THE UNIVERSITY OF MELBOURNE

Department of Economics

Working Paper Series

GMM Estimation of Mixtures from Grouped Data: An Application to Income Distributions

William E. Griffiths and Gholamreza Hajargasht

June 2012

Research Paper Number 1148

ISSN: 0819 2642
ISBN: 978 073 4044 983

Department of Economics
The University of Melbourne
Parkville VIC 3010
www.economics.unimelb.edu.au

GMM Estimation of Mixtures from Grouped Data: An Application to Income Distributions

William E. Griffiths and Gholamreza Hajargasht

Department of Economics
University of Melbourne, Australia

June 3, 2012

Corresponding author:
William Griffiths
Department of Economics
University of Melbourne
Vic 3010
Australia
Phone: +613 8344 3622
Fax: +613 8344 6899
Email: wegrif@unimelb.edu.au

GMM Estimation of Mixtures from Grouped Data: An Application to Income Distributions

ABSTRACT

We show how the generalized method of moments (GMM) framework developed in Hajargasht et al. (2012) for estimating income distributions from grouped data can be adapted for estimating mixtures. This approach can be used to estimate a mixture of any distributions where the moments and moment distribution functions of the mixture components can be expressed in terms of the parameters of those components. The required expressions for mixtures of lognormal and gamma densities are provided; in our empirical work we focus on estimation of mixtures of lognormal distributions. Two- and three-component lognormal mixtures are estimated for the income distributions of China rural, China urban, India rural, India urban, Pakistan, Russia, South Africa, Brazil and Indonesia. Their performance, in terms of goodness-of-fit and validity of moment conditions, is compared with that of a generalized beta (GB2) distribution. We find that the three-component lognormal mixture always outperforms the GB2 distribution, but the two-component mixture does not. For Brazil and Indonesia we have single observations, making it possible to compare maximum likelihood estimation of the mixtures from a complete set of single observations with GMM estimates obtained after grouping the data. Estimates from both procedures are found to be comparable, lending support to the usefulness of the GMM approach.

Keywords: Lognormal distribution, Generalized beta distribution, Inequality measures

1. Introduction

Data used for international comparisons of income, inequality and poverty are typically available in grouped form, comprising income and population shares or population shares coupled with class mean incomes. Although such data are collected from household surveys, they are aggregated by institutions such as the World Bank and the World Institute for Development Economics Research (WIDER) and published in grouped form that facilitates large scale investigations that involve many countries, different time periods, and the estimation of regional and global income distributions. See, for example, Milanovic (2002) and Chotikapanich et al. (2012). Building on the work by Chotikapanich et al. (2007, 2012), Hajarghast et al. (2012) develop a general GMM approach for estimating parametric income distributions from grouped data. Using eight example countries/regions, they illustrate how their methodology can be used to estimate the generalized beta distribution of the second kind (GB2), and some of its special cases, the lognormal, generalized gamma, beta-2, Singh-Maddala and Dagum distributions. While many of these distributions were good fits in the sense that they were able to accurately predict observed income shares, they can be criticized on the grounds that their parametric assumptions make them too inflexible, a point that was reinforced by test rejections of excess moment conditions for some of the distributions. Since mixtures of distributions are known to be able to approximate any distribution (see, for example, Ghosal and Van der Vaart, 2001), and can model fat tails and multi-modality (see, for example, Haas and Pigorsch, 2009), one way to overcome the inflexibility of parametric income distributions is to extend the grouped-data GMM methodology to mixtures of distributions. Income distribution studies which have estimated mixtures using data in the form of single observations include Bakker and Creedy (1999), Flachaire and Nunez (2007) and Lubrano and Ndoyye (2011), who used mixtures of lognormals, Chotikapanich and

Griffiths (2008) who used a mixture of gamma densities, and Hasegawa and Kozumi (2003) who used a lognormal mixture combined with a Dirichlet process prior.

The purpose of this paper is to show how the GMM framework developed in Hajargasht et al. (2012) for estimating single distributions from grouped data can be adapted for estimating mixtures The results we give in terms of moments and moment distribution functions can be used for a mixture of any distributions, but we focus particularly on lognormal and gamma densities and restrict the discussion of our empirical work to mixtures of lognormal distributions. Using data from China rural, China urban, India rural, India urban, Pakistan, Russia, South Africa, Brazil and Indonesia, we estimate 2- and 3-component lognormal mixtures and compare their performance, in terms of goodness-of-fit and validity of moment conditions, with that of a single generalized beta (GB2) distribution. We find that the 3-component lognormal mixture always outperforms the GB2 distribution, but the 2 component mixture does not. For Brazil and Indonesia, we have single observations, making it possible to compare maximum likelihood estimation of the mixtures from a complete set of single observations with GMM estimates obtained after grouping the data. Estimates from both procedures are found to be comparable, lending support to the usefulness of the GMM approach.

The paper is organized as follows. In Section 2 we briefly review the GMM methodology for estimating the parameters of a general income distribution developed in Hajargasht et al. (2012). Expressions needed for GMM estimation of mixtures are provided in Section 3, with the moments, distribution functions and first and second-moment distribution functions explicitly specified for mixtures with lognormal and gamma components. For completeness, we also include the quantities required for estimation of the GB2 distribution against which the lognormal mixtures are compared in the empirical work. Section 4 contains a description of the data used to illustrate the theoretical framework. The results presented in

Section 5 include parameter estimates and their standard errors, test results for excess moment conditions, mean-square-error comparisons for goodness-of-fit, and Gini coefficients. Some example plots are presented to illustrate both the precision of density estimation, and the high level of correspondence between maximum likelihood estimation using the complete set of single observations, GMM estimation from grouped data, and kernel density estimation. Concluding remarks are provided in Section 6.

2. The GMM Estimator

To describe the general form of the GMM estimator that can be used to estimate any income distribution, we begin with a sample of *T* observations $(y_1, y_2, \ldots, y_T)$, assumed to be randomly drawn from a parametric income distribution $f(y;\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a vector of unknown parameters. These observations have been grouped into *N* income classes (z_0, z_1), $(z_1, z_2), \ldots, (z_{N-1}, z_N)$, with $z_0 = 0$ and $z_N = \infty$. The available data are the mean class incomes $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N$ and the proportions of observations in each class $c_1, c_2, \ldots, c_N$. The estimation problem is to estimate $\boldsymbol{\phi}$, along with the unknown class limits $z_1, z_2, \ldots, z_{N-1}$, which are typically not published by the World bank and WIDER. To tackle this problem, Hajargasht et al. (2012) proposed a GMM estimator given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbf{H}(\boldsymbol{\theta})' \boldsymbol{\Omega}\, \mathbf{H}(\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta} = (z_1, z_2, \ldots, z_{N-1}, \boldsymbol{\phi}')'$,

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}(y_t, \boldsymbol{\theta}) \tag{2}$$

is a set of moments constructed for c_i and $\bar{y}_i$ such that the moment conditions $E\left[\mathbf{H}(\boldsymbol{\theta})\right] = \mathbf{0}$ are suitable for estimating $\boldsymbol{\theta}$, and $\boldsymbol{\Omega}$ is the weight matrix

$$\mathbf{\Omega} = \left[\operatorname{plim} \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(y_t, \boldsymbol{\theta}) \mathbf{h}(y_t, \boldsymbol{\theta})' \right]^{-1} \tag{3}$$

The first N elements in the $(2N \times 1)$ vector $\mathbf{h}(y_t, \boldsymbol{\theta})$ are

$$g_i(y_t) - k_i(\boldsymbol{\theta}) \qquad i = 1, 2, \ldots, N \tag{4}$$

where $g_i(y_t)$ is an indicator function such that

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

and

$$k_i(\boldsymbol{\theta}) = \int_{z_{i-1}}^{z_i} f(y; \boldsymbol{\phi}) dy = \int_{0}^{\infty} g_i(y) f(y; \boldsymbol{\phi}) dy = E\left[g_i(y) \right] \tag{5}$$

The second set of N elements in the $(2N \times 1)$ vector $\mathbf{h}(y_t, \boldsymbol{\theta})$ are

$$y_t g_i(y_t) - m_i(\boldsymbol{\theta}) \qquad i = 1, 2, \ldots, N \tag{6}$$

where

$$m_i(\boldsymbol{\theta}) = \int_{z_{i-1}}^{z_i} y f(y; \boldsymbol{\phi}) dy = \int_{0}^{\infty} y g_i(y) f(y; \boldsymbol{\phi}) dy = E\left[y g_i(y) \right] \tag{7}$$

