Looking for the Standard Shape of Income Distributions[1)]

Hang Keun Ryu[2)]

Abstract

Income distributions are derived from the given Gini coefficients using the entropy maximization method (Ryu, 1993). To check the accuracy of this method, the approximated income distribution using the Gini coefficient is compared with the true distribution (U.S. 1983 CPS data). All income distributions will be shown to have relatively the same shapes except the starting points given by the Gini coefficients. Using the derived share function, the Lorenz dominance effects and Wolfson's scalar polarization index are analyzed.

JEL Classification Number: D31, D63
Keywords: Gini coefficient, maximum entropy estimation of probability density functions, Lorenz dominance effects, scalar polarization index.

1. Introduction

The choice of functional form of an income distribution is important for income inequality analysis. Once the underlying income distributions are well approximated with the maximum entropy method, three suggestions can be made. 1) The Lorenz dominance effects (two different Lorenz curves cross at one point) may not be observed for the cross sectional comparison of 57 countries. 2) Estaban and Ray (1994) and Wolfson (1994) introduced the scalar polarization index. If the Gini coefficient and polarization index move along the opposite directions for certain income changes then the Gini coefficient will

1) This Research was supported by the Chung-Ang University Research Grants in 2011.
2) Department of Economics, Chung-Ang University, 221 Heuk Seok Dong, Dong Jak Ku, Seoul, Korea, 156-756. Tel) 82-11-253-6500, Fax) 82-2-515-3256, Email: hangryu@cau.ac.kr.

indicate improvement while the scalar polarization index will indicate a worsening of income distributions. Such opposite direction movements and contradictory implications are possible only if the income share function has two or more maximum values; however, the standard shape of income share functions derived in this paper has only one maximum value and both indices will move along the same direction. Thus, the scalar polarization index is a theoretical proposition but not needed to explain real world income distribution changes. 3) The Gini coefficient was considered as a poor inequality measure to describe the income changes of the poorest group; however, the income distribution derived in this paper using only the Gini coefficient described the income shares of the poorest group fairly well. Thus, the Gini coefficient is a more powerful tool than previously considered.

Mathematical procedures taken in this paper are:

1) The first moment of income distribution can be derived from the given Gini coefficient.
2) The income distribution can be derived from the given first moment using the maximum entropy method.
3) Higher order moments can be estimated from the first moment with very high R^2 values using the cross sectional quintile data of 57 countries.
4) Given higher moments, a better approximation of the income distribution function is possible using the maximum entropy method.
5) Compare the income distribution functions derived using only the Gini coefficient and derived higher moments with the true distribution derived from population data.
6) Lorenz dominance effect (two Lorenz curves are crossing at one point) may not be observed if all income distributions have relatively the same standard shape except the different starting points described by the Gini coefficients.
7) The scalar polarization index introduced by Wolfson (1994) is good only for the theoretical imagination. The Lorenz curves derived in this paper do not meet or come closer near the center point. Therefore, the scalar polarization index and the Gini coefficient move along the same direction.
8) The share function derived by the maximum entropy method describes income shares of the poorest group fairly well.

2. The first moment of income distribution can be derived from the given Gini coefficient.

The Lorenz curve is defined as

$$L \equiv \int_0^z s(z')\,dz' \tag{1}$$

where $s(z)$ is the share function and the coordinate z is the population income coordinate with z=0.005 (the center point of the domain between zero and 0.01) for the poorest 1% group and z=0.995 for the richest 1% group (the center point of the domain between 0.99 and 1.00).

Consider the partial integration of

$$\int_0^1 z\,dL = zL(z)\Big|_0^1 - \int_0^1 L(z)\,dz = 1-g$$

where $g \equiv \int_0^1 L(z)\,dz$

Since

$$dL(z) = s(z)\,dz$$

The mean of the share function is

$$\mu_1 = \int_0^1 z s(z)\,dz = 1-g = \frac{1+GINI}{2} \tag{2}$$

Knowledge of the GINI is equivalent to knowledge of the first moment of the true share function. This result is also reported in Yitzhaki (1998). It means $\mu_1=0.5$ if GINI=0 and $\mu_1=1$ if GINI=1.

3. The income distribution can be derived from the given first moment.

Solving an entropy maximization problem as stated in Ryu (1993)

$$Max_s\, W \equiv -\int s(z)\log s(z)\,dz \tag{3}$$

satisfying

$$\int z s(z)\,dz = \mu_1, \tag{4}$$

the Lagrangian method produces

$$s(z)=\exp[a+bz]=\left[\frac{b}{e^b-1}\right]\cdot\exp[bz] \tag{5}$$

where the normalization condition of the share function is used to remove a. Now the first moment condition (4) produces,

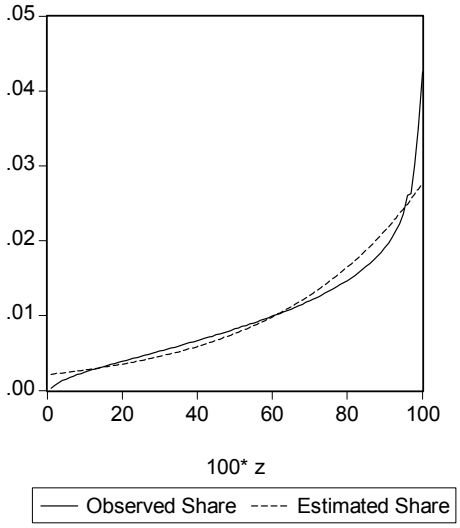
$$\mu_1=\left[\frac{b}{e^b-1}\right]\int_0^1 z\exp[bz]\,dz=\frac{1+GINI}{2} \tag{6}$$

Since the integration is a function of b, $h(b)$ is used to label μ_1.

$$h(b)\equiv-\frac{1}{b}+\frac{e^b}{e^b-1}=\frac{1+GINI}{2} \tag{7}$$

Then b approaches zero if the GINI=0 and b approaches infinity if the GINI=1. Since the LHS of (7) is a monotonic increasing function, a given Gini coefficient uniquely determines b and the share function $s(z)$.

Fig.1 ME shares from Gini only and the Observed Shares

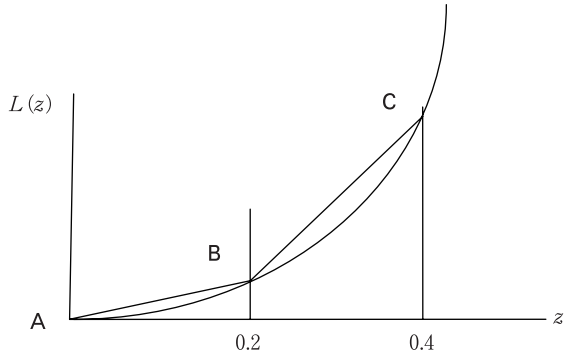


The above Fig. 1 compares empirical U.S. 1983 CPS data with approximated shares function derived using (5) and the Gini coefficient only. For the poor and rich groups, the approximation was inaccurate.

