Economics Bulletin

Volume 34, Issue 3

A flexible descriptive model for the size distribution of incomes

Masato Okamoto
Statistical Research and Training Institute, Ministry of Internal Affairs and Communications

Abstract

There are four-parameter income distribution models that have good reputation for goodness-of-fit. In contrast, the existing models with five or more parameters fail to achieve a satisfactory level of goodness-of-fit. We propose a new seven-parameter model that is empirically shown to be substantially better fitted than the existing models by imposing an appropriate restriction on the parameter domain prior to the maximum likelihood estimation.

The author would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The usual disclaimers apply.
Citation: Masato Okamoto, (2014) "A flexible descriptive model for the size distribution of incomes", *Economics Bulletin*, Vol. 34 No. 3 pp. 1600-1610.
Contact: Masato Okamoto - mokamoto2@soumu.go.jp.
Submitted: April 30, 2014. **Published:** July 25, 2014.

1. Introduction

Parametric income distribution models (PIDMs) are used for various objectives such as estimation of income distribution and inequality/poverty indices from grouped data when survey micro data are unavailable, and construction of regression models for economic analysis, e.g. the Mincer-type equation. Typically, the lognormal (LN) distribution is used as the error term for the regression analysis; however, the goodness-of-fit of the LN is generally not perfect. To attain better fits, various models have been proposed. The generalized beta distribution of the first and second kinds (GB1 and GB2, McDonald, 1984) is popular as four-parameter PIDMs. The double-Pareto lognormal (dPLN) distribution (Reed 2003) has been shown to be fitted to income distributions better in several countries than the GB1 and GB2. It is notable that the dPLN has a parameter corresponding to the location parameter μ for the LN. Thus, the dPLN can possibly displace the LN from the regression models.

To seek further improvement, these four-parameter PIDMs were generalized to five-parameter PIDMs, the GB (McDonald and Xu 1995) and the generalized dPLN (denoted by 'GdPLN' hereafter, Reed and Wu, 2008), respectively. However, the GB fails to clearly improve the goodness-of-fit (*cf*. Kleiber and Kotz 2003, p.232), and the GdPLN is impractical because the probability density function (PDF) and cumulative distribution function (CDF) need to be obtained by difficult and slow numerical computation for inverting its characteristic function (Reed and Wu 2008). Furthermore, the GdPLN does not necessarily produce better fits, as shown in Section 3. So far, no PIDM with five or more parameters achieve a satisfactory level of goodness-of-fit. This paper generalizes the dPLN in a different way to break the deadlock.

2. New model

The dPLN is defined as the probability distribution for a random variable when the logarithm of the variable follows the normal-Laplace distribution (convolution of the Laplace and normal distributions). Its PDF is expressed as follows:

$$\begin{aligned} f_{dPLN}(x) &= \frac{\alpha}{\alpha+\beta} f_L(x|\mu,\sigma,\beta) + \frac{\beta}{\alpha+\beta} f_R(x|\mu,\sigma,\alpha) \\ &= \frac{\alpha\beta}{\alpha+\beta} x^{\beta-1} e^{-\beta\mu+\frac{1}{2}\beta^2\sigma^2} \Phi^c\left(\frac{\log x-\mu+\beta\sigma^2}{\sigma}\right) \\ &\quad + \frac{\alpha\beta}{\alpha+\beta} x^{-\alpha-1} e^{\alpha\mu+\frac{1}{2}\alpha^2\sigma^2} \Phi\left(\frac{\log x-\mu-\alpha\sigma^2}{\sigma}\right), \end{aligned} \tag{1}$$

where $\sigma, \alpha, \beta > 0$; Φ denotes the CDF of the standard normal distribution; $\Phi^c := 1-\Phi$. Parameters μ and σ correspond to the location and scale parameters for the LN. The dPLN can be regarded as a mixture of two distributions that have PDFs f_L and f_R by a ratio $\alpha:\beta$.

A function $f_L(x|\mu,\sigma,\beta)$ is the PDF of the product of two independent variables: a lognormal random variable with parameters μ and σ and an inverse Pareto variable which has the PDF $\beta x^{\beta-1}$, $0<x\leq 1$. Another function $f_R(x|\mu,\sigma,\alpha)$ is the PDF of the product of two independent variables: a lognormal random variable with parameters μ and σ and a Pareto variable which has the PDF $\alpha x^{-\alpha-1}$, $1\leq x$.