With these definitions, we can write

$$\mathbf{H}(\boldsymbol{\theta}) \;=\; \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(y_t; \boldsymbol{\theta}) \;==\; \begin{bmatrix} \mathbf{c} - \mathbf{k} \\ \tilde{\mathbf{y}} - \mathbf{m} \end{bmatrix} \tag{8}$$

where $\mathbf{c}$, $\mathbf{k}$, and $\mathbf{m}$ are N-dimensional vectors containing the elements c_i, k_i and m_i, respectively, and, using $T_i = \sum_{t=1}^{T} g_i(y_t)$, the i-th element of $\tilde{\mathbf{y}}$ is given by

$$\tilde{y}_i = c_i \bar{y}_i = \frac{T_i}{T} \frac{1}{T_i} \sum_{t=1}^{T} y_t g_i(y_t) = \frac{1}{T} \sum_{t=1}^{T} y_t g_i(y_t) \tag{9}$$

Hajargasht et al. (2012) show that the weight matrix can be written as

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{D}(\boldsymbol{\omega}_1) & -\mathbf{D}(\boldsymbol{\omega}_3) \\ -\mathbf{D}(\boldsymbol{\omega}_3) & \mathbf{D}(\boldsymbol{\omega}_2) \end{bmatrix} \tag{10}$$

where $\mathbf{D}(\boldsymbol{\omega})$ denotes a diagonal matrix with elements of the vector $\boldsymbol{\omega}$ on the diagonal. The elements in the vectors $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_3$ are

$\omega_{1i} = m_i^{(2)}/v_i$, $\omega_{2i} = k_i/v_i$, and $\omega_{3i} = m_i/v_i$, where

$$m_i^{(2)}(\boldsymbol{\theta}) = \int_{z_{i-1}}^{z_i} y^2 f(y;\boldsymbol{\phi})dy = \int_0^{\infty} y^2 g_i(y) f(y;\boldsymbol{\phi})dy = E\left[y^2 g_i(y)\right] \tag{11}$$

and $v_i = k_i m_i^{(2)} - m_i^2$. Collecting all these various terms, the GMM objective function in (1) can be written as

$$\begin{aligned} GMM = \mathbf{H}'\boldsymbol{\Omega}\mathbf{H} &= \begin{bmatrix} \mathbf{c}-\mathbf{k} \\ \tilde{\mathbf{y}}-\mathbf{m} \end{bmatrix}' \begin{bmatrix} \mathbf{D}(\boldsymbol{\omega}_1) & -\mathbf{D}(\boldsymbol{\omega}_3) \\ -\mathbf{D}(\boldsymbol{\omega}_3) & \mathbf{D}(\boldsymbol{\omega}_2) \end{bmatrix} \begin{bmatrix} \mathbf{c}-\mathbf{k} \\ \tilde{\mathbf{y}}-\mathbf{m} \end{bmatrix} \\ &= \sum_{i=1}^{N} \omega_{1i}(c_i - k_i)^2 + \sum_{i=1}^{N} \omega_{2i}(\tilde{y}_i - m_i)^2 - 2\sum_{i=1}^{N} \omega_{3i}(c_i - k_i)(\tilde{y}_i - m_i) \end{aligned} \tag{12}$$

Equations (1) to (12) are a useful summary of the results in Hajargasht et al. (2012), and equation (12) is a computationally convenient expression for finding the GMM estimator for $\boldsymbol{\theta}$. However, the above results mask much of the development that led to the final result in (12). Because $\sum_{i=1}^{N} k_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} c_i = 1$, one of the moments in $\mathbf{H}(\boldsymbol{\theta})$ (see equation (8)) is redundant, and, unless we consider a generalized inverse, the inverse defined in (3) does not exist. Hajargasht et al. set up $(2N-1)$ non-redundant moment conditions, derived the corresponding matrix $\boldsymbol{\Omega}^{-1}$, and found its inverse $\boldsymbol{\Omega}$. They then showed that the relatively complicated objective function, expressed in terms of a $[(2N-1)\times 1]$ vector $\mathbf{H}$, and a $[(2N-1)\times(2N-1)]$ matrix $\boldsymbol{\Omega}$, can be written much more simply in terms of the $2N$-dimensional versions of $\mathbf{H}$ and $\boldsymbol{\Omega}$ given in equation (12). If K is the dimension of $\boldsymbol{\phi}$ (the number of unknown parameters in the income density), there are a total of $(N-1+K)$

unknown parameters. Given there are $(2N-1)$ non-redundant moment conditions, the number of excess moment conditions is $(N-K)$.

The quantities k_i, m_i and $m_i^{(2)}$ are all functions of the unknown parameters $\boldsymbol{\theta}$, and will depend on the assumed form of the income distribution. For most distributions it is convenient to compute these quantities by expressing them in terms of the distribution function and the first and second moment distribution functions of the assumed distribution. Specifically,

$$k_i(\boldsymbol{\theta}) = F^{(0)}(z_i;\boldsymbol{\phi}) - F^{(0)}(z_{i-1};\boldsymbol{\phi}) \tag{13}$$

$$m_i(\boldsymbol{\theta}) = \mu\left(F^{(1)}(z_i;\boldsymbol{\phi}) - F^{(1)}(z_{i-1};\boldsymbol{\phi})\right) \tag{14}$$

and

$$m_i^{(2)}(\boldsymbol{\theta}) = \mu^{(2)}\left(F^{(2)}(z_i;\boldsymbol{\phi}) - F^{(2)}(z_{i-1};\boldsymbol{\phi})\right) \tag{15}$$

where $\mu = E(y) = \int_0^\infty y f(y;\boldsymbol{\phi})dy$ and $\mu^{(2)} = E\left(y^2\right) = \int_0^\infty y^2 f(y;\boldsymbol{\phi})dy$ are the first and second moments for *y*, and $F^{(\ell)}(z_i;\boldsymbol{\phi})$, $\ell = 0,1,2$ denote the (moment) distribution functions for *y* evaluated at z_i. Adopting the conventions $\mu^{(0)} = 1$ and $\mu^{(1)} = \mu$, these functions are defined as

$$F^{(\ell)}(z_i;\boldsymbol{\phi}) = \frac{\int_0^{z_i} y^\ell f(y;\boldsymbol{\phi})dy}{\mu^{(\ell)}} \qquad \ell = 0,1,2 \tag{16}$$

Hajargasht et al. (2012) provide the required expressions for $\mu^{(\ell)}$ and $F^{(\ell)}(z_i;\boldsymbol{\phi})$ for the GB2, beta-2, generalized gamma, Singh-Maddala, Dagum and lognormal distributions. In the next section we show how the general expressions are modified to accommodate mixtures of distributions and provide specific information for mixtures of lognormal and gamma distributions. Before doing so, it is convenient to note three more things: our estimation strategy, the covariance matrix for the GMM estimator $\hat{\boldsymbol{\theta}}$, and the test statistic for testing the validity of excess moment conditions.

In our empirical work we employed an iterative two-step GMM estimator. In the first stage we find $\hat{\boldsymbol{\theta}}_1 = \arg\min_{\boldsymbol{\theta}} \mathbf{H}(\boldsymbol{\theta})' \boldsymbol{\Omega}_0 \mathbf{H}(\boldsymbol{\theta})$ where $\boldsymbol{\Omega}_0 = \mathbf{D}(c_1^{-2}, c_2^{-2}, \ldots, c_N^{-2}, \tilde{y}_1^{-2}, \tilde{y}_2^{-2}, \ldots, \tilde{y}_N^{-2})$. In the second stage we find $\hat{\boldsymbol{\theta}}_2 = \arg\min_{\boldsymbol{\theta}} \mathbf{H}(\boldsymbol{\theta})' \boldsymbol{\Omega}(\hat{\boldsymbol{\theta}}_1) \mathbf{H}(\boldsymbol{\theta})$, and then we iterate until convergence. The rationale behind using $\boldsymbol{\Omega}_0$ in the first step is that it leads to an estimator that minimizes the sum of squares of the percentage errors in the moment conditions.

The asymptotic covariance for $\hat{\boldsymbol{\theta}}$ can be shown to be

$$\text{var}(\hat{\boldsymbol{\theta}}) = \frac{1}{T}\left(\begin{bmatrix} \frac{\partial \mathbf{k}'}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{m}'}{\partial \boldsymbol{\theta}} \end{bmatrix} \begin{bmatrix} \mathbf{D}(\boldsymbol{\omega}_1) & -\mathbf{D}(\boldsymbol{\omega}_3) \\ -\mathbf{D}(\boldsymbol{\omega}_3) & \mathbf{D}(\boldsymbol{\omega}_2) \end{bmatrix} \begin{bmatrix} \partial \mathbf{k} / \partial \boldsymbol{\theta}' \\ \partial \mathbf{m} / \partial \boldsymbol{\theta}' \end{bmatrix} \right)^{-1} \tag{17}$$

This expression can be used to compute standard errors for the elements in $\hat{\boldsymbol{\theta}}$ and functions of them. In our empirical work we successfully used both analytical and numerical derivatives to compute (17).

