Sains Malaysiana 53(2)(2024): 461-476
http://doi.org/10.17576/jsm-2024-5302-18

Examining Tail Index Estimators in New Pareto Distribution: Monte Carlo Simulations and Income Data Applications

(Menyemak Penganggar Indeks Ekor dalam Taburan Pareto Baharu: Simulasi Monte Carlo dan Aplikasi Data Pendapatan)

Muhammad Aslam Mohd Safari[1,2,*], Nurulkamal Masseran[3] & Mohd Azmi Haron[4]

[1]Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[3]Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[4]Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Received: 22 July 2023/Accepted: 16 January 2024

ABSTRACT

An evolved form of Pareto distribution, the new Pareto-type distribution, offers an alternative model for data with heavy-tailed characteristics. This investigation examines and discusses fourteen diverse estimators for the tail index of the new Pareto-type, including estimators such as maximum likelihood, method of moments, maximum product of spacing, its modified version, ordinary least squares, weighted least squares, percentile, Kolmogorov-Smirnov, Anderson-Darling, its modified version, Cramér-von Mises, and Zhang's variants of the previous three. Using Monte Carlo simulations, the effectiveness of these estimators is compared both with and without the presence of outliers. The findings show that, without outliers, the maximum product of spacing, its modified version, and maximum likelihood are the most effective estimators. In contrast, with outliers present, the top performers are Cramér-von Mises, ordinary least squares, and weighted least squares. The study further introduces a graphical method called the new Pareto-type quantile plot for validating the new Pareto-type assumptions and outlines a stepwise process to identify the optimal threshold for this distribution. Concluding the study, the new Pareto-type distribution is employed to model the high-end household income data from Italy and Malaysia, leveraging all the methodologies proposed.

Keywords: Estimation techniques; heavy-tailed data; income data modelling; Monte Carlo analysis; Pareto distribution; robustness

ABSTRAK

Satu taburan Pareto yang berkembang iaitu taburan jenis Pareto baharu, menawarkan model alternatif untuk data dengan ciri ekor berat. Kajian ini meneliti dan membincangkan empat belas penganggar yang pelbagai bagi indeks ekor jenis Pareto baharu, termasuk penganggar seperti kebolehjadian maksimum, kaedah momen, produk jarak maksimum bersama versi yang diubah suai, kuasa dua terkecil biasa, kuasa dua terkecil berwajaran, persentil, Kolmogorov-Smirnov, Anderson-Darling Bersama versi yang diubah suai, Cramér-von Mises, dan varian Zhang bagi Kolmogorov-Smirnov, Anderson-Darling serta Cramér-von Mises. Dengan menggunakan simulasi Monte Carlo, keberkesanan penganggar ini dibandingkan dengan kehadiran dan tanpa kehadiran titik terpencil. Hasil kajian menunjukkan bahawa, tanpa titik terpencil, produk jarak maksimum bersama versi yang diubah suai dan kebolehjadian maksimum adalah penganggar yang paling berkesan. Sebaliknya, dengan kehadiran titik terpencil, penganggar terbaik adalah Cramér-von Mises, kuasa dua terkecil biasa dan kuasa dua terkecil berwajaran. Kajian ini seterusnya memperkenalkan kaedah grafik yang disebut sebagai plot kuantil jenis Pareto baharu untuk mengesahkan andaian jenis Pareto baharu dan menggariskan proses bertahap untuk mengenal pasti ambang optimum untuk taburan ini. Mengakhiri kajian, taburan jenis Pareto baharu digunakan untuk memodelkan data pendapatan isi rumah kelas atas dari Itali dan Malaysia, memanfaatkan semua kaedah yang dicadangkan.