By allowing μ and σ for f_L and f_R to take different values and the mixing ratio to be independent from the parameters, the dPLN is generalized to a seven-parameter model that has the following PDF:

$$f(x)=rf_L(x|\mu_L,\sigma_L,\beta)+(1-r)f_R(x|\mu_R,\sigma_R,\alpha), \tag{2}$$

where $\sigma_L,\sigma_R,\alpha,\beta>0$; $0<r<1$. This new model shall be called the dual-parameterized dPLN, denoted as 'dP^{2}LN'. $\mu=r\mu_L+(1-r)\mu_R$ can be regarded as a parameter corresponding to the location parameter for the LN. ($\Delta\mu=\mu_R-\mu_L$ should be regarded as another parameter if replacing μ_L and μ_R using μ). The dP^{2}LN as well as the dPLN follows power laws in both tails, i.e. for constants c_1 and c_2,

$$f(x)\sim c_1x^{-\alpha-1}\ (x\to\infty);\ f(x)\sim c_2x^{\beta-1}\ (x\to 0). \tag{3}$$

Thus, the left tail behavior is determined by f_L, while the right tail behavior is determined by f_R. The *h*–th ($<\alpha$) order moments are expressed as follows:

$$\mu_h=E(X^h)=r\frac{\beta}{\beta+h}e^{\mu_Lh+\frac{1}{2}h^2\sigma_L^2}+(1-r)\frac{\alpha}{\alpha-h}e^{\mu_Rh+\frac{1}{2}h^2\sigma_R^2}. \tag{4}$$

The mean (first order moment) converges as follows, when $\alpha,\ \beta\to\infty$:

$$\mu_1\to re^{\mu_L+\frac{1}{2}\sigma_L^2}+(1-r)e^{\mu_R+\frac{1}{2}\sigma_R^2}. \tag{5}$$

The value equals to a mix of the means of two lognormal distributions with parameters μ_L, σ_L and μ_R, σ_R by a ratio $r:1-r$. The normalized incomplete moments are expressed as follows:

$$\begin{aligned}F_{(h)}(x)&=\frac{\int_0^x t^hf(t)dt}{\mu_h}\\&=\frac{r}{\mu_h}\frac{\beta}{\beta+h}\left[x^{\beta+h}e^{-\beta\mu_L+\frac{1}{2}\beta^2\sigma_L^2}\Phi^c\left(\frac{\log x-\mu_L+\beta\sigma_L^2}{\sigma_L}\right)\right.\\&\qquad\left.+e^{\mu_Lh+\frac{1}{2}h^2\sigma_L^2}\Phi\left(\frac{\log x-\mu_L-h\sigma_L^2}{\sigma_L}\right)\right]+\\&\frac{1-r}{\mu_h}\frac{\alpha}{\alpha-h}\left[-x^{-\alpha+h}e^{\alpha\mu_R+\frac{1}{2}\alpha^2\sigma_R^2}\Phi\left(\frac{\log x-\mu_R-\alpha\sigma_R^2}{\sigma_R}\right)\right.\\&\qquad\left.+e^{\mu_Rh+\frac{1}{2}h^2\sigma_R^2}\Phi\left(\frac{\log x-\mu_R-h\sigma_R^2}{\sigma_R}\right)\right].\end{aligned} \tag{6}$$

The CDF of the dP^{2}LN corresponds to $F_{(0)}(x)$. The Lorenz curve is implicitly expressed as $L(\theta)=F_{(1)}\left(F_{(0)}^{-1}(\theta)\right)$. An analytic expression of the Gini index listed in Appendix is derived

in the same way as that of a mixture of dPLNs (Okamoto 2012a, 2013a). The existence of the formula is one of the advantages of the dP^{2}LN over the existing five-parameter PIDMs.

Many special cases with five or more parameters are conceivable. In the case that $\mu_L = \mu_R$ and $r = \alpha/(\alpha+\beta)$, the restricted model shall be denoted as 'dP^{2}LN$_\sigma$'. In the case that $r = \alpha/(\alpha+\beta)$, the restricted model shall be denoted as 'dP^{2}LN$_{\mu\sigma}$'. Additional special cases are defined in the next section. The other special cases we examined are omitted because their fits were inferior to or approximately the same as these models. The PDFs of the dP^{2}LN and its special cases, unlike that of the dPLN, do not generally satisfy the unimodality condition. However, regarding the income data studied in the next section, the PDF is single-peaked along with a superior fit by selecting an appropriate special case.

3. Empirical comparison

The new PIDM was compared with the existing PIDMs by fitting them to gross income data from the seven waves between 1992 and 2010 of the US Survey of Consumer Finances (SCF), conducted by the US Federal Reserve Board, and to disposable income data from the seven waves between 2000 and 2012 of the Italian Survey of Household Income and Wealth (SHIW), conducted by the Bank of Italy. The results for unadjusted household income are presented here. Similar results were obtained for equivalized income, as adjusted for household size. The Gini index ranged from 0.50 to 0.57 for the seven waves of the SCF and stayed approximately 0.35 for the seven waves of the SHIW.

