Sains Malaysiana 50(7)(2021): 2047-2058
http://doi.org/10.17576/jsm-2021-5007-19

Composite Pareto Distributions for Modelling Household Income Distribution in Malaysia

(Taburan Komposit Pareto untuk Pemodelan Taburan Pendapatan Isi Rumah di Malaysia)

MUHAMMAD HILMI ABDUL MAJID* & KAMARULZAMAN IBRAHIM

ABSTRACT

Composite Pareto distributions are flexible as the models allow for data to be described by two distributions: a Pareto distribution for the data above a threshold value and another separate distribution for data below the threshold value. It is noted in some previous literatures that the Paretian tail behaviour can be observed in the distribution of Malaysian household income. In this paper, the composite Pareto models are fitted to the Malaysian household income data of several years. These fitted composite Pareto models are then compared to several univariate models for describing income distribution using pseudo-likelihood based AIC, BIC and Kolmogorov-Smirnov goodness-of-fit test. It is found that the income distributions in Malaysia can be best described by the lognormal-Pareto (II) model as compared to other candidate models.

Keywords: Composite model; goodness-of-fit; income distribution; Pareto distribution; pseudo-likelihood

ABSTRAK

Taburan komposit Pareto adalah luwes kerana model ini boleh menerangkan sesuatu data menggunakan dua taburan: taburan Pareto untuk data di atas suatu nilai ambang dan taburan yang berasingan untuk data di bawah nilai ambang tersebut. Kajian sebelum ini telah menyatakan bahawa ciri-ciri ekor Pareto dapat diperhatikan pada taburan pendapatan isi rumah di Malaysia. Dalam kajian ini, model komposit Pareto disuaikan ke atas data pendapatan isi rumah di Malaysia. Model komposit Pareto ini akan dibandingkan dengan model univariat lain untuk menerangkan taburan pendapatan dengan menggunakan AIC, BIC dan ujian kebagusan penyuaian Kolmogorov-Smirnov berasaskan pseudo-kebolehjadian. Kajian mendapati taburan pendapatan di Malaysia boleh diterangkan menggunakan model lognormal-Pareto (II) lebih baik berbanding calon model lain.

Kata kunci: Kebagusan penyuaian; model komposit; pseudo-kebolehjadian; taburan Pareto; taburan pendapatan

INTRODUCTION

Various parametric models for income distributions have been proposed by many authors, including lognormal, gamma, Weibull, generalized beta, and Pareto distributions. These models are useful for assessing income inequality and other economic indicators. Amongst these models, Pareto type I distribution is found to be able to fit the upper tail of the income distribution very well since it has been found that the number of individuals with income above a given value can be approximated by $Cx^{-\alpha}$ for some positive values C and α (Arnold 2008). However, Pareto type I distribution can only be fitted to the upper tail of the income data and cannot be used for the whole income data.

Alternatively, composite Pareto distribution model can be used to model the whole distribution by combining Pareto distribution for the upper tail and a separate distribution for the lower tail. The composite Pareto model has a probability density function (pdf) of the form

$$f(x|\theta) = \begin{cases} \rho_1 f_1(x|\theta), & x \leq \tau \\ \rho_2 f_2(x|\theta), & x > \tau \end{cases},$$

where the model separates the distribution at the threshold value τ. The observation in the lower tail is modelled using the pdf $f_1(x|\theta)$ while the upper tail observations are modelled using $f_2(x|\theta)$, which is the pdf of Pareto distribution. Composite Pareto distribution was first introduced by Cooray and Ananda (2005) using lognormal distribution for the lower tail. Then the model was improved by Scollnik (2007) by freeing the mixing weight. However, in the literature, composite Pareto distributions have been mainly used to model insurance data and limited applications are found for the income data (Bakar et al. 2015; Cooray & Ananda 2005; Scollnik 2007; Scollnik & Sun 2012; Teodorescu & Vernic 2009).

In this paper the composite Pareto models are applied to Malaysian income distribution from the Malaysia Household Income Survey (HIS) for the year 2007, 2009, 2012, 2014 and 2016. The composite Pareto models are compared to other parametric models used for income distribution, namely lognormal, gamma, Weibull, Dagum, beta 2 (also called beta prime), Singh-Maddala and generalized beta of the second kind distributions. The performance of these distributions in model fitting are compared using AIC, BIC, and Kolmogorov-Smirnov goodness-of-fit test. Since the survey data contains sample weights, the weights must be included in the analysis to avoid biased parameter estimates (Pfeffermann 1993). For that reason, a pseudo-likelihood based approach is used.

This paper is organized as follows. In the next section, a brief description on the Malaysia Household Income Survey is given. After that, the composite Pareto models are described in detail for both models with Pareto type I and Pareto type II for the upper tail data. The statistical methods used for analysis including pseudo-likelihood approach and model selection criteria are described in the following section. Subsequently, the results of the application of composite Pareto models on the data are given and discussed. Last section concludes the paper.

MALAYSIA HOUSEHOLD INCOME SURVEY (HIS)

Twice every five years, the Department of Statistics Malaysia (DOSM) conducted surveys to collect information on the income distribution and identify the accessibility of basic amenities for citizens in Malaysia in the Household Income and Basic Amenities Survey (HIS&BA). To gather the data, personal interviews are conducted by trained officers and staffs of the department. Then, data quality is checked to detect and rectify errors by experienced officers. There are several variables and information gathered in the survey including household's annual income, location, size, and the head of household's age, gender, education, marital status, and occupation. Each household is given weight based on its location (state and urbanity).

Five HIS datasets are considered: HIS year 2007, 2009, 2012, 2014, and 2016. The data used in our analysis is a subset of the total survey data provided by DOSM and Bank Data UKM (These datasets are confidential but can be requested online from https://www.dosm.gov.my/v1/index.php). To equivalise the income data, the household monthly gross income in Ringgit Malaysia (RM) is divided by the square root of the number of household size. This square root equivalence scale is used in many studies including Congressional Budget Office (2019) and OECD (2015) and is used since two households with the same monthly gross income may not be in the same economic position due to the difference in the number of people in the households. Since no comparison is made between the different years, inflation adjustment is not required.