Kata kunci: Analisis Monte Carlo; data ekor berat; kaedah penganggaran; keteguhan; pemodelan data pendapatan; taburan Pareto

INTRODUCTION

Introduced by Vilfredo Pareto, an Italian civil engineer, economist, and sociologist, the Pareto distribution was initially conceptualized as a means to illustrate the distribution of income or wealth within a society (Amoroso 1938). Due to its inherent properties of heavy-tailed behavior, the Pareto distribution has found frequent application in modelling the upper tail, or top income, of income distribution (Díaz, Cubillos & Griñen 2021; García & Caballero 2021; Majid & Ibrahim 2021; Majid, Ibrahim & Masseran 2023; Safari, Masseran & Ibrahim 2018a). Its suitability in this context is further reinforced by the observed transition in the shape of empirical income distribution from the middle segment, which demonstrates exponential decay, to the upper tail where decay is relative to power, hence exhibiting power-law behavior (Safari et al. 2021). Additionally, the Pareto model's ability to accurately depict top income data has made it a key tool in calculating economic indicators, like the Gini coefficient (Alfons, Templ & Filzmoser 2013; Giorgi & Gigliarano 2017; Hlasny & Verme 2018). Over time, its application has extended beyond economics into a variety of fields including the sciences, medicine, social sciences, and finance (Bee, Riccaboni & Schiavo 2019; Coronel-Brizio & Hernandez-Montoya 2005; Filimonov & Sornette 2015; Gabaix 2009; Giesen, Zimmermann & Suedekum 2010; Lux & Alfarano 2016; Meyer & Held 2014; Newman 2005; Pinto, Lopes & Machado 2012; Xu et al. 2017).

In more recent years, a new Pareto-type (NP) distribution has been developed by Bourguignon, Saulo and Fernandez (2016). They expounded upon the probabilistic and inferential properties of this distribution and demonstrated its applicability in modeling income and reliability data. The NP distribution, a transformation of the half logistic distribution and a generalization of the Pareto distribution, has had its additional significant properties further explored by Sarabia, Jordá and Prieto (2019). The NP distribution is characterized by the transformation of the half logistic distribution and is a generalization of the well-known Pareto distribution. The NP distribution is increasingly being recognized as a superior alternative to the traditional Pareto distribution for a broader range of data.

One notable aspect of the traditional Pareto distribution is the 80/20 rule, a principle where approximately 80% of outcomes result from 20% of causes. In the context of the conventional Pareto distribution, this rule is typically observed when the tail index is approximately 1.16 (Dunford, Su & Tamang 2014). Interestingly, this rule also applies to the NP distribution, with a slight modification in the tail index value. For the NP distribution, the 80/20 rule is observed when the tail index is around 1.2062. This subtle yet significant variation in the tail index value for the NP distribution highlights its unique characteristics and potential applicability, especially in the analysis of income distribution and other areas where Pareto's principle is traditionally applied.

If we assume that a random variable X is associated with an NP distribution, its probability density function (PDF), cumulative distribution function (CDF), and quantile function are given as follows:

$$f(x;\alpha,x_0)=\frac{2\alpha(x_0/x)^{\alpha}}{x[1+(x_0/x)^{\alpha}]^2}=\frac{2\alpha x_0^{\alpha}x^{\alpha-1}}{(x^{\alpha}+x_0^{\alpha})^2},\quad x>x_0,\quad (1)$$

$$F(x;\alpha,x_0)=\frac{1-(x_0/x)^{\alpha}}{1+(x_0/x)^{\alpha}}=1-\frac{2x_0^{\alpha}}{x^{\alpha}+x_0^{\alpha}},\quad x>x_0, (2)$$

and

$$Q(y;a;x_0)=x_0\left(\frac{1+y}{1-y}\right)^{1/\alpha},\quad 0<y<1.\qquad (3)$$

Herein, α signifies the shape parameter or tail index, and x_0 refers to the scale parameter or threshold. The random variable with the CDF illustrated in Equation (2) can be denoted as $X \sim NP(\alpha, x_0)$.