The existing PIDMs compared include three-parameter models such as the Singh-Maddala (SM, Singh and Maddala 1975), Dagum (Da, Dagum 1977) and κ-generalized distribution (κG, Clementi *et al*. 2007) and four-parameter models such as the extended κG distribution of the first and second kinds (EκG1 and EκG2, Okamoto 2013b) in addition to the GB1, GB2 and dPLN. The EκG2 is a new kind of generalized beta distribution which tends to outperform the GB2 in terms of both likelihood and accuracy of inequality estimates. The PIDMs were fitted to the income data using the maximum likelihood (ML) criterion and a Nelder-Mead simplex algorithm. Because the parameters may possibly converge to a local maximal point in the cases of the dP^{2}LN and its special cases, plural sets of initial values were used for the ML estimation (MLE). As for the full model of the dP^{2}LN, the procedure was performed 35 times in total without constraints on the parameters or under constraint $\sigma_L < \sigma_R$, $\sigma_L > \sigma_R$, $\mu_L < \mu_R$ or $\mu_L > \mu_R$, using initial values created based on the ML parameters of the special cases.

The goodness-of-fit of PIDMs with different numbers of parameters were compared using the AICs. Because a better fit in terms of frequency-based measures, such as the likelihood, does not necessarily imply a better fit in terms of the accuracy of inequality estimates, the estimation error of the Lorenz curve, as defined below, and the absolute error of the Gini index

(absolute deviation from the sample estimate) were also adopted as evaluation measures representing the accuracy of inequality estimates.[1]

$$\text{L-RSSE} := \sqrt{\sum_{i=1}^{n}\left(\hat{L}_i - L(\theta_i)\right)^2};\ \hat{L}_i = \sum_{j\le i} w_j x_j / \sum_{j=1}^{n} w_j x_j;\ \theta_i = \sum_{j\le i} w_j / \sum_{j=1}^{n} w_j, \quad (7)$$

where θ_i and $\hat{L}_i$ are the cumulative population and income share of households up to household i; $\left(\theta_i, \hat{L}_i\right)_i$ represents the empirical Lorenz curve. In the above formula, it is assumed that each household is assigned an aggregation weight w_j and is sorted in ascending order, according to its income x_j. Simulation studies show that a model fitted by the MLE tends to worsen the L-RSSE, as opposed to the effect on the likelihood when a sample is randomly generated from its special case model. Thus, it can be said that no penalty to the L-RSSE is required for PIDMs with a larger number of parameters.

Table 1 shows the goodness-of-fit of the PIDMs to the income data from the seven waves in an aggregated form.[2] The GB failed to achieve better fits than the GB2, similar to the GdPLN in the case of the SCF. In the case of the SHIW, the GdPLN substantially lowered the AIC; however, the L-RSSE and accuracy of the Gini index rather deteriorated. As for the goodness-of-fit of the existing PIDMs, it also should be noted that several three-parameter PIDMs tended to yield more accurate inequality estimates than four-parameter PIDMs for the SCF. Unfortunately, such phenomenon is not unusual.[3]

The results of the dP^{2}LN in Table 1 are those of the model with constraint $\sigma_L < \sigma_R$ for the SCF because its expectation was infinite in wave 1992 if no constraints were applied. The dP^{2}LN$_{\mu\sigma}$ and dP^{2}LN improved the AIC relative to the existing PIDMs for both SCF and SHIW, whereas both PIDMs did not improve the accuracy of inequality estimates except the dP^{2}LN for the SHIW. Furthermore, the dP^{2}LN did not satisfy the unimodality condition in one or two waves of either survey; the same was true for the dP^{2}LN$_{\mu\sigma}$ in the case of the SHIW. Table 2 reveals that the ML parameters of the dP^{2}LN, as fitted to the SCF data, were unstable across the seven waves. Furthermore, the PDF for wave 1995 was bimodal. To address this instability, the imposition of constraints $\mu_L > \mu_R$ and $\beta = 1.01$ prior to the MLE was considered (the constraint model is denoted as 'dP^{2}LN″). This constraint made the ML parameters relatively stable and made the PDF unimodal along with a substantial improvement of both AIC and accuracy of inequality estimates, relative to the existing PIDMs including the three-parameter PIDMs. When looking at the results for each wave, the dP^{2}LN′ was better fitted than the existing PIDMs for all seven waves in terms of the AIC, for four waves in terms of the L-RSSE and for three waves in terms of the accuracy of the Gini index. The inverse distribution of the

[1] The absolute errors of the Theil index and mean log deviation were also examined. The results (omitted here) are similar to those for the L-RSSE and absolute error of the Gini index.