The HIS datasets have been used extensively in studying the income distribution in Malaysia and some literatures have discussed and used Pareto tail for the upper income data (Masseran et al. 2019; Ragayah 2008; Razak & Shahabuddin 2018; Safari et al. 2018a, 2018b). For example, the Paretian tail behaviour of the Malaysian household income has been noted by Razak and Shahabuddin (2018). Safari et al. (2018b) on the other hand, have used a semi-parametric approach to measure income inequality by combining nonparametric distribution for the lower data distribution and Pareto distribution for the upper tail. It is of interest in this paper to check if the composite Pareto models can provide a better fit as compared to other model when applied to income distribution.

COMPOSITE PARETO MODELS

The composite Pareto model is a mixture model in which the distribution is spliced at a certain threshold value, τ. Any observations above the value τ follows the Pareto distribution and observations below the value τ follow another model such as lognormal, gamma or Weibull. We call this the 'lower data distribution model'.

In the introduction of composite Pareto model, Cooray and Ananda (2005) have used Pareto type I distribution for the upper tail data. However, in the model proposed by them, the proportion of data coming from the Pareto tail is fixed and deemed to be restrictive (Scollnik 2007). Scollnik then improved the model by removing the restriction on the mixing weight in the model and

considering the use of generalized Pareto distribution (GPD), also known as Pareto type II distribution, for the upper tail data. Since then, a variety of composite Pareto models have been introduced including Weibull-Pareto composite model (Ciumara 2006; Scollnik & Sun 2012), exponential-Pareto composite model (Teodorescu & Vernic 2009), and inverse gamma-Pareto model (Aminzadeh & Deng 2019). In these literatures, composite Pareto models were found to be able to fit real life data better compared to single distribution models.

Figure 1 illustrates the pdf of the composite Pareto models. In this figure, the pdf of three distribution models is plotted: lognormal, lognormal-Pareto (I), and lognormal-Pareto (II). The data follows similar lower data distribution model (the lognormal distribution) up to the threshold τ. However, data above τ. are modelled using Pareto type I and Pareto type II distributions for lognormal-Pareto (I) and lognormal-Pareto (II), respectively. Note that the properties of the pdfs including heaviness of the upper tail may differ based on the parameter values.

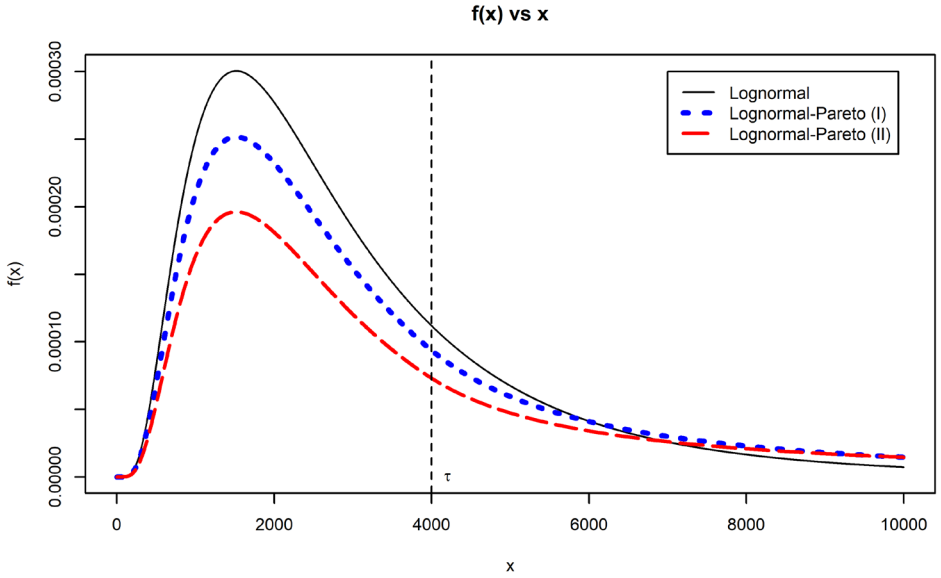

FIGURE 1. The pdf of lognormal ($\mu = 7.8$, $\sigma = 0.69$), lognormal-Pareto (I) ($\mu = 7.8$, $\sigma = 0.69$, $\tau = 4000$, $\rho = 0.3611$, $\alpha = 1.0377$), and lognormal-Pareto (II) ($\mu = 7.8$, $\sigma = 0.69$, $\tau = 4000$, $\rho = 0.5018$, $\alpha = 0.4$, $\beta = 6870.486$)

Various R packages are available to assist statistical analysis when using the composite Pareto model. The **CompLognormal** package specializes in composite lognormal models with the upper tail can be specified to be Pareto distribution (Nadarajah & Bakar 2013). The **gendist** package computes the pdf, cdf, and quantile function as well as handles random value generation for composite models (Bakar et al. 2016). The **mistr** package provides some computational framework for composite models (Sablica & Hornik 2020). Other useful packages include **evmix** (Hu & Scarrott 2018) and **ReIns** (Reynkens et al. 2020).

COMPOSITE PARETO I MODEL

The composite Pareto I model has the following pdf

$$f(x|\theta) = \begin{cases} (1-\rho)\frac{f_1(x|\eta)}{F_1(\tau|\eta)}, & x \le \tau \\ \rho f_{P1}(x|\tau,\alpha), & x > \tau \end{cases}, \quad (1)$$

where $\theta = (\rho, \eta, \tau, \alpha)$ is a collection of all the parameters in the model, ρ is the weight of the Pareto type I distribution, $f_1(x|\eta)$ is the pdf of lower data distribution model with parameter η and cumulative distribution function (cdf) $F_1(x|\eta)$, and $f_{P1}(x|\tau,\alpha)$ is the pdf of Pareto type I distribution with threshold $\tau > 0$ and tail index $\alpha > 0$. The lower data distribution model is not specified to any distribution model as to enable a more general distribution model. But for simplicity, it is assumed that $f_1(x|\eta)$ is continuous and differentiable.

