J Indian Soc Probab Stat (2018) 19:345–358
https://doi.org/10.1007/s41096-018-0048-x

RESEARCH ARTICLE

Indirect Inference for Fitting Income Distributions

S. Rao Jammalamadaka[1] · Gaoyuan Tian[1]

Accepted: 20 June 2018 / Published online: 2 August 2018

Abstract
We suggest a novel application of indirect inference method for estimating income distribution using limited data. Generalized method of moments (GMM) method is the classical way to estimate a parametric income distribution in this case. To use GMM, explicit expression for moments are needed. In this paper, indirect inference method allows us to estimate it without needing to find explicit analytical expression. Theoretical properties of this estimator and a goodness-of-fit test are provided.

Keywords Indirect inference · Estimation · Income distribution · Lorenz curve

1 Introduction

The distribution of incomes and wealth play an important role in the measurement of inequality and poverty among people as well as nations. Various methods and different parametric models for income distribution have been developed in a number of articles by many economists— see e.g. Chotikapanich et al. (2007), McDonald and Xu (1995), Hajargasht et al. (2012) and the references contained therein. In most of this work, a specific parametric form is assumed for the income distribution, and the generalized method of moments (GMM) is the usually preferred technique for estimating the parameters involved. To use the GMM, one needs explicit expressions for the expected values of the moments, or of the estimating functions used. The calculation of estimating functions for a given model in terms of its parameters can be complicated and sometimes not readily available. In this article, we employ a general method of fitting these models, using what is known as the "indirect inference" method which allows us to estimate these quantities for a given model without needing to

✉ Gaoyuan Tian
tiangaoyuan@gmail.com

S. Rao Jammalamadaka
sreenivas@ucsb.edu

1 Department of Statistics and Applied Probability, University of California, Santa Barbara, USA

find explicit analytical expressions, and thus provides what is essentially a flexible computational technique for fitting income distributions.

While typically estimation of parameters by methods such as maximum likelihood, the method of moments, and the GMM proceeds by having a random sample from the given model, our main goal here is to introduce the indirect inference technique, and to demonstrate how such parametric inference can be handled when only limited data or just some statistics or empirical values of estimating functions based on it, are available

This article is organized as follows. In Sect. 2, we give a brief introduction to some measures of inequality including the Gini index and the Lorenz Curve (LC), because our estimation of the income distribution is based on certain values of the empirical Lorenz Curve as given by World Bank (see Table 2). We also briefly review here some parametric models that are commonly used for income distributions. In Sect. 3, we describe indirect inference method as a suitable approach for estimating parameters from such limited information, and describe some theoretical properties that these estimators enjoy. In Sect. 4, we test the optimization algorithm used in our method. Also a Monte Carlo study is conducted to compare and evaluate these estimators. As a demonstration of the power of this method, in Sect. 5, we illustrate it by comparing the income distributions as well as inequality indices for India, China and the USA over the past 30 years. We end with brief concluding remarks in Sect. 6.

2 Introduction to Some Inequality Measures

2.1 Lorenz Curve

Let $\{x_i\}$ denote data drawn from a probability distribution with the distribution function $F(x)$, probability density function $f(x)$, and mean μ. Let z_p denote the quantile corresponding to a proportion $0 \leq p \leq 1$ i.e.

$$p = F(z_p) = \int_0^{z_p} f(t)\,dt \tag{1}$$

and then the theoretical Lorenz Curve is defined as

$$L(p) = \mu^{-1} \int_0^{z} tf(t)\,dt = \frac{\int_0^z tf(t)\,dt}{\int_0^\infty tf(t)\,dt}. \tag{2}$$

The numerator corresponds to the total income of the bottom p proportion of the population, while the denominator represents the total income for all the population.

Assuming that F is continuous, one may write $z = F^{-1}(p)$ and a change of variable to write the LC in a direct way as

$$L(p) = \mu^{-1} \int_0^{p} F^{-1}(t)\,dt \tag{3}$$

Table 1 Lorenz Curve for some distributions

Distribution	CDF	Lorenz Curve
Exponential	$F(x) = 1 - \exp^{-\lambda x}, x > 0$	$p + (1-p)\log(1-p)$
General uniform	$F(x) = \dfrac{x-a}{\theta}, a < x < a+\theta$	$\dfrac{ap + \theta p^2/2}{a + \theta/2}$
Pareto	$F(x) = 1 - (a/x)^a, x > a, a > 1$	$1 - (1-p)^{(a-1)/a}$
Lognormal	$F(x) = 1/2 + 1/2\,\mathrm{erf}\left[\dfrac{\log x - \mu}{\sqrt{2}\sigma}\right]$	$\Phi(\Phi^{-1}(p) - \sigma)$