Bourguignon, Saulo and Fernandez (2016) proposed and employed the maximum likelihood (ML) estimator for evaluating the tail index of the NP model. The tail index's ML estimate can be derived by maximizing the log-likelihood function of the NP distribution with respect to α. Typically, the ML estimator is regarded as efficient and is frequently used in parametric distribution for parameter estimation. However, the ML estimator's efficacy diminishes in the presence of outliers, often leading to significant bias. There exists a range of alternative estimators to estimate the NP model's tail index. Aligning with our primary objective, this study contemplates and introduces 14 diverse estimators for the NP tail index. Subsequently, the performance of all 14 estimators is evaluated through a simulation study, in settings both with and without outliers. The findings from this study would enable us to identify the most efficient estimator for the NP tail index under varying conditions, thereby offering a practical guide for choosing the right methodology for real-world data applications.

This investigation also proposes a graphical instrument, termed the NP quantile plot, employed for examining the NP model's assumption in the data. This tool serves as an initial analytical instrument before conducting further statistical analysis. Additionally, we present a straightforward stepwise approach to establish the NP model's threshold. Essentially, the process involves choosing a threshold value that minimizes the Kolmogorov-Smirnov statistic. This method enables optimal determination of the threshold parameter. In terms of real-world data application, we utilize all the methodologies proposed in this paper to model the top-end data of household income in Italy and Malaysia.

The remainder of this paper unfolds as follows: Next section presents the 14 unique estimators for the NP tail index. The NP quantile plot is introduced in the following section. Subsequently, we outline a procedure for establishing the optimal threshold of the NP model, followed by the contrasts performances of the 14 NP tail index estimators through a Monte Carlo simulation. In the section that follows, we employ the NP distribution to model the top-end data of household income. Finally, in the last section, we summarize our findings and conclusions.

TAIL INDEX ESTIMATORS FOR THE NP MODEL

This section introduces 14 estimators designed to assess the tail index of the NP distribution. These estimators include the ML, Method of Moments (MoM), Maximum Product of Spacing (MPS), Modified Maximum Product of Spacing (MMPS), Ordinary Least Squares (OLS), Weighted Least Squares (WLS), Percentile (PC), Kolmogorov-Smirnov (KS), Anderson-Darling (AD), Modified Anderson-Darling (MAD), Cramér-von Mises (CVM), Zhang's Kolmogorov-Smirnov (ZKS), Zhang's Anderson-Darling (ZAD), and Zhang's Cramér-von Mises (ZCVM). For all estimators mentioned, it is assumed that the threshold parameter x_0 is already known. In the Threshold Selection section, we will provide a straightforward process for choosing an appropriate value for the parameter x_0.

MAXIMUM LIKELIHOOD

Consider a random sample of size n, denoted as $X_1, X_2, \ldots, X_n$ from the NP(α, x_0) distribution. The NP's log-likelihood function can be expressed as follows:

$$\ell(\alpha, x_0) = n\,log(2\alpha) - n\,log(x_0) \\ +(\alpha+1)\sum_{i=1}^{n} log\left(\frac{x_0}{x_i}\right) - 2\sum_{i=1}^{n} log\left[1+\left(\frac{x_0}{x_i}\right)^{\alpha}\right]. \quad (4)$$

The ML estimations of the paramter α, designated as $\hat{\alpha}$ $\hat{}_{\text{ML}}$, are derived by maximizing the log-likelihood function with respect to α. In other terms, $\hat{\alpha}_{\text{ML}}$ can also be found by resolving the following non-linear equation:

$$\frac{\partial \ell(\alpha, x_0)}{\partial \alpha} = -2\sum_{i=1}^{n} \frac{\left(\frac{x_0}{x_i}\right)^{\alpha} log\left(\frac{x_0}{x_i}\right)}{1+\left(\frac{x_0}{x_i}\right)^{\alpha}} \\ +\sum_{i=1}^{n} log\left(\frac{x_0}{x_i}\right) + \frac{n}{\alpha} = 0. \quad (5)$$