Under the null hypothesis that the moment conditions are correct $\left(E\left[\boldsymbol{H}(\boldsymbol{\theta}) = \mathbf{0}\right]\right)$, the *J* statistic

$$J = T\, \mathbf{H}(\hat{\boldsymbol{\theta}})' \mathbf{W}(\hat{\boldsymbol{\theta}}) \mathbf{H}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{N-K} \tag{18}$$

where $N-K$ is the number of excess moment conditions. In traditional GMM estimation this test statistic is used to assess whether excess moment conditions are valid. In our case, since we assume a particular form of parametric income distribution, and use its properties to construct the moment conditions and weight matrix, the *J* statistic can be used to test the validity of the assumed income distribution.

3. Mixtures and their Moment Distributions

There are several approaches to the estimation of mixture models using a set of single observations. They include maximum likelihood and Bayesian estimation, use of the EM algorithm and method of moments. See, for example, Lindsay and Basak (1993), McLachlan

and Peel (2000), and Mengersen et al. (2011). Here we extend the grouped-data GMM framework described in the previous section to the estimation of mixture distributions and provide the necessary expressions for estimating mixtures of lognormal and gamma densities.

Consider a finite mixture distribution consisting of *n* components, with density and distribution functions for the *j*-th component given by $f_j(y;\boldsymbol{\phi}_j)$, and $F_j(y;\boldsymbol{\phi}_j)$, respectively. The parameter vector for the *j*-th component is $\boldsymbol{\phi}_j$. If the *j*-th component is sampled with probability w_j, then $\boldsymbol{\phi}' = (\boldsymbol{\phi}_1', \boldsymbol{\phi}_2', \ldots, \boldsymbol{\phi}_n', w_1, w_2, \ldots, w_{n-1})$ is the complete set of unknown parameters and the density and distribution functions for the mixture distribution can be written respectively as

$$f(y;\boldsymbol{\phi}) = \sum_{j=1}^{n} w_j f_j(y;\boldsymbol{\phi}_j) \tag{19}$$

and

$$F(y;\boldsymbol{\phi}) = \sum_{j=1}^{n} w_j F_j(y;\boldsymbol{\phi}_j) \tag{20}$$

For GMM estimation we need expressions for $k_i(\boldsymbol{\theta})$, $m_i(\boldsymbol{\theta})$ and $m_i^{(2)}(\boldsymbol{\theta})$. From (20), we have

$$k_i(\boldsymbol{\theta}) = F(z_i;\boldsymbol{\phi}) - F(z_{i-1};\boldsymbol{\phi}) = \sum_{j=1}^{n} w_j \left(F_j(z_i;\boldsymbol{\phi}_j) - F_j(z_{i-1};\boldsymbol{\phi}_j) \right) \tag{21}$$

For $m_i(\boldsymbol{\theta}) = m_i^{(1)}(\boldsymbol{\theta})$ and $m_i^{(2)}(\boldsymbol{\theta})$, we first note that, from (16),

$$\begin{aligned} \mu^{(\ell)} F^{(\ell)}(z_i;\boldsymbol{\phi}) &= \int_0^{z_i} y^{\ell} f(y;\boldsymbol{\phi}) dy \\ &= \int_0^{z_i} y^{\ell} \sum_{j=1}^{n} w_j f_j(y;\boldsymbol{\phi}_j)\, dy \\ &= \sum_{j=1}^{n} w_j \int_0^{z_i} y^{\ell} f_j(y;\boldsymbol{\phi}_j)\, dy \\ &= \sum_{j=1}^{n} w_j \mu_j^{(\ell)} F_j^{(\ell)}(z_i;\boldsymbol{\phi}_j) \qquad \ell = 1,2 \end{aligned} \tag{22}$$

where $\mu_j^{(\ell)} = \int_0^\infty y^\ell f_j(y;\boldsymbol{\phi}_j)\,dy$ is the ℓ-th moment of the j-th component, and $F_j^{(\ell)}(y;\boldsymbol{\phi}_j)$ is the ℓ-th moment distribution function of the j-th component. It then follows that

$$\begin{aligned} m_i^{(\ell)}(\boldsymbol{\theta}) &= \mu^{(\ell)}\left(F^{(\ell)}(z_i;\boldsymbol{\phi}) - F^{(\ell)}(z_{i-1};\boldsymbol{\phi})\right) \\ &= \sum_{j=1}^{n} w_j \mu_j^{(\ell)}\left(F_j^{(\ell)}(z_i;\boldsymbol{\phi}_j) - F_j^{(\ell)}(z_{i-1};\boldsymbol{\phi}_j)\right) \qquad \ell = 1,2 \end{aligned} \tag{23}$$

Given expressions for the means, distribution functions and moment distribution functions for each of the components, equations (21) and (23) can be used to set up the moment conditions and weight matrix for any mixture distribution. The single-distribution expressions given in Hajargasht et al. (2012) for the GB2, beta-2, Singh-Maddala, Dagum, generalized gamma and lognormal distributions can be used for each of the components in (21) and (23). Also, if it is considered desirable, different distributions could be used for different components. In this paper we focused on mixtures of lognormal distributions and mixtures of gamma densities. In the results section we restrict our discussion to the lognormal mixture because it provided a superior fit for all data sets.

Expressions for the density functions, moments and moment distribution functions for both lognormal and gamma components are given in Table 1. We also include those for the GB2 distribution which was estimated as a single distribution – not a mixture – and used as a basis for comparison. These results can be found in Kleiber and Kotz (2003), and, for the lognormal distribution, in Aitchison and Brown (1957). We use the following notation: $B(\cdot,\cdot)$ is the beta function, $\Gamma(\cdot)$ is the gamma function, $B_u(p,q) = \int_0^u t^{p-1}(1-t)^{q-1}dt \Big/ B(p,q)$ is the distribution function for a standard beta random variable defined on the (0,1) interval, $G_u(p,b) = \int_0^{u/b} t^{p-1}e^{-t}dt \Big/ \Gamma(p)$ is the distribution function for a standard gamma random variable with parameters *p* and *b*, and $\Phi(\cdot)$ is the distribution function for a standard normal random variable.

A quantity that is almost always of interest in income distribution studies is the Gini coefficient that is given by

$$G = -1 + \frac{2}{\mu}\int_0^\infty yF(y;\phi)\,f(y;\phi)dy \tag{24}$$

In the context of a mixture distribution it becomes

$$\begin{aligned} G &= -1 + \frac{2}{\mu}\int_0^\infty y\left(\sum_{j=1}^{n} w_j F_j(y;\phi_j)\right)\left(\sum_{i=1}^{n} w_i f_i(y;\phi_i)\right)dy \\ &= -1 + \frac{2}{\mu}\sum_{j=1}^{n}\sum_{i=1}^{n} w_j w_i \int_0^\infty y F_j(y;\phi_j)\,f_i(y;\phi_i)dy \end{aligned} \tag{25}$$

For the regions in our study and the mixture distributions, Gini coefficients were estimated using (25). For the GB2 distribution, equation (24) was used. In both cases the integrals were evaluated numerically.

4. Description of Data and Sources

We illustrate the methodology using grouped data from the "regions" China urban, China rural, India urban, India rural, Pakistan, Russia and South Africa, and single observation data (that we used in both single observation and grouped form) from Brazil and Indonesia. The grouped data were downloaded from the World Bank web site http://go.worldbank.org/WE8P1I8250. The version we used was updated in August 2008 to incorporate 2005 purchasing power parity estimated by the World Bank International Comparison Program. Population shares c_i and the corresponding income shares s_i were available for 20 groups for China urban, Pakistan, Russia and South Africa, for 17 groups for China rural, and for 12 groups for India rural and urban. Also available from the World Bank website is each region's mean monthly income $\bar{y}$, found from surveys and then converted using a 2005 purchasing-power-parity exchange rate. The class mean incomes $\bar{y}_i$ needed for

the methodology described in Sections 2 and 3, are found from $\overline{y}_i = s_i \overline{y}/c_i$; and then we can obtain $\tilde{y}_i$ from $\tilde{y}_i = c_i \overline{y}_i = s_i \overline{y}$.