[2] The results for each wave and all special cases of the dP^{2}LN we examined are listed in the supplementary tables.

[3] For example, the SM tends to yield better inequality estimates than the GB2 and dPLN for Japanese income data (Okamoto 2012b).

κG (IκG) produced the smallest values for three and two waves in terms of the L-RSSE and absolute error of the Gini index, respectively; however, the accuracy of the estimates was unstable relative to the dP^2LN'.

Table 1. Goodness-of-fit of PIDMs

Model	No. of para.	USA (SCF, 1992-2010)			Italy (SHIW, 2000-2012)		
		AIC#,*	L-RSSE*	Abs. error of * the Gini index	AIC#,*	L-RSSE*	Abs. error of * the Gini index
SM	3	6.9	1.695	0.0249	19.0	0.185	0.0022
Da	3	4.1	1.350	0.0198	33.6	0.515	0.0080
κG	3	14.7	1.248	0.0198	68.5	0.377	0.0054
IκG ##	3	45.7	1.079	0.0170	121.0	1.376	0.0222
GB1	4	152.3	3.153	0.0395	273.1	0.510	0.0021
GB2	4	0.9	1.717	0.0256	15.1	0.174	0.0024
EκG1	4	33.3	1.519	0.0277	16.4	0.162	0.0019
EκG2	4	-5.6	1.767	0.0274	25.3	0.295	0.0044
IEκG1 ##	4	-6.1	1.614	0.0247	25.5	0.188	0.0027
IEκG2 ##	4	9.2	10.148	0.1929	13.4	2.332	0.0441
dPLN	4	0.0	1.490	0.0215	0.0	0.159	0.0022
GB	5	2.9	1.717	0.0256	16.4	0.166	0.0022
GdPLN	5	1.2	1.590	0.0245	-50.2	0.228	0.0033
dP^2LN_{σ}	5	-20.3	1.098	0.0179	0.4	0.158	0.0022
$dP^2LN_{\mu\sigma}$	6	-47.8	4.241	0.0582	-60.4	3.314	0.0581
dP^2LN **	7	<u>-49.1</u>	1.691	0.0233	<u>-88.7</u>	<u>0.107</u>	<u>0.0014</u>
dP^2LN'	6	-42.6	<u>0.500</u>	<u>0.0060</u>	-86.2	0.127	0.0016
dP^2LN''	6	-43.1	0.534	0.0063	-84.6	0.114	0.0014

Differences from the corresponding values of the dPLN.

IκG: the inverse distribution of the κG. IEκG1 and IEκG2: the inverse distribution of the EκG1 and EκG2 (Okamoto 2013b).

* Averages over the seven waves are taken for the AIC. The root-of-mean-square errors over the seven waves are taken for the L-RSSE and absolute error of the Gini index.

** Constraint $\sigma_L < \sigma_R$ is imposed for the SCF because of the infinite expectation in wave 1992 if no constraints are applied.

Table 2. ML parameters of the dP^{2}LN and dP^{2}LN′, USA [4]

Year	dP^{2}LN (with constraint $\sigma_L < \sigma_R$)							dP^{2}LN′					
	μ_L	μ_R	σ_L	σ_R	α	β	r	μ_L	μ_R	σ_L	σ_R	α	r
1992	11.5	10.0	0.33	0.87	2.1	1.1	0.36	11.5	10.1	0.31	0.87	2.2	0.34
	(0.03)	(0.08)	(0.052)	(0.035)	(0.18)	(0.05)	(0.043)	(0.03)	(0.05)	(0.042)	(0.024)	(0.16)	(0.032)
1995	10.4	10.6	0.12	0.32	1.9	1.3	0.38	11.4	10.2	0.35	0.88	2.1	0.46
	(0.53)	(0.25)	(0.140)	(0.283)	(0.09)	(0.11)	(0.070)	(0.03)	(0.05)	(0.026)	(0.033)	(0.13)	(0.031)
1998	11.6	10.0	0.31	0.80	1.6	1.1	0.41	11.6	10.1	0.29	0.80	1.7	0.38
	(0.04)	(0.06)	(0.051)	(0.026)	(0.09)	(0.05)	(0.044)	(0.03)	(0.05)	(0.034)	(0.022)	(0.08)	(0.026)
2001	11.7	10.0	0.36	0.79	1.4	1.2	0.40	11.7	10.1	0.29	0.80	1.5	0.33
	(0.04)	(0.09)	(0.046)	(0.040)	(0.10)	(0.06)	(0.056)	(0.03)	(0.04)	(0.030)	(0.018)	(0.06)	(0.031)
2004	11.5	10.0	0.52	0.84	1.5	1.4	0.51	11.7	10.2	0.36	0.83	1.8	0.30
	(0.05)	(0.08)	(0.031)	(0.039)	(0.09)	(0.08)	(0.043)	(0.04)	(0.04)	(0.030)	(0.018)	(0.07)	(0.028)
2007	11.6	9.9	0.44	0.76	1.3	1.5	0.40	11.7	10.2	0.00	0.81	1.7	0.16
	(0.05)	(0.08)	(0.053)	(0.037)	(0.07)	(0.37)	(0.053)	(0.01)	(0.03)	(0.094)	(0.015)	(0.07)	(0.032)
2010	11.6	10.0	0.41	0.73	1.4	1.4	0.29	11.7	10.1	0.27	0.75	1.7	0.16
	(0.04)	(0.05)	(0.036)	(0.023)	(0.06)	(0.07)	(0.038)	(0.05)	(0.02)	(0.025)	(0.012)	(0.05)	(0.017)