4. Higher order moments of share function can be estimated from the first moment with high R^2 values.

The quintile income shares of 57 countries are reported in the World Development Report (2011). The higher moments of the share function are calculated using the quintile shares. The domain of the discrete share function is represented with the following notation. The domain of the first quintile share s_1 of $z=[0, 0.2]$ is represented with its center point $(z=0.1)$. Similarly, the second share s_2 of $z=[0.2, 0.4]$ is represented with its center point $(z=0.3)$ and, s_5 of $z=[0.8, 1]$ is represented with its center point $(z=0.9)$. The Lorenz curves can be made of quintile shares with kinked lines or with a quadratic approximation.

$$\begin{aligned}
&L_0=L(z=0)=0, && : \text{point A}\\
&L_1=L(z=0.2)=s_1, && : \text{point B}\\
&L_2=L(z=0.4)=s_1+s_2, && : \text{point C}\\
&L_3=L(z=0.6)=s_1+s_2+s_3, && : \text{point D}\\
&L_4=L(z=0.8)=s_1+s_2+s_3+s_4, && : \text{point E}\\
&L_5=L(z=1.0)=s_1+s_2+s_3+s_4+s_5 && : \text{point F}
\end{aligned} \tag{8}$$



Now approximate the Lorenz curve with a quadratic function that passes the points A, B, and C. Introduce a second order polynomial series,

$$L(z)=a_1z+a_2z^2.$$

Let $\Delta z \equiv 0.2$.

$$L_1 - L_0 = s_1 = a_1(\Delta z) + a_2(\Delta z)^2$$
$$L_2 - L_0 = s_1 + s_2 = a_1(2\Delta z) + a_2(2\Delta z)^2 \tag{9}$$

Similarly, for points B, C, and D, use

$$L(z) = L_1 + b_1(z - 0.2) + b_2(z - 0.2)^2$$

where point B is considered as a new origin,

$$L_2 - L_1 = s_2 = b_1(\Delta z) + b_2(\Delta z)^2$$
$$L_3 - L_1 = s_2 + s_3 = b_1(2\Delta z) + b_2(2\Delta z)^2 \tag{10}$$

Similarly, for C, D, and E, use

$$L(z) = L_2 + c_1(z - 0.4) + c_2(z - 0.4)^2$$

$$L_3 - L_2 = s_3 = c_1(\Delta z) + c_2(\Delta z)^2$$
$$L_4 - L_2 = s_3 + s_4 = c_1(2\Delta z) + c_2(2\Delta z)^2 \tag{11}$$

Similarly, for D, E, and F, use

$$L(z) = L_3 + d_1(z - 0.6) + d_2(z - 0.6)^2$$

$$L_4 - L_3 = s_3 = d_1(\Delta z) + d_2(\Delta z)^2$$
$$L_5 - L_3 = s_4 + s_5 = d_1(2\Delta z) + d_2(2\Delta z)^2 \tag{12}$$

Find a_1 and a_2 from points A, B, and C. From (9),

$$\begin{bmatrix} \Delta z & (\Delta z)^2 \\ 2\Delta z & (2\Delta z)^2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} s_1 \\ s_1 + s_2 \end{bmatrix} \tag{13}$$

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \frac{1}{2(\Delta z)^2} \begin{bmatrix} 4\Delta z\, s_1 - \Delta z(s_1 + s_2) \\ -2s_1 + (s_1 + s_2) \end{bmatrix} = \begin{bmatrix} 7.5s_1 - 2.5s_2 \\ -12.5s_1 + 12.5s_2 \end{bmatrix} \tag{14}$$

Similarly,

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 7.5s_2 - 2.5s_3 \\ -12.5s_2 + 12.5s_3 \end{bmatrix}, \quad \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 7.5s_3 - 2.5s_4 \\ -12.5s_3 + 12.5s_4 \end{bmatrix},$$

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 7.5s_4 - 2.5s_5 \\ -12.5s_4 + 12.5s_5 \end{bmatrix}$$

For $z_i = 0.005,\ 0.015, \ldots, 0.195,\ L(z) = a_1 z + a_2 z^2$,

$$s(z_i) \equiv L(z_i + 0.005) - L(z_i - 0.005) = 0.01 a_1 + 0.02 a_2 z_i \tag{15}$$

Similarly for $z_i = 0.205,\ 0.215, \ldots, 0.395$,

$$L(z) = L(z = 0.2) + b_1(z - 0.2) + b_2(z - 0.2)^2,$$

$$S(z) \equiv L(z_i + 0.005) - L(z_i - 0.005) = 0.01 b_1 + 0.02 b_2 (z_i - 0.2) \tag{16}$$

A similar calculation also allows for the share functions of remaining positions to be found as well as for moments to be derived. For notational convenience, let z_1 mean the domain of $z = [0, 0.01]$ which is represented with $z_1 = 0.005$, z_2 means the domain of $z = [0.01, 0.02]$ that is represented with $z_2 = 0.015$. Let $s_1 = s(z_1)$ and $s_2 = s(z_2)$ means shares for these regions.

$$\mu_1 \equiv \sum_{i=1}^{100} z_i s_i \quad \mu_2 \equiv \sum_{i=1}^{100} z_i^2 s_i \quad \mu_3 \equiv \sum_{i=1}^{100} z_i^3 s_i \quad \mu_4 \equiv \sum_{i=1}^{100} z_i^4 s_i$$

$$\mu_5 \equiv \sum_{i=1}^{100} z_i^5 s_i \quad \mu_6 \equiv \sum_{i=1}^{100} z_i^6 s_i \quad \mu_7 \equiv \sum_{i=1}^{100} z_i^7 s_i \quad \mu_8 \equiv \sum_{i=1}^{100} z_i^8 s_i \tag{17}$$

Least squares estimations are possible for the higher moments with respect to the first moment for the 57 countries.

$$\begin{aligned}
\mu_2 &= -0.28463_{(-57.97)} + 1.2006_{(168.5)}\mu_1 && R^2 = 0.9981 \\
\mu_3 &= -0.38878_{(-49.06)} + 1.2215_{(106.2)}\mu_1 && R^2 = 0.9952 \\
\mu_4 &= -0.42606_{(-44.58)} + 1.1854_{(85.49)}\mu_1 && R^2 = 0.9925 \\
\mu_5 &= -0.43414_{(-41.91)} + 1.1300_{(75.17)}\mu_1 && R^2 = 0.9904 \\
\mu_6 &= -0.42845_{(-40.14)} + 1.0694_{(69.05)}\mu_1 && R^2 = 0.9886 \\
\mu_7 &= -0.41617_{(-38.91)} + 1.0094_{(65.03)}\mu_1 && R^2 = 0.9872 \\
\mu_8 &= -0.40088_{(-38.01)} + 0.95235_{(62.22)}\mu_1 && R^2 = 0.9860
\end{aligned} \tag{18}$$

The numbers inside the parentheses are the standard deviations. To check the accuracy of the above regression of higher moments with respect to the first moment, compare predicted higher moments with the estimated moments from the percentile data