The pdf of Pareto type I is

$$f_{P1}(x|\tau,\alpha) = \frac{\alpha\tau^{\alpha}}{x^{\alpha+1}}, \quad x > \tau. \tag{1}$$

To make the pdf (1) continuous and differentiable at τ, we let $f(\tau^-|\theta) = f(\tau^+|\theta)$ and $f'(\tau^-|\theta) = f'(\tau^+|\theta)$ by specifying

$$\alpha = -1 - \frac{\tau f_1'(\tau|\eta)}{f_1(\tau|\eta)} \text{ and } \rho = \frac{\tau f_1(\tau|\eta)^2}{\tau f_1(\tau|\eta)^2 - F_1(\tau|\eta)[f_1(\tau|\eta) + \tau f_1'(\tau|\eta)]},$$

condition on $\tau f_1'(\tau|\eta) < -f_1(\tau|\eta)$ so that $\alpha > 0$. For example, if $f_1(\tau|\eta)$ is a lognormal distribution with logmean μ and logvariance σ^2, let $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cdf of the standard normal distribution, respectively, and let $z = (log\ \tau - \mu)/\sigma$, then the values of α and ρ are as follows

$$\alpha = \frac{log\ \tau - \mu}{\sigma^2} \text{ and } \rho = \frac{\varphi(z)}{\varphi(z) + z\Phi(z)}.$$

COMPOSITE PARETO II MODEL

The composite Pareto II model is similar to composite Pareto I model but with Pareto type II distribution model for the upper tail. It has the pdf of the form

$$f(x|\theta) = \begin{cases} (1-\rho)\frac{f_1(x|\eta)}{F_1(\tau|\eta)}, & x \le \tau \\ \rho f_{P2}(x|\tau,\alpha,\beta), & x > \tau \end{cases}, \tag{2}$$

where $\theta = (\rho, \eta, \tau, \alpha, \beta)$ is the collection of all the parameters, ρ is the weight of Pareto type II distribution in the model, $f_1(x|\eta)$ is the pdf of lower data distribution model with parameter η and cdf $F_1(x|\eta)$, and $f_{P2}(x|\tau, \alpha, \beta)$ is the pdf of Pareto type II distribution with threshold $\tau > 0$, tail index $\alpha > 0$, and scale parameter $\beta > 0$. Again, the lower data distribution model is not specified but it is assumed that $f_1(x|\eta)$ is continuous and differentiable.

The pdf of Pareto type II is

$$f_{P2}(x|\tau,\alpha,\beta) = \frac{1}{\beta}\left(1 + \frac{x-\tau}{\alpha\beta}\right)^{-\alpha-1}, \qquad x > \tau.$$

To satisfy the continuity and differentiability conditions of pdf (2), we must have $f(\tau^-|\theta) = f(\tau^+|\theta)$ and $f'(\tau^-|\theta) = f'(\tau^+|\theta)$ which leads to

$$\beta = \frac{-(\alpha+1)f_1(\tau|\eta)}{\alpha f_1'(\tau|\eta)} \text{ and } \rho = \frac{(\alpha+1)f_1(\tau|\eta)^2}{(\alpha+1)f_1(\tau|\eta)^2 - \alpha f_1'(\tau|\eta)F_1(\tau|\eta)},$$

subject to $f_1'(\tau|\eta) < 0$ so that $\beta > 0$. As an example, if the lower data distribution is a lognormal distribution with logmean μ and logvariance σ^2, let $z = (log\ \tau - \mu)/\sigma$, then the values of β and ρ are as follows

$$\beta = \frac{\sigma^2(\alpha\tau + \tau)}{\alpha\sigma(z+\sigma)} \text{ and } \rho = \frac{\tau\sigma(\alpha+1)\varphi(z)}{\tau\sigma(\alpha+1)\varphi(z) + \alpha\sigma(z+\sigma)\Phi(z)}.$$

STATISTICAL METHODS

PSEUDO-LIKELIHOOD APPROACH

The survey data contains sampling weights which must be considered for unbiased statistical analysis. One approach is by using pseudo-likelihood to replace the likelihood used in analysis. Other approaches that can be used for parameter estimation for data with sampling weights include the Bayesian finite population inference (Skinner et al. 1989) and using representative samples created from the samples and its weight (Gunawan et al. 2020).

Suppose we observe the households income $\boldsymbol{x} = (x_1, ..., x_n)$ and the households' weight $\boldsymbol{w} = (w_1, ..., w_n)$ where n is the sample size. The weights are assumed to be scaled such that $\sum w_i = n$. This can be done by dividing the original (unscaled) weights by its empirical mean. The pseudo-pdf for a household i is the pdf raised to the power of its weight w_i. The pseudo-likelihood can then be written as

$$\tilde{L} = \prod_{i=1}^{n} [f(x_i|\theta)]^{w_i}.$$

If the weights are uninformative, then $w_i = 1$ for all i and the pseudo-likelihood will be the regular likelihood. The pseudo-maximum likelihood estimate is defined as the parameter that maximises the pseudo-likelihood:

$$\hat{\theta} = arg\,\max_{\theta} \{\prod_{i=1}^{n} [f(x_i|\theta)]^{w_i}\}.$$

Computing maximum likelihood estimates for composite Pareto models can be difficult. There are some algorithms discussed in the literature, but in general, the algorithms are computationally costly when sample size is large (Cooray & Ananda 2005; Teodorescu & Vernic 2013, 2009). In our implementation, the maximum pseudo-likelihood estimates are computed numerically using nlminb function in R. It requires us to only specify the likelihood function for the composite Pareto models and the boundary of the parameters.