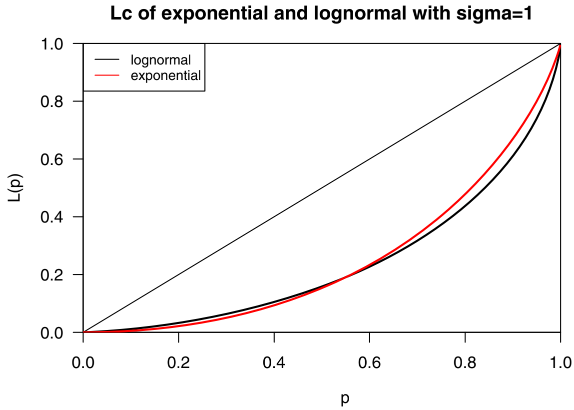


Fig. 1 Lorenz Curve of lognormal and exponential

Table 1 shows LC expression for some common distributions. Notice that, for exponential distributions, LC does not depend on the scale-parameter, a property that could be used for goodness of fit tests (see Gail and Gastwirth 1978). Figure 1 compares LC for lognormal and exponential.

On the other hand, data-based empirical Lorenz curve is defined as follows: Let $0 < x_1 \leq x_2 \leq \cdots \leq x_n$ be ordered data, say on incomes. Then the empirical Lorenz Curve is defined as

$$L(i/n) = s_i/s_n \tag{4}$$

where $s_i = x_1 + x_2 + \cdots + x_i$, $L(0) = 0, i = 0, \ldots, n$.

2.2 Gini Index

Gini index is a commonly used a measure of income inequality in a country, and takes values between 0 and 1. As a U-statistic, it is also widely used in goodness of fit tests. See e.g. Jammalamadaka and Goria (2004) who discuss a test of goodness of fit based on Gini index of spacings. Recently, Noughabi (2014) introduced a general test of goodness of fit based on the Gini index of data. One way to define Gini index is through the expected mean difference.

Definition 1 If X, Y denote two non-negative random variables drawn independently from the distribution F, then the Gini index is defined as

$$Gini := \frac{E|X-Y|}{2 \cdot E(X)}$$

The corresponding sample version can be written in the following way:

$$Gini(S) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2(n-1)\sum_{i=1}^{n} x_i} \tag{5}$$

It can also be calculated via LC (see e.g. Gastwirth 1972):

$$G(t) = 2 \cdot \int_0^1 (t - L(t))\, dt \tag{6}$$

2.3 Some Commonly Used Parametric Models for Income

The income distribution is heavily positively skewed and has a long right tail. The popular income distribution models include Generalized Beta-2 distribution, Gamma distribution, and the lognormal distribution.

Generalized Beta-2 distribution (7), which is widely used for modeling income distributions, has the density

$$f(x; a, b, p, q) = \frac{a x^{ap-1}}{b^{ap} B(p,q)(1 + (x/b)^a)^{p+q}}, x > 0. \tag{7}$$

Beta-2 (with $a = 1$), Singh–Maddala (with $p = 1$), Dagum (with $q = 1$) and Generalized gamma (with $q \to \infty$) are some of the special cases of this Generalized Beta-2 distribution (see McDonald and Xu 1995).

Lognormal distribution (8) is another popular model for income distributions, with pdf derived from $log(X) = Y$ which has a normal distribution

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}}, x > 0,\ \sigma > 0. \tag{8}$$

There are many alternate models for income, but as an extensive analysis by Cowell (1995) reveals, the more complicated four-parameter densities are not particularly good choices because their parameters are hard to interpret and may have an over-fitting problem. He argues in favor of lognormal and gamma densities which have two parameters. Among distributions with two parameters, he argues for the Pareto density for modeling high incomes, while gamma and lognormal are more appropriate for modeling middle range incomes. In this article, our goal is not to get into the extensive literature on income distributions, but chose the lognormal distribution for illustrative purposes.

3 Indirect Inference Method

3.1 Indirect Inference Framework

Gourierou et al. (1993) first introduced indirect inference as a simulation based method for estimating the parameters of an extensive class of models. This method is particularly useful when, the likelihood function is analytically intractable or considerably difficult to evaluate. Indirect inference method greatly simplifies the estimation procedure from the theoretical point of view because all we need is to simulate the required moments or estimating functions for a candidate model.

