Economics Bulletin

Volume 32, Issue 4

Evaluation of the goodness of fit of new statistical size distributions with consideration of accurate income inequality estimation

Masato Okamoto
Ministry of Internal Affairs and Communications

Abstract

This paper compares the goodness-of-fit of two new types of parametric income distribution models (PIDMs), the kappa-generalized (kG) and double-Pareto lognormal (dPLN) distributions, with that of beta-type PIDMs using US and Italian data for the 2000s. A three-parameter model kG tends to estimate the Lorenz curve and income inequality indices more accurately when the likelihood value is similar to that of the beta-type PIDMs. For the first half of the 2000s in the USA, the kG outperforms the other PIDMs in goodness-of-fit evaluated by both frequency-based criteria (such as the maximum likelihood value) and money-amount-based criteria (such as accuracy of estimation of the Lorenz curve). A four-parameter model dPLN generally outperforms the GB2 in both criteria. Furthermore, when the overall income distribution is approximated by a mixture of distributions fitted separately for each age class of the household heads (the 'MLE-by-Age' method), the goodness-of-fit of the dPLN mixture model is found to be comparable to or higher than that of all of the PIDMs in the ordinary MLE fit, and, in the overall evaluation, this mixture model outperforms all of the single PIDMs in the sense that it is better fitted according to at least in one of the two criteria in almost all cases. The dPLN and its mixture model are found to have an explicit analytic expression for the Gini coefficient. The dPLN is therefore also suitable for the MLE-by-Age method in this respect.

Citation: Masato Okamoto, (2012) "Evaluation of the goodness of fit of new statistical size distributions with consideration of accurate income inequality estimation", *Economics Bulletin*, Vol. 32 No. 4 pp. 2969-2982.
Contact: Masato Okamoto - mokamoto2@soumu.go.jp.
Submitted: August 31, 2012. **Published:** October 25, 2012.

1. Introduction

Parametric income distribution models (PIDMs) are used for various objectives. Examples are estimation of income distribution and inequality/poverty indices from grouped data when survey micro data are unavailable, construction of regression models for economic analysis using, e.g., the Mincer equation (typically with a lognormal error distribution), and study of the mechanisms that generate income distributions. Various models have been proposed, but the search for models that provide a better fit is ongoing. The novelties of this paper are as follows: i) Two promising new PIDMs – the κ-generalized (κG) distribution (Clementi *et al*. 2007) and the double-Pareto lognormal (dPLN) distribution (Reed 2003) – are compared with the existing beta-type PIDMs simultaneously in terms of goodness-of-fit; ii) not only frequency-based (FB) measures (such as the likelihood value) but also money-amount-based (MAB) measures (such as the accuracy of estimation of the Lorenz curve) are employed in the goodness-of-fit evaluation; and iii) the combined effect, using a mixture model approach in which the overall income distribution is approximated by a mixture of distributions separately fitted by subgroup (age class of household heads) is also investigated to address the issue of heterogeneity in population composition. When PIDMs are fitted to empirical data for the purpose of estimating income inequality, a sufficiently high goodness-of-fit is required in terms of MAB criteria. Even if the estimation of income inequality is not an explicit objective, better MAB evaluation is desirable in addition to better FB evaluation. However, the goodness-of-fit is evaluated using only FB measures in most cases in the literature. The empirical results in this paper demonstrate that superior FB evaluation does not necessarily imply higher accuracy in the inequality indices when new types of PIDMs are introduced. Some PIDMs such as the dPLN are derived under specific assumptions regarding heterogeneity in population composition. If the assumptions are not valid, then the heterogeneity may adversely affect the goodness-of-fit of the models. Use of a mixture model approach is one way of overcoming the heterogeneity issue; however, the ordinary fitting method for mixture models is generally difficult to apply, and the results are often unstable. Instead, a PIDM is simply fitted to each subgroup using the maximum likelihood estimation (MLE), and the overall income distribution is approximated by a mixture of distributions fitted to the subgroups in this paper. The empirical results show that, when separately fitted by the age class of household heads (the 'MLE-by-Age' method) to US and Italian income data, the dPLN mixture model attains a fit comparable to or better fitting than that obtained by fitting all of the PIDMs in the ordinary MLE in terms of both FB and MAB criteria. Furthermore, in the overall evaluation, the empirical results show that the dPLN mixture model outperforms all of the PIDMs in the ordinary MLE in the sense that the dPLN mixture model is better fitted according to at least one of the two criteria in almost all cases. By contrast, (single) four-parameter models, the dPLN and the generalized beta distribution of

the second kind (GB2, McDonald 1984), and their generalizations with five parameters, the generalized dPLN (GdPLN) distribution (Reed and Wu 2008) and the generalized beta (GB) distribution (McDonald 1995) show no clear improvement over three-parameter models in the overall evaluation of the goodness-of-fit. The dPLN is also suitable for the MLE-by-Age method in that its mixture model has an explicit analytic expression for the Gini coefficient.

2. Statistical size distributions to be compared

The GB2 and special cases thereof are popular as well-fitted PIDMs. The probability density function (pdf) of the GB2 is given by:

$$f_{\mathrm{GB2}}(x;a,b,p,q)=\frac{ax^{ap-1}}{b^{ap}B(p,q)[1+(x/b)^{a}]^{p+q}}. \tag{1}$$

The GB2 is identical to the Singh-Maddala (SM) distribution (Singh and Maddala 1975) when $p=1$ and to the Dagum (Da) distribution (Dagum 1977) when $q=1$.