MODEL SELECTION CRITERIA

To select the best distribution model for the income data, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used. The AIC and BIC values measure the trade-off between the fit of the model and its complexity. Model with the lowest AIC or BIC values are said to be the best model to represent the data out of the candidate models. In cases where the model with the lowest AIC value differs with the model with the lowest BIC value, the simpler model is preferred. However, since sample weights are used in the analysis, the formulae for AIC and BIC must reflect on that as well. Let k be the number of parameters in the distribution model and $\tilde{L}$ be the pseudo-likelihood under the pseudo-maximum likelihood estimates. The AIC and BIC values are modified such that

$$\text{AIC} = 2k - 2\log(\hat{L})$$

$$\text{BIC} = k\log(n) - 2\log(\hat{L})$$

While AIC and BIC values are useful when comparing different distribution models, they do not measure the goodness-of-fit of the distribution model. It is possible that the model with the lowest AIC or BIC values may not have a good fit with the data. To test whether the observed data follows a distribution model, Kolmogorov-Smirnov goodness-of-fit test can be used. However, similar to AIC and BIC formulae, the goodness-of-fit test must be modified to allow for sample weights in the analysis. To do this, we use a theorem proven by Janczura and Weron (2010) below.

Theorem 1. *If $X_1, X_2, ..., X_n$ are independent, $Var(X_i)<\infty$, $0 \le w_i \le M$ for some positive value M for all i =1, ..., n, $\lim_{n\to\infty} \sum w_i = \infty$, and the theoretical distribution F(t) is continuous, then*

$$\frac{\sum_{i=1}^{n} w_i}{\sqrt{\sum_{i=1}^{n} w_i^2}} sup\,|F_n(t) - F(t)|$$

converges weakly to Kolmogorov-Smirnov distribution as $n \to \infty$ where

$$F_n(t) = \frac{\sum_{i=1}^{n} w_i \mathbb{I}(X_i < t)}{\sum_{i=1}^{n} w_i}$$

is the weighted empirical cumulative distribution function and is the indicator function.

Under Theorem 1, suppose $F(t)$ is the target distribution from the model, then

$$D_n = \frac{\sum_{i=1}^{n} w_i}{\sqrt{\sum_{i=1}^{n} w_i^2}} max\,|F_n(t) - F(t)| \quad (3)$$

can be used as a test statistic for the goodness-of-fit test to compare the empirical cdf with the target distribution. The p-value in this case is the probability $P(\kappa > D_n)$ where k follows the Kolmogorov distribution. Note that in the case of uninformative sampling weights where $w_i = 1$ for $i = 1, 2, ..., n$, the test statistic (3) becomes the test statistic for a regular Kolmogorov-Smirnov goodness-of-fit test.

RESULTS AND DISCUSSION

MODEL PERFORMANCE FOR EACH DATASET

The composite Pareto models are applied to the HIS datasets and the model fits are measured using AIC, BIC, and Kolmogorov-Smirnov goodness-of-fit test. We consider 20 candidate models, seven of which are composite Pareto I models, and six are composite Pareto II models. The details of the candidate models are shown in Table 1. The lower data distribution models considered here are commonly used to model income distribution and comparison will be made when Pareto distribution is added to the upper tail. Table 2 shows the AIC, BIC and p-values for all the candidate models when applied to the dataset. The bolded values are the lowest AIC or BIC values, respectively, for each dataset.

TABLE 1. Description of all the candidate models considered in analysis

Code	Lower data distribution model	Upper data distribution model	No. of parameters
LN	Lognormal	None	2
G	Gamma	None	2
W	Weibull	None	2
LN-P1	Lognormal	Pareto type I	3
G-P1	Gamma	Pareto type I	3
W-P1	Weibull	Pareto type I	3
D	Dagum	None	3
B2	Beta 2	None	3
SM	Singh–Maddala	None	3
LN-P2	Lognormal	Pareto type II	4
G-P2	Gamma	Pareto type II	4
W-P2	Weibull	Pareto type II	4
D-P1	Dagum	Pareto type I	4
B2-P1	Beta 2	Pareto type I	4
SM-P1	Singh–Maddala	Pareto type I	4
GB2	Generalized beta of the second kind	None	4
D-P2	Dagum	Pareto type II	5
B2-P2	Beta 2	Pareto type II	5
SM-P2	Singh–Maddala	Pareto type II	5
GB2-P1	Generalized beta of the second kind	Pareto type I	5

TABLE 2. AIC, BIC, and p-values for the candidate models. The bolded figures indicate the best fitted models based on AIC or BIC values

Dataset		2-parameter			3-parameter					
		LN	G	W	LN-P1	G-P1	W-P1	D	B2	SM
MY07	AIC	202863.0	205090.0	206052.2	202775.6	203130.7	203395.2	202853.2	202733.0	202967.5
	BIC	202877.8	205104.8	206067.0	202797.8	203152.9	203417.4	202875.4	202755.2	202989.7
	p-value	0.0001	< 0.0001	< 0.0001	0.0032	< 0.0001	< 0.0001	0.0011	0.1848	0.0001
MY09	AIC	219069.1	221067.9	222026.3	219037.3	219506.6	219843.2	219225.8	219019.7	219353.6
	BIC	219084.1	221082.8	222041.2	219059.7	219529.0	219865.6	219248.2	219042.1	219376.0
	p-value	0.0061	< 0.0001	< 0.0001	0.0036	< 0.0001	< 0.0001	0.0001	0.0115	< 0.0001
MY12	AIC	229973.0	231974.8	233097.7	229923.8	230373.9	230721.5	230146.3	229929.3	230246.5
	BIC	229988.0	231989.7	233112.7	229946.2	230396.4	230744.0	230168.8	229951.8	230269.0
	p-value	0.0728	< 0.0001	< 0.0001	0.0295	< 0.0001	< 0.0001	0.0001	0.0208	< 0.0001
MY14	AIC	433311.5	437617.7	440307.2	432932.2	433294.2	433793.5	433047.6	432921.0	433186.5
	BIC	433327.7	437633.9	440323.4	432956.5	433318.5	433817.8	433071.9	432945.3	433210.9
	p-value	< 0.0001	< 0.0001	< 0.0001	0.1717	< 0.0001	< 0.0001	0.0111	0.1061	0.0006
MY16	AIC	422919.5	426853.4	429774.8	422667.7	423163.2	423733.1	422890.5	422665.7	423032.0
	BIC	422935.6	426869.5	429790.9	422691.9	423187.4	423757.3	422914.7	**422689.9**	423056.2
	p-value	0.0008	< 0.0001	< 0.0001	0.1582	< 0.0001	< 0.0001	0.0003	0.5289	< 0.0001