Single observations on monthly per capita income for Brazil for the year 2009 were kindly provided by João Pedro Azevedo of the World Bank. It is a large sample of over 390,000 observations and includes over 8,000 zeros. To illustrate our methods, we randomly selected 100,000 of the nonzero observations. The single observations for Indonesia are monthly household expenditure data in urban areas for 2005, adjusted using household equivalence scales. They were kindly provided by Ari Handayani of Monash University. In this case we used 20,000 observations randomly selected from a larger sample of size 24,687. For using the Brazilian and Indonesian data in grouped form, we first sorted the data, then divided the sample into 20 groups of equal size ($c_i = 0.05$), and calculated $\tilde{y}_i$ as the sum of income for *i*-th group divided by the total number of observations.

Sample sizes T for each of the surveys from which the World Bank grouped data were computed are not available. For calculating standard errors, we used $T = 20,000$. This is a conservative value since most of surveys have sample sizes which are much larger. If standard errors for other sample sizes are of interest, they can be obtained from our results by multiplying by the appropriate scaling factor.

5. Empirical Analysis

The results from our empirical analysis are considered under a number of subsections. First, we examine the GMM estimates and standard errors for the parameters and class limits estimated using 3-component lognormal mixtures, restricted versions of 3-component lognormal mixtures, and single component GB2 distributions. In all cases we computed standard errors using numerical derivatives although it is straightforward to use the analytical derivatives provided in the appendix of Hajargasht et al. (2012). Second, we consider the

results of J tests for the validity of the excess moment conditions, comparing the GB2 distribution with 2- and 3-component lognormal mixtures. Then, a goodness-of-fit comparison of these distributions is made on the basis of their ability to predict the observed income shares for each group. We then compare GMM estimates from grouped data with those from maximum likelihood estimation from single observations, using the data from Brazil and Indonesia. Finally, we compare the Gini coefficient estimates and standard errors obtained under the alternative distributional assumptions. We chose not to report the results for the mixtures of gamma densities because, in all cases, the lognormal mixtures provided a better fit.

Estimates of the parameters and class limits

Both 2-component and 3-component lognormal mixtures were estimated using MATLAB. Generally, the "two-step iterative estimator" converged in less than 20 steps, and, at each step, the minimization algorithm converged in 50-100 iterations, except in the first step, where it took approximately 500 iterations. Estimates for the 3-component mixtures are reported in Table 2, with both "restricted" and "unrestricted" estimates reported for some of the regions (China rural and urban, India rural and urban, Brazil and South Africa). The restricted versions for China rural and Brazil are such that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, and for China urban, India rural and urban, and South Africa, we set $\sigma_2^2 = \sigma_3^2$. We chose to estimate these restricted versions in regions where standard errors for $\hat{\sigma}_j$ and $\hat{w}_j$ were relatively large, suggesting that it may be difficult to identify completely different components. Imposing the restrictions had the desired effect; it led to a dramatic reduction in the standard errors for estimates of all the parameters (σ_j, μ_j and w_j). In regions where we report unrestricted estimates only (Indonesia, Pakistan and Russia), the standard errors were already relatively small. Also, the standard errors for the class limits (z_i) were small, in both the restricted and unrestricted

estimations. As we see later, Gini coefficients are precisely estimated and confidence bands for density estimation are narrow, irrespective of whether we use restricted or unrestricted estimates of the coefficients. Estimates of 2-component mixtures are not reported to save space; they are available from the authors upon request.

Table 3 contains the estimates and standard errors for parameters and class limits for the GB2 distributions. In this case all standard errors are relatively small. Reassuringly, estimates of the class limits are almost identical to those estimated using log-normal mixtures. A potential problem is the non-existence of the second moment. For the existence of the *k*-th moment, the GB2 distribution requires $aq > k$. If, in the first step of estimation using the weighting matrix $\boldsymbol{\Omega}_0$, estimates are such that $\hat{a}\hat{q} < 2$, we are unable to proceed to the second step whose optimal weighting matrix requires the existence of second order moments. This problem occurs with Brazil and South Africa where there is a relatively high level of inequality. Hajargasht et al. (2012) encountered the same problem when using earlier data (2005) from Brazil. For Brazil, we overcame the problem by minimizing the objective function subject to the constraint $aq > 2$. For South Africa, where inequality is more extreme, and the product $\hat{a}\hat{q}$ was much less than that for Brazil, we were unable to get constrained estimation to converge. Thus, no GB2 estimates are reported for South Africa in Table 3. The same problem does not arise with a mixture of lognormals where the second moments always exist.

J tests

When sample sizes are very large, the probability of rejecting a false null hypothesis can be close to one, even when the difference between an actual parameter value and the value hypothesized under the null is so small as to be meaningless. In the context of testing excess moment conditions to assess the validity for a particular income distribution, this means that a particular income distribution can be rejected even when it fits the data well,

both visually and in terms of predicting income shares. Given these circumstances, and given that sample sizes are not available for the seven regions for which we have only grouped data, rather than report *J* statistics or *p*-values for some arbitrary sample size such as $T = 20,000$, we have chosen to report the smallest sample size at which the excess moment conditions would be rejected. Table 4 contains those sample sizes for both 0.05 and 0.01 significance levels. If we use the results of the *J* test as a comparative measure of goodness of fit, then the larger the sample size, the better the fit of the distribution. In this sense, the 3-component lognormal mixture is vastly superior to both its 2-component counterpart and the GB2 distribution for all regions except Brazil and South Africa. For South Africa, the GB2 was not estimated, and the 3-component mixture is only marginally better than that with 2 components. For Brazil, the 3-component mixture is only marginally better than both the 2-component mixture and the GB2 distribution. Moreover, all of the Brazilian distributions are a relatively poor fit. We tried 4 components for Brazil, which more than doubled its minimum sample sizes in Table 4, but even then its sample sizes were far smaller than those for all other regions.

Comparing the results from restricted and unrestricted estimation, we find that, when the standard errors are high, and it is hard to distinguish between components, restricting two or more of the variances to be equal does not have a large impact on the fit. Also, a comparison of the sample sizes for the GB2 and 2-component lognormal mixtures does not lead to any firm conclusions. The GB2 was better in approximately 50% of the cases. Finally, this discussion suggests that the *J* statistics, or the minimum sample sizes as we have presented them, could be used as a criterion for deciding on the number of components. If adding another component has little effect on the minimum sample size, this can be taken as evidence that the number of components is adequate.

Predicting income shares

Goodness-of-fit in terms of predicting income shares was carried out by comparing the observed income shares s_i with the predicted income shares derived from the estimated distributions. The income shares were predicted in the following way. Beginning with the original population shares c_i, and corresponding cumulative proportions $\pi_i = \sum_{j=1}^{i} c_j$, we found class limits z_i (not necessarily the same as the previously-estimated class limits) by solving the equations $F\left(z_i;\hat{\boldsymbol{\phi}}\right) = \pi_i$. Then, predicted cumulative income shares $\hat{\eta}_i$ were found from the first moment distribution function $\hat{\eta}_i = F^{(1)}\left(z_i;\hat{\boldsymbol{\phi}}\right)$, giving the predicted income shares $\hat{s}_i = \hat{\eta}_i - \hat{\eta}_{i-1}$.

Table 5 contains the percentage root-mean-squared errors (PRMSE), calculated from $\sqrt{N^{-1}\sum_{i=1}^{N}\left[100\left(\hat{s}_i - s_i\right)\right]^2}$; predictions from unrestricted estimation were used for regions where both restricted and unrestricted estimates of the 3-component lognormal mixture were computed. The most impressive thing to report about these values is that they are extremely small. For example, the PRMSE for China rural, using the 3-component lognormal mixture, is approximately 0.04%. That means that the average error (in the RMSE sense) from predicting the income shares for this region is 0.0004, a very accurate prediction indeed. One of the worst performers using this measure is for India rural, using the GB2 distribution, but even here the average error is only 0.0029. A comparison of the PRMSEs for the three distributions yields similar conclusions to those reached by examining the *J*-statistic sample sizes. The 3-component lognormal mixture is clearly the best. With the exception of China urban and Brazil, the 2-component lognormal mixture is not an improvement over the GB2 distribution.