METHOD OF MOMENTS

The MoM estimates of parameter α, denoted as $\hat{\alpha}_{\text{MoM}}$, are achieved by equating the first theoretical moment of the NP distribution to the first empirical moment. As per Sarabia, Jordá and Prieto (2019), if the random variable X follows NP(α, x_0), the rth moment of X is represented as:

$$E(X^r) = 2x_0^r B\left(\frac{1}{2}; 1-\frac{r}{\alpha}, 1+\frac{r}{\alpha}\right), \quad r > \alpha, \quad (6)$$

where $B(\cdot;\ \cdot,\cdot)$ represents the incomplete beta function defined as:

$$B(y; p, q) = \int_0^y t^{p-1}(1-t)^{q-1}dt, \\ 0 \le y \le 1,\ p > 0,\ q > 0. \quad (7)$$

By aligning the first theoretical moment with the first sample moment, we derive:

$$2x_0 B\left(\frac{1}{2}; 1-\frac{1}{\alpha}, 1+\frac{1}{\alpha}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (8)$$

The estimate $\hat{\alpha}_{\text{MoM}}$ can be obtained numerically by resolving the nonlinear equation expressed in Equation (8) for α.

MAXIMUM AND MODIFIED MAXIMUM PRODUCT OF SPACING

Parameter α estimates, denoted $\hat{\alpha}_{MPS}$, are calculated using the MPS methodology, suggested by Cheng and Amin (1983). This technique revolves around the concept of discrepancies between consecutive data points' CDF values. The uniform spacing of a random sample from the NP(α, x_0) distribution is characterized by:

$$D_i = F(x_{(i)}; \alpha, x_0) - F(x_{(i-1)}; \alpha, x_0), \quad (9)$$

where $x_{(i)}$ stands for an ordered sample observation for i = 1, 2,…, n. Here, $F\ (x_{(0)});\ \alpha,\ x_0) = 0$, $F\ (x_{(n+1)};\ \alpha,\ x_0) = 1$ and $\sum_{i=1}^{n+1} D_i = 1$. Using the MPS, the parameter estimates $\hat{\alpha}_{MPS}$ are procured by maximizing the geometric mean of the spacing:

$$G = \left[\prod_{i=1}^{n+1} D_i\right]^{1/(n+1)}, \tag{10}$$

in relation to α, or equivalently, by optimizing the function

$$log(G) = \frac{1}{n+1}\sum_{i=1}^{n+1} log(D_i). \tag{11}$$

A variation of MPS, the MMPS, was suggested by Jiang (2013). In this method, the square roots of the smallest and largest spacings were computed to produce a product of n effective spacings. In MMPS, the parameter estimates $\hat{\alpha}_{MMPS}$ are procured by maximizing the following function:

$$H = D_1^{1/2}\left(\prod_{i=2}^{n} D_i\right) D_{n+1}^{1/2}, \tag{12}$$

or by maximizing the function

$$log(H) = \frac{1}{2}log(D_1) + \sum_{i=2}^{n} log(D_i) + \frac{1}{2}log(D_{n+1}). \tag{13}$$

Notably, the MPS and MMPS estimators are sensitive to closely situated observations and especially duplicates (Cheng & Stephens 1989). In case of duplications resulting from multiple observations, the repeated spacing should be substituted by the corresponding likelihood.

ORDINARY AND WEIGHTED LEAST SQUARES

Consider $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ to be the ordered statistics from a random sample of size n taken from the NP(α, x_0) distribution. We know that:

$$E[F(X_{(i)})] = \frac{i}{n+1},$$

$$Var[F(X_{(i)})] = \frac{i(n-i+1)}{(n+1)^2(n+2)}, \quad i = 1,2,\ldots,n.$$

Parameter α is estimated as $\hat{\alpha}_{OLS}$ via OLS by minimizing the function:

$$L(\alpha) = \sum_{i=1}^{n}\left[F(x_{(i)};\alpha,x_0) - \frac{i}{n+1}\right]^2, \tag{14}$$

WLS estimation, $\hat{\alpha}_{WLS}$ minimizes the function:

$$W(\alpha) = \sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left[F(x_{(i)};\alpha,x_0) - \frac{i}{n+1}\right]^2, \tag{15}$$

with respect to α.