Dataset		4-parameter							5-parameter			
		LN-P2	G-P2	W-P2	D-P1	B2-P1	SM-P1	GB2	D-P2	B2-P2	SM-P2	GB2-P1
MY07	AIC	202724.0	202728.1	202757.8	202855.2	202735.0	202969.5	202727.6	**202719.8**	202724.9	202729.1	202767.8
	BIC	**202753.6**	202757.7	202787.4	202884.8	202764.6	202999.1	202757.2	202756.9	202761.9	202766.1	202804.8
	p-value	0.7994	0.7370	0.0630	0.0011	0.1847	0.0001	0.4770	0.7299	0.7128	0.6318	0.0119
MY09	AIC	218967.9	218981.7	219014.7	219227.8	219021.8	219355.6	218970.2	**218958.6**	218959.2	218978.8	218971.8
	BIC	218997.8	219011.6	219044.6	219257.7	219051.6	219385.5	219000.1	**218995.9**	218996.5	219016.1	219009.2
	p-value	0.1343	0.0371	0.0055	0.0001	0.0114	< 0.0001	0.1169	0.3419	0.4845	0.0628	0.1143
MY12	AIC	229906.7	229951.4	230003.8	230148.3	229931.4	230248.5	**229886.2**	229897.7	229891.5	229951.0	229894.1
	BIC	229936.6	229981.3	230033.7	230178.3	229961.3	230278.5	**229916.2**	229935.1	229928.9	229988.4	229931.6
	p-value	0.4129	0.0021	0.0002	0.0001	0.0208	< 0.0001	0.2441	0.1256	0.3073	0.0033	0.1040
MY14	AIC	**432901.0**	433014.2	433182.0	433049.6	432923.3	433188.5	432921.1	432972.0	432912.2	433063.1	432923.6
	BIC	**432933.4**	433046.6	433214.4	433082.0	432955.7	433221.0	432953.5	433012.5	432952.7	433103.6	432964.1
	p-value	0.5248	0.0140	0.0015	0.0111	0.1764	0.0006	0.2132	0.0985	0.4793	0.0397	0.2402
MY16	AIC	422665.5	422802.2	422950.8	422892.5	**422659.0**	423034.0	422666.2	422740.5	422659.6	422833.5	422664.3
	BIC	422697.8	422834.4	422983.1	422924.8	422691.3	423066.3	422698.4	422780.9	422700.0	422873.8	422704.6
	p-value	0.3603	0.0009	0.0001	0.0003	0.4043	< 0.0001	0.6218	0.0329	0.4119	0.0019	0.2550

From Table 2, none of the 2-parameter candidate models fit the data based on the low *p*-values. Therefore, lognormal, gamma or Weibull distributions alone are not enough to model the data and should not be used to model the income distribution. Some improvements can be seen when Pareto distribution is added to the 2-parameter models and the *p*-values increase. For example, LN-P1 model has *p*-values greater than 0.1 for the year 2014 and 2016 and LN-P2 model has *p*-values greater than 0.1 for all the datasets. Hence, using Pareto distributions to model the upper income data improves the model fit.

Table 2 also shows that D-P2, GB2, and LN-P2 are the models which have both the lowest AIC and BIC values for the year 2009, 2012, and 2014, respectively. These models are then considered as the best models to describe the income distribution for the respective year, out of all the candidate models. For the year 2007, D-P2 gives the lowest AIC value but LN-P2 gives the lowest BIC value. Since LN-P2 is a simpler model compared to D-P2 due to lower number of parameters, LN-P2 is considered as the best model for the year 2007. Similarly, B2 model is considered as the best model for the year 2016 when B2-P1 gives the lowest AIC value and B2 gives the lowest BIC value. Therefore, the best models to describe the Malaysian household income data for the year 2007, 2009, 2012, 2014, and 2016 are LN-P2, D-P2, GB2, LN-P2 and B2 models, respectively. Additionally, all the best models have high *p*-values indicating that they have good fit for the data.

OVERALL MODEL PERFORMANCE

To identify the overall performance for the models, the models are ranked based on the BIC values as shown in Table 3. In the table, lower ranking of BIC values indicates a better model, and the average rank over the five datasets is given for all the candidate models. The model with the lowest average rank is LN-P2 (2.8), followed by GB2 (3.6) and B2-P2 (3.8). This indicates that LN-P2 is the best model out of all the candidate models for describing the household income distribution in Malaysia. The *p*-values for LN-P2 for each dataset are also found to be high indicating a good fit for all the datasets.

TABLE 3. Ranks for each distribution models based on BIC values

Dataset	LN	G	W	LN-P1	G-P1	W-P1	D	B2	SM	LN-P2
MY07	13	19	20	10	17	18	12	2	15	1
MY09	12	19	20	11	17	18	13	8	15	3
MY12	10	19	20	6	17	18	13	7	15	5
MY14	17	19	20	6	16	18	10	2	13	1
MY16	13	19	20	3	17	18	11	1	15	4
Average	13	19	20	7.2	16.8	18	11.8	4	14.6	2.8

Dataset	G-P2	W-P2	D-P1	B2-P1	SM-P1	GB2	D-P2	B2-P2	SM-P2	GB2-P1
MY07	5	9	14	7	16	4	3	6	8	11
MY09	6	9	14	10	16	4	1	2	7	5
MY12	9	12	14	8	16	1	4	2	11	3
MY14	9	14	11	5	15	4	8	3	12	7
MY16	9	14	12	2	16	5	8	6	10	7
Average	7.6	11.6	13	6.4	15.8	3.6	4.8	3.8	9.6	6.6

APPLICATION TO OTHER COUNTRIES

The same analyses are performed to the income data from Italy and United Kingdom to study the performance of composite Pareto models for datasets from these two countries. For Italy, the net disposable income divided by the square root of number of household members for the year 2014 and 2016 are used. For United Kingdom, the equivalised disposable income using OECD scale for the year 2016 and 2018 are used. Since the distribution models considered can only be fitted to positive incomes, households with zero or negative income are removed. Table 4 shows the AIC, BIC and *p*-values of the candidate models when applied to Italy and UK datasets.