The relatively poor performance of Brazil that was noted when considering the *J* statistics is also evident in Table 5. Initially, we thought it might be attributable to the high

level of inequality and the inability of the lognormal mixture to capture the heavy tail. However, the PRMSE prediction for the 3-component mixture for South Africa, a country with even higher inequality, is lower than that for all other regions with the exception of China rural. The problem with Brazil turned out to be a hard-to-predict spike in incomes just above 500, and lesser spikes at approximate multiples of this value, as illustrated in the finely divided histogram in Figure 1. The major spike led to a pronounced bump in the kernel density estimate – see Figure 2 – a bump that we failed to capture with up to 5 components in the lognormal mixture. João Pedro Azevedo of the World Bank suggested that the large spike corresponds to the minimum wage and the lesser spikes at higher incomes occur because wage contracts tend to be written in terms of multiples of the minimum wage.

Comparing GMM grouped data estimates with ML single observation estimates

For Brazil and Indonesia we have the complete sample and can therefore compare GMM estimates from grouped data with maximum likelihood (ML) estimates based on the complete sample. ML estimates were obtained using MATLAB's EM algorithm. Because this algorithm does not report standard errors, we computed standard errors using 500 bootstrapped samples. Convergence was very slow. It was common for the likelihood function to still be increasing after 10,000 iterations, despite there being no discernible change in the estimates at that stage. Convergence with GMM estimation was quicker and more decisive. To confirm the accuracy of the two sets of estimates, GMM estimates were used as ML starting values and vice versa. The results did not change.

A comparison of the ML and GMM estimates is presented in Table 6 for the 3-component lognormal mixture. The density functions from these estimates are graphed in Figures 2 and 3, along with nonparametric kernel density estimates. For Indonesia the parameter estimates and standard errors in Table 6 are of a similar order of magnitude. For Brazil there are some differences between the unrestricted GMM estimates and the ML

estimates, but these differences tend to disappear when the ML estimates are compared with the restricted GMM estimates. The GMM-estimated density in Figure 2 is based on unrestricted estimates, and is almost identical to that from ML estimation, despite the difference in parameters estimates. From the histogram in Figure 4, we note that the Indonesian data is much more regular than that from Brazil. As a consequence, the GMM, ML and kernel density plots in Figure 3 are almost visually indistinguishable.

In Figure 3 we present the density function along with 95% confidence bounds for Indonesia, estimated from the GMM grouped data estimates for the 3-component mixture. To find the confidence bounds, standard errors were computed for the estimated density at a number of income levels using the covariance matrix of the parameter estimates, the delta rule, and numerical derivatives. The narrowness of the confidence bounds suggests we are accurately estimating the density, despite relatively large confidence intervals for some of the parameters.

In general we can conclude that application of the GMM grouped data approach to mixtures is a promising procedure for estimating income distributions. It produces reliable estimates of the density, estimates that are comparable to ML estimation when single observations are available, and estimates that are comparable to kernel density estimation when no parametric assumptions about the income distribution are made.

Inequality measures

In Table 7 we report the Gini coefficients and their corresponding standard errors estimated from the unrestricted 3-component lognormal mixtures and the GB2 distributions, along with nonparametric estimates obtained from the complete data sets for Indonesia and Brazil. Standard errors for the latter estimates were obtained using the method developed in Davidson (2009). All estimates lie between 0.30 and 0.38 except for the high inequality countries of Brazil (0.55) and South Africa (0.67). The different distributions produce

estimates that are very similar, and, for Brazil and Indonesia, they are similar to the nonparametric estimates. There are some differences in the standard errors, although in most cases these are not dramatic.

6. Concluding Remarks

We have shown how to find GMM estimates of a mixture of distributions using grouped data, with emphasis on using a mixture of lognormals to estimate income distributions. We find that a 3-component lognormal mixture (a) outperforms the GB2 distribution estimated previously by Hajargasht et al. (2012), (b) fits the data extremely well in terms of its ability to predict income shares, (c) is equally well estimated by ML applied to single observations or GMM applied to grouped data, and (d) yields a density estimate almost identical to nonparametric kernel density estimation. When only grouped data are available, estimating a parametric distribution such as lognormal mixture has several advantages. Within-group inequality can be taken into account when measuring inequality and poverty, and the distributions can be conveniently combined to estimate income distributions for larger regions as well as globally, as in Chotikapanich et al. (2012).

Acknowledgments

This research was supported by Australian Research Council Grant DP1094632.

References

Aitchison, J., Brown, J. A. C. 1957. The lognormal distribution. Cambridge University Press.

Bakker, A., Creedy, J., 1999. Macroeconomic variables and income inequality in New Zealand: an exploration using conditional mixture distributions. New Zealand Economic Papers 33, 59-79.

Chotikapanich, D. (ed.), 2008. Modeling income distributions and lorenz curves. Springer.

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.

Chotikapanich, D., Griffiths, W. E., 2008. Estimating income distributions using a mixture of gamma densities. In Chotikapanich, D., editor, Modeling Income distributions and Lorenz Curves, 285–302. Springer.

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

Davidson, R., 2009. Reliable inference for the Gini index. Journal of Econometrics 150, 30–40.

Flachaire, E., Nunez, O., 2007. Estimation of the income distribution and detection of subpopulations: An explanatory model, Computational Statistics and Data Analysis 51, 3368–3380.

Ghosal, S., Van der Vaart, A. W., 2001. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Annals of Statistics 29, 1233–1263.

Haas, M., Pigorsch, C., 2009. Financial economics: fat-tailed distributions. In Encyclopedia of Complexity and Systems Science 4, 3404-3435. Springer-Verlag.

Hajargasht, G., Griffiths, W., Brice, J., Rao. D.S.P, Chotikapanich, D., 2012. Inference for income distributions using grouped data. University of Melbourne Department of Economics Working Paper No. 1140. [Minor revision under review at Journal of Business and Economic Statistics].

Hasegawa, H., Kozumi, H., 2003. Estimation of Lorenz curves: a Bayesian nonparametric approach. Journal of Econometrics 115, 277-291.

Kleiber, C., Kotz, S., 2003. Statistical size distributions in economics and actuarial sciences. John Wiley.

Lindsay, B. G., Basak, P., 1993. Multivariate normal mixtures: A fast consistent method of moments. Journal of the American Statistical Association 88, 468-476.

Lubrano, M., Ndoye, A.A.J., 2011. Inequality decomposition using the Gibbs output of a mixture of lognormal distributions. Universités d'Aix-Marseille GREQAM Working Paper No. 2011-19.

McLachlan, G.J., Peel, D., 2000. Finite mixture models. John Wiley.

Mengersen, K.L., Robert, C.P., Titterington, D.M., (editors), 2011. Mixtures: estimation and applications. John Wiley.

Milanovic, B., 2002. True world income distribution, 1988 and 1993: first calculations based on household surveys alone. The Economic Journal 112, 51-92.

Table 1 Moments and Distribution Functions for Mixture Components and the GB2 Distribution

	Density Function	Moments	Distribution Function	Moment Distribution Functions
Lognormal (j-th component in mixture)	$f_j\left(y;\boldsymbol{\phi}_j\right) = \frac{1}{y\sqrt{2\pi}\sigma_j}\exp\left(-\frac{(\ln y-\beta_j)^2}{2\sigma_j^2}\right)$	$\mu_j^{(\ell)} = \exp\left(\ell\beta_j+\frac{\ell^2\sigma_j^2}{2}\right)$	$F_j\left(y;\boldsymbol{\phi}_j\right) = \Phi\left(\frac{\ln(y)-\beta_j}{\sigma_j}\right)$	$F_j^{(\ell)}\left(y;\boldsymbol{\phi}_j\right) = \Phi\left(\frac{\ln(y)-\beta_j-\ell\sigma_j^2}{\sigma_j}\right)$
Gamma (j-th component in mixture)	$f_j(y;\boldsymbol{\phi}_j) = \frac{y^{p_j-1}}{\beta_j^{p_j}\Gamma(p_j)}\exp\left(-\left(\frac{y}{\beta_j}\right)\right)$	$\mu_j^{(\ell)} = \frac{\beta_j^{\ell}\Gamma\left(p_j+\ell\right)}{\Gamma(p_j)}$	$F_j(y;\boldsymbol{\phi}_j) = G_y\left(p_j,\beta_j\right)$	$F_j^{(\ell)}(y;\boldsymbol{\phi}_j) = G_y\left(p_j+\ell_j,\beta_j\right)$
GB2	$f(y;\boldsymbol{\phi}) = \frac{ay^{ap-1}}{b^{ap}B\left(p,q\right)\left(1+\left(y/b\right)^a\right)^{p+q}}$	$\mu^{(\ell)} = \frac{b^{\ell}B\left(p+\ell/a\,,q-\ell/a\right)}{B\left(p,q\right)}$	$F(y;\boldsymbol{\phi}) = B_u(p,q)$ with $u = \left(y/b\right)^a\Big/\left[1+\left(y/b\right)^a\right]$	$F^{(\ell)}(y;\boldsymbol{\phi}) = B_u\left(p+\ell/a\,,q-\ell/a\right)$ with $u = \left(y/b\right)^a\Big/\left[1+\left(y/b\right)^a\right]$