Reed (2003) demonstrates that if each individual/household log income follows Brownian motion, $d\log X_t=\mu dt+\sigma dB_t$, where B_t denotes standard Brownian motion, and its elapsed time from birth (entry into the labor market) to observation follows the exponential distribution, then the income distribution follows the dPLN. The pdf of the dPLN can be expressed as follows:

$$\begin{aligned} f_{\mathrm{dPLN}}(x;\mu,\sigma^2,\alpha,\beta) &= \frac{\alpha\beta}{\alpha+\beta}\left[x^{\beta-1}e^{-\beta\mu+\beta^2\sigma^2/2}\Phi^{c}\left(\frac{\log x-\mu+\beta\sigma^2}{\sigma}\right)\right. \\ &\left.+\,x^{-\alpha-1}e^{\alpha\mu+\alpha^2\sigma^2/2}\Phi\left(\frac{\log x-\mu-\alpha\sigma^2}{\sigma}\right)\right], \end{aligned} \tag{2}$$

where $\alpha,\beta,\sigma>0$, Φ denotes the cumulative distribution function (cdf) of the standard normal distribution and $\Phi^{c}:=1-\Phi$. The dPLN is equivalent to the distribution of the product of two mutually independent random variables that follow the double-Pareto and lognormal distribution, respectively. The log-transformed random variable of the dPLN therefore follows the normal Laplace (NL) distribution, given by the convolution of the Laplace and normal distributions. The lower and upper tails of the dPLN follow the Pareto law,

$$f_{\mathrm{dPLN}}(x)\sim c_1x^{-\alpha-1}\ (x\to\infty),\qquad f_{\mathrm{dPLN}}(x)\sim c_2x^{\beta-1}\ (x\to 0), \tag{3}$$

where c_1 and c_2 are positive constants. It should be noted that the assumption of regularity of the cohort's birth and death is not necessarily required for emergence of the dPLN. If an independent transitory component following an (asymmetric) zero-reverting diffusion process is introduced, then the income distribution in any homogenous group approaches the dPLN after sufficiently long time (see Toda 2012).

Reed and Wu (2008) demonstrate empirically that the dPLN outperforms the GB2 in

various countries using FB criteria. They mention that the Gini coefficient of the dPLN has no closed-form expression; however, the closed-form expression does, in fact, exist. Explicit formulae for the mean log deviation (MLD) and Theil index of the dPLN can be derived from the formula for the higher moments (Reed and Wu 2008), applying the procedure employed for the respective indices of the GB2 suggested by Jenkins (2007). These formulae are presented in the Appendix with an implicit expression for the Lorenz curve of the dPLN.

Reed and Wu (2008) also demonstrate empirically that a generalization of the dPLN substantially improves the goodness-of-fit evaluated using FB measures in various countries. The log-transformed random variable of the generalized dPLN (GdPLN) is equivalent to the generalized NL distribution which has the following characteristic function (cf):

$$\phi_{\mathrm{GNL}}(s) = \left[\frac{\alpha\beta}{(\alpha - is)(\beta + is)} e^{i\mu s - \sigma^2 s^2/2}\right]^{\rho}. \tag{4}$$

In formula (4), the cf of the NL distribution is derived when $\rho = 1$. As neither the pdf nor the cdf of the GdPLN has a closed-form expression, the inverse transformation of the cf in (4) must be performed numerically when fitting the GdPLN to empirical data.

Clementi *et al*. (2007) derive the κG distribution by 'Weibullizing' the κ-exponential function, $\exp_\kappa(x) \coloneqq \left(\sqrt{1+\kappa^2 x^2} + \kappa x\right)^{1/\kappa}$. This deformed exponential function is a product of the generalized entropy studies in thermostatistics. Clementi *et al*. show empirically that, in terms of FB criteria, the κG is fitted better than the SM and Da to equivalized disposable incomes obtained from the US panel survey PSID and the German panel survey GSOEP. The κG has the following pdf:

$$f_{\kappa\mathrm{G}}(x; a, b, \kappa) = \frac{a x^{a-1} \exp_\kappa(-(x/b)^a)}{b^a \sqrt{1+\kappa^2 (x/b)^{2a}}}, \tag{5}$$

where $a, b > 0$ and $0 \le \kappa < 1$. The cdf is given by $F_{\kappa\mathrm{G}}(x) = 1 - \exp_\kappa(-(x/b)^a)$. Note that $\lim_{\kappa\to 0} \exp_\kappa(x) = e^x$. The lower and upper tails of the κG follow the Pareto law,

$$f_{\kappa\mathrm{G}}(x) \sim k_1 x^{-\frac{a}{\kappa}-1} \ (x \to \infty), \qquad f_{\kappa\mathrm{G}}(x) \sim k_2 \left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a} \ (x \to 0), \tag{6}$$

where k_1 and k_2 are positive constants. The κG has explicit analytic expressions for its MLD, Theil index, Gini coefficient and Lorenz curve (Clementi *et al.* 2009).

3. Data and Methods

The PIDMs mentioned in the previous section are fitted to micro data from the US Survey of Consumer Finances (SCF) and the Italian Survey of Household and Wealth (SHIW), both conducted in 2000 – 2010. Owing to limited data availability, the gross income before taxation is used for the SCF, and the disposable income after taxation is used for the SHIW. The results for the equivalized personal incomes, calculated by dividing each household

income by the square root of the number of household members are presented in this paper. The corresponding results for unadjusted household income generally display similar trends. Points of differences between both results are described in the next section.[1] The corresponding results for the Japanese income distribution, obtained using grouped data (Okamoto 2012b), are also briefly described in footnote in the next section for reference.