TABLE 4. AIC, BIC, and p-values for the candidate models using Italy and UK datasets. The bolded figures indicate the best fitted models based on AIC or BIC values

Dataset		2-parameter			3-parameter					
		LN	G	W	LN-P1	G-P1	W-P1	D	B2	SM
IT14	AIC	134801.5	133308.4	133903.4	134802.9	133003.9	132745.8	132720.2	133208.1	132826.1
	BIC	134815.5	133322.4	133917.4	134823.9	133024.9	132766.8	**132741.2**	133229.1	132847.1
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.0007	0.1266	0.1758	< 0.0001	0.1840
IT16	AIC	123511.6	121538.9	122015.2	123513.7	121300.6	121015.5	120991.4	121471.0	121100.1
	BIC	123525.4	121552.8	122029.0	123534.5	121321.3	121036.2	121012.1	121491.7	121120.8
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.0009	0.1847	0.2115	< 0.0001	0.0888
UK16	AIC	66494.0	66270.5	66688.0	66497.8	66095.2	66107.9	66066.1	66150.9	**66060.5**
	BIC	66507.0	66283.5	66700.9	66517.3	66114.6	66127.4	66085.6	66170.4	**66080.0**
	p-value	0.0001	0.0001	< 0.0001	0.0001	0.1858	0.0063	0.4382	0.0980	0.5056
UK18	AIC	74287.8	73983.1	74461.2	74292.1	73766.9	73762.1	73717.7	73826.8	73715.6
	BIC	74301.0	73996.3	74474.4	74311.8	73786.7	73781.9	73737.4	73846.6	**73735.4**
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.1794	0.0585	0.4151	0.0352	0.4451

Dataset		4-parameter							5-parameter			
		LN-P2	G-P2	W-P2	D-P1	B2-P1	SM-P1	GB2	D-P2	B2-P2	SM-P2	GB2-P1
IT14	AIC	134677.1	133002.8	132747.6	**132715.8**	133020.7	132747.8	132716.0	132717.8	133005.7	132749.6	132717.4
	BIC	134705.1	133030.8	132775.6	132743.8	133048.7	132775.8	132744.0	132752.8	133040.7	132784.6	132752.4
	p-value	< 0.0001	0.0007	0.1312	0.1654	0.0087	0.1267	0.1780	0.1654	0.0004	0.1312	0.1721
IT16	AIC	123303.2	121300.0	121015.3	120984.9	121309.4	121017.5	120983.6	**120982.9**	121302.0	121017.3	120986.6
	BIC	123330.8	121327.6	121042.9	121012.5	121337.0	121045.2	**121011.2**	121017.5	121336.5	121051.9	121021.1
	p-value	< 0.0001	0.0008	0.1877	0.2244	0.0045	0.1840	0.2442	0.2267	0.0009	0.1873	0.2143
UK16	AIC	66495.1	66094.4	66076.4	66068.2	66097.2	66062.7	66062.0	66063.6	66096.7	66063.2	66064.2
	BIC	66521.0	66120.7	66102.4	66094.2	66123.2	66088.7	66088.0	66096.0	66129.2	66095.7	66096.7
	p-value	0.0001	0.1976	0.3549	0.4380	0.2154	0.5069	0.4694	0.4986	0.1913	0.4915	0.4691
UK18	AIC	74276.1	73759.5	73722.7	73719.8	73768.9	73717.8	73716.6	**73714.3**	73762.6	73716.3	73718.8
	BIC	74302.5	73785.8	73749.1	73746.2	73795.3	73744.2	73743.0	73747.3	73795.6	73749.26	73751.7
	p-value	< 0.0001	0.2114	0.2957	0.4150	0.1761	0.4466	0.4012	0.4411	0.3440	0.4336	0.4015

From Table 4, it is clear that none of the 2-parameter models have a good fit for any of the datasets since the *p*-values are found to be very small. When the Pareto distributions are added to the 2-parameter models, the models improve significantly and some of them achieve good fit. For example, W-P1 model has good fit with both Italy datasets and G-P1 model has good fit with both UK datasets. However, the performance of the composite Pareto models are still worse when compared to some of the non-composite Pareto models based on the AIC and BIC values. The best models to describe the income distributions for Italian data for the year 2014 and 2016, UK's data for the year 2016 and 2018, are D, GB2, SM and SM models, respectively. In this case, for these two countries, none of the composite Pareto models provide the best fit, unlike in the case of Malaysian income distribution.

This example shows that composite Pareto models are not always better than other models in fitting income data. While the composite models are useful in describing income distribution in Malaysia, they are not as useful for the income distribution in Italy and UK, even though some of the composite Pareto models have good fit with the datasets. There is no known absolute rule on when or where the composite Pareto models would be better for the data. In practice, all candidate models must be applied to the data and comparisons between the models can be made using AIC and BIC values, as well as goodness-of-fit tests, as what have been done in this paper.

LORENZ CURVE AND GINI COEFFICIENT

The Lorenz curve and Gini coefficient are useful for assessing the income inequality of a population. The Lorenz curve $LC(u)$ is a graphical representation of the proportion of wealth accumulated by the bottom $100u\%$ of the population. When the income distribution is given as a continuous probability distribution with pdf $f(x|\theta)$ and quantile function $F^{-1}(u|\theta)$, the Lorenz curve can be defined as

$$LC(u) = \frac{1}{\mu_X}\int_0^{F^{-1}(u)} x\, f(x|\theta)dx,$$

where μ_X is the mean of income. For LN-P2 model with the lower data distributed by a lognormal distribution with logmean μ and logvariance σ^2, the overall mean is

$$\mu_X = (1-\rho)e^{\mu+\frac{\sigma^2}{2}}\frac{\Phi(z-\sigma)}{\Phi(z)} + \rho\left(\tau + \frac{\alpha\beta}{\alpha-1}\right),$$