Let $\boldsymbol{\pi}(\theta)$ be an auxiliary parameter vector which is a function of $\boldsymbol{\theta}$ corresponding to $F_{\boldsymbol{\theta}}$, and let $\hat{\boldsymbol{\pi}}$ be an easy-to-compute and available empirical estimator of it. An exact or explicit analytical expression for the estimating function $\boldsymbol{\pi}(\boldsymbol{\theta})$ is not required, in contrast with the generalized method of moment (GMM) proposed by Hansen (1982). We then find an estimator of $\boldsymbol{\theta}$ as the solution to the following optimization problem:

$$\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \ (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}(\boldsymbol{\theta}))^T \boldsymbol{\Omega}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}(\boldsymbol{\theta}))$$

where $\boldsymbol{\Theta}$ is the parameter space, and $\boldsymbol{\Omega}$ is a positive definite weight matrix (discussed later). The idea here is to find the parameter vector $\boldsymbol{\theta}$ such that $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}(\boldsymbol{\theta})$ are as close as possible. If $\boldsymbol{\pi}(\boldsymbol{\theta})$ is not readily available as an explicit function of $\boldsymbol{\theta}$, which is usually what calls for indirect inference, it is calculated starting with a reasonable initial value of $\boldsymbol{\theta}$, and thereafter an iterative process is triggered to search the optimal θ until a specified convergence criterion is satisfied. Or it is estimated through the software and approximated by a parametric bootstrap through the following steps:

- Step 1 H samples of sample size N are simulated from $F_{\boldsymbol{\theta}}$. Since $\boldsymbol{\theta}$ is unknown, we start with a reasonable initial value of $\boldsymbol{\theta}$ and iteratively search for the value of $\boldsymbol{\theta}$ that best fits the data, via the BFGS algorithm mentioned later on (see also Fig. 2).
- Step 2 For each sample h, $h = 1, 2, \ldots, H$, its $\boldsymbol{\pi}^{*h}(\boldsymbol{\theta})$ is calculated based on its empirical distribution function.
- Step 3 $\boldsymbol{\pi}(\boldsymbol{\theta})$ could be approximated by $\boldsymbol{\pi}^*(\boldsymbol{\theta}) = \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{\pi}^{*h}(\boldsymbol{\theta})$.

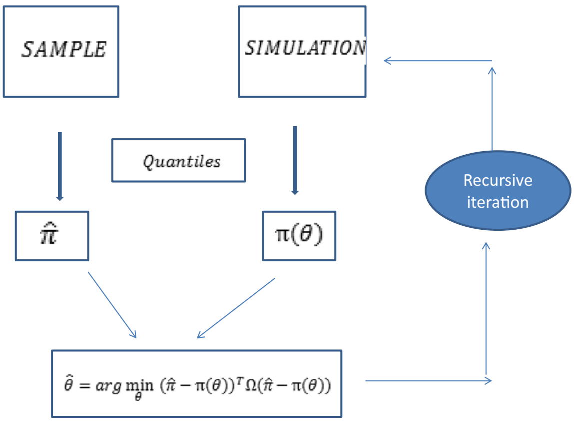
Then the indirect inference estimator $\hat{\boldsymbol{\theta}}$ is obtained as follows:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \ (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*(\boldsymbol{\theta}))^T \boldsymbol{\Omega}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*(\boldsymbol{\theta}))$$

Here it can be seen that the optimal choice of the weight matrix is given by $\boldsymbol{\Omega} = (Var(\hat{\boldsymbol{\pi}}))^{-1}$, which ensures that the estimator $\hat{\boldsymbol{\theta}}$ is asymptotically efficient. However, since $(Var(\hat{\boldsymbol{\pi}}))^{-1}$ is itself a function of $\boldsymbol{\theta}$, we obtain an estimate for this matrix through an iterative two-stage procedure (similar to the two-step GMM) as follows:

- Stage 1 We start by taking $\boldsymbol{\Omega} = \boldsymbol{I}$ with $\boldsymbol{I}$ denoting the identity matrix, to solve the optimization problem and obtain an initial estimate $\boldsymbol{\theta}_1$.

Fig. 2 Estimation algorithm



Stage 2 The weighting matrix is then estimated with $\hat{\mathbf{\Omega}} = (Var(\hat{\boldsymbol{\pi}}(\boldsymbol{\theta}_1)))^{-1}$.

Again the expression for $Var(\hat{\boldsymbol{\pi}}(\boldsymbol{\theta}_1))$ is hard to derive and hence evaluated through parametric bootstrap with simulations using the initial estimate $\boldsymbol{\theta}_1$. The optimization algorithm used in this case is the so-called Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm which is an iterative method that solves non-linear optimization problems. This estimation procedure could be described by the genetic algorithm shown in Fig. 2.