Fig.2 Comparison of estimated second moment with observed second moment

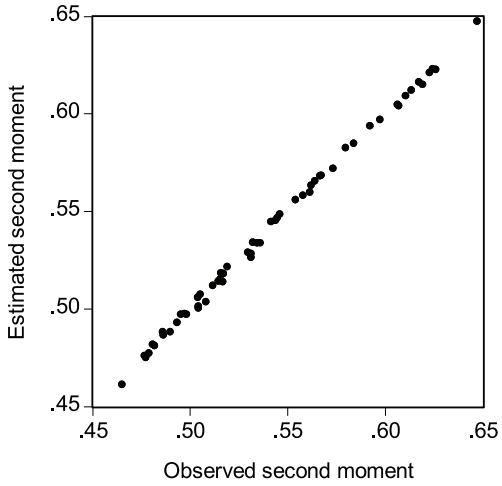


Fig3. Comparison of estimated thrid moment with observed third moment

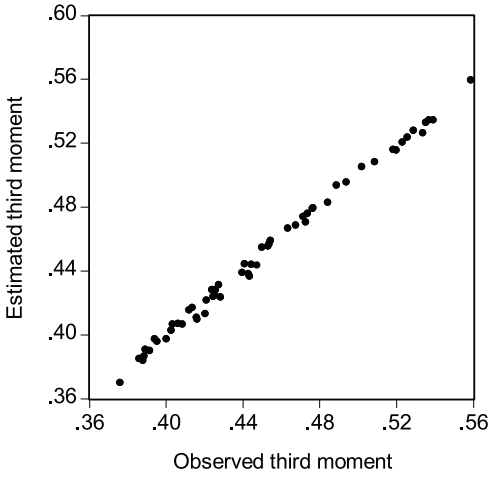


Fig.4 Comparison of estimated fourth moment with observed fourth moment

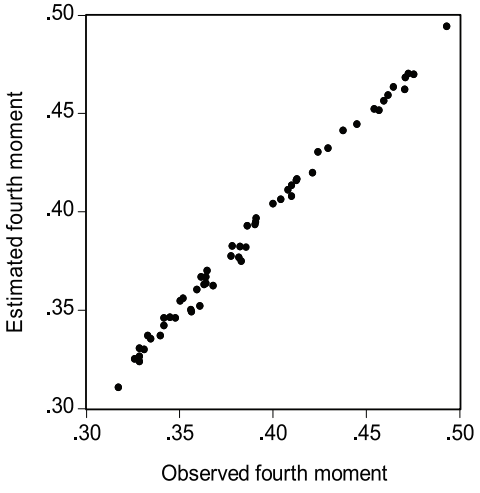


Fig.5 Comparison of estimated fifth moment with observed fifth moment

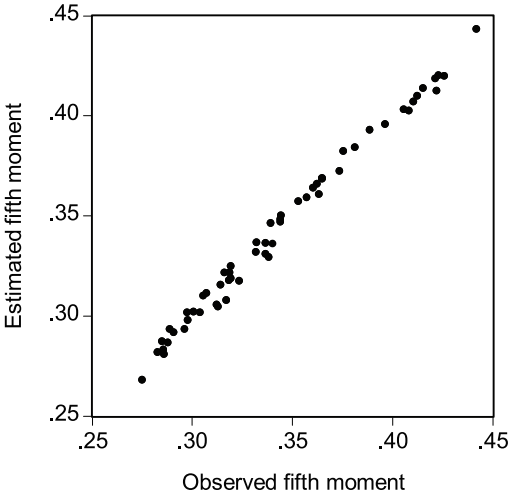


Fig. 6 Comparion of estimated sixth moment with observed sixth moment

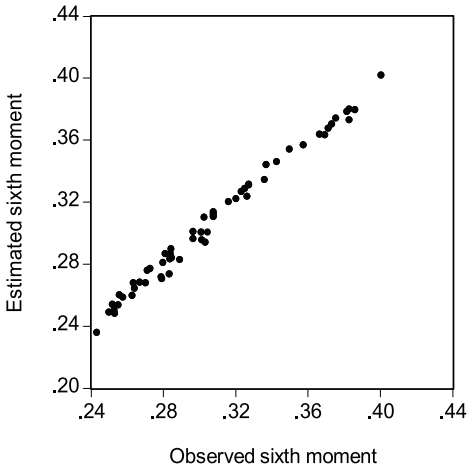


Fig. 7 Comparison of estimated seventh moment with observed seventh moment

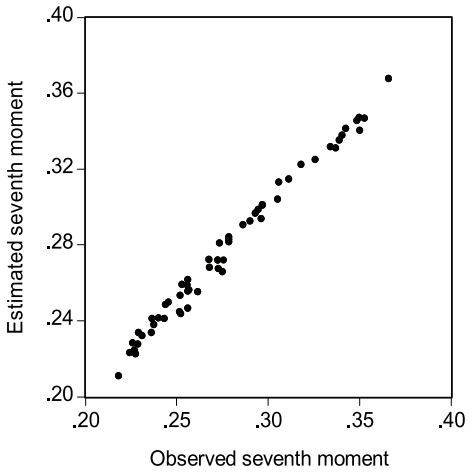
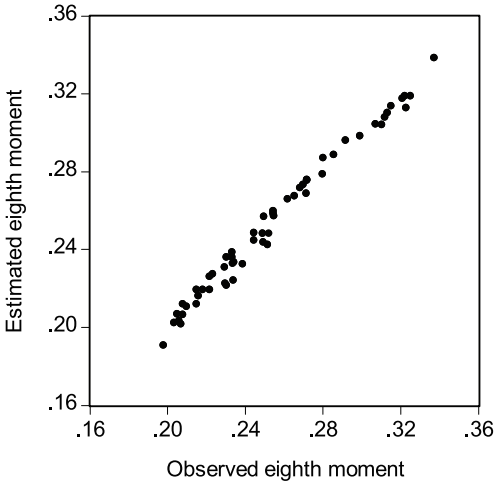


Fig.8 Comparison of estimated eighth moment with observed eighth moment



To check the accuracy of the above estimation method, compare the observed moments of U.S. 1983 CPS data with estimated moments derived with the first moment and regression equations (18). In Table 1 below, the second column shows observed moments based on the whole U.S. 1983 CPS data. Third column is estimated using the observed first moment of U.S. 1983 CPS data and the regression method (18).

Table 1: Comparison of estimated moments with the observed moment

Moments	Observed Moments	Estimated Moments Use (18) and Gini
First Moment	0.6947837	0.6947837
Second Moment	0.5447280	0.5495273
Third Moment	0.4533559	0.4598983
Fourth Moment	0.3910719	0.3975366
Fifth Moment	0.3455222	0.3509656
Sixth Moment	0.3105631	0.3145517
Seventh Moment	0.2827690	0.2851447
Eighth Moment	0.2600679	0.2607973

5. Estimation of the Gini coefficient using only the quintile data.

Suppose the Gini coefficient is not known for a certain country, but assume the quintile income share data is given. Then the Gini coefficient can be estimated with the quadratic approximation of Lorenz curve. Thus, the above share function can be derived with a good starting point though the Gini coefficient is not known.