Table 2. Estimated Coefficients from 3-Component Lognormal Distributions

	China Rural				China Urban				Brazil				Indonesia		Pakistan	
	Unrestricted		Restricted		Unrestricted		Restricted		Unrestricted		Restricted					
	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE
z1	11.254	0.193	11.230	0.177	52.310	0.266	52.349	0.217	79.652	0.464	79.540	0.302	147.630	0.494	26.877	0.072
z2	14.000	0.102	14.003	0.100	64.995	0.138	64.985	0.132	118.490	0.204	118.510	0.191	175.390	0.304	31.088	0.044
z3	16.733	0.081	16.731	0.081	75.503	0.111	75.505	0.111	149.960	0.168	149.960	0.167	198.350	0.264	34.370	0.036
z4	22.360	0.066	22.361	0.066	84.598	0.102	84.598	0.102	179.190	0.166	179.190	0.166	219.680	0.252	37.213	0.032
z5	27.898	0.055	27.898	0.055	93.065	0.099	93.065	0.099	212.630	0.171	212.630	0.171	240.460	0.255	39.842	0.031
z6	33.473	0.043	33.473	0.043	101.340	0.099	101.340	0.099	245.280	0.174	245.280	0.174	262.520	0.262	42.358	0.030
z7	36.263	0.043	36.264	0.043	109.860	0.102	109.860	0.102	278.280	0.183	278.280	0.183	284.430	0.271	44.775	0.030
z8	41.828	0.051	41.828	0.051	118.610	0.106	118.610	0.106	313.720	0.199	313.720	0.199	307.100	0.290	47.336	0.031
z9	47.403	0.059	47.403	0.059	127.650	0.112	127.650	0.111	353.780	0.222	353.780	0.222	332.500	0.314	49.961	0.033
z10	55.743	0.078	55.742	0.078	137.290	0.119	137.290	0.119	398.130	0.251	398.130	0.252	359.730	0.342	52.808	0.036
z11	69.715	0.106	69.716	0.106	147.350	0.129	147.350	0.129	452.040	0.280	452.050	0.280	389.960	0.376	55.909	0.039
z12	83.580	0.129	83.581	0.129	158.510	0.144	158.510	0.144	501.840	0.303	501.850	0.304	422.790	0.417	59.366	0.045
z13	97.633	0.156	97.633	0.156	171.010	0.164	171.010	0.164	556.200	0.361	556.200	0.362	459.500	0.475	63.540	0.054
z14	111.410	0.190	111.410	0.189	185.260	0.192	185.260	0.192	634.380	0.470	634.380	0.470	502.670	0.557	68.579	0.067
z15	125.680	0.243	125.670	0.241	202.680	0.234	202.680	0.234	736.940	0.614	736.940	0.615	552.970	0.675	74.804	0.085
z16	139.600	0.452	139.640	0.430	223.310	0.297	223.310	0.297	867.100	0.848	867.110	0.848	615.900	0.880	82.810	0.109
z17					250.940	0.428	250.920	0.422	1067.800	1.353	1067.800	1.345	704.140	1.288	93.300	0.151
z18					294.220	0.854	294.340	0.735	1412.000	2.747	1412.100	2.602	843.760	2.376	108.530	0.286
z19					381.610	5.500	380.350	2.303	2241.200	11.666	2239.000	8.322	1153.000	8.467	143.980	1.144
mu1	3.576	1.001	3.768	0.050	3.738	0.279	3.867	0.367	5.6340	0.1512	3.7599	0.2625	5.399	0.059	4.341	0.023
sig1	0.418	0.379	0.464	0.014	0.282	0.256	0.365	0.185	1.7563	0.1284	0.8199	0.0210	0.356	0.021	0.716	0.015
mu2	4.672	0.760	5.482	0.101	6.573	0.449	6.274	0.093	5.9629	0.0049	5.9262	0.0186	6.051	0.080	4.558	0.024
sig2	0.743	0.058			0.337	0.657	0.547	0.023	0.8331	0.0122			0.442	0.036	0.176	0.031
mu3	4.087	1.440	4.375	0.069	4.869	0.012	4.872	0.026	7.8326	0.1982	7.6241	0.0824	6.422	0.050	3.865	0.008
sig3	0.502	0.535			0.560	0.019			0.2965	0.2716			0.823	0.021	0.342	0.007
w1	0.216	2.374	0.593	0.076	0.014	0.018	0.026	0.039	0.1060	0.0253	0.0314	0.0097	0.294	0.093	0.258	0.019
w2	0.136	0.286	0.038	0.011	0.018	0.009	0.028	0.008	0.8634	0.0246	0.8892	0.0177	0.508	0.094	0.058	0.014

Table 2 (continued) Estimated Coefficients from 3-Component Lognormal Distributions

	Russia		India Rural				India Urban				South Africa			
			Unrestricted		Restricted		Unrestricted		Restricted		Unrestricted		Restricted	
	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE
z1	80.213	0.337	20.919	0.080	20.883	0.053	19.960	0.081	19.946	0.058	24.126	0.105	24.127	0.104
z2	102.410	0.233	24.177	0.041	24.186	0.038	23.486	0.047	23.489	0.044	31.460	0.081	31.460	0.081
z3	119.980	0.198	28.550	0.035	28.549	0.035	28.859	0.045	28.858	0.045	37.760	0.074	37.760	0.074
z4	135.570	0.185	32.603	0.033	32.604	0.033	34.401	0.047	34.401	0.047	43.957	0.073	43.957	0.073
z5	150.850	0.182	36.623	0.034	36.623	0.034	40.029	0.053	40.029	0.053	49.796	0.074	49.796	0.074
z6	166.180	0.183	40.613	0.038	40.613	0.038	46.812	0.064	46.812	0.064	56.142	0.080	56.142	0.080
z7	181.390	0.189	45.491	0.047	45.491	0.047	55.321	0.079	55.321	0.079	63.351	0.090	63.351	0.090
z8	197.570	0.200	51.835	0.064	51.834	0.064	65.513	0.105	65.512	0.105	71.520	0.101	71.520	0.101
z9	214.720	0.215	61.534	0.104	61.542	0.104	81.492	0.171	81.505	0.171	80.597	0.112	80.597	0.112
z10	233.530	0.236	79.385	0.231	79.345	0.233	111.750	0.364	111.680	0.356	89.922	0.130	89.921	0.130
z11	254.210	0.261	103.440	0.852	103.770	0.883	151.200	1.024	151.640	0.982	102.090	0.162	102.090	0.162
z12	277.110	0.292									117.270	0.206	117.270	0.206
z13	303.180	0.334									136.820	0.278	136.820	0.278
z14	333.010	0.390									165.340	0.407	165.340	0.407
z15	368.150	0.476									207.370	0.631	207.370	0.630
z16	414.010	0.624									275.830	1.060	275.850	1.051
z17	475.590	0.881									392.910	2.014	392.810	1.872
z18	565.700	1.462									615.180	5.348	615.730	3.968
z19	751.050	3.709									1130.200	29.336	1126.600	13.006
mu1	5.211	0.109	3.481	0.275	3.502	0.036	3.341	0.031	3.358	0.022	4.156	0.191	4.172	0.041
sig1	0.414	0.063	0.311	0.252	0.287	0.031	0.285	0.101	0.288	0.023	0.668	0.073	0.672	0.015
mu2	5.521	0.014	4.247	0.373	5.145	0.067	4.432	0.773	5.313	0.056	5.670	0.444	6.979	0.203
sig2	0.738	0.013	0.787	0.034	0.446	0.015	0.820	0.143	0.550	0.007	0.924	2.486	0.740	0.078
mu3	5.861	0.152	3.833	1.518	3.778	0.035	3.904	0.167	3.933	0.020	7.294	2.768	5.616	0.254
sig3	0.287	0.103	0.395	0.444			0.518	0.121			0.650	1.084		
w1	0.221	0.070	0.366	2.468	0.248	0.101	0.155	0.195	0.155	0.025	0.677	0.524	0.709	0.046
w2	0.726	0.059	0.127	0.114	0.036	0.007	0.163	0.225	0.043	0.006	0.276	0.953	0.093	0.023