Having defined the proposed indirect inference estimator and described the practical procedure for obtaining it, we now provide some asymptotic and robustness properties of such an estimator in the following section, just for completeness.

3.2 Some Theoretical Properties

We quote the regularity conditions and the asymptotic properties for such indirect estimators from Gourierou et al. (1993). Denoting $\boldsymbol{\theta}_0$ as the true parameter vector, we need the following assumptions for the main theorem, Theorem 1 to hold:

(A1) $\boldsymbol{\xi}_n = \sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}(\boldsymbol{\theta}_0)) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{V})$ where $\boldsymbol{V} = \lim_{n\to\infty} Var(\boldsymbol{\xi}_n)$
(A2) There is a unique $\boldsymbol{\theta}_0$ such that auxiliary estimator equals the auxiliary parameter: $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\boldsymbol{\theta}_0) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}_0$
(A3) If $\mathbf{\Omega}$ is estimated by $\hat{\mathbf{\Omega}}$, then $\hat{\mathbf{\Omega}} \xrightarrow{P} \mathbf{\Omega}$, where $\mathbf{\Omega} > 0$
(A4) $\boldsymbol{\pi}(\boldsymbol{\theta})$ is a differentiable function with $\mathbf{D}(\boldsymbol{\theta}) = \partial\boldsymbol{\pi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$.
(A5) The matrix $\mathbf{D}^T(\boldsymbol{\theta})\,\mathbf{\Omega}\mathbf{D}(\boldsymbol{\theta})$ is full rank.
(A6) $\mathbf{\Theta}$ is compact.
(A7) The choice of the initial value of $\boldsymbol{\theta}$ is independent of the estimation algorithm.

Theorem 1 (Gourierou et al. 1993) *Under the Assumptions (A1)–(A7), the indirect estimator is asymptotically normal, for any fixed H (the number of samples used in Steps 2 and 3 above), and as n goes to infinity:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xmapsto{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$$

$$\textit{with}\ \mathbf{\Lambda} = (1+\frac{1}{H})\mathbf{\Gamma}\mathbf{V}\mathbf{\Gamma}^T\ \textit{where}\ \mathbf{\Gamma} = \left(\mathbf{D}^T\left(\boldsymbol{\theta}_0\right)\mathbf{\Omega}\mathbf{D}\left(\boldsymbol{\theta}_0\right)\right)^{-1}\mathbf{D}^T\left(\boldsymbol{\theta}_0\right)\mathbf{\Omega}.$$

This theorem provides the asymptotic normality of the estimator $\hat{\boldsymbol{\theta}}$ from that of the auxiliary estimator $\hat{\boldsymbol{\pi}}$. Consistency follows from this asymptotic normality. Notice, the factor $(1+\frac{1}{H})$ is what distinguishes the asymptotic variance of the indirect inference estimator from that of the GMM: when H goes to infinity, they have the same expression. H is set to be 100 in this article.

Now, a bit of explanation about how these conditions (A1) to (A7) relate to our problem of using indirect inference using the empirical LC.

3.3 Lorenz Curve Based Indirect Inference

In our particular case, all we are given are the values of the empirical Lorenz curve and we wish to determine the parameters of the given model that match these. We take the empirical estimator $\hat{\boldsymbol{\pi}}$ to be the sample mean, along with the 9 points on empirical LC, as given in Table 3. i.e.

$$\hat{\boldsymbol{\pi}} = (\bar{X}, \hat{L}(0.1), \ldots, \hat{L}(0.9))^T.$$

Thus the auxiliary parameters $\boldsymbol{\pi}(\boldsymbol{\theta})$ correspond to the theoretical mean and 9 points on theoretical LC of the lognormal distribution. For Condition (A1), we need to check the asymptotic normality of $\hat{\boldsymbol{\pi}} = (\bar{X}, \hat{L}(0.1), \ldots, \hat{L}(0.9))^T$. By central limit theorem, $\bar{X}$ is asymptotic normal. Under some mild regularity conditions, Goldie (1977) proved the weak convergence of the Lorenz process $l_n(\boldsymbol{p}) = \sqrt{n}[L_n(\boldsymbol{p}) - L(\boldsymbol{p})], 0 \le \boldsymbol{p} \le 1$, to a Gaussian process if $L(\boldsymbol{p})$ is continuous at the empirical points. Thus the asymptotic normality of $\hat{\boldsymbol{\pi}} = (\bar{X}, \hat{L}(0.1), \cdots, \hat{L}(0.9))^T$ is established.