The ML parameter values listed are those of the respective models fitted to the fourth of the five datasets in each wave, which were officially generated by the multiple imputation procedure for income and wealth variables.

Figures in parenthesis are the standard errors due to sampling and imputation (see footnote 4). The large error estimates in the dP^{2}LN's parameters for the wave 1995 are produced by the imputation error calculation.

Table 3. ML parameters of the dP^{2}LN and dP^{2}LN″, Italy [4]

Year	dP^{2}LN							dP^{2}LN″					
	μ_L	μ_R	σ_L	σ_R	α	β	r	μ_L	μ_R	σ_L	σ_R	α	r
2000	11.4	10.4	0.12	0.60	3.7	1.1	0.18	11.4	10.4	0.09	0.60	3.9	0.15
2002	10.7	9.8	0.15	0.60	4.1	1.1	0.15	10.8	9.8	0.12	0.60	4.3	0.13
2004	10.9	9.7	0.24	0.50	2.7	1.4	0.23	11.0	9.8	0.08	0.54	3.3	0.11
2006	10.9	10.2	1.30	0.59	292.4*	0.9	0.03	10.9	9.9	0.14	0.54	3.5	0.11
2008	10.9	9.9	0.21	0.57	3.6	1.3	0.12	10.9	9.9	0.00	0.59	4.2	0.06
	(0.08)	(0.04)	(0.060)	(0.017)	(0.38)	(0.13)	(0.039)	(0.01)	(0.03)	(0.000)	(0.015)	(0.62)	(0.014)
2010	10.9	9.9	0.19	0.57	3.7	1.1	0.18	10.9	10.0	0.15	0.57	3.8	0.15
	(0.06)	(0.03)	(0.066)	(0.017)	(0.34)	(0.09)	(0.035)	(0.05)	(0.03)	(0.043)	(0.015)	(0.35)	(0.016)
2012	11.0	9.9	0.14	0.58	3.9	0.8	0.11	11.0	9.8	0.20	0.56	3.4	0.16
	(0.07)	(0.04)	(0.042)	(0.019)	(0.56)	(0.12)	(0.038)	(0.06)	(0.03)	(0.061)	(0.019)	(0.33)	(0.027)

* The value 292.4 is not a typo.

Figures in parenthesis are the standard errors due to sampling (see footnote 4).

[4] The standard errors (SEs) in the ML parameters were added in the tables, following a suggestion from one of the anonymous reviewers. The (observed) Fisher information matrix (FIM) is often used for the variance estimation; however, the FIM does not produce correct estimates when the survey data are obtained from non-simple random sampling procedures. To make the variance estimation consistent with the survey design, some statistical agencies provide micro data together with replicate weights generated from resampling procedures such as bootstrapping and jackknifing (from wave 2008 in the case of the SHIW). The respective figures in the tables were calculated using the replicate weights although the calculation was time-consuming. The imputation errors were also taken into account in the case of the SCF. The square root of the design effect factor ($\sqrt{D^2}$), i.e. the ratio of the SE derived from the replicate weights relative to the corresponding value derived from the FIMs, ranges from 0.0001 to 3.21 (about 0.85 on average) in the case of the SCF (if excluding the imputation errors), and ranges from 0.01 to 2.10 (about 1.4 on average) in the case of the SHIW. $\sqrt{D^2}$s are, in many cases, below unity in the former due to the very effective stratified sampling; however, $\sqrt{D^2}$ tends to substantially vary among the parameters. In contrast, $\sqrt{D^2}$s are, in most cases, above unity in the latter case (and cases of ordinary sample surveys). Roughly speaking, the confidence level needs to be set to 99.4% (to attain 95% confidence level in practice) if inferences are made from the FIMs.