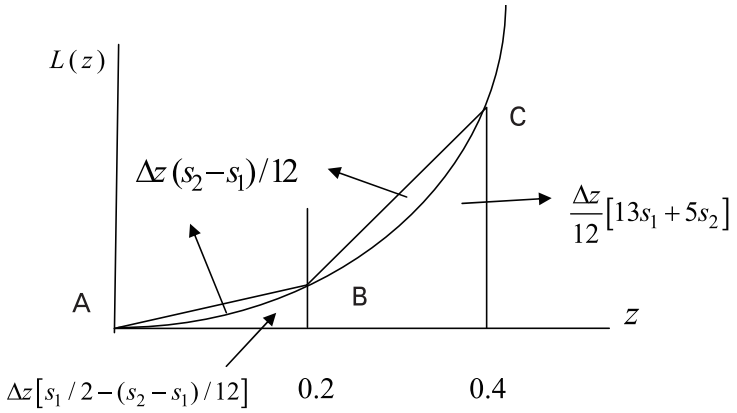
The area under the Lorenz curves for the intervals $z=[0.0, 0.2]$ and $z=[0.2, 0.4]$ are

$$\int_0^{0.2} L(z)\,dz=\int_0^{0.2}\left(a_1z+a_2z^2\right)dz=\Delta z\left[s_1/2-\left(s_2-s_1\right)/12\right] \tag{19}$$

$$\int_{0.2}^{0.4} L(z)\,dz=\int_{0.2}^{0.4}\left(a_1z+a_2z^2\right)dz=\frac{\Delta z}{12}\left[13s_1+5s_2\right] \tag{20}$$

where parameters a_1 and a_2 are estimated in (14) and $\Delta z=0.2$.

Fig. 9 Continuous approximation for the given quintile data



The difference between the kinked Lorenz curve and curved Lorenz curve for points A, B, and C is $\Delta z\left(s_2-s_1\right)/12$ for both intervals $z=[0, 0.2]$ and $z=[0.2, 0.4]$.

Similar calculation shows the difference between the kinked Lorenz curve and curved Lorenz curve for points B, C and D is $\Delta z\left(s_3-s_2\right)/12$ for both intervals $z=[0.2, 0.4]$ and $z=[0.4, 0.6]$. The area difference is different depending on which quadratic equation was chosen for the interval $z=[0.2, 0.4]$. There are two quadratic equations, one from the points A, B, C and another one from points B, D, C. The average value of two area differences will be taken for actual calculation of departure for BC. Similar calculations can be done for intervals CD and DE. The points D, E, and F are defined in (8).

Therefore the difference in area for interval AB is

$$D_1=\Delta z\left(s_2-s_1\right)/12$$

The difference in area for interval BC is

$$D_2=\frac{\Delta z}{12}\left[\frac{\left(s_2-s_1\right)+\left(s_3-s_2\right)}{2}\right] \tag{21}$$

Therefore the difference in area for interval CD is

$$D_3=\frac{\Delta z}{12}\left[\frac{\left(s_3-s_2\right)+\left(s_4-s_3\right)}{2}\right] \tag{22}$$

Therefore the difference in area for interval DE is

$$D_4=\frac{\Delta z}{12}\left[\frac{\left(s_4-s_3\right)+\left(s_5-s_4\right)}{2}\right] \tag{23}$$

Therefore the difference in area for interval EF is

$$D_5=\frac{\Delta z}{12}\left(s_5-s_4\right) \tag{24}$$

The addition of corrected area is

$$D_1+D_2+D_3+D_4+D_5=\frac{\Delta z}{12}\left[-1.5s_1+0.5s_2-0.5s_4+1.5s_5\right] \tag{25}$$

The Gini coefficient with the kinked Lorenz curve is

$$\text{Gini}=2(0.2s_1+0.4s_2+0.6s_3+0.8s_4+s_5)-1.2 \tag{26}$$

The Gini coefficient with the smooth corrected Lorenz curve with correction (25) is

$$\begin{aligned}\text{Gini}&=0.4s_1+0.8s_2+1.2s_3+1.6s_4+2s_5-1.2+\frac{1}{60}\left[-3s_1+s_2-s_4+3s_5\right]\\&=\frac{1}{60}\left[21s_1+49s_2+72s_3+95s_4+123s_5\right]-1.2\end{aligned} \tag{27}$$

For U.S. 1983 CPS data with the 100 percentile, true Gini coefficient with 100 data is 0.389567, with quintile data using the above smoothed curve produced Gini=0.382340, and the kinked Lorenz curve produced Gini=0.3657320. The quadratic approximation produced a good result.

6. Given higher moments, a better approximation of income distribution function can be derived

Suppose we obtain approximating densities by choosing a density that maximizes entropy (subject to moment side conditions) and solves the following problem. This section is a reproduction of Ryu (1993).

$$\max_f W=-\int f(x)\log f(x)\,dx \tag{28}$$

satisfying

$$\int x^m f(x)\,dx=\mu_m,\ m=0,1,\ldots,N \tag{29}$$

with the μ_m having known values. Once the model moments $\mu_0,\ldots,\mu_N$ are known, the problem of (28) becomes a mathematical optimization problem subject to

given side conditions. We shall assume that a solution exists to this problem. See Mead and Papanicolaou (1984) for the existence conditions.

The Lagrangian method produces a maximum entropy distribution

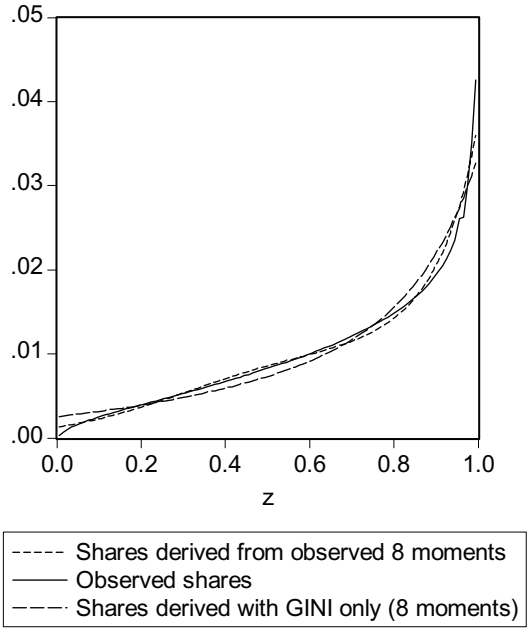
$$f(x)=\exp\Big[\sum_{n=0}^{N} c_n x^n\Big] \quad \text{satisfying} \quad \int x^m f(x)\,dx=\mu_m,\ m=0,1,\ldots,N. \quad (30)$$

The parameters of $f(x)$ can be determined from the moment restriction conditions. There is another interpretation for (30). We can think of it as a polynomial expansion of the logarithmic pdf and we determine the coefficients c_n 's mechanically from the given moments μ_m.