The maximum likelihood estimation (MLE) method is employed to fit the PIDMs. The following log likelihood value is maximized:

$$l(\boldsymbol{\theta}; x) = \sum_{i=1}^{n} w_i \log f(x_i; \boldsymbol{\theta}), \tag{7}$$

where f denotes the pdf of the PIDM, $\boldsymbol{\theta}$ is the set of PIDM parameters, n is the sample size. Each sample household is assumed to earn an (equivalized) income x_i and be assigned a weight w_i for tabulation purposes (multiplied by the number of household members in the case of equivalized personal income), which is normalized to unity on average. In addition to the MLE, a variant method allowing for the exclusion of effects owing to age heterogeneity is employed. This method uses a mixture model approach in which the PIDMs are fitted for each age class of the household heads and the overall income distribution is approximated by a mixture of the fitted distributions (the 'MLE-by-Age' method). Fitting by household size and region were also evaluated, but the MLE-by-Age method outperforms those alternatives.

To compare the goodness-of-fit of the three- and four-parameter PIDMs, the likelihood value is converted into a value of Schwarz's Bayesian Information Criterion (BIC):

$$\mathrm{BIC} = -2l + \log n \cdot \#\boldsymbol{\theta}, \tag{8}$$

where $\#\boldsymbol{\theta}$ denotes the number of parameters in the PIDM of interest. The value of $\log n$ ranges from 8.0 to 8.8 for the SCF and is approximately 9.0 for the SHIW. For comparisons between PIDMs with the same number of parameters, the bootstrap method is applied to test the significance of the differences at the one-sided 5 percent level. [2]

As for the MLE-by-Age method, comparisons are performed using the likelihood value of the mixture distribution,

$$\tilde{f}\left(x; \{\widehat{\boldsymbol{\theta}}_g\}_g\right) = \sum_g p_g f\left(x; \widehat{\boldsymbol{\theta}}_g\right), \tag{9}$$

called the 'synthetic' value in this paper and its corresponding BIC value. In formula (9), $p_g = \sum_{i \in G_g} w_i / \sum_{i=1}^{n} w_i$ is the population share of age class g, and $\widehat{\boldsymbol{\theta}}_g$ denotes the set of

[1] Strictly speaking, the survey units of the SCF are 'primary economic units' (PEUs), economically independent subunits within households (see Bricker *et al*. 2012). Equivalized personal income is defined as the income of the PEU divided by the square root of the number of household members belonging to the PEU.

[2] In the SCF, 999 sets of replicate weights are available. Iterative calculations with different sets of replicate weights are equivalent to the bootstrap procedure consistent with the survey design. The simple bootstrap method is performed for the SHIW. Approximately 300 sets of replicate weights are made available from the 2008 SHIW. The replicate weights enable us to perform the Jackknife procedure consistent with the survey design. The results of the comparison do not appear significantly different from those based on the simple bootstrap method.

parameter values for class g estimated by the MLE. The sample households are classified into four age classes containing roughly equal numbers of households: 44 years or younger, 45 – 54, 55 – 64 and 65 years or older. In the BIC calculation of the MLE-by-Age results, the synthetic likelihood value is penalized based on the number of PIDM parameters. The penalty is not multiplied by the number of age classes because no biases in the synthetic likelihood value derived using the MLE-by-Age method relative to that derived using the ordinary MLE is observed in simulation owing to the difference in the total number of parameters. A notable difference between the ordinary likelihood and synthetic one is that the value of latter for the GB2 may be lower than that for its special cases such as the SM.

The reason for the choice of the BIC is that, according to the bootstrap simulation, there are cases where the synthetic likelihood value of a three-parameter PIDM can be higher than that of a four-parameter PIDM with a probability above 5 percent if the latter is judged to be better fitted by Akaike's Information Criterion (AIC). Other FB measures, such as the sum of squared errors of the cdf, display similar results with the (synthetic) likelihood value and therefore omitted from the description.

To incorporate accuracy in income inequality estimation into the goodness-of-fit evaluation, the square root of the sum of the squared errors of the Lorenz curve (L-RSE)[3],

$$\sqrt{\sum_{i=1}^{n}\left(\hat{L}_i - L\left(c_i;\widehat{\boldsymbol{\theta}}\right)\right)^2} \tag{10}$$

and the estimation errors of four major inequality indices (the Gini coefficient, coefficient of variation (CV), MLD and Theil index) are used as MAB measures. In formula (10), the incomes x_i are assumed to be arranged in ascending order, and the empirical Lorenz curve, $\hat{L}_i = \sum_{j\le i} w_j x_j / \sum_{j=1}^{n} w_j x_j$, and Lorenz curve of the PIDM, $L\left(c;\widehat{\boldsymbol{\theta}}\right) = \int_0^{F^{-1}(c)} yf\left(y;\widehat{\boldsymbol{\theta}}\right)dy \Big/ \int_0^{\infty} yf\left(y;\widehat{\boldsymbol{\theta}}\right)dy$, are compared at cumulative population shares $c_i = \sum_{j\le i} w_j / \sum_{j=1}^{n} w_j$, $i = 1, \cdots, n$. The term 'error' is used in this paper to refer to the deviation from the empirical respective statistic directly calculated from the survey data (Table 1). Although the L-RSE obtained using the MLE and MLE-by-Age methods may be affected by the number of parameters in the PIMDs, as is the likelihood value, the proper penalty for the number of parameters is not clear. In practice, it is frequently observed that the L-RSE of the GB2 is larger than that of the SM and Da. The L-RSEs of the three- and four-parameter PIDMs are therefore directly compared without adjustment in this paper, considering that it is meaningful to test statistically whether the L-RSEs of the

[3] The errors at different points on the Lorenz curve are not independent of each other. It may be argued that the correlations should be taken into account in measuring the estimation accuracy. However, estimating the correlations is nontrivial, particularly near either ends of the Lorenz curve. Parametric models for the Lorenz curve are therefore fitted using the simple OLS in many cases (*cf*. Kakwani 1980). No problems with this method have been identified in empirical applications thus far.

three-parameter PIDM is equal or smaller than that of the four-parameter PIDM for the MAB evaluation. Significance tests of the accuracy of the inequality index estimates are omitted because those estimates are presented as supplements to demonstrate the appropriateness of the L-RSE as a MAB measure.