Condition (A2) is often called the “global identifiability” problem in econometrics and is often hard to prove and such, is assumed in many cases. In condition (A3), our 2-step matrix $\hat{\mathbf{\Omega}}$ is estimated through the 2-step GMM procedure described above and thus is consistent. The rest of the conditions are standard conditions for indirect estimators such as the one put forward in this paper. We therefore have that the estimator $\hat{\boldsymbol{\theta}}$ proposed here is consistent and asymptotically normal.

3.4 Data

The data comes from the Website of the World Bank, and takes the form of summary statistics including mean income, measures of inequality and 9 points on the empirical LC. In Table 2, the poverty line is the minimum level of income deemed adequate in a particular country. The head-count ratio is the proportion of a population lives below the poverty line. The first part of Table 2 shows the data in the following way: the first 10% of the population owns 1.7% of the total income, the second 10% of the population owns 3.4% of the total income, etc. Since the sum of these 10 numbers equals 1, only the numbers of the first 9 groups need to be included in the moment

Table 2 Original data

USA's income share by deciles (%)										
Year	Lowest	2nd	3rd	4th	5th	6th	7th	8th	9th	Highest
2010	1.70	3.40	4.56	5.73	7.00	8.44	10.19	12.52	16.25	30.19
USA's poverty index										
Year	Mean ($/month)	Pov. line	Headcount (%)	Gini index (%)						
2010	1917.38	1.90	1.00	41.06						

Table 3 Transformed data (cumulative share)

By deciles (%)									
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\hat{L}(p)$	1.70	5.10	9.66	15.39	22.39	30.83	41.02	53.54	69.77

Table 4 Goodness of fit assessment

USA 2010									
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\hat{L}(\boldsymbol{p})$	1.70	5.10	9.66	15.39	22.39	30.83	41.02	53.54	69.77
$L(\boldsymbol{p}, \hat{\boldsymbol{\theta}})(lognormal)$	2.15	5.63	10.14	15.63	22.31	30.59	40.66	52.98	69.17
$L(\boldsymbol{p}, \hat{\boldsymbol{\theta}})(Gamma)$	1.29	4.22	8.56	14.34	21.55	30.49	41.39	54.99	72.39

conditions. The cumulation of these 9 numbers yields the 9 points on the empirical LC $\hat{L}(\boldsymbol{p})$ in Table 3.

With our indirect inference estimator $\hat{\boldsymbol{\theta}}$, $\hat{L}(\boldsymbol{p})$ and $L(\boldsymbol{p}, \hat{\boldsymbol{\theta}})$ are compared as shown at Table 4. This table contains lognormal and gamma, and may be extended to other models to assess how well these distributions fit the data. The test statistic J_n given below can be used to test how well a given income distribution fits the data

$$J_n = n \left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}) \right)^T \hat{\boldsymbol{\Omega}} \left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}) \right) \xrightarrow{D} \chi^2_{M-K} \tag{9}$$

where n is the number of individuals surveyed, M the dimension of auxiliary parameters, K is the number of parameters in the parametric model with $\hat{\boldsymbol{\Omega}}$ representing the two-step weight matrix. One problem with this statistic is that n is usually unknown and we will not attempt to use it here.

4 A Simulation Study

4.1 Numerical Optimization

The default optimization algorithm used in R is Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. Similar to the Newton-Raphson method, it is an iterative method

Fig. 3 Objective function versus (θ_1, θ_2)

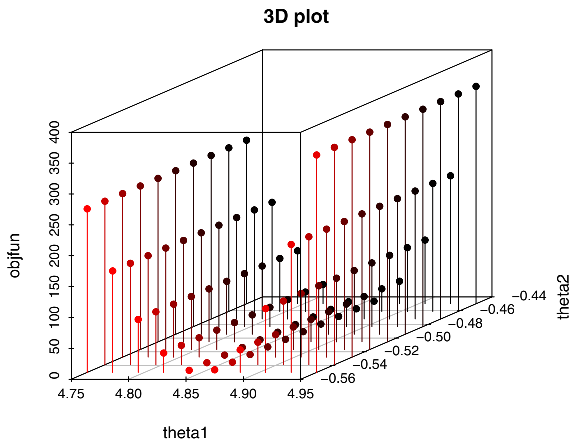
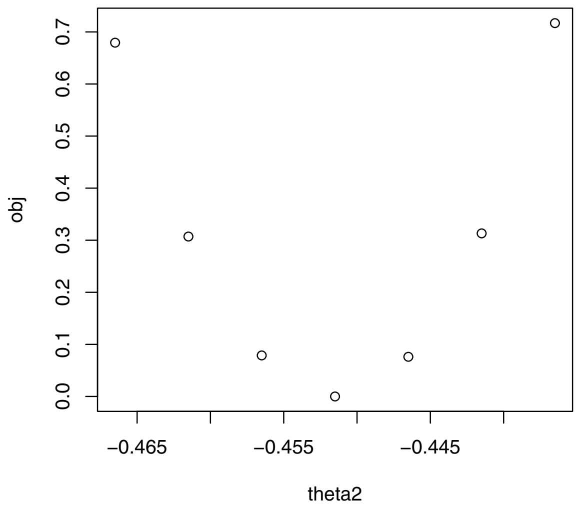