A similar effect emerged for the SHIW by imposing constraints $\beta = 1.01$ and $\sigma_L < \sigma_R$ (the constraint model is denoted as 'dP^{2}LN‴'), as shown in Table 3, although the accuracy of inequality estimates grew slightly worse. This constraint made the AIC, L-RSSE and accuracy of the Gini index superior to those of the existing PIDMs for all seven, four and two waves, respectively. Some existing PIDMs attained the smallest in one or two waves; however, these models failed to stably produce accurate inequality estimates.

These results suggest that selections from the special cases should be made by referring to smaller L-RSSEs and the parameters' stability across several waves (in addition to unimodality, if required) rather than based solely on the AIC/likelihood. Figures 1 and 2 illustrate the PDFs of the PIDMs fitted to the US and Italian data.[5] Much better fits of the dP^{2}LN′ and dP^{2}LN″ relative to the existing PIDMs, particularly around the peaks, can be visually observed. Both figures clearly show that any existing PIDM, in reality, is not able to reproduce the original PDFs accurately.

4. Discussion and concluding remarks

A unique feature of the emerging procedure for fitting and selecting special cases of the dP^{2}LN, as described in the previous section, is that parameter constraints are selected as prior information by referring to the L-RSSEs and parameters' stability (as well as unimodality, if required). Appropriate choices of the constraint empirically make the parameters stable and the goodness-of-fit of the new PIDM superior to the existing PIDMs not only in terms of the likelihood but also in terms of the accurate inequality estimation.

The superiority of our new model over the existing PIDMs for many other countries and the generality of the procedure for choosing appropriate parameter constraints appear to be well-anticipated.[6] Further investigation is required for confirming the anticipation and establishing a variety of parameter constraints. Exploring an efficient MLE procedure is also an issue. Although possible bimodality of the PDF is not necessarily a disadvantage of the new model, ways to avoid multimodal PDF need to be studied when unimodality is a prerequisite. As stated above, future tasks remain. Nevertheless, this new model looks promising, as we have empirically shown here that it outperforms the existing models.

[5] One of the anonymous reviewers suggests it may be better to use kernel density estimations instead of histograms; however, the kernel method is known to have a difficulty to accurately recover densities bounded and coarse at lower ends (at least when applying the standard method). Furthermore, the original densities are not necessarily completely smooth because the sample data contain several point masses. The author considers it is preferable that the actual densities in the samples are presented by the histograms. It is not so easy to appropriately apply the kernel method, despite of its popularity. The results of the kernel density estimations are presented in the supplementary charts.

[6] As for Japan, the anticipation was confirmed by applying the dP^{2}LN to (grouped) income data for 1984 - 2009. The selected constraint $\sigma_L = \sigma_R$ is different from those for USA and Italy.