The explicit analytic expression given in the Appendix can be used for the calculation of the Gini coefficient of the dPLN mixture distribution derived using the MLE-by-Age method, whereas numerical computation is required for the other PIDM mixture distributions.[4]

Table 1. Empirical inequality indices

Survey	Year	Gini	CV	MLD	Theil
SCF	2001	0.538	3.243	0.542	0.757
	2004	0.516	2.924	0.495	0.660
	2007	0.559	4.170	0.573	0.843
	2010	0.534	3.448	0.511	0.695
SHIW	2000	0.328	0.729	0.197	0.192
	2002	0.322	0.695	0.193	0.184
	2004	0.329	0.823	0.194	0.204
	2006	0.323	0.840	0.190	0.200
	2008	0.324	0.726	0.187	0.188
	2010	0.327	0.682	0.197	0.186

4. Results

The goodness-of-fit of the PIDMs in the MLE analysis is presented in Table 2. The κG attains the minimum BIC value for all of the PIDMs in the first half of the decade 2000–2010 for the SCF and attains the minimum or near-the-minimum L-RSE[5] for the entire period for both surveys. On average, the κG also attains the minimum errors in estimation of the major inequality indices. For the similar likelihood/BIC values, the κG tends to yield smaller L-RSEs and errors in inequality measures than the Da and SM.

The GB2 suffers from larger BIC values than the Da for both surveys and larger L-RSEs than the κG for the SCF. The dPLN displays a goodness-of-fit superior to that of the GB2. Nevertheless, its BIC values and L-RSEs are several times larger than those of the Da or κG. Both PIDMs therefore fail to attain goodness-of-fit that is clearly superior to that of the three-parameter PIDMs.

More extensive models with five parameters, the GB and GdPLN, were also evaluated;

[4] For the SCF, five sets of income data are provided because the multiple imputation method is adopted. The (synthetic) likelihood value, Gini coefficient, MLD and Theil index are calculated for each set and then averaged. As for the corresponding five calculations for the L-RSE and CV, the square root of the mean-squared values is adopted.

[5] For simplicity, the term 'near the minimum' is used to mean that the value is not significantly different from the minimum in this paper.

however, the GB does not improve the BIC values for both surveys, and the GdPLN does not improve the BIC values for the SCF. The GdPLN improves the BIC values substantially for the SHIW, but its L-RSEs are unstable, as shown in Table 3. The GdPLN therefore cannot be regarded as superior to the dPLN in the overall evaluation. The results imply that the estimation stability of the Lorenz curve and inequality indices must be ensured even if the GdPLN displays remarkable improvement in goodness-of-fit over the dPLN in terms of the FB criteria.

The dPLN becomes the best-fit model when combined with the MLE-by-Age method, as shown in Table 2. The BIC values significantly decrease relative to those in the ordinary MLE except for the 2010 SCF. [6] As for the SCF, the L-RSEs also decrease significantly relative to those for the MLE except for the 2010 survey. If including insignificant changes, the L-RSEs decrease in all cases except for the 2002 SHIW. The BIC value and L-RSE for the dPLN mixture model are significantly below or near the minimums of all of the PIDMs in the MLE except for the BIC value for the 2001 SCF, and the value of which (for at least one of the two measures) is significantly below the minimum in the MLE in all cases.

Table 2(a). Goodness-of-fit of the PIDMs – SCF

Stat.	Year	MLE					MLE-by-Age				
		SM	Da	κG	GB2	dPLN	SM	Da	κG	GB2	dPLN
BIC [a]	2001	5.3	-3.9	-16.7*	-0.1	0.0	4.6	-9.6	-19.0#	-14.4	-13.6
	2004	-8.1*	-9.5*	-10.2*	-1.2	0.0	-2.8	-6.4	-12.8	-15.2##	-14.8##
	2007	-2.6*	-2.6*	5.8*	5.7	0.0	-11.4	-7.4	2.6	-24.5##	-26.3##
	2010	17.4	5.6	83.9	11.4	0.0*	20.6	12.3	77.8	6.7	-0.2#
L-RSE	2001	2.51	2.55	1.93*	2.03	1.84*	1.88	2.39	1.88	1.50	1.34##
	2004	1.81	1.59	1.25*	1.66	1.50	1.59	1.75	1.48	1.27	1.14#
	2007	2.14	2.17	1.51*	2.19	1.85	1.19#	1.98	2.21	1.12	0.76##
	2010	0.30*	0.45*	0.59*	0.99	0.64*	0.50#	0.31#	0.88	0.71#	0.34#
Gini	2001	-0.039	-0.042	-0.032	-0.033	-0.030	-0.028	-0.039	-0.032	-0.025	-0.023
	2004	-0.029	-0.026	-0.021	-0.027	-0.024	-0.024	-0.029	-0.026	-0.021	-0.019
	2007	-0.034	-0.034	-0.026	-0.034	-0.028	-0.018	-0.032	-0.041	-0.024	-0.012
	2010	-0.002	-0.004	-0.011	-0.014	-0.009	0.006	0.001	-0.016	-0.010	-0.002
MLD	2001	-0.083	-0.084	-0.063	-0.065	-0.059	-0.064	-0.080	-0.063	-0.048	-0.042
	2004	-0.058	-0.049	-0.036	-0.052	-0.047	-0.052	-0.056	-0.046	-0.039	-0.035
	2007	-0.071	-0.073	-0.047	-0.073	-0.063	-0.040	-0.067	-0.073	-0.037	-0.023
	2010	-0.003	-0.010	-0.008	-0.029	-0.019	0.015	0.002	-0.017	-0.020	-0.006
Theil	2001	-0.239	-0.239	-0.179	-0.190	-0.172	-0.163	-0.218	-0.165	-0.126	-0.105
	2004	-0.176	-0.155	-0.120	-0.162	-0.147	-0.150	-0.166	-0.139	-0.116	-0.101
	2007	-0.211	-0.215	-0.135	-0.217	-0.180	-0.046	-0.181	-0.194	-0.066	-0.006
	2010	0.001	-0.023	-0.011	-0.080	-0.039	0.081	0.017	-0.032	-0.041	0.021

