i = 1, 2, \ldots, k\}$ is ancending order. Thus, we have selected order statistics $\{X(n_1) \leq X(n_2) \leq \ldots \leq X(n_k)\}$ for $(1 \leq n_1 \leq n_2 \leq \ldots \leq n_k \leq n)$. Let the observed values of selected order statistics be $(x_1, x_2, \ldots, x_k)$. Figure 1 shows the example of the quintile data, that is, $k = 4$, $\frac{n_1}{n} = \frac{n_2 - n_1}{n} = \frac{n_3 - n_2}{n} = \frac{n_4 - n_3}{n} = \frac{n - n_4}{n} = 0.2$. From these information, we can draw the histogram in the figure. Then, we can find that the definiton of x_i is same and $n_1^* = n_1$, $n_2^* = n_2 - n_1$, ..., $n_k^* = n_k - n_{k-1}$ and $n_{k+1}^* = n - n_k$. Moreover, it is required to estimate the true distribution (dotted line) in the figure assuming some distribution.

As is shown in McDonald and Xu (1995), the generalized beta distribution (hereafter referred to as GB) includes several kinds of standard size distributions as special or limiting cases. Even if we focus on the generalized beta distribution of the second kind (GB2), which is the special case of GB distribution, as is shown in Kleiber and Kotz (2003) and McDonald (1984), it contains several kinds of size distributions including such as the lognormal, generalized gamma and Singh-Maddala (1976) distributions, which are thought to be suitable to the real data in Japan (see Atoda *et al*. (1988), Nishino and Kakamu (2011), Tachibanaki *et al*. (1997) and so on). Therefore, it is reasonable to estimate the parameters of the GB2 distribution. However, it is difficult to estimate the parameters of GB2 distribution from quintile data.

Since the seminal work by McDonald (1984), GB2 distribution is applied to the real data. However, as far as we know, there is no research which examined the GB2 distribution from quintile data using the existing methods such as maximum likelihood, minimum chi-square and so on, because it is difficult to estimate four parameters from quintile data. Therefore, we propose a Bayesian method to estimate the parameters of the distribution. Our method is applied to the Japanese quintile data from 1969 to 2007. From the empirical results, we can capture the mobility of income distribution, which is consistent with the history of postwar Japan.

This paper is organized as follows. In the next section, we will introduce the features of GB2 distribution and the Bayesian Markov chain Monte Carlo estimation procedure. In Section 3, we will examine the Japanese data and state the empirical findings. In Section 4, we conclude the discussion and state the remaining issues.

2 The Generalized Beta Distribution of the Second Kind (GB2)

2.1 The Density and Cumulative Distribution Functions

While various probability distributions are used to estimate an income distribution, the generalized beta distribution of the second kind (GB2) includes the various probability distributions as the special or limiting cases.

The GB2 distribution has 4 parameters $(a,\ b,\ p,\ q)$ and its probability density function is written by

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

where $B(p,q)$ is a beta function. For example, if we set $p=1$, the distribution is reduced to the probability density function of the Singh-Maddala distribution, which is thought as a desirable distribution in many empirical applications.

To introduce the cumulative distribution function, let me introduce the following function

$$I_x(p,q)=\frac{B_x(p,q)}{B(p,q)},$$

where $B_x(p,q)$ is an incomplete beta function. Then, the cumulative distribution function is written by

$$F(x)=I_z(p,q),\ \text{where } z=\frac{\left(\frac{x}{b}\right)^{a}}{1+\left(\frac{x}{b}\right)^{a}}. \tag{2}$$

To explain the features of GB2 distribution, we introduce the mode and the moments. The mode of the GB2 distribution occurs at

$$x_{mode}=b\left(\frac{ap-1}{aq+1}\right)^{1/a},\ \text{if } ap>1$$

and at zero otherwise. The moments exist only for $-ap<k<ap$ with

$$E[X^k]=\frac{b^kB(p+k/a,q-k/a)}{B(p,q)}.$$

The more details for the GB2 distribution are well written in Kleiber and Kotz (2003).