PERCENTILE

The PC estimates of the parameter α are calculated using Kao's (1958) method, leveraging the NP distribution's defined CDF. In this context, p_i estimates $F(x_{(i)};\ \alpha,\ x_0)$. The Euclidean distance, as defined herewith, is the measure between the population and sample percentiles:

$$E(\alpha) = \sum_{i=1}^{n}\left[x_{(i)} - Q(p_i;\alpha,x_0)\right]^2, \tag{16}$$

where $x_{(i)}$ signifies an ordered sample observation for $i = 1$, 2,..., n and $p_i = i/(n+1)$. The PC estimates, represented as $\hat{\alpha}_{PC}$, are determined by minimizing the Euclidean distance $E(\alpha)$ concerning α.

EMPIRICAL DISTRIBUTION FUNCTION STATISTICS

The parameter α estimates can also be derived by minimizing the empirical distribution function (EDF) statistics, a set of statistics predicated on the discrepancy between the CDF estimates and the EDF (Luceño 2008, 2006). Luceño (2008) also refers to these estimators as maximum goodness-of-fit estimators to differentiate from unrelated minimum distance methods. This section presents seven EDF statistics estimators for the NP distribution's tail index, including KS, AD, MAD, CVM, ZKS, ZAD, and ZCVM.

The estimates for each parameter α, symbolized as $\hat{\alpha}_{KS}$, $\hat{\alpha}_{AD}$, $\hat{\alpha}_{MAD}$, $\hat{\alpha}_{CVM}$, $\hat{\alpha}_{ZKS}$, $\hat{\alpha}_{ZAD}$, and $\hat{\alpha}_{ZCVM}$, are attained by minimizing the following EDF statistics with respect to α:

$$K(\alpha) = \max_{1\leq i\leq n}\left(F(x_{(i)};\alpha,x_0) - \frac{i-1}{n}, \frac{i}{n} - F(x_{(i)};\alpha,x_0)\right), \tag{17}$$

$$A(\alpha) = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[log\left(F(x_{(i)};\alpha,x_0)\right) + log\left(1 - F(x_{(n+1-i)};\alpha,x_0)\right)\right], \tag{18}$$

$$MA(\alpha) = \frac{n}{2} - 2\sum_{i=1}^{n} F(x_{(i)};\alpha, x_0) \tag{19}$$

$$-\sum_{i=1}^{n}\left[2 - \frac{2i-1}{n}\right] log\big(1 - F(x_{(i)};\alpha, x_0)\big),$$

$$C(\alpha) = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_{(i)};\alpha, x_0) - \frac{2i-1}{2n}\right]^2, \tag{20}$$

$$ZK(\alpha) = \max_{1\le i\le n}\left(\left(i - \frac{1}{2}\right) log\left(\frac{i - 1/2}{nF(x_{(i)};\alpha, x_0)}\right)\right. \tag{21}$$

$$\left. + \left(n - i + \frac{1}{2}\right) log\left(\frac{n - i + 1/2}{n\big(1 - F(x_{(i)};\alpha, x_0)\big)}\right)\right),$$

$$ZA(\alpha) =$$

$$-\sum_{i=1}^{n}\left[\frac{log\big(F(x_{(i)};\alpha, x_0)\big)}{n - i + 1/2} + \frac{log\big(1 - F(x_{(i)};\alpha, x_0)\big)}{i - 1/2}\right], \tag{22}$$

$$ZC(\alpha) = \sum_{i=1}^{n}\left[log\left(\frac{F(x_{(i)};\alpha, x_0)^{-1} - 1}{(n - 1/2)/(i - 3/4) - 1}\right)\right]^2, \tag{23}$$

where $x_{(i)}$ refers to an ordered sample observation for $i = 1, 2,\ldots, n$.