[6] The ordinary log maximum likelihood value in the MLE-by-Age method, i.e. the sum of the log maximum likelihood values for all age classes, is significantly higher than the log maximum likelihood value in the ordinary MLE in terms of the likelihood ratio test and BIC in all cases.

Table 2(b). Goodness-of-fit of the PIDMs – SHIW

Stat.	Year	MLE					MLE-by-Age				
		SM	Da	κG	GB2	dPLN	SM	Da	κG	GB2	dPLN
BIC [a]	2000	31.7	-1.4*	3.0*	7.4	0.0	19.0	-26.5##	-3.3	-23.0†	-26.2†
	2002	44.4	-0.9*	9.8	6.2	0.0	29.4	-26.8##	0.5	-22.4†	-26.7†
	2004	8.7	0.6	6.7	9.0	0.0*	6.1	-13.9##	3.0	-7.7†	-13.8†
	2006	24.3	2.8	11.1	11.7	0.0*	5.2	-15.9†	-7.0	-11.6	-19.4##
	2008	3.8	1.9	8.3	7.7	0.0*	-6.7	-14.7†	6.5	-24.4##	-32.2##
	2010	35.1	-2.1*	4.8*	6.9	0.0	3.2	-27.6##	-2.5	-20.7†	-25.3†
L-RSE	2000	0.251*	0.118*	0.098*	0.104*	0.095*	0.183	0.140#	0.171#	0.095#	0.080#
	2002	0.199*	0.116*	0.135*	0.133*	0.123*	0.175#	0.108#	0.145#	0.162#	0.154#
	2004	0.323*	0.221*	0.224*	0.264*	0.238*	0.291	0.257#	0.292#	0.257	0.218#
	2006	0.379*	0.302*	0.254*	0.285*	0.287*	0.356#	0.388#	0.238#	0.230#	0.225#
	2008	0.174*	0.123*	0.125*	0.136*	0.125*	0.136#	0.193#	0.150#	0.139#	0.086#
	2010	0.164*	0.094*	0.100*	0.100*	0.103*	0.142	0.130#	0.091#	0.082#	0.084#
Gini	2000	-0.003	0.000	0.000	0.000	0.000	0.000	-0.001	-0.002	0.000	0.000
	2002	0.000	0.001	0.003	0.003	0.002	0.002	0.001	0.003	0.003	0.003
	2004	-0.005	-0.003	-0.003	-0.004	-0.003	-0.004	-0.004	-0.004	-0.004	-0.003
	2006	-0.005	-0.004	-0.003	-0.004	-0.004	-0.005	-0.006	-0.003	-0.003	-0.003
	2008	-0.002	0.001	0.000	-0.001	-0.001	0.001	0.003	-0.001	-0.002	0.000
	2010	0.000	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
MLD	2000	-0.009	-0.003	-0.002	-0.003	-0.003	-0.008	-0.004	-0.006	-0.002	-0.002
	2002	-0.009	-0.003	-0.002	-0.001	-0.002	-0.007	-0.004	-0.002	0.000	0.000
	2004	-0.010	-0.006	-0.004	-0.007	-0.007	-0.009	-0.006	-0.005	-0.006	-0.006
	2006	-0.014	-0.009	-0.007	-0.009	-0.010	-0.013	-0.011	-0.007	-0.007	-0.008
	2008	-0.005	0.000	0.001	-0.003	-0.003	-0.002	0.002	0.000	-0.003	-0.001
	2010	-0.007	-0.001	-0.001	0.000	-0.001	-0.003	0.001	0.000	0.000	0.000
Theil	2000	-0.012	-0.004	-0.002	-0.003	-0.003	-0.007	-0.005	-0.006	-0.002	-0.001
	2002	-0.006	0.000	0.004	0.004	0.003	-0.003	0.000	0.005	0.007	0.006
	2004	-0.020	-0.014	-0.012	-0.016	-0.016	-0.019	-0.016	-0.015	-0.016	-0.014
	2006	-0.026	-0.020	-0.017	-0.020	-0.020	-0.024	-0.023	-0.015	-0.016	-0.015
	2008	-0.009	-0.001	-0.001	-0.006	-0.005	0.000	0.003	-0.001	-0.005	-0.001
	2010	-0.005	0.003	0.003	0.003	0.003	0.002	0.005	0.004	0.003	0.004
CV	2000	-0.062	-0.024	-0.011	-0.018	-0.015	-0.022	-0.020	-0.021	-0.004	0.009
	2002	-0.035	-0.003	0.018	0.017	0.013	-0.017	-0.002	0.029	0.040	0.041
	2004	-0.130	-0.098	-0.079	-0.109	-0.103	-0.121	-0.104	-0.091	-0.100	-0.092
	2006	-0.175	-0.148	-0.125	-0.144	-0.145	-0.161	-0.156	-0.102	-0.097	-0.095
	2008	-0.050	-0.011	-0.003	-0.034	-0.029	0.019	0.026	0.012	-0.020	0.000
	2010	-0.018	0.020	0.026	0.022	0.022	0.022	0.034	0.036	0.028	0.034