Theorem: Suppose a solution exists for the problem stated in (30). The following method provides an analytic solution for the model parameters $\{c_0,\ldots,c_N\}$ in terms of the known model moments $\mu_0,\ldots,\mu_{2N}$. First, we shall find an N by 1 vector $\mathbf{c}=\{c_1,\ldots,c_N\}$ from the following relationship and c_0 will be determined by the normalization of the density functions.

$$B\mathbf{c}=\mathbf{d} \quad \rightarrow \quad \mathbf{c}=B^{-1}\mathbf{d}$$

Fig.10 Comparison of observed with approximated share functions



where the N x N square matrix B, and the N x 1 vectors **c** and **d** are defined as follows. If the domain of x is finite, we can always transform it to $[0, 1]$.

$$B_{mn} \equiv -mn\left[\mu_{m+n} - \mu_{m+n-1}\right] \quad \text{where} \quad m,n = 1,\ldots,N$$

$$\mathbf{c}' = (c_1,\ldots,c_N), \qquad \text{and}$$

$$\mathbf{d}_m \equiv \left[m(m+1)\mu_m - m^2\mu_{m-1}\right] \quad \text{where} \quad m = 1,\ldots,N.$$

Proof of theorem: See Ryu (1993) and Aroian (1948).

Table 2: Comparison of estimated Lorenz curves with observed Lorenz curve

z	Observed Lorenz	Lorenz from Gini only use (18)	Lorenz from est. moments (17)	Singh-Maddala	Kakwani-Podder
0.01	0.26E-03	0.26E-02	0.13E-02	0.68E-03	0.19E-02
0.02	0.98E-03	0.52E-02	0.27E-02	0.19E-02	0.39E-02
0.03	0.20E-02	0.79E-02	0.42E-02	0.34E-02	0.60E-02
0.04	0.33E-02	0.11E-01	0.58E-02	0.52E-02	0.81E-02
0.05	0.48E-02	0.13E-01	0.75E-02	0.72E-02	0.10E-01
0.06	0.66E-02	0.16E-01	0.92E-02	0.94E-02	0.13E-01
0.07	0.85E-02	0.19E-01	0.11E-01	0.12E-01	0.15E-01
0.08	0.11E-01	0.22E-01	0.13E-01	0.15E-01	0.17E-01
0.09	0.13E-01	0.25E-01	0.15E-01	0.17E-01	0.20E-01
0.10	0.15E-01	0.28E-01	0.17E-01	0.20E-01	0.22E-01
0.20	0.48E-01	0.63E-01	0.47E-01	0.57E-01	0.53E-01
0.30	0.94E-01	0.11	0.92E-01	0.11	0.94E-01
0.40	0.15	0.16	0.15	0.17	0.15
0.50	0.23	0.23	0.23	0.24	0.22
0.60	0.32	0.31	0.32	0.33	0.31
0.70	0.43	0.41	0.43	0.43	0.43
0.80	0.56	0.55	0.56	0.55	0.57
0.90	0.73	0.73	0.73	0.71	0.76
0.91	0.75	0.75	0.75	0.73	0.78
0.92	0.77	0.78	0.77	0.75	0.81
0.93	0.79	0.80	0.79	0.77	0.83
0.94	0.82	0.82	0.82	0.79	0.85
0.95	0.84	0.85	0.84	0.82	0.87
0.96	0.87	0.88	0.87	0.84	0.90
0.97	0.89	0.91	0.90	0.87	0.92
0.98	0.92	0.94	0.93	0.90	0.95
0.99	0.96	0.97	0.96	0.94	0.97
1.00	1	1	1	1	1

If observed 8 moments are available, it produces better accuracy; however, using only the Gini coefficient and the regression equation (18) also produced a good approximation. To check the performance of the above Lorenz curve, derived result is compared with those of the other well known Lorenz curve derivations.

The Singh and Maddala (1976) Lorenz curve is

$$L(u)=\left[1-(1-z)^{\frac{a}{a+1}}\right]^{\frac{a+1}{a}} \qquad 0<z<1,\ a>0$$

The Kakwani and Podder (1973) Lorenz curve is

$$L(z)=z\exp[-h(1-z)], \qquad h>0$$

The parameters a and h are determined such that the sum of the squared residuals are minimized.

Singh-Maddala (1976) and Kakwani-Podder (1973) methods produced relatively poor results for the medium and high income groups, but showed good result for the poorest group ($z=0.01$). The methods of (17) and (18) are based on the polynomial series expansion of the log share function and their performance will not be good for the very end points.

7. Lorenz dominance effect may not be observed

For the Gini coefficients beginning from 0.2, 0.25,...., 0.55, corresponding share functions can be derived using the above ME estimation method of pdf. The Gini range corresponds to the observed range of Gini coefficients for 57 countries of the World Development Report. All income distributions will be shown to have relatively the same standard shape. Fig. 11 and Fig. 12 show the share functions and the Lorenz curves. The two Lorenz curves do not cross at the center points. The gap between two different Lorenz curves initially increases and then decreases near the end point.

The validity of the above claim (two different Lorenz curves do not meet at the center point) depends on the accuracy of the estimation method of the share functions and corresponding Lorenz curves. However, the first 8 moments estimated from the Gini coefficient can be inaccurate and the neglect of higher order moments (higher than 8^{th} moment) can lead to inaccuracy.

Fig.11 Comparison of share functions derived from various Ginies

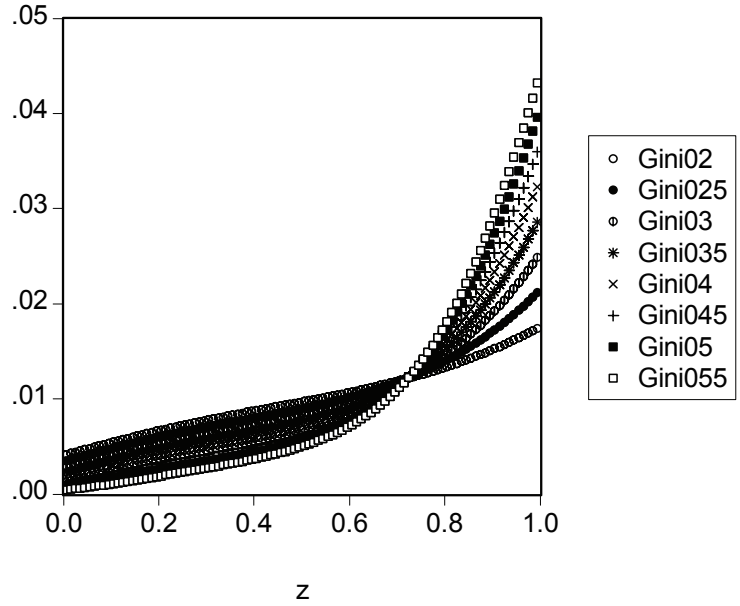
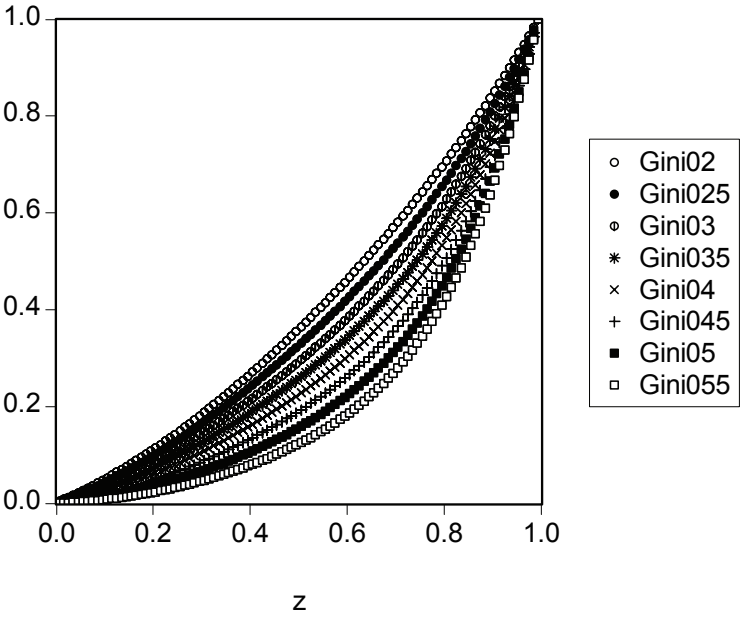