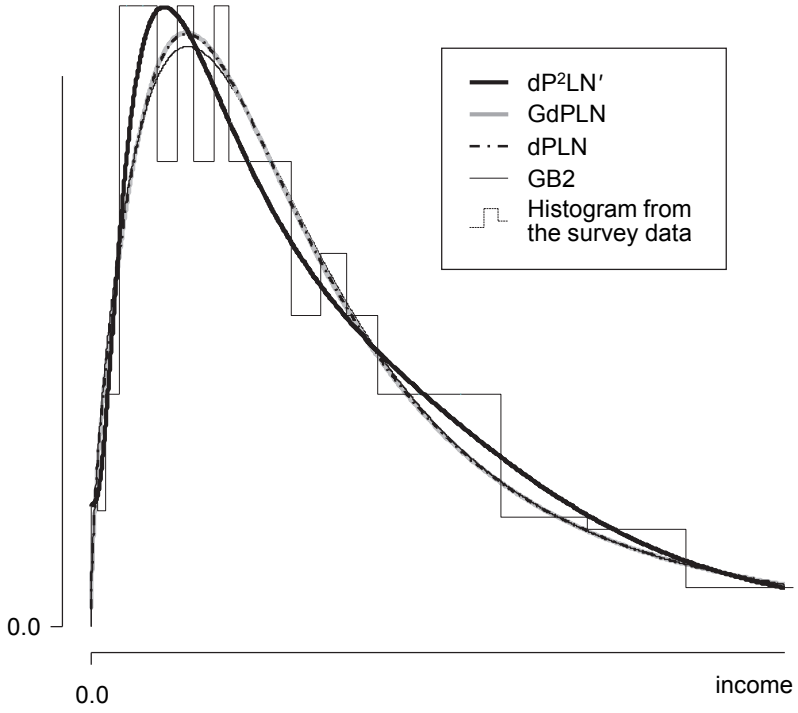

Figure 1. PDFs of fitted PIDMs, USA, 2001

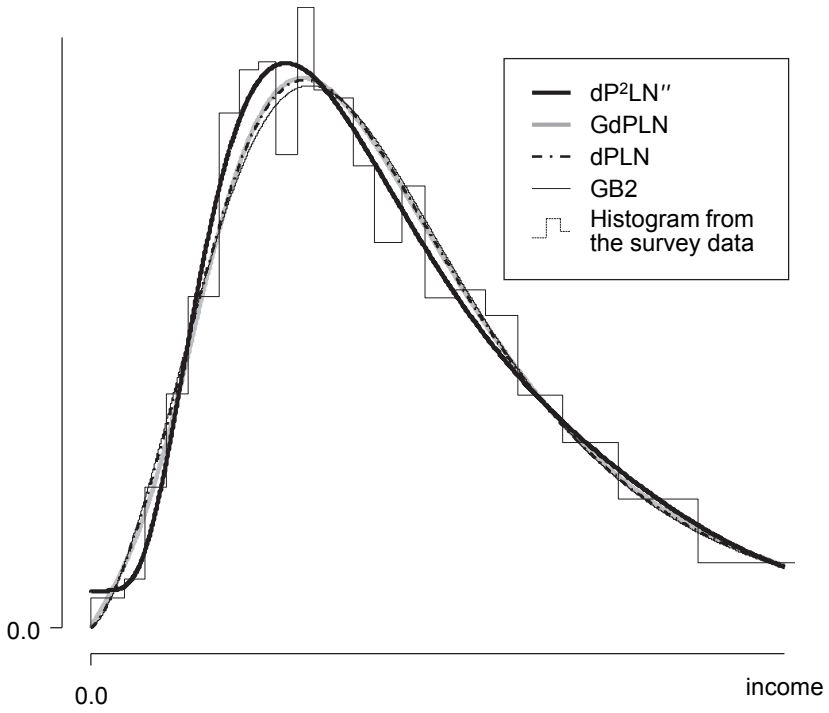

Figure 2. PDFs of fitted PIDMs, Italy, 2006

References

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

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

Kleiber, C. and S. Kotz (2003) *Statistical Size Distributions in Economics and Actuarial Sciences*, Hoboken: Wiley.

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**, 133–152.

Okamoto, M. (2012a) "Evaluation of the goodness of fit of new statistical size distributions with consideration of accurate income inequality estimation," *Economics Bulletin* **32(4)**, 2969–2982.

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

Okamoto, M. (2013a) "Erratum to "Evaluation of the goodness of fit of new statistical size distributions with consideration of accurate income inequality estimation"," *Economics Bulletin* **33(3)**, 2443–2444.

Okamoto, M. (2013b) "Extension of the κ-generalized distribution: new four-parameter models for the size distribution of income and consumption," LIS working paper 600.

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,. New York: Springer, 211–223.

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

Appendix. The Gini index of the dP^{2}LN

$$G = 2\frac{r^2 I_{LL}+(1-r)^2 I_{RR}+r(1-r)I_{RL}+r(1-r)I_{LR}}{\mu_1} - 1,$$

where $I_{LL} = \frac{1}{\beta(\beta+1)} e^{\mu_L+\frac{1}{2}\sigma_L^2}\Phi\left(\frac{1}{\sqrt{2}}\sigma_L\right) + \frac{1}{\beta} e^{\mu_L+\left(\beta^2+\beta+\frac{1}{2}\right)\sigma_L^2}\left[\frac{2}{2\beta+1} - \frac{1}{\beta+1}\right]\Phi\left(-\frac{2\beta+1}{\sqrt{2}}\sigma_L\right)$;

$I_{RR} = \frac{1}{\alpha(\alpha-1)} e^{\mu_R+\frac{1}{2}\sigma_R^2}\Phi\left(\frac{1}{\sqrt{2}}\sigma_R\right) + \frac{1}{\alpha} e^{\mu_R+\left(\alpha^2-\alpha+\frac{1}{2}\right)\sigma_R^2}\left[\frac{1}{\alpha-1} - \frac{2}{2\alpha-1}\right]\Phi\left(-\frac{2\alpha-1}{\sqrt{2}}\sigma_R\right)$;

$$I_{RL} = \frac{1}{\alpha}\frac{1}{\beta+1}\exp\left(\frac{1}{2}\sigma_L^2 + \mu_L\right)\Phi\left(\frac{-\mu_R+\mu_L+\sigma_L^2}{\sqrt{2}\sigma_{RL}}\right)$$