Note: The errors in the CV estimates for the SCF are omitted because the CVs of the fitted PIDMs are infinite in most cases. [a] The deviations from the corresponding BIC values of the dPLN in the MLE fitting are presented. * The minimum among the PIDMs in the MLE fitting or (statistically) near the minimum. # The minimum among the PIDMs in the MLE-by-Age fitting or near the minimum (excluding the cases corresponding to '##'). ## In addition to satisfy the condition '#', significantly lower than the minimum in the MLE fitting in cases where the minimum in the MLE is attained by a PIDM with the same number of parameters, or lower than the minimum in the MLE fitting in cases where the minimum in the MLE is attained by a PIDM with a larger number of parameters. † Of those not the minimum/near-the-minimum in the MLE-by-Age fitting but significantly lower than the minimum in the MLE fitting in cases where the minimum in the MLE is attained by a PIDM with the same number of parameters, or lower than the minimum in the MLE fitting in cases where the minimum in the MLE is attained by a PIDM with a larger number of parameters.

Table 3. Goodness-of-fit of the four- and five-parameter PIDMs – SHIW

Year	BIC [a]				L-RSE			
	GB2	dPLN	GB	GdPLN	GB2	dPLN	GB	GdPLN
2000	7.4	0.0	15.3	-27.6	0.104	0.095	0.103	0.155
2002	6.2	0.0	8.7	-70.1	0.133	0.123	0.113	0.076
2004	9.0	0.0	18.0	2.9	0.264	0.238	0.267	0.328
2006	11.7	0.0	18.9	-13.7	0.285	0.287	0.272	0.285
2008	7.7	0.0	16.7	-34.7	0.136	0.125	0.137	0.132
2010	6.9	0.0	15.5	-24.7	0.100	0.103	0.090	0.096

Note: The Lorenz curve of the GB is computed numerically. [a] The deviations from the corresponding BIC values of the dPLN in the MLE fitting are presented.

As for the SHIW, the MLE-by-Age method also substantially improves the BIC values for the Da. In particular, the BIC values are lower than those of the dPLN mixture model for 2000 – 2004 and 2010. However, averaged over the entire period, the BIC value for the Da mixture model is higher than that of the dPLN mixture model, and the L-RSEs for the Da mixture model are larger than those for the Da in the MLE and significantly inferior to those of the dPLN mixture model on average. The GB2 is also improved using the MLE-by-Age; however, the dPLN still maintains its superiority to the GB2. When comparing the goodness-of-fit of all of the PIDMs across the MLE and MLE-by-Age methods, on average over the entire period, the dPLN mixture model attains the minimum BIC and L-RSE, as well as the minimum errors or nearly minimum errors in the inequality indices.

In the case of the unadjusted household income, the dPLN similarly improves using the MLE-by-Age, whereas the κG suffers from poor goodness-of-fit in terms of the BIC. Nevertheless, for the SCF, the κG attains the minimum L-RSE for all PIDMs in the MLE.[7]

5. Concluding Remarks

Most of the existing major PIDMs belong to the GB family. The emergence of new types of PIDMs such as the κG and dPLN raises the question of the appropriateness of goodness-of-fit evaluation performed using FB measures (such as the maximum likelihood value) alone. This paper proposes the L-RSE, an accuracy measure based on the Lorenz curve, as a MAB measure. The empirical results indicate that the L-RSE complements the FB measures well for a more comprehensive goodness-of-fit evaluation.

Although the search for better-fitted PIDMs should continue, heterogeneity in the

[7] As for the Japanese size distribution of unadjusted household gross incomes of two-or-more-person households, the κG is inferior to the SM and Da in terms of both likelihood value and L-RSE. The SM is better fitted to the Japanese income distribution than the Da. This fact may relate to the poor performance of the κG, which appears relatively close to the Da. The same speculation holds for unadjusted household incomes from the SHIW. The dPLN is so remarkably improved by the MLE-by-Age method in Japan, at least around 2000 and later, that it attains the highest goodness-of-fit in the overall evaluation. (Okamoto 2012b)

population may render such models difficult to obtain. This paper suggests that fitting-by-subgroup approaches, such as the MLE-by-Age method, or more sophisticated methods incorporating regression techniques can provide alternative ways to obtain better-fitted models.

References

Bricker, J., A. B. Kennickell, K. B. Moore and J. Sabelhaus (2012) "Changes in U. S. family finances from 2007 to 2010: evidence from the Survey of Consumer Finances" Federal Reserve Bulletin 98.

Clementi, F., M. Gallegati and G. Kaniadakis (2007) "κ-generalized statistics in personal income distribution" *European Physical Journal B* **52**, 187–193.

Clementi, F., M. Gallegati and G. Kaniadakis (2009) "A κ-generalized statistical mechanics approach to income analysis" *Journal of Statistical Mechanics* **2009**, P02037.

Dagum, C. (1977) "A new model of personal income distribution: specification and estimation" *Economie Appliquée* **30**, 413–437.

Jenkins, S. P. (2007) "Inequality and the GB2 income distribution" ISER Working Paper No 2007-12, University of Essex, Essex.

Kakwani, N. C. (1980) "One class of poverty measures" *Econometrica* **48**, 437–446.

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

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

Okamoto, M. (2012a) "The relationship between the equivalence scale and the inequality index and its impact on the measurement of income inequality" LIS working paper 575.

Okamoto, M. (2012b) "Comparison in goodness of fit between the double-Pareto lognormal distribution and the generalized beta distribution of the second type" mimeo.

Reed, W. J. (2003) "The Pareto law of incomes – an explanation and an extension" *Physica A*, **319**, 579–597.