where $z = (log\ \tau-\mu)/\sigma$ and $\alpha > 1$. Let

$$A(u) = \frac{(1-\rho)e^{\mu+\frac{\sigma^2}{2}}}{\mu_X\Phi(z)}\Phi\left[\Phi^{-1}\left(\frac{u\Phi(z)}{1-\rho}\right) - \sigma\right],$$

where $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the cdf and quantile function of the standard normal distribution, respectively. Then the Lorenz curve under LN-P2 model can be written as follows

$$LC(u) = \begin{cases} A(u), & u \le 1-\rho \\ A(1-\rho) + \frac{\rho\alpha^2\beta}{\mu_X(\alpha-1)}\left[1-\left(\frac{1-u}{\rho}\right)^{1-\frac{1}{\alpha}}\right] + \frac{(\tau-\alpha\beta)(u+\rho-1)}{\mu_X}, & u > 1-\rho, \end{cases}$$

and the corresponding Gini coefficient is

$$\begin{aligned} \text{Gini} &= 1-2\int_0^1 LC(u)du \\ &= 1-2\int_0^{1-\rho} A(u)du - 2\rho A(1-\rho) - \frac{2\rho\alpha^2\beta}{\mu_X(\alpha-1)}\left(1-\alpha\rho^{\frac{1}{\alpha}-1}\right) - \frac{(\tau-\alpha\beta)(2\rho+1)}{\mu_X}. \end{aligned}$$

Unfortunately, the integral in this equation cannot be solved analytically and numerical method, for example the trapezoidal rule, is required to approximate $\int_0^{1-\rho} A(u)du$ and the Gini coefficient.

The Lorenz curves for all five datasets of the Malaysian income are shown in Figure 2. From the figure, the Lorenz curve approaches the line of equality from 2007 to 2016. Additionally, the Gini coefficients for the year 2007, 2009, 2012, 2014, and 2016 are 0.438, 0.433, 0.421, 0.404, and 0.390, respectively, which shows a decreasing trend. These indicate that the income inequality is reduced over the period of study and wealth is shared more evenly in the population.

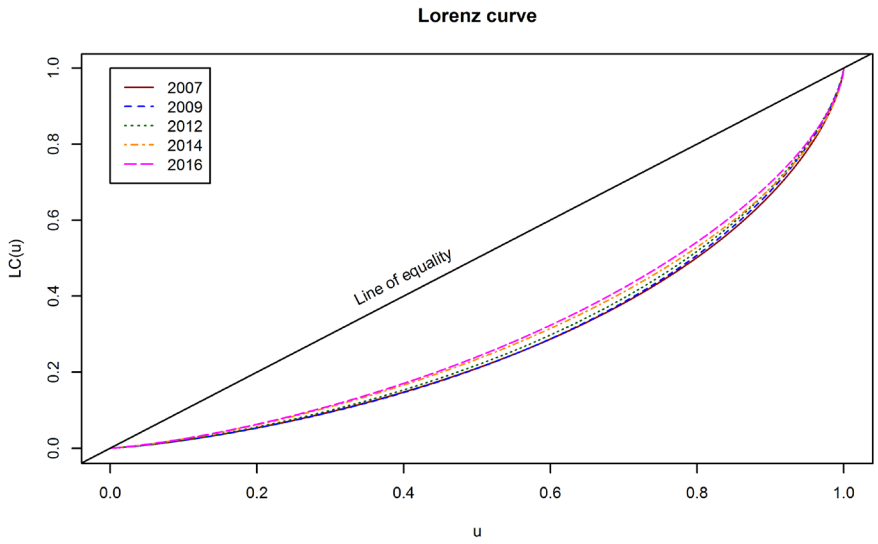

FIGURE 2. Lorenz curves for the Malaysian income data

CONCERNS AND LIMITATIONS

While composite Pareto models may be useful in describing income distribution in Malaysia, there are some concerns that arise from the application and should be discussed. Firstly, the pseudo-maximum likelihood estimates are computed numerically, not analytically. Our application requires initial values for the parameters to compute the pseudo-maximum likelihood estimates. In our application, for some models and initial estimates, the method has failed to converge to a solution and forced us to use several other initial estimates. There is no guarantee that the estimates given by the method are the values that maximise the pseudo-maximum likelihood.

There have been other methods or algorithms proposed in literature. Cooray and Ananda (2005) and Teodorescu and Vernic (2009) have proposed algorithm to find the maximum likelihood by finding the interval at which the threshold parameter τ lies. Suppose the

observed data are ordered such that $x_1 \leq x_2 \leq ... \leq x_n$. The proposed algorithm attempts to find the value of m such that $x_m \leq \tau \leq x_{m+1}$ by maximising the likelihood for every $m = 1, 2, ..., n$.. As one would expect, this is very costly for large n as every n -1 intervals have to be checked. Alternatively, Teodorescu and Vernic (2013) and Bee (2015) have proposed to use the method of moments, instead of maximum likelihood estimator, for parameter estimation. Another possible workaround is by estimating the threshold and the tail index first, for example by using Kolmogorov-Smirnov statistic as done by Safari et al. (2018a). Then, using the estimated threshold and tail index values, the rest of the parameters can be estimated easily by using maximum likelihood estimator. The proposed algorithms and methods are not used in our application due to the large number of sample size and the variety of lower data distributions considered.

Another concern is regarding the interpretation of the composite Pareto models. In extreme value theory for example, the threshold for the Pareto type II distribution is often regarded as the cut-off for extreme values. However, when the composite Pareto models are applied to the datasets in the study, in many cases, the estimated thresholds are found to be too low and the proportions of data coming from Pareto distribution are found to be too high to be interpreted as the high income earners. For example, under LN-P2 model for the Malaysian household income year 2007, the pseudo-maximum likelihood estimate for the proportion of data under Pareto distribution is $\hat{\rho}$.= 0.4926. It is not sensible to conclude that 49.26% of households are high income earners as this proportion is deemed too high. Clearly in this case the parameters cannot be used as the cut-off points for the upper income earners. Nevertheless, since the composite Pareto model provide a better fit for some of the datasets, to some extent, the underlying features of the income distribution can be reliably explained based on this model.