NEW PARETO-TYPE QUANTILE PLOT

Visual exploration is an essential preliminary step in applied data analysis. To facilitate this, we propose the NP quantile plot, a graphical method for verifying the NP distribution assumption in upper-tail data. Based on transformation techniques, it can be proven that NP random variable X's logarithms follow an exponential-type distribution (Appendix). Therefore, we can construct the NP quantile plot by mapping the observed values' logarithms, log (x_i) for $i = 1, 2,\ldots, n$, against the theoretical quantiles of the standard exponential-type distribution, i.e.,

$$log\left(\frac{p_i + 1}{1 - p_i}\right), \quad i = 1,2, \ldots, n, \tag{24}$$

where $p_i = i/(n + 1)$. If the upper-tail data align with the NP distribution, the NP quantile plot's observations will appear nearly linear. We can use the fitted line's leftmost point to estimate the threshold, i.e., x_0. An additional benefit of the NP quantile plot is its utility as a graphical tool to detect outliers or extreme observations in the upper-tail data by identifying points that deviate from the fitted line.

THRESHOLD SELECTION

Choosing the correct threshold parameter is vital when working with the Pareto family of distributions. The accuracy of the threshold affects sample size of the upper-tail data and in turn, impacts the bias of the estimated tail index and variance of parameter estimates. Simple methods like selecting a fixed top proportion of the distribution (10%, 5% or 1%) have been suggested (Gabaix 2009). Graphical tools like the Pareto quantile plot, Zipf plot, and mean excess function plot can also be used to determine the threshold (Beirlant, Vynckier & Teugels 1996; Cirillo 2013; Cirillo & Hüsler 2009). However, these methods can be subjective and may not yield an optimal threshold.

In our study, we use the KS statistic to determine the optimal NP distribution threshold, i.e., we select the threshold value that minimizes the KS statistic (Equation 17). This approach has been widely used in the Pareto family of distributions (Safari et al. 2020, 2019; Soriano-Hernández et al. 2017) and has been applied in income data to identify high earners (Oancea, Andrei & Pirjol 2017; Safari, Masseran & Ibrahim 2018b; Soriano-Hernández et al. 2017). The procedure to ascertain the optimal threshold for the NP model leveraging the KS statistic unfolds as follows: Step 1 Use the NP quantile plot to identify candidates for the optimal threshold, say, $x_{01}, x_{02},\ldots, x_{0n}$. Step 2 Estimate the shape parameter of the NP model, say, $\hat{\alpha}_1$, using $\hat{x}_{01}$ as the estimated threshold parameter. Step 3 Compute the KS statistic for the estimated NP model, i.e., $\hat{F}(x;\ \hat{\alpha}_1, \hat{x}_{01})$. Step 4 Repeat steps 2 and 3 for other parameters of the NP model, i.e., $(x_{02}, \alpha_2), (x_{03}, \alpha_3), \ldots, (x_{0n}, \alpha_n)$. Step 5 Finally, choose the threshold that yields the minimum KS statistic value as the optimal threshold.

MONTE CARLO SIMULATION

Monte Carlo simulation is employed to evaluate the efficacy of the various methods used for estimating the tail index of the NP distribution, both in scenarios with and without outliers. This approach allows us to determine which methods remain robust in the face of outlier data. The subsequent subsections detail the design and outcomes of the simulation.

SIMULATION DESIGN

Data sets are generated from the NP(α, x_0) distribution distribution with a known threshold value, $x_0 = 1$, and varied shape parameter values, $\alpha = 1.5, 2, 2.5$, and 3. The simulation study uses two categories of sample sizes: small (n = 30, 50, and 70) and large (n = 300, 500, and 1000). Observations from the simulated data are then randomly chosen and replaced with outliers. For small sample sizes, we introduce outliers in fixed numbers, m = 0, 2, and 5, while for large samples, outliers are incorporated at fixed proportions, ε = 0%, 2%, and 5%. The outliers stem from a normal distribution N(μ, σ) with mean μ = 736.78, 141.42, 52.53, and 27.14, and standard deviation $\sigma = 1$. They correspond to the 99.99% quantile of the NP distribution for the true NP model with shape parameter α = 1.5, 2, 2.5, and 3 and $x_0 = 1$. This simulation is executed for 10000 pseudo-random samples over 10000 simulation runs.