2.2 The Likelihood Function

Let $\boldsymbol{\theta}=(a,\ b,\ p,\ q)$ be the vector of parameters and let $\mathbf{x}=(x_1,\ x_2,\ldots,\ x_k)$ be the vector of observations. Then, the likelihood function of the model is obtained from a joint distribution of order statistics. From David and Nagaraja (2003), we have a joint distribution of order statistics as follows:

$$L(\mathbf{x}|\boldsymbol{\theta})=n!\frac{F(x_1)^{n_1-1}}{(n_1-1)!}f(x_1)\left\{\prod_{i=2}^{k}\frac{\left(F(x_i-F(x_{i-1})^{n_i-n_{i-1}-1}\right)}{(n_i-n_{i-1}-1)!}f(x_i)\right\}\frac{(1-F(x_k))^{n-n_k}}{(n-n_k)!}. \tag{3}$$

If we substitute (1) and (2) for (3), it becomes the likelihood function for GB2. Nishino and Kakamu (2011) applied the likelihood to the lognormal distribution. If the probability density and cumulative distribution functions for the concerning distribution are available, we can apply this likelihood to the concerning distribution.

2.3 Posterior Analysis

Since we adopt a Bayesian approach, we complete the model by specifying the prior distribution over the parameters [1]. Therefore, we apply the following prior:

$$\pi(\boldsymbol{\theta}) = \pi(a)\pi(b)\pi(p)\pi(q)$$

Given a prior density $\pi(\boldsymbol{\theta})$ and the likelihood function in (3), the joint posterior distribution can be expressed as

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta})L(\mathbf{x}|\boldsymbol{\theta}) \tag{4}$$

Finally, we assume the following prior distributions:

$$a \sim \mathcal{IG}(\alpha_0, \beta_0),\ b \sim \mathcal{IG}(\gamma_0, \delta_0),\ p \sim \mathcal{IG}(\epsilon_0, \zeta_0),\ q \sim \mathcal{IG}(\eta_0, \nu_0),$$

where $\mathcal{IG}(a, b)$ denotes the inverse gamma distribution.

Since the full conditional distributions for each parameter are not standard forms, we adopt the random walk Metropolis-Hastings algorithm to each parameter (see Tierney (1994), Chib and Greenberg (1995) for details). We implement the following MCMC algorithm:

1. Initialize a, b, p, q.
2. Generate $a|b, p, q, \mathbf{x}$.
3. Generate $b|a, p, q, \mathbf{x}$.
4. Generate $p|a, b, q, \mathbf{x}$.
5. Generate $q|a, b, p, \mathbf{x}$.
6. Go to Step 2.

3 Empirical Results

Before examining empirics, we explain the data set, which is used in this section. We use the two quintile data from Family Income and Expenditure Survey in Japan from 1969 to 2007. One is the data of two-or-more person households and the other is that of workers' household. The sample size is $10,000$ households

[1] There are some Bayesian approach using grouped data such as Chotikapanich and Griffiths (2000, 2002) and our approach is similar to the one of Chotikapanich and Griffiths (2000). However, our approach is completely different from the one because our approach use the exact likelihood based upon the selected order statistics.

($n = 10,000$) and the sample is divided into five groups, that is, each group has $2,000$ households ($n_1 = 2,000$, $n_2 = 4,000$, $n_3 = 6,000$ and $n_4 = 8,000$).

We set the hyper-parameters as $\alpha_0 = \beta_0 = \gamma_0 = \delta_0 = \epsilon_0 = \zeta_0 = \eta_0 = \nu_0 = 1.0$. We perform the MCMC procedure by generating $1,000,000$ draws in a single sample path and discard the first $400,000$ draws as the initial burn-in. Out of the remaining draws, we keep every 100 draw to obtain the posterior statistics for the parameters. All computational results were obtained using Ox 6.21 (see Doornik (2006)).

Figure 2 shows the mobility of income distribution and the trends of mean and mode for two-or-more person households. From the figures, we can observe that the household income concentrated at the mode first. After that, the levels of mean and mode of the distribution moved to higher income level until 1991 and the level of mean increased higher than that of mode. In addition, the tail of higher income became fat at the same time. After 1991, the levels of mean and mode for the household income remained stagnant or decreased and the movements of them are parallel. It means that the household income increased until collapse of bubbles and higher income households decreased after collapse of bubbles.

Figure 3 displays the mobility of income distribution and the trends of mean and mode for workers' household. From the figures, we find that the moblity and trends are similar to those for two-or-more person households. However, the difference appears in the trends of mean and mode after 1991. The level of mean moved similar to that of two-or-more person households. On the other hand, the mode decreased more drastically than that for two-or-more person households, that is, the distribution of the income moved to lower income but some of the higher income households remain in higher income. It means that the effect of collapse of bubbles is smaller for higher income households for workers' household. Therefore, from thsese figures, we can conclude that the wealthy people without job damaged by the collapse of bubbles. This results are consistent with the history of postwar Japan.