Conclusion

In this paper, the composite Pareto models are applied to the Malaysian household income data for the year 2007, 2009, 2012, 2014, and 2016. Using pseudo-likelihood based AIC, BIC, and Kolmogorov-Smirnov goodness-of-fit test, the LN-P2 model, which consists of lognormal distribution for lower data distribution and Pareto type II distribution model for the upper data distribution, is found to have the best overall performance when compared to other candidate models. Therefore, LN-P2 model can describe the income distribution in Malaysia better compared to other candidate models.

However, composite Pareto models are found to be less useful to describe the income distribution for Italy and UK, as it is found that models without Pareto distribution have better fit compared to the composite Pareto models. This indicates that the composite Pareto models may not be suitable for every dataset. There is no known rule to indicate when composite Pareto models would work better. Thus, practitioners may have to test every possible candidate model and compare their performance for example by using AIC and BIC values.

Some concerns are found in the application of composite Pareto models to these income datasets. Particularly, the numerical method used to compute the pseudo-maximum likelihood estimates may be problematic even though it is the easiest method to find the estimates. The composite Pareto model interpretation on the income data is also unclear. More works are required for these parts.

Acknowledgements

The authors would like to thank the Department of Statistics Malaysia (DOSM), Bank Data UKM, UK Office for National Statistics and Bank of Italy for providing datasets for this study.

References

Aminzadeh, M.S. & Deng, M. 2019. Bayesian predictive modeling for Inverse Gamma-Pareto composite distribution. *Communications in Statistics-Theory and Methods* 48(8): 1938-1954.

Arnold, B.C. 2008. Pareto and generalized Pareto distributions. In *Modeling Income Distributions and Lorenz Curves,* edited by Chotikapanich, D. New York: Springer Science & Business Media. pp. 119-145.

Bakar, S.A.A., Nadarajah, S., Adzhar, Z.A.A.K. & Mohamed, I. 2016. Gendist: An R package for generated probability distribution models. *PLoS ONE* 11(6): e0156537.

Bakar, S.A.A., Hamzah, N.A., Maghsoudi, M. & Nadarajah, S. 2015. Modeling loss data using composite models. *Insurance: Mathematics and Economics* 61: 146-154.

Bee, M. 2015. Estimation of the lognormal-Pareto distribution using probability weighted moments and maximum likelihood. *Communications in Statistics-Simulation and Computation* 44(8): 2040-2060.

Ciumara, R. 2006. An actuarial model based on the composite Weibull-Pareto distribution. *Mathematical Reports* 8(4): 401-414.

Congressional Budget Office. 2019. Projected changes in the distribution of household income, 2016 to 2021.

Cooray, K. & Ananda, M. 2005. Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal* 2005(5): 321-334.

Gunawan, D., Panagiotelis, A., Griffiths, W. & Chotikapanich, D. 2020. Bayesian weighted inference from surveys. *Australian & New Zealand Journal of Statistics* 62(1): 71-94.

Hu, Y. & Scarrott, C. 2018. evmix: An R package for extreme value mixture modeling, threshold estimation and boundary corrected kernel density estimation. *Journal of Statistical Software* 84(5): 1-18.

Janczura, J. & Weron, R. 2010. Goodness-of-fit testing for regime-switching models. MPRA Paper. https://mpra.ub.uni-muenchen.de/id/eprint/22871.

Masseran, N., Yee, L.H., Safari, M.A.M. & Ibrahim, K. 2019. Power law behavior and tail modeling on low income distribution. *Mathematics and Statistics* 7(3): 70-77.

Nadarajah, S. & Bakar, S.A.A. 2013. CompLognormal: An R package for composite lognormal distributions. *The R Journal* 5(2): 97-103.

OECD. 2015. *In it Together: Why Less Inequality Benefits All*. Paris: OECD Publishing. pp. 19-58.

Pfeffermann, D. 1993. The role of sampling weights when modeling survey data. *International Statistical Review/ Revue Internationale de Statistique* 61(2): 317-337.

Ragayah, H.M.Z. 2008. Income inequality in Malaysia. *Asian Economic Policy Review* 3(1): 114-132.

Razak, F.A. & Shahabuddin, F.A. 2018. Malaysian household income distribution: A fractal point of view. *Sains Malaysiana* 47(9): 2187-2194.

Reynkens, T., Verbelen, R., Bardoutsos, A., Cornilly, D., Geogebeur, Y. & Herrmann, K. 2020. ReIns: Functions from "reinsurance: actuarial and statistical aspects." https://CRAN.R-project.org/package=ReIns.

Sablica, L. & Hornik, K. 2020. mistr: A Computational framework for mixture and composite distributions. *The R Journal* 12(1): 283-299.

Safari, M.A.M., Masseran, N. & Ibrahim, K. 2018a. Optimal threshold for Pareto tail modelling in the presence of outliers. *Physica A: Statistical Mechanics and its Applications* 509: 169-180.

Safari, M.A.M., Masseran, N. & Ibrahim, K. 2018b. A robust semi-parametric approach for measuring income inequality in Malaysia. *Physica A: Statistical Mechanics and its Applications* 512: 1-13.

Scollnik, D.P.M. 2007. On composite lognormal-Pareto models. *Scandinavian Actuarial Journal* 2007(1): 20-33.

Scollnik, D.P.M. & Sun, C. 2012. Modeling with Weibull-Pareto models. *North American Actuarial Journal* 16(2): 260-272.

Skinner, C.J., Holt, D. & Smith, T.M.F. 1989. *Analysis of Complex Surveys.* Chichester, New York: Wiley.

Teodorescu, S. & Vernic, R. 2013. On composite Pareto models. *Mathematical Reports* 15(65): 11-29.

Teodorescu, S. & Vernic, R. 2009. Some composite exponential-Pareto models for actuarial prediction. *Journal for Economic Forecasting* 12(4): 82-100.

Department of Mathematical Sciences
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: hilmi.majid@ukm.edu.my

Received: 19 June 2020
Accepted: 19 November 2020