Fig. 4 Objective function versus θ_2



for solving non-linear optimization problems. In this case, the parameter space for σ is $(0, \infty)$. Since it has a lower bound, sometimes this optimization algorithm breaks down when searching for points larger than but close to 0. Instead, we would estimate the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where $(\mu, \sigma) = (\theta_1, \exp(\theta_2))$. The estimated parameter $\hat{\sigma}$ approximately equals $\log(\hat{\theta_2})$.

Here we want to verify that the estimated point is the local minimum. The true parameters $\boldsymbol{\theta} = (4.8276, -0.4963)$ is obtained from the estimate value of data in Table 2. The data (9 points on Lorenz curve and mean) is simulated from lognormal distribution with above parameters with sample size $N = 1000$. The estimated value $\hat{\boldsymbol{\theta}} = (4.8381, -0.4515)$. It has a local minimum as can be seen from Figs. 3 and 4.

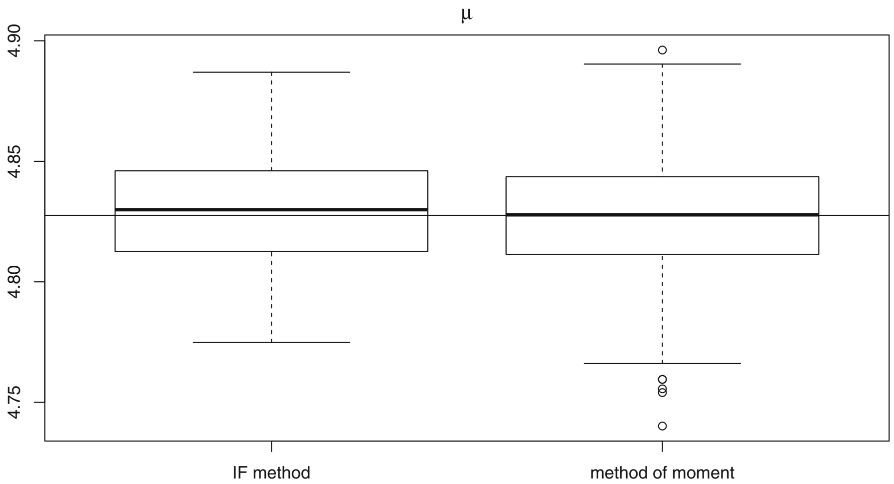


Fig. 5 Boxplot of $\hat{\mu}$

4.2 Indirect Inference Compared to the Method of Moments: A Monte Carlo Study

Instead of the 9 points on the empirical LC curve, suppose we only have the sample mean and sample median. For lognormal distribution, the mean $EX = \exp(\mu+\sigma^2/2)$, and the Median $m = \exp(\mu)$. By setting these equal to their empirical parts, the method of moment estimators are given by:

$$\hat{\mu} = \log(m), \hat{\sigma} = \sqrt{2(\log(\bar{x}) - \log(m))} \tag{10}$$

Suppose the true parameters $(\mu, \sigma) = (4.8276, \exp(-0.4963))$. Box-plots to compare these two estimators are obtained by Monte Carlo study with sample size $N = 1000$ and Monte Carlo replication $B = 1000$ in Figs. 5 and 6. Our indirect inference method has smaller variance especially for σ.

5 Case Study

Greenwood and Jovanovic (1990) found a positive correlation between growth and income inequality in a cross-section of international data. Here we are interested to see the change of India's income distribution and inequality over the past 30 years. In addition, income distribution and inequality of India and China are compared with USA, the largest economy.

5.1 Data

Data is collected every 3 years by the World Bank. It takes the form of summary statistics as shown at Table 5 e.g. for India.