4 Conclusions

This paper considered a Bayesian Markov chain Monte Carlo method to estimate the parameters of the GB2 distribution and applied to the Japanese quintile data from 1969 to 2007. Based upon the concept of the selected order statistics, the likelihood function was constructed and the posterior distribution was derived from the function and the prior distributions. From the empirical results, we captured the mobility of income in Japan. In the beginning, the distribution concentrated at the mode and the mode lay on lower income. Toward the collapse of bubbles in 1991, the mode moved to higher income and the height of the distribution around the mode became lower. The tail of higher income became fat, that is, the rich people increased at that time. After the collapse of bubbles, the mode moved to lower income and the height of the mode became higher. It means that the rich people decreased and the household income concentrated on mode. Especially, such phenomena

can be captured in two-or-more person households. This result is consistent with the history of postwar Japan.

We will state the remaining issues. First, we utilize a random walk Metropolis-Hastings algorithm. However, this algorithm is inefficient and requires a lot of random draws. Thus, we need more efficient algorithm to estimate the parameters of the distribution. Second, GB distribution, which has five parameters, includes more size distribution than GB2 distribution (see McDonald and Xu (1995)). Thus, it is important to extend the model to GB distribution. Finally, it is also important to calculate the Gini coefficient, because the Gini coefficient is sometimes used to formulate an income distribution policy. Therefore, the trend of Gini coefficient is another concern.

References

[1] Atoda, N., T. Suruga and T. Tachibanaki (1988) "Statistical inference of functional forms for income distribution," *The Economic Studies Quarterly*, **39**, 14–40.

[2] Chib, S. and E. Greenberg (1995) "Understanding the Metropolis-Hastings argorithm," *American Statistician*, **49**, 327–335.

[3] Chotikapanich, D. and W.E. Griffiths (2000) "Posterior distributions for the Gini coefficient using grouped data," *Australian and New Zealand Journal of Statistics*, **42**, 383–392.

[4] Chotikapanich, D. and W.E. Griffiths (2002) "Estimating Lorenz curves using a Dirichlet distribution," *Journal of Business & Economic Statistics*, **20**, 290–295.

[5] David, H.A. and H.N. Nagaraja (2003) *Order Statistics*, 3rd ed., Wiley: New York.

[6] Doornik, J.A. (2008) *Ox: An Object Oriented Matrix Programming Language*, Timberlake: London.

[7] Kleiber, C. and S. Kotz (2003) *Statistical Size Distributions in Economics and Actuarial Science*, Wiley: New York.

[8] McDonald, J.B. (1984) "Some generalized functions for the size distribution of income," *Econometrica*, **52**, 647–663.

[9] McDonald, J.B. and Y.J. Xu (1995) "A generalization of the beta distribution with applications," *Journal of Econometrics*, **66**, 133–152.

[10] Nishino, H. and K. Kakamu (2011) "Grouped data estimation and testing of Gini coefficients using log-normal distributions," *Sankhya Series B*, **73**, 193–210.

[11] Singh, S.K. and G.S. Maddala (1976) "A function for size distribution of income," *Econometrica*, **47**, 1513–1525.

[12] Tachibanaki, T., T. Suruga and N. Atoda (1997) "Estimations of income distribution parameters for individual observations by maximum likelihood method," *Journal of the Japan Statistical Society*, **27**, 191–203.

[13] Tierney L. (1994) "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, **22**, 1701–1762.