Reed, W. J. and F. Wu (2008) "New four- and five-parameter models for income distributions" in *Modeling Income Distributions and Lorenz Curves* by D. Chotikapanich, Ed, Springer-Verlag, pp. 211–223.

Singh, S. K. and G. S. Maddala (1975) "A function for the size distribution of incomes" *Econometrica* **44**, 963–970.

Toda, A. A. (2012) "The double power law in income distribution: explanations and evidence" *Journal of Economic Behavior and Organization*. (in press)

Appendix. Inequality indices and Lorenz curve of the double-Pareto lognormal (dPLN) distribution

The derivation of the analytic expressions for the Gini coefficient of the dPLN and dPLN mixture distributions requires the following formulae: the pdf, cdf and mean of the dPLN in (2), (A1) and (A2), a general formula for the Gini coefficient in (A3), and useful integral formulae in (A4) and (A5).

$$F_{\mathrm{dPLN}}(x;\mu,\sigma^2,\alpha,\beta) \tag{A1}$$
$$= \frac{\alpha\beta}{\alpha+\beta}\left[\frac{1}{\beta}x^{\beta}e^{\beta\mu+\beta^2\sigma^2/2}\Phi^c\left(\frac{\log x-\mu+\beta\sigma^2}{\sigma}\right)+\frac{1}{\beta}\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right.$$
$$\left.-\frac{1}{\alpha}x^{-\alpha}e^{\alpha\mu+\alpha^2\sigma^2/2}\Phi\left(\frac{\log x-\mu-\alpha\sigma^2}{\sigma}\right)+\frac{1}{\alpha}\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right],$$

$$M_{\mathrm{dPLN}} = \int_0^{\infty} x f_{\mathrm{dPLN}}(x)dx = \frac{\alpha\beta}{(\alpha-1)(\beta+1)}e^{\mu+\sigma^2/2}, \tag{A2}$$

$$G = 2\int_0^{\infty} x(F(x)-1/2)f(x)dx \Big/ \int_0^{\infty} xf(x)dx, \tag{A3}$$

$$\int x^{-a-1}\Phi\left(\frac{\log x-\mu}{\sigma}\right)dx = \frac{1}{a}\exp\left(-\mu a+\frac{\alpha^2\sigma^2}{2}\right), \tag{A4}$$

$$\int x^{-a-1}\Phi\left(\frac{\log x-\mu_1}{\sigma_1}\right)\Phi\left(\frac{\log x-\mu_2}{\sigma_2}\right)dx \tag{A5}$$
$$= \frac{1}{a}\left[\exp\left(\frac{a^2\sigma_1^2}{2}-a\mu_1\right)\Phi\left(-\frac{-\mu_1+\mu_2+a\sigma_1^2}{\sqrt{2}\sigma_{12}}\right)\right.$$
$$\left.+\exp\left(\frac{a^2\sigma_2^2}{2}-a\mu_2\right)\Phi\left(\frac{-\mu_1+\mu_2-a\sigma_2^2}{\sqrt{2}\sigma_{12}}\right)\right],$$

where $a>0$, $\sigma_{12}^2 := \frac{\sigma_1^2+\sigma_2^2}{2}$. The identity $\Phi^c(x)=\Phi(-x)$, and variable transformation $z=1/x$ are also employed for the derivation.

The Gini coefficient of the dPLN is given by:

$$G_{\mathrm{dPLN}} = \left[2\Phi\left(\sigma/\sqrt{2}\right)-1\right]+I, \tag{A6}$$

where I is defined as follows:

$$2\frac{(\alpha-1)(\beta+1)}{(\alpha+\beta)(\alpha-\beta-1)}\left[\begin{array}{l}-\frac{\beta}{(\alpha-1)(2\alpha-1)}e^{\alpha(\alpha-1)\sigma^2}\Phi\left(-\frac{2\alpha-1}{\sqrt{2}}\sigma\right)\\+\frac{\alpha}{(\beta+1)(2\beta+1)}e^{\beta(\beta+1)\sigma^2}\Phi\left(-\frac{2\beta+1}{\sqrt{2}}\sigma\right)\end{array}\right] \text{ if } \alpha\neq\beta+1,$$

$$2\frac{\alpha(\alpha-1)}{(2\alpha-1)^2}e^{\alpha(\alpha-1)\sigma^2}\left[\begin{array}{c}\left(\frac{1}{\alpha}+\frac{1}{\alpha-1}+\frac{2}{2\alpha-1}-(2\alpha-1)\sigma^2\right)\Phi\left(-\frac{2\alpha-1}{\sqrt{2}}\sigma\right)\\+\sqrt{2}\sigma\phi\left(-\frac{2\alpha-1}{\sqrt{2}}\sigma\right)\end{array}\right] \text{ if } \alpha=\beta+1,$$

where ϕ denotes the pdf of the standard normal distribution. The Gini coefficient of the dPLN mixture distribution, with the pdf $f(x)=\sum_g p_g f_{\mathrm{dPLN}}\left(x;\mu_g,\sigma_g^2,\alpha_g,\beta_g\right)$, can be

expressed as follows:

$$G_{\mathrm{MdPLN}} = \left(2\frac{J}{M_{\mathrm{MdPLN}}} - 1\right) + 2\frac{K}{M_{\mathrm{MdPLN}}}, \tag{A7}$$

$$M_{\mathrm{MdPLN}} = \sum\nolimits_g p_g \frac{\alpha_g \beta_g}{(\alpha_g - 1)(\beta_g + 1)} e^{\mu_g + \sigma_g^2/2},$$

$$J = \sum\nolimits_{g,h} p_g p_h \frac{\alpha_g \beta_g}{(\alpha_g - 1)(\beta_g + 1)} e^{\mu_g + \sigma_g^2/2} \Phi\left(\frac{\mu_g - \mu_h + \sigma_g^2}{\sqrt{2}\sigma_{gh}}\right),$$