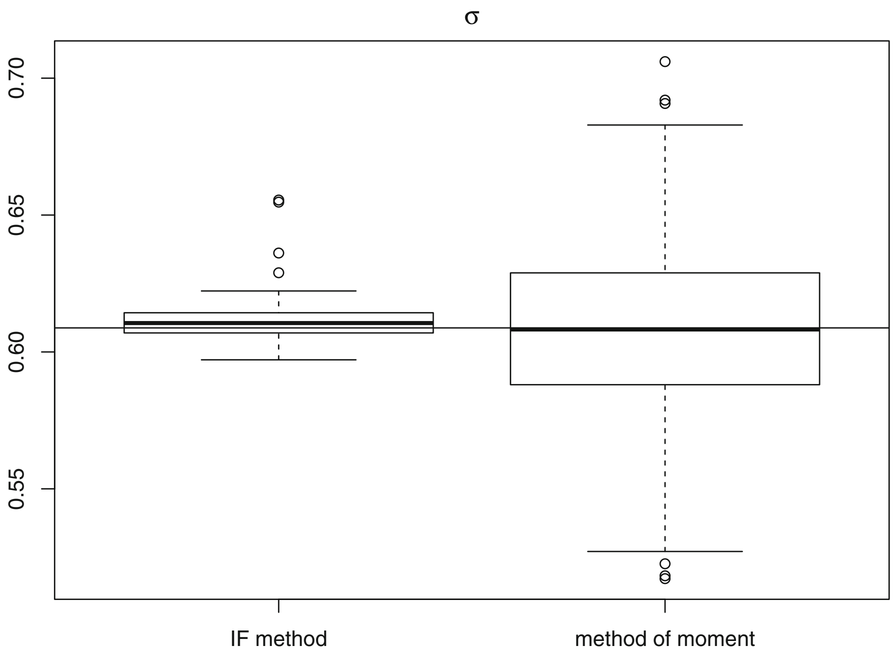


Fig. 6 Boxplot of $\hat{\sigma}$

Table 5 Income inequality of India: 1983 versus 2010

India (Urban)'s income share by deciles (%)										
Year	Lowest	2nd	3rd	4th	5th	6th	7th	8th	9th	Highest
2010	2.92	4.04	4.87	5.76	6.76	7.95	9.45	11.49	15.01	31.75
1983	3.59	4.61	5.57	6.51	7.50	8.60	9.94	11.75	14.80	27.12
Poverty Index										
Year	Mean ($/month)	Pov. line ($/day)	Headcount (%)	Gini index (%)						
2010	129.75	1.90	19.85	39.35						
1983	89.36	1.90	34.20	33.33						

5.2 Some Results

With the 9 points on the empirical LC, a smooth empirical LC is estimated by the non-parametric spline technique in R. The income distributions are assumed to be lognormal and are estimated by above indirect inference method. The results are illustrated in Figs. 7, 8, 9, 10 and 11.

Although India's Gini index slightly increased in the last 30 years, the population proportion of low income class decreases (Fig. 8). The population proportion of low income class of India is significantly larger than USA'S, but smaller than China's (Figs. 9, 10).