Table 3. Estimated Coefficients from GB2 Distributions

	China Rural		China Urban		India Rural		India Urban	
Par	Par	SE	Par	SE	Par	SE	Par	SE
z1	11.255	0.250	52.589	0.257	20.897	0.075	19.871	0.079
z2	14.001	0.143	64.931	0.181	24.183	0.054	23.503	0.064
z3	16.732	0.116	75.517	0.155	28.550	0.049	28.857	0.065
z4	22.361	0.094	84.595	0.143	32.603	0.047	34.404	0.066
z5	27.898	0.078	93.066	0.138	36.623	0.048	40.028	0.073
z6	33.473	0.061	101.342	0.139	40.613	0.053	46.811	0.088
z7	36.264	0.060	109.857	0.143	45.490	0.066	55.319	0.111
z8	41.828	0.072	118.610	0.148	51.836	0.091	65.511	0.151
z9	47.403	0.082	127.647	0.157	61.518	0.149	81.477	0.249
z10	55.742	0.110	137.291	0.169	79.464	0.297	111.857	0.499
z11	69.715	0.151	147.345	0.184	103.036	0.708	150.739	1.157
z12	83.581	0.185	158.507	0.206				
z13	97.633	0.224	171.005	0.235				
z14	111.411	0.269	185.265	0.277				
z15	125.671	0.333	202.684	0.337				
z16	139.656	0.531	223.312	0.428				
z17			250.918	0.603				
z18			294.324	1.014				
z19			380.707	2.601				
b	22.704	4.771	107.345	4.964	28.454	1.094	2.296	5.598
p	7.084	2.261	2.429	0.406	2.046	0.308	59.286	131.390
q	2.366	0.427	1.807	0.259	0.814	0.083	2.902	0.797
a	1.341	0.161	1.820	0.168	3.404	0.262	1.039	0.204
	Pakistan		**Russia**		**Brazil**		**Indonesia**	
Par	Par	SE	Par	SE	Par	SE	Par	SE
z1	26.814	0.098	80.077	0.462	78.286	0.238	147.220	0.436
z2	31.099	0.063	102.434	0.328	118.390	0.187	175.070	0.304
z3	34.368	0.051	119.970	0.282	149.970	0.168	197.930	0.266
z4	37.213	0.046	135.573	0.265	178.740	0.168	219.010	0.250
z5	39.842	0.044	150.848	0.261	212.290	0.174	239.060	0.252
z6	42.358	0.042	166.179	0.262	244.570	0.177	261.190	0.259
z7	44.775	0.043	181.390	0.269	277.920	0.186	283.560	0.268
z8	47.336	0.044	197.571	0.284	313.210	0.202	306.550	0.286
z9	49.961	0.047	214.716	0.304	353.310	0.223	332.770	0.309
z10	52.808	0.051	233.525	0.331	397.270	0.251	360.590	0.334
z11	55.910	0.056	254.210	0.365	449.690	0.280	391.060	0.370
z12	59.366	0.064	277.109	0.410	500.730	0.305	423.500	0.414
z13	63.541	0.075	303.176	0.470	556.110	0.361	459.410	0.479
z14	68.580	0.092	333.013	0.550	633.230	0.463	502.810	0.573
z15	74.803	0.116	368.141	0.675	735.570	0.604	553.420	0.702
z16	82.818	0.154	414.018	0.883	864.990	0.828	616.700	0.919
z17	93.235	0.219	475.546	1.242	1066.000	1.309	705.550	1.326
z18	108.808	0.371	566.108	2.042	1410.900	2.530	845.230	2.284
z19	141.959	1.021	747.929	5.182	2210.000	8.690	1142.100	6.245
b	39.152	1.135	175.292	19.054	393.350	6.240	62.374	32.855
p	1.555	0.204	4.738	1.334	1.454	0.068	17.099	9.257
q	0.689	0.067	3.525	0.834	1.441	0.074	2.639	0.466
a	3.729	0.277	1.103	0.156	1.388	0.043	1.136	0.139

Table 4. Smallest Sample Sizes that Lead to Significant *J*-Statistics

J stat α=.05				
	3 Mixture Unrestricted	**3 Mixture Restricted**	**2 Mixture**	**GB2**
China R	102738	94607	21601	38543
China U	45660	49912	28874	12980
India R	85770	72777	8890	19417
India U	41389	38013	2762	13268
Pakistan	33178		4230	13606
Russia	29493		22843	20621
Brazil	3606	3581	2610	2447
Indonesia	47330		21601	12079
Sth Africa	12020	12755	10744	

J stat α=.01				
	3 Mixture Unrestricted	**3 Mixture Restricted**	**2 Mixture**	**GB2**
China R	128093	116371	26750	46896
China U	56929	61394	35757	15794
India R	106937	89520	11010	23625
India U	51604	46759	3421	16143
Pakistan	41366		5238	16555
Russia	36772		28288	25090
Brazil	4496	4963	3437	3724
Indonesia	59010		26750	14697
Sth Africa	14987	15793	13144	

Table 5. Root-Mean-Square Errors for Income Share Predictions

Mean Square Errors			
	3 Mixture	**2 Mixture**	**GB2**
China R	0.039883	0.14102	0.0798
China U	0.066146	0.11421	0.1409
India R	0.069353	0.33447	0.2921
India U	0.10431	0.51482	0.1192
Pakistan	0.048279	0.35816	0.1336
Russia	0.087484	0.11097	0.1022
Poland	0.065097	0.24293	0.1145
Brazil	0.18522	0.28321	0.1976
Indonesia	0.054414	0.14102	0.21402
Sth Africa	0.041783	0.13272	

Table 6. Comparison of GMM and ML Estimates

	Brazil						**Indonesia**			
	GMM-Unrestricted		**GMM-Restricted**		**ML**		**GMM**		**ML**	
	Par	SE	Par	SE	Par	SE	Par	SE	Par	SE
Mu1	5.6340	0.1512	3.7599	0.2625	4.1126	0.1161	5.3994	0.0588	5.3745	0.0444
Sig1	1.7563	0.1284	0.8199	0.0210	0.7939	0.0384	0.3562	0.0210	0.3424	0.0272
Mu2	5.9629	0.0049	5.9262	0.0186	5.9083	0.0142	6.0506	0.0804	5.9820	0.0669
Sig2	0.8331	0.0122	0.8199	0.0210	0.7783	0.0112	0.4417	0.0361	0.5248	0.0307
Mu3	7.8326	0.1982	7.6241	0.0824	7.3001	0.1162	6.4224	0.0498	6.5343	0.0905
Sig3	0.2965	0.2716	0.8199	0.0210	0.8904	0.0337	0.8231	0.0209	0.8593	0.0257
W1	0.1060	0.0253	0.0314	0.0097	0.0512	0.0082	0.2942	0.0931	0.1957	0.0798
W2	0.8634	0.0246	0.8892	0.0177	0.8207	0.0233	0.5083	0.0936	0.6836	0.0898

Table 7. Gini Coefficients

Gini Coefficient						
	Mixture	**SE**	**GB2**	**SE**	**Davidson**	**SE**
China R	0.3587	0.0046	0.3572	0.0038		
China U	0.3476	0.0040	0.3452	0.0046		
India R	0.3043	0.0045	0.2982	0.0042		
India U	0.3758	0.0052	0.3743	0.0045		
Pakistan	0.3117	0.0040	0.3153	0.0056		
Russia	0.3765	0.0039	0.3777	0.0036		
Brazil	0.5468	0.0043	0.5468	0.0024	0.5456	0.0019
Indonesia	0.3792	0.0054	0.3753	0.0033	0.3794	0.0038
Sth Africa	0.6695	0.0129				