Fig.12 Comparison of Lorenz curves derived from various Ginies



8. Redundancy of the Scalar Polarization Index

The scalar polarization index introduced by Estaban and Ray (1994) and Wolfson (1994) is useful only for the theoretical imagination. If the underlying density function has a bimodal form (two maximum values), then the Gini coefficient and the scalar polarization index can move along opposite directions. However, if the underlying density function has a single maximum, then both the Gini coefficient and scalar index will move along the same direction. Therefore, the scalar polarization index may not be necessary to explain the changes of the real world income distributions if the observed income distribution has a single maximum point. It is proved by showing two Lorenz curves do not meet or do not come closer at the center point.

The following example shows a case where the scalar polarization index is necessary because the Gini coefficient is insufficient to describe the income changes of a group. Suppose there are six persons inside the group and each person has income of \$1, \$2, \$3, \$4, \$5, and \$6. As a way of income redistribution the third person gives \$1 to the first person and similarly sixth person gives one dollar to the fourth person. The Gini coefficient decreased from 0.278 to 0.214 and it seems as if the income distribution improves; however, the scalar polarization index (which will be defined later) increased from 0.3 to 0.428. Therefore, society is divided into two groups and income distribution is polarized.

Fig13. Income Distribution before Income Transfer

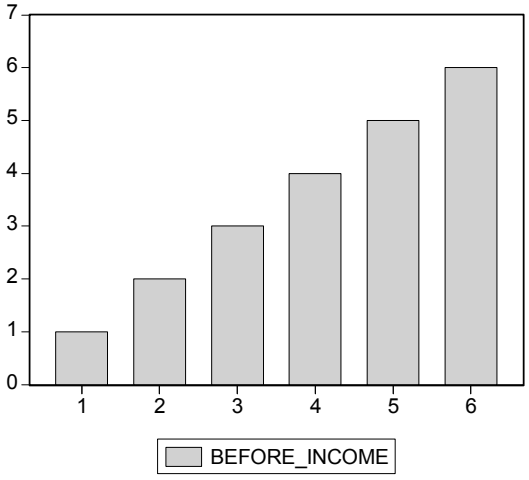


Fig14. Income Distribution after Income Transfer

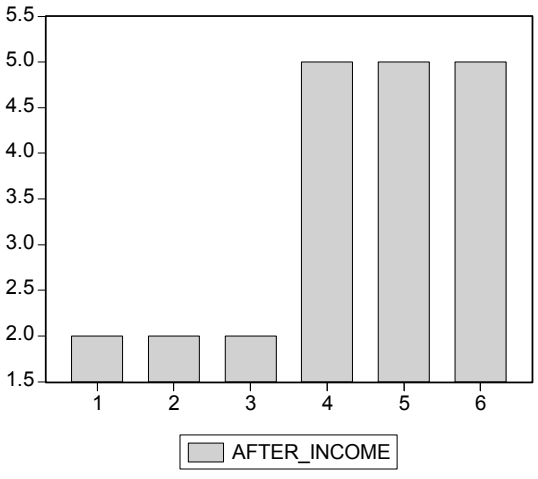


Fig.15 Income Frequency before Income Shift

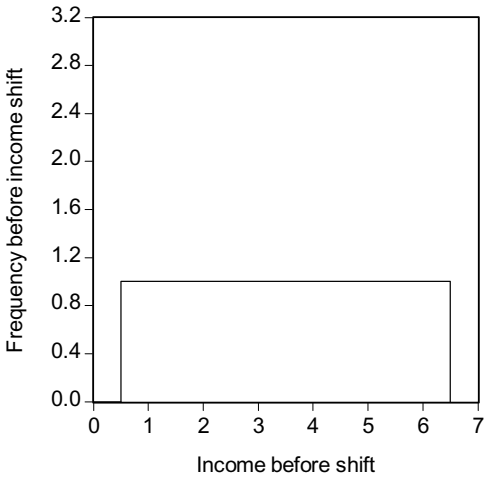
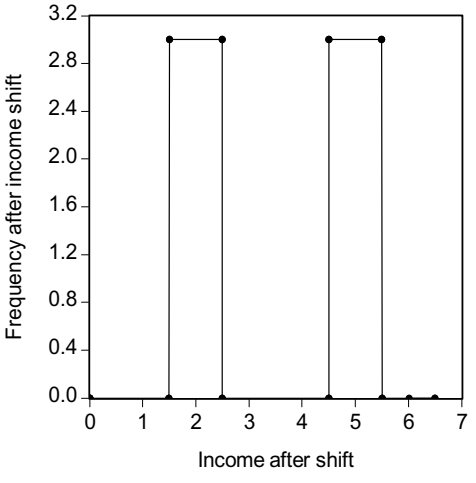


Fig.16 Income Frequency after Income Shift



Figures 13 and 14 show the amount of income before and after the income shift. Figures 15 and 16 show the income shares before and after the income shift. Figure 17 shows the Lorenz curves made of income shares before and after the income shift.

Fig.17 Lorenz Curves Before and After Income Shift

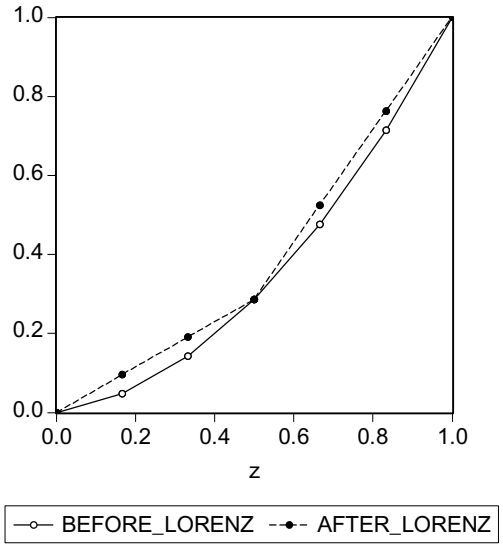
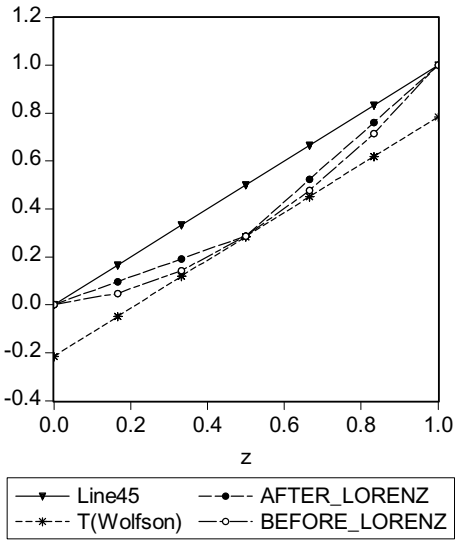