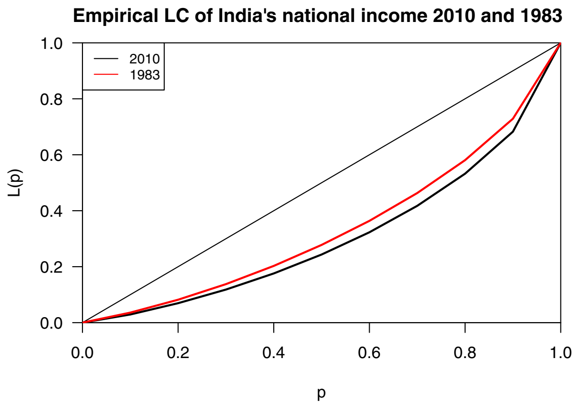


Fig. 7 Lorenz Curve of India 1983 versus 2010

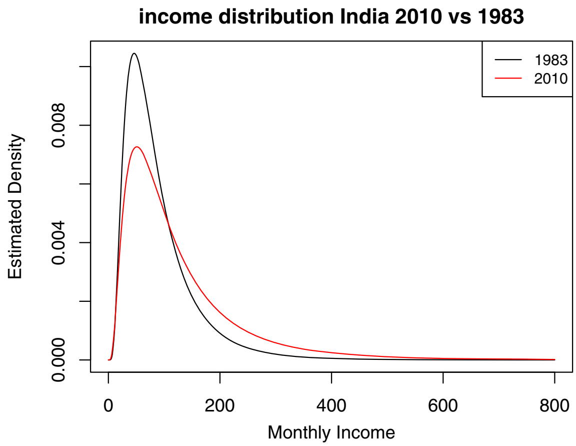


Fig. 8 Income distribution of India 2010 versus 1983

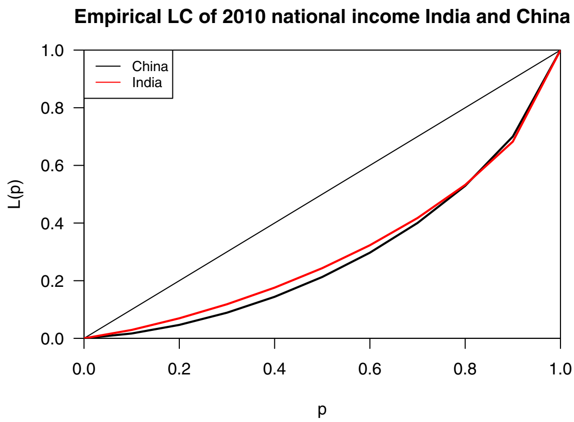


Fig. 9 Lorenz Curve of 2010 India versus China

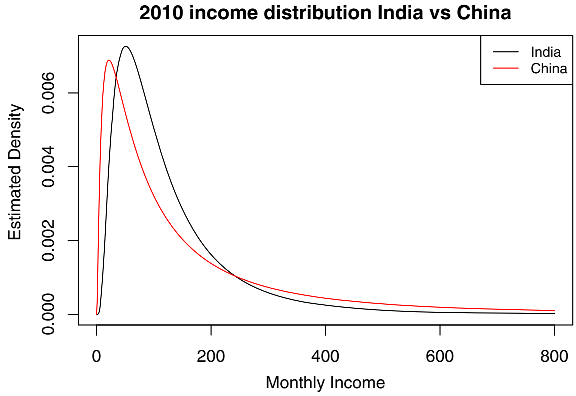


Fig. 10 Income distribution of 2010 India versus China

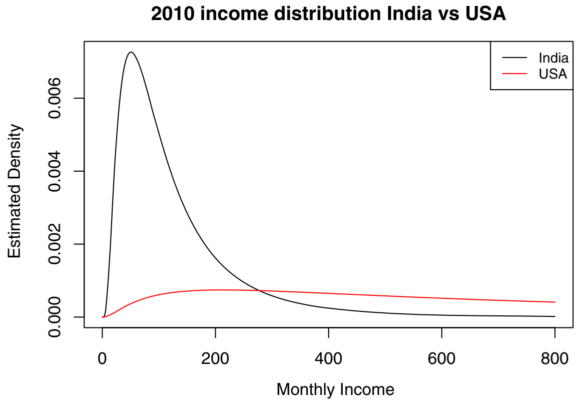


Fig. 11 Income distribution of 2010 India versus USA

6 Conclusions

We develop a practical estimation framework using indirect inference for fitting and analyzing income distributions and income inequality, based on a limited amount of data, such as the empirical Lorenz curve. This simulation based method is very flexible, and allows the parametric models and the auxiliary parameters to be adjusted adaptively, and is a practical tool in many other contexts. Interested readers may obtain the R-code for implementing indirect inference in this context, from the authors.

References

Chotikapanich D, Griffiths WE, Prasada Rao DS (2007) Estimating and combining national income distribution using limited data. J Bus Econ Stat 25:97–109

Cowell FA (1995) Measuring inequality. Harvester Wheatsheaf, Hemel Hempstead

Gail MH, Gastwirth JL (1978) A scale-free goodness of fit test for the exponential distribution based on the Lorenz curve. J Am Stat Assoc 73:787–793

Gastwirth JL (1972) The estimation of the Lorenz curve and Gini Index. Rev Econ Stat 54:306–316

Goldie CM (1977) Convergence theorems for empirical Lorenz Curve and their inverses. Adv Appl Probab 9:765–791

Gourierou C, Monfort A, Renault E (1993) Indirect inference. J Appl Econom 8:1–10
Greenwood J, Jovanovic B (1990) Financial development, growth, and the distribution of income. J Polit Econ 98:1076–1107
Hajargasht G, Griffiths WE, Brice J, Prasada Rao DS, Chotikapanich D (2012) Inference for income distribution using grouped data. J Bus Econ Stat 30:563–575
Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50:1029–1054
Jammalamadaka S Rao, Goria MN (2004) A test of goodness of Fit based on Gini's Index of spacing. Stat Probab Lett 68:177–187
McDonald JB, Xu YJ (1995) A generalization of the beta distribution with applications. J Econom 66:133–152
Noughabi HA (2014) A test of goodness of fit based on Gini Index. J Turk Stat Assoc 7:23–32