Figure 1 Histogram of Observations for Brazil

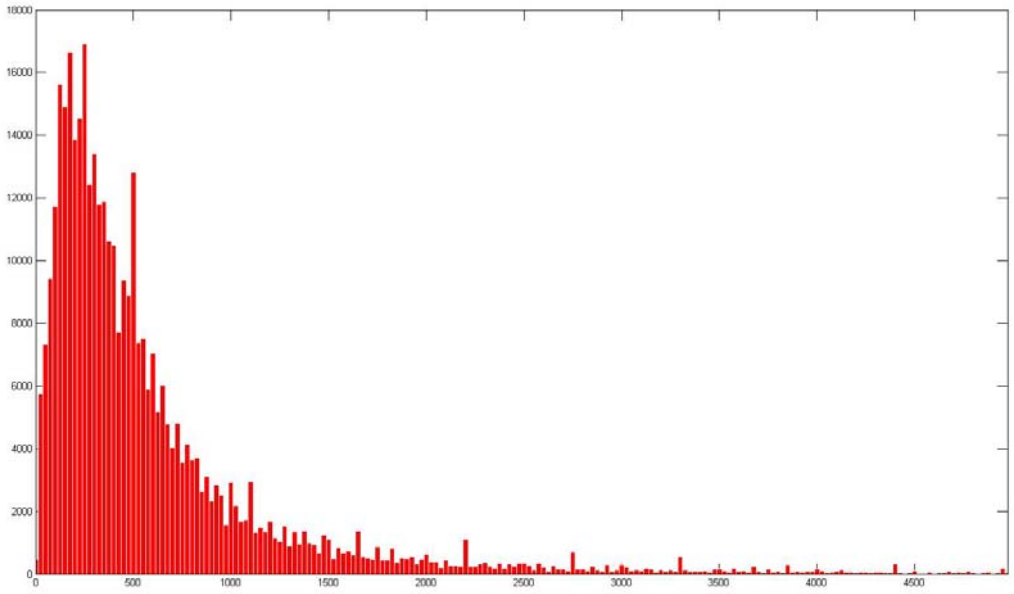

Figure 2. GMM vs ML vs Kernel Density Plot for Brazil

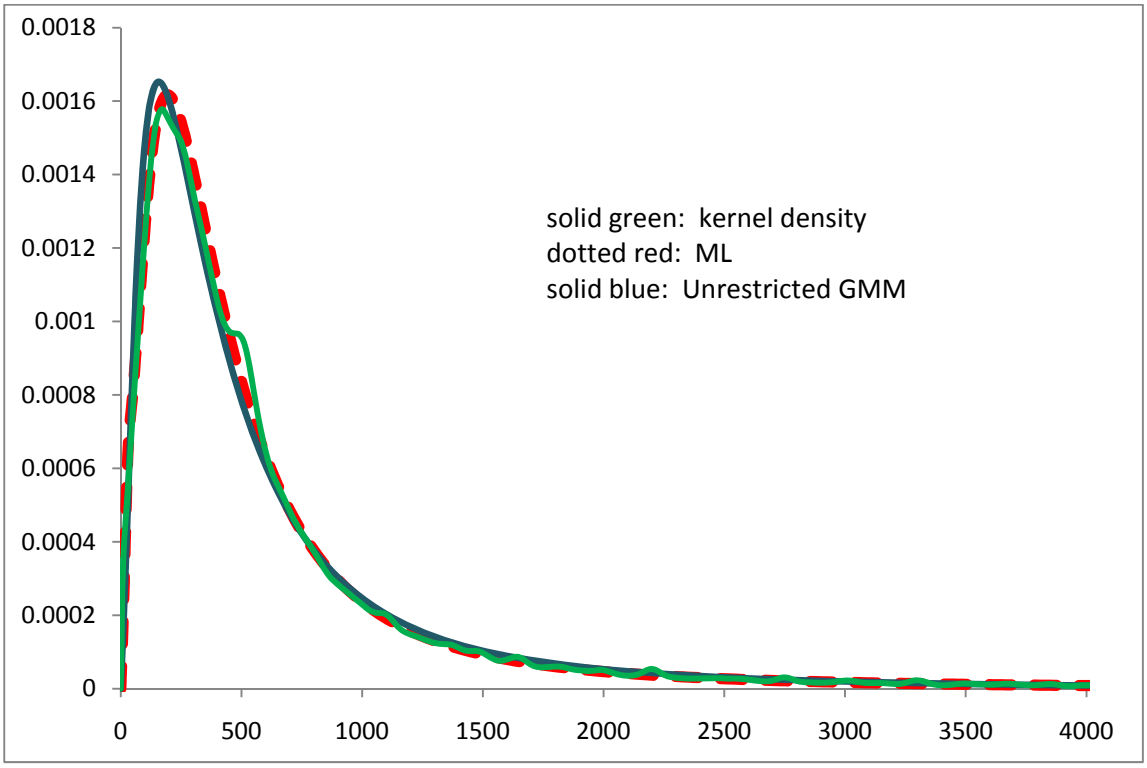

Figure 3. GMM vs ML vs Kernel Density Plot for Indonesia

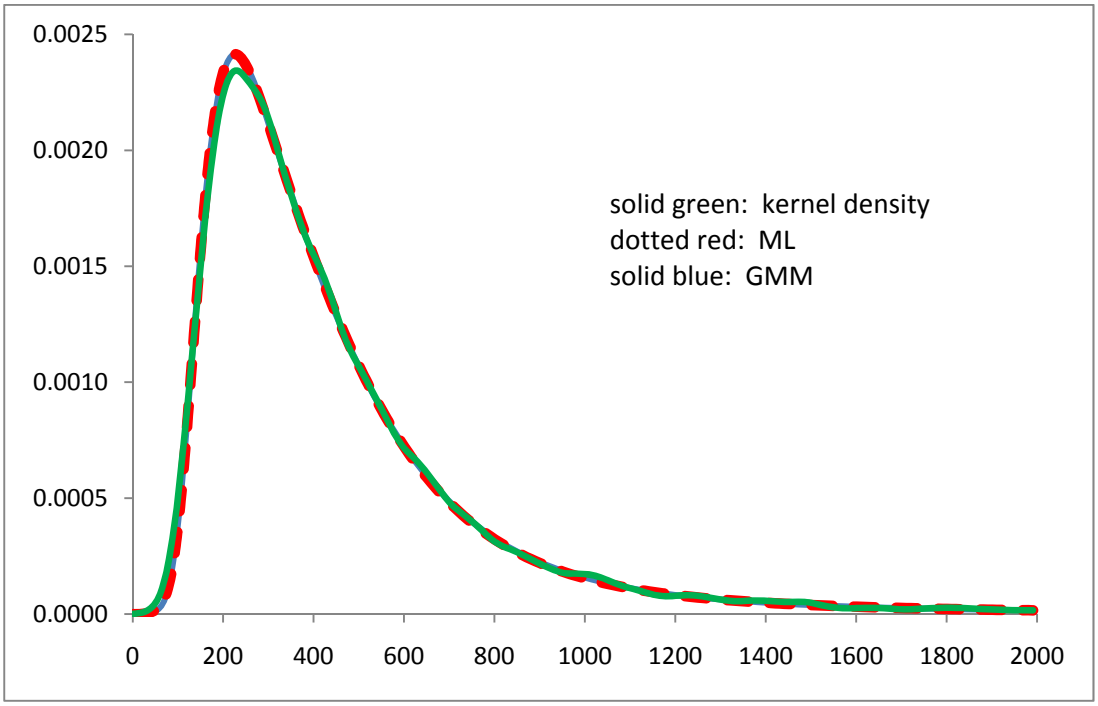

Figure 4 Histogram of Observations for Indonesia

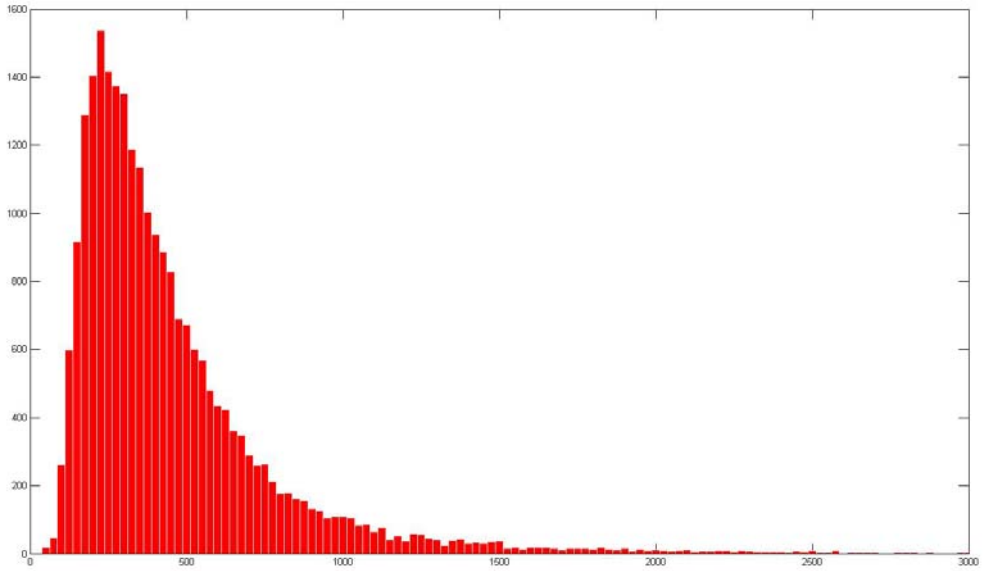

Figure 5. Density and Confidence Bands for Indonesia