Fig.18 Wolfson Index



Wolfson (1994) defined the scalar polarization index arbitrarily. The word

arbitrarily is used to modify the index to have a range between zero and one.

$$P=4P^{*} \text{ and } P^{*}=\frac{T-0.5^{*}\text{Gini}}{\text{mtan}}, \qquad \text{where } \text{mtan}=\frac{m}{\mu} \tag{31}$$

Rewrite it as

$$P=\frac{4\mu}{m}|0.5-L(0.5)-0.5\text{Gini}| \tag{32}$$

where $T-0.5^{*}\text{Gini}$ is the two triangular area between the Lorenz curve and stared broken line $(^{***})$ in Fig. 18. Note m is the median income and μ is the mean income. The median tangent m/μ is the slope of tangent to the Lorenz curve at the 50th population percentile. Let there be N persons and $\Delta z=1/N$. Let $M=N\mu$ be the total income of the society.

The median income $= m$

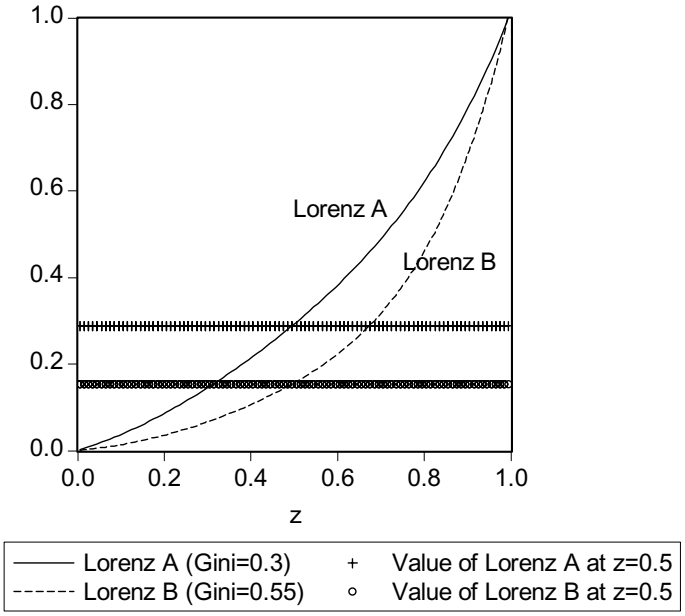
$= M^{*}$(the slope of tangent to the Lorenz curve at the 50th population percentile)$^{*}\Delta z$

$= M^{*}$(share of person at the 50th population percentile)

where $M\Delta z=M/N=\mu$ is used.

Hence, the slope of the tangent to the Lorenz curve at the 50th population percentile$= m/\mu$

Fig. 19 Comparison of Lorenz curves A (Gini =0.3) and B (Gini =0.55)

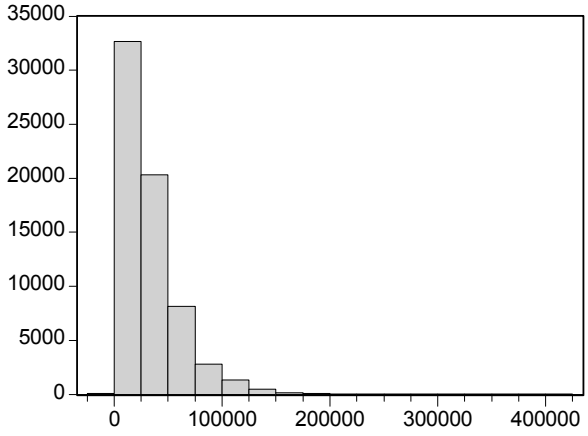


The gap between the Lorenz curves A and B initially increases and then decreases for $z=[0, 1]$. The scalar polarization index is

$$P=\frac{4\mu}{m}|0.5-L(0.5)-0.5\mathrm{Gini}|.$$

When the Gini coefficient increases from 0.3 to 0.55, $[0.5-L(0.5)]$ increases by the square area made of two lines (+++ and 000) in Fig. 19. In addition, $[0.5-L(0.5)-0.5\mathrm{Gini}]$ increases because the starred square is bigger than the gap between two Lorenz curves. If the median value is smaller for Gini 0.55 case, then tangent value (μ/m) is bigger for Gini 0.55 and P increases as Gini has increased. The assumption for this conclusion was that the gap between the Lorenz curves increases initially and decreases monotonically. However, the scalar polarization index and the Gini coefficient will move along the opposite direction if two Lorenz curves meet at the center point as shown in Fig. 17. The square made of two lines (+++ and 000) will be zero in this case. If two Lorenz curves meet at the center, the slope of the inner Lorenz curve (which corresponds to the share function) should change rapidly near the meeting point as in Fig. 17 and the number of persons belonging to this share (or income) is small. It means the income distribution will have a minimum near the center point and few persons are positioned around the medium income range. Such extreme distribution is hard to find in the real world because observed income distributions have a fat right tail distribution with a single maximum, but with no minimum near the center point. The following Fig. 20 shows U.S. CPS data of 1990 family income. It has a single maximum value.

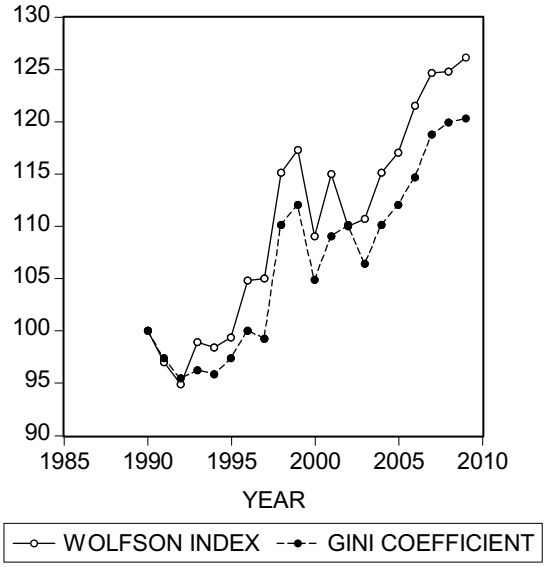
Fig. 20 U.S. Income Histogram



Series: CPS90	
Sample 1 66227	
Observations 66227	
Mean	32264.34
Median	25050.00
Maximum	410998.0
Minimum	-21898.00
Std. Dev.	28372.49
Skewness	2.152119
Kurtosis	11.73837
Jarque-Bera	261832.6
Probability	0.000000

If no minimum value is observed near the medium income point, then the scalar polarization index will move along the same direction as the Gini coefficient changes. Fig. 21 shows the movement of Korean family income from 1990 to 2009. Both indices move along the same direction and the scalar polarization index showed no particular extra information.