$$+\frac{1}{\alpha}\frac{1}{\alpha-\beta-1}\exp\left(\alpha^2\sigma_{RL}^2 - \frac{2\alpha-1}{2}\sigma_L^2 + \alpha\mu_R - (\alpha-1)\mu_L\right)\Phi\left(\frac{-\mu_R+\mu_L-2\alpha\sigma_{RL}^2+\sigma_L^2}{\sqrt{2}\sigma_{RL}}\right)$$

$$-\frac{1}{(\beta+1)(\alpha-\beta-1)}\exp\left(\beta^2\sigma_{RL}^2 + \frac{2\beta+1}{2}\sigma_R^2 + (\beta+1)\mu_R - \beta\mu_L\right)\Phi\left(\frac{-\mu_R+\mu_L-2\beta\sigma_{RL}^2-\sigma_R^2}{\sqrt{2}\sigma_{RL}}\right) \quad \text{if } \alpha \neq \beta+1,$$

$$I_{RL} = \frac{1}{(\beta+1)^2}\exp\left(\frac{1}{2}\sigma_L^2 + \mu_L\right)\Phi\left(\frac{-\mu_R+\mu_L+\sigma_L^2}{\sqrt{2}\sigma_{RL}}\right)$$

$$-\frac{1}{\beta+1}\sqrt{2}\sigma_{RL}\exp\left(\frac{1}{2}\sigma_R^2 + \mu_R\right)\phi\left(\frac{-\mu_R+\mu_L-\sigma_R^2}{\sqrt{2}\sigma_{RL}}\right)$$

$$+\frac{1}{\beta+1}\left[(2\beta\sigma_{RL}^2 + \sigma_R^2 + \mu_R - \mu_L) - \frac{1}{\beta+1}\right]$$

$$\cdot\exp\left(\beta^2\sigma_{RL}^2 + \frac{2\beta+1}{2}\sigma_R^2 + (\beta+1)\mu_R - \beta\mu_L\right)\Phi\left(\frac{-\mu_R+\mu_L-2\beta\sigma_{RL}^2-\sigma_R^2}{\sqrt{2}\sigma_{RL}}\right) \text{ if } \alpha = \beta+1;$$

$$I_{RL} = \frac{1}{\beta}\frac{1}{\alpha-1}\exp\left(\frac{1}{2}\sigma_R^2 + \mu_R\right)\Phi\left(\frac{-\mu_L+\mu_R+\sigma_R^2}{\sqrt{2}\sigma_{LR}}\right)$$

$$+\frac{1}{\beta}\frac{1}{\alpha-\beta-1}\exp\left(\beta^2\sigma_{LR}^2 + \frac{2\beta+1}{2}\sigma_R^2 - \beta\mu_L + (\beta+1)\mu_R\right)\Phi\left(-\frac{-\mu_L+\mu_R+2\beta\sigma_{LR}^2+\sigma_R^2}{\sqrt{2}\sigma_{LR}}\right)$$

$$-\frac{1}{(\alpha-1)(\alpha-\beta-1)}\exp\left(\alpha^2\sigma_{LR}^2 - \frac{2\alpha-1}{2}\sigma_L^2 - (\alpha-1)\mu_L + \alpha\mu_R\right)\Phi\left(-\frac{-\mu_L+\mu_R+2\alpha\sigma_{LR}^2-\sigma_L^2}{\sqrt{2}\sigma_{LR}}\right) \quad \text{if } \alpha \neq \beta+1,$$

$$I_{RL} = \frac{1}{(\alpha-1)^2}\exp\left(\frac{1}{2}\sigma_R^2 + \mu_R\right)\Phi\left(\frac{-\mu_L+\mu_R+\sigma_R^2}{\sqrt{2}\sigma_{LR}}\right)$$

$$+\frac{1}{\alpha-1}\sqrt{2}\sigma_{LR}\exp\left(\frac{1}{2}\sigma_L^2 + \mu_L\right)\phi\left(-\frac{-\mu_L+\mu_R-\sigma_L^2}{\sqrt{2}\sigma_{LR}}\right)$$

$$-\frac{1}{\alpha-1}\left[(2\alpha\sigma_{LR}^2 - \sigma_L^2 - \mu_L + \mu_R) - \frac{1}{\alpha-1}\right]$$

$$\cdot\exp\left(\alpha^2\sigma_{LR}^2 - \frac{2\alpha-1}{2}\sigma_L^2 - (\alpha-1)\mu_L + \alpha\mu_R\right)\Phi\left(-\frac{-\mu_L+\mu_R+2\alpha\sigma_{LR}^2-\sigma_L^2}{\sqrt{2}\sigma_{LR}}\right) \text{ if } \alpha = \beta+1.$$

Note that ϕ denotes the PDF of the standard normal distribution.