$$K = \sum\nolimits_{g,h} p_g p_h \frac{\alpha_g \beta_g \alpha_h \beta_h}{(\alpha_g + \beta_g)(\alpha_h + \beta_h)} K_{gh},$$

where $\sigma_{gh}^2 := \frac{\sigma_g^2 + \sigma_h^2}{2}$ and K_{gh} is defined by:

$$-\frac{(\alpha_g + \beta_g)}{\alpha_h(\alpha_h - 1)(\alpha_g + \alpha_h - 1)(\alpha_h - \beta_g - 1)} e^{-(\alpha_h - 1)\mu_g + \alpha_h \mu_h + \alpha_h^2 \sigma_{gh}^2 - \frac{2\alpha_h - 1}{2}\sigma_g^2} \Phi\left(\frac{\mu_g - \mu_h - 2\alpha_h \sigma_{gh}^2 + \sigma_g^2}{\sqrt{2}\sigma_{gh}}\right) +$$

$$\frac{(\alpha_h + \beta_h)}{\beta_g(\beta_g + 1)(\beta_g + \beta_h + 1)(\alpha_h - \beta_g - 1)} e^{-\beta_g \mu_g + (\beta_g + 1)\mu_h + \beta_g^2 \sigma_{gh}^2 + \frac{2\beta_g + 1}{2}\sigma_h^2} \Phi\left(\frac{\mu_g - \mu_h - 2\beta_g \sigma_{gh}^2 - \sigma_h^2}{\sqrt{2}\sigma_{gh}}\right) \quad \text{if } \alpha_h \neq \beta_g + 1,$$

$$\frac{1}{\alpha_h(\alpha_h - 1)} e^{-(\alpha_h - 1)\mu_g + \alpha_h \mu_h + \alpha_h^2 \sigma_{gh}^2 - \frac{2\alpha_h - 1}{2}\sigma_g^2} \left[\left(\frac{1}{\alpha_h} + \frac{1}{\alpha_h - 1} + \frac{1}{\alpha_g + \alpha_h - 1} + \frac{1}{\alpha_h + \beta_h} + \mu_g - \mu_h - 2\alpha_h \sigma_{gh}^2 + \right.\right.$$

$$\left.\left.\sigma_g^2\right) \Phi\left(\frac{\mu_g - \mu_h - 2\alpha_h \sigma_{gh}^2 + \sigma_g^2}{\sqrt{2}\sigma_{gh}}\right) + \sqrt{2}\sigma_{gh} \phi\left(\frac{\mu_g - \mu_h - 2\alpha_h \sigma_{gh}^2 + \sigma_g^2}{\sqrt{2}\sigma_{gh}}\right)\right] \quad \text{if } \alpha_h = \beta_g + 1.$$

The MLD and Theil index of the dPLN are givn by:

$$\mathrm{MLD} = \frac{\sigma^2}{2} + \log\left[\frac{\alpha\beta}{(\alpha - 1)(\beta + 1)}\right] - \frac{\beta - \alpha}{\alpha\beta}, \tag{A8}$$

$$\mathrm{Theil} = \frac{\sigma^2}{2} - \log\left[\frac{\alpha\beta}{(\alpha - 1)(\beta + 1)}\right] + \frac{\beta - \alpha + 2}{(\alpha - 1)(\beta + 1)}. \tag{A9}$$

Note that the condition $\alpha > 1$ is required for the mean and inequality indices to be defined. The second term in equation (A6) vanishes when $\alpha, \beta \to \infty$. The first term is equal to the Gini coefficient of the lognormal distribution with the same dispersion parameter. The second term in equation (A7) vanishes and the first term converges to the Gini coefficient of the lognormal mixture distribution, with the pdf $f(x) = \sum_g p_g f_{\mathrm{LN}}(x; \mu_g, \sigma_g^2)$, when $\alpha_g, \beta_g \to \infty$ for all g, as follows:

$$G_{\mathrm{MLN}} = 2\sum\nolimits_{g,h} p_g p_h e^{\mu_g + \sigma_g^2/2} \Phi\left(\frac{\mu_g - \mu_h + \sigma_g^2}{\sqrt{2}\sigma_{gh}}\right) \Big/ \sum\nolimits_g p_g e^{\mu_g + \sigma_g^2/2} - 1. \tag{A10}$$

The analytic expression for the Gini coefficient of the lognormal mixture model was used to investigate the conditions for U-shaped relation between the size elasticity θ and the Gini coefficient of the income distributions when the equivalized personal income is calculated as $x = y/m^\theta$, where y is the amount of household income and m is the number of household members (Okamoto 2012a). To the author's knowledge, the formula in (A10) do not appear elsewhere in the literature.

The Lorenz curve of the dPLN can be expressed implicitly as follows:

$$L_{\mathrm{dPLN}}(c) = \Phi\left(\frac{\log y - \sigma^2}{\sigma}\right) - \frac{\beta+1}{\alpha+\beta} y^{-\alpha+1} e^{(\alpha^2-1)\sigma^2/2} \Phi\left(\frac{\log y - \alpha\sigma^2}{\sigma}\right) + \frac{\alpha-1}{\alpha+\beta} y^{\beta+1} e^{(\beta^2-1)\sigma^2/2} \Phi^c\left(\frac{\log y + \beta\sigma^2}{\sigma}\right), \quad 0 < c < 1, \tag{A11}$$

$L_{\mathrm{dPLN}}(0) = 0$, and $L_{\mathrm{dPLN}}(1) = 1$,

where $y = F^{-1}_{\mathrm{dPLN}}(c)$.