Fig. 21 Comparison of Korean Gini and Wolfson indices (Year 1990=100)



9. The share function derived by the maximum entropy method describes income shares of the poorest group fairly well.

The Gini coefficient was often criticized for the inability to describe income share changes of the poorest group accurately. See Ryu and Slottje (1996, 1998). However, Bonferroni (1930) and Ryu (2008) explain the Bonferroni index that describes the changes of the poorest group relatively well.

The area contribution to the Gini coefficient from the accumulated income shares of the poorest group is so small that the changes of the income shares of the poorest group derived from the small changes of the Gini coefficient may be difficult to catch. However, the share functions derived in this paper clearly show how income shares of the poorest decrease when the Gini coefficient increases.

Table 3: Estimated share functions for various values of Gini coefficients

	Gini＝0.2	Gini＝0.25	Gini＝0.3	Gini＝0.35
Z＝0.005	0.004131156	0.003505265	0.002913465	0.002356813
Z＝0.015	0.004266775	0.003643756	0.00304936	0.002484685
Z＝0.025	0.004402431	0.003782277	0.003185563	0.002613357
Z＝0.035	0.004537975	0.003920607	0.003321787	0.002742495
Z＝0.045	0.004673267	0.004058536	0.003457756	0.002871773
Z＝0.055	0.004808169	0.004195864	0.003593205	0.003000883
Z＝0.065	0.004942553	0.004332401	0.003727888	0.003129529
Z＝0.075	0.005076296	0.004467972	0.003861575	0.003257436
Z＝0.085	0.005209283	0.004602416	0.003994052	0.003384349
Z＝0.095	0.005341407	0.004735583	0.004125128	0.003510037

	Gini＝0.4	Gini＝0.45	Gini＝0.5	Gini＝0.55
Z＝0.005	0.001834136	0.001343614	0.0008846485	0.0004633178
Z＝0.015	0.001949074	0.00144138	0.0009615506	0.0005154733
Z＝0.025	0.002065422	0.001541206	0.001041132	0.0005707808
Z＝0.035	0.002182833	0.001642773	0.001123152	0.0006291421
Z＝0.045	0.002300966	0.001745762	0.001207356	0.0006904352
Z＝0.055	0.00241949	0.001849856	0.001293484	0.0007545172
Z＝0.065	0.002538088	0.001954744	0.00138127	0.0008212273
Z＝0.075	0.002656459	0.002060127	0.00147045	0.0008903898
Z＝0.085	0.002774321	0.002165719	0.001560762	0.0009618175
Z＝0.095	0.002891415	0.002271251	0.001651954	0.001035315

10. Conclusion and Summarizing Remarks

In this paper, the usefulness of the Gini coefficient is reemphasized. The knowledge of the Gini coefficient is equivalent to the knowledge of the first moment of the share function. Beginning from the first moment, higher moments can be estimated with a high accuracy with regression equations derived for the quintile data of 57 countries reported in the World Development Report (2011). Once higher moments (up to 8^{th} moments) are known, underlying income distributions can be estimated with the maximum entropy method of Ryu (1993). The derived functional forms follow a relatively systematic pattern since income distributions are derived only from the given

Gini coefficients. Two Lorenz curves corresponding to different Gini coefficients will not meet at the center that the Lorenz dominance effect will not be observed. Furthermore, the Gini coefficient and the scalar polarization index of Wolfson (1994) will move toward the same direction for the 57 countries reported in the World Development Report (2011) that the scalar polarization index may be a redundant measure once the Gini coefficient is given. These results came from a comparison of approximated Lorenz curves but not from a comparison of empirical Lorenz curves.

The motivation for deriving the share function and corresponding Lorenz curves from the Gini coefficients are following. If raw data is available, no need to consider the functional form of the Lorenz curves. If quintile data is available then use the quadratic extension method described in (9)−(12) or use the higher moments described in (17) and the maximum entropy method of (28) to derive the share functions and the Lorenz curves. However, if we want to summarize the raw data with one number (inequality measure) then the Gini coefficient is an excellent measure because this measure carries most of the raw data information such that reasonable reproduction of the share function is possible. Though the Lorenz curve derived from the Gini coefficient will not be the most accurate one because only one summary measure is used, it will have some standard shape. If two numbers are allowed to summarize the raw data, Yitzhaki (1998) showed that approximation of poor or rich income groups can be improved. If many numbers are allowed to summarize the raw data, Ryu and Slottje (1996) used orthonormal basis coefficient to summarize the raw data. Quintile data is another way to summarize the raw data. The whole point is that if a good approximation can be derived from one summary measure (the Gini coefficient in this case), then it will have the standard shape of income distributions. Any correction due to extra information will be minor correction to the standard shape.

There are several limitations in applying the method of this paper for practical purposes. 1) The data reported in the World Development Report (2011) is not current for some unknown reason. The reported years are different for different countries. 2) It will be convenient if the income distribution can be derived directly from the given Gini coefficient; however, the method introduced in this paper requires a sequence of calculations. a) The higher moments of distribution are estimated from the given Gini coefficient, b)

the parameters of the share functions are estimated (subject to the given moments) and then c) income distribution is derived with the estimated parameters.

In the future, several remaining questions can be analyzed. Different countries are at different development stages. Why do income distributions of different countries have the similar shapes that depend only on the Gini coefficients? Why do higher moments of the income share function depend only on the first moment (which is equivalent to knowledge of the Gini coefficients)?

References

Aroian L. (1948), "The fourth degree exponential distribution function", *Annals of Mathematical statistics* 19, pp. 589-592.

Bonferroni, C. (1930), *Elemente di Statistica Generale*, libereria Seber, Firenze.

Esteban, J. and D. Ray (1994), "On the Measurement of Polarization", *Econometrica* vol. 62, no.4, pp. 819-851.

Kakwani, N. and N. Podder, (1973), On the estimation of Lorenz curves from grouped observations", *International Economic Review* 14, No. 2, pp. 278-292.

Mead, L. and N. Papanicolaou, (1984), "Maximum entropy in the problem of moments", *Journal of Mathematical Physics* 25, pp. 2404-2417.

Ryu, H. (1993), "Maximum entropy estimation of density and regression functions", *Journal of Econometrics* 56, pp. 397-440.

Ryu, H. (2008), "Maximum Entropy Estimation of Income Distributions from Bonferroni Indices", pp. 193-211, *in Modeling Income Distributions and Lorenz Curves* edited by Duangkamon Chotikapanich, Springer.

Ryu, H. and D. Slottje, (1998), *Measuring Trends in U.S. Income Inequality*, Theory and Applications, Springer, New York.

Ryu, H. and D. Slottje, (1996), "Two Flexible Functional Form Approaches for Approximating the Lorenz Curve", *Journal of Econometrics* 72, pp. 251-274.

Singh, S. and G. Maddala, (1976), "A function for size distribution of incomes", *Econometrica* 44, No. 5, pp. 963-970.

Wolfson, M. (1994), "Conceptual issues in Normative Measurement: When Inequalities Diverge", *American Economic Review* vol. 84, no. 2, pp. 353-358.

World Development Report, (2011), The World Bank.

Yitzhaki, S. (1998), "More than a dozen alternative ways of spelling Gini", *Research on Economic Inequality* 8, pp. 13-20.