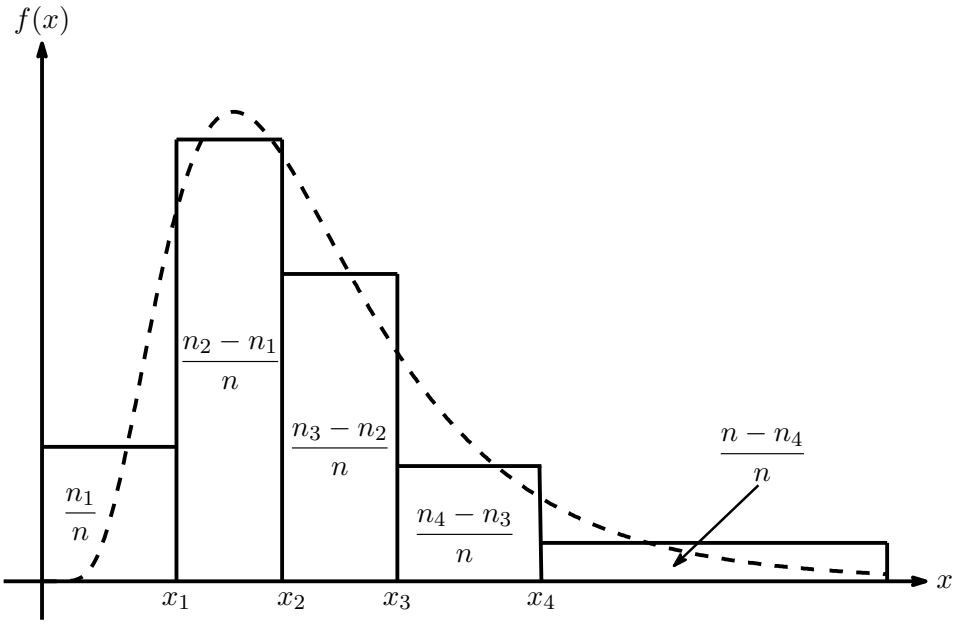

Figure 1: Quintile Data

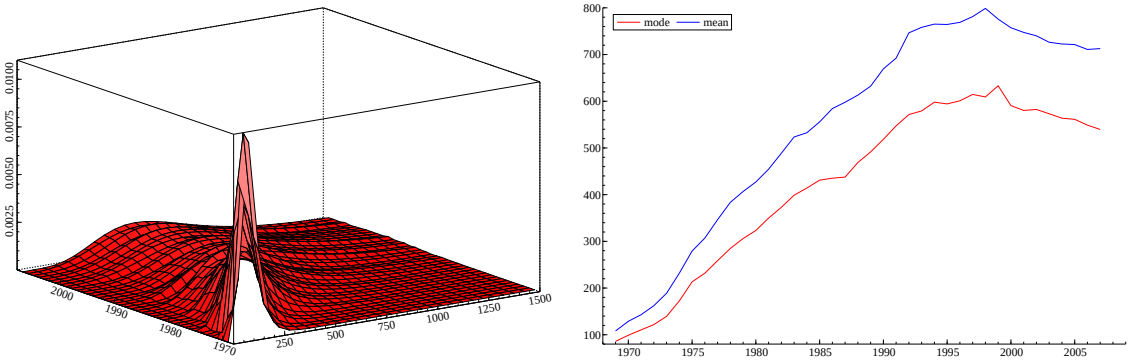

Figure 2: Mobility of Income Distribution for Two-or-More Person Households

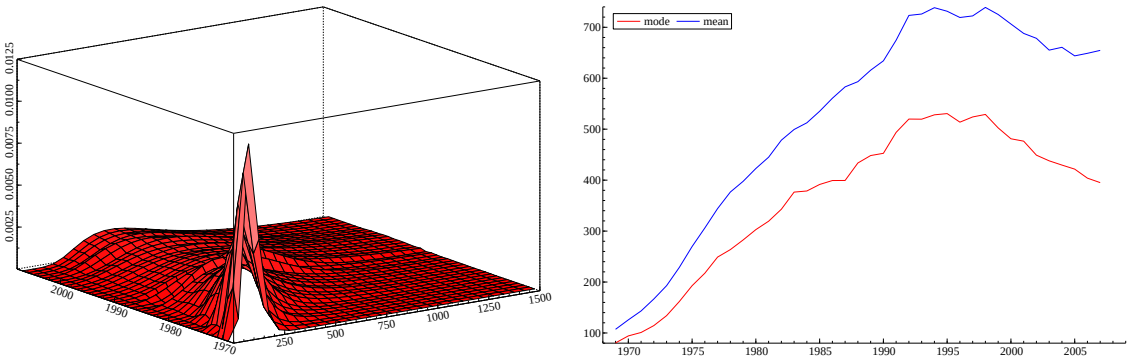

Figure 3: Mobility of Income Distribution for Workers' Households

range	frequency
$x \leq x_1$	n_1^*
$x_1 < x \leq x_2$	n_2^*
$\vdots$	$\vdots$
$x_{k-1} < x \leq x_k$	n_k^*
$x_k < x$	n_{k+1}^*

Table 1: Example of Grouped Data