The performance of each estimation method is evaluated based on the percentage relative root mean square error (RRMSE). For a known value of the NP tail index α, the RRMSE is calculated as:

$$\mathrm{RRMSE} = \frac{100}{\alpha}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i - \alpha)^2}, \tag{25}$$

where $\hat{\alpha}_i$ is the estimated NP tail index for the ith (i = 1, 2,…, N) simulated sample, and N represents the number of simulation runs or simulated samples. A smaller RRMSE signifies greater accuracy and precision.

SIMULATION RESULTS

The Monte Carlo simulation findings are illustrated in Figures 1 and 2, and can be summarized as follows: 1). All methods' performances enhance as the sample size n increases, resulting in a decrease in RRMSEs, 2). No significant pattern of RRMSE change is observable for most estimators relative to α. Only the MoM estimator displays an improvement as α increases, 3). In both small and large sample size scenarios, the PC estimator is least effective at estimating the α parameter, 4). The ML, MOM, MPS, MMPS, PC, ZKS, and ZAD ZCVM estimators are not resilient to outliers. As outlier contamination increases, these estimators' performances noticeably decline, 5). With a small sample size (Figure 1), we observe: a) Without outliers (m = 0), MPS, MMPS, and ML estimators are top-tier at estimating parameter α, with MPS performing slightly better than MMPS and ML, b) For a small number of outliers (m = 2), robust estimators include CVM, OLS, WLS, KS, AD, and MAD, with CVM having a marginal edge over others in this scenario, and c) When the outlier count is high (m = 5), CVM, OLS, and WLS prove robust, with CVM slightly outperforming the rest. 6) For large sample sizes (Figure 2): a) Without outliers (ε = 0%), MPS, MMPS, and ML estimators are most effective, b) For a small outlier proportion (ε = 2%), CVM, OLS, WLS, KS, AD, and MAD hold robustness, with CVM and OLS exhibiting slightly better performances than the others, c) When the outlier proportion is high (ε = 5%), CVM, OLS, and WLS remain robust, with CVM and OLS performing slightly better than WLS.

APPLICATION TO HOUSEHOLD INCOME DATA

This section focuses on employing all methodologies to model the upper-tail data for household income in Italy and Malaysia.

DATA DESCRIPTION

Our first dataset comprises the annual net disposable income of Italian households for 2014 and 2016, sourced from the Bank of Italy's Survey on Household Income and Wealth (Banca d'Italia 2008). This survey, initiated in the 1960s, collects information on Italian households' income and savings. We note that this data includes minor instances of zero and negative incomes, representing less than 0.8% of total samples per year. Due to our study's NP model which only accounts for positive random variables, these instances have been excluded, focusing the analysis on positive income data. Table 1 presents the descriptive statistics of this data.

The second dataset features 2014 and 2016 monthly net incomes of Malaysian households, derived from the Household Income Surveys (HIS) conducted by the Malaysia's Department of Statistics (DOSM 2017). The HIS serves three main purposes: collecting data on households' income distribution, gathering statistics on impoverished households, and determining households' accessibility to basic amenities. This data informs governmental policy-making, especially in poverty eradication and income distribution strategies. Descriptive statistics of this dataset are provided in Table 2.

MODELING THE UPPER-TAIL DATA USING THE NEW PARETO-TYPE DISTRIBUTION

We start by building an NP quantile plot to validate the suitability of the NP model for the upper sections of Italian and Malaysian household income distributions. The NP quantile plots for Italian household income data from 2014 and 2016 are displayed in Figure 3, while those for Malaysian household income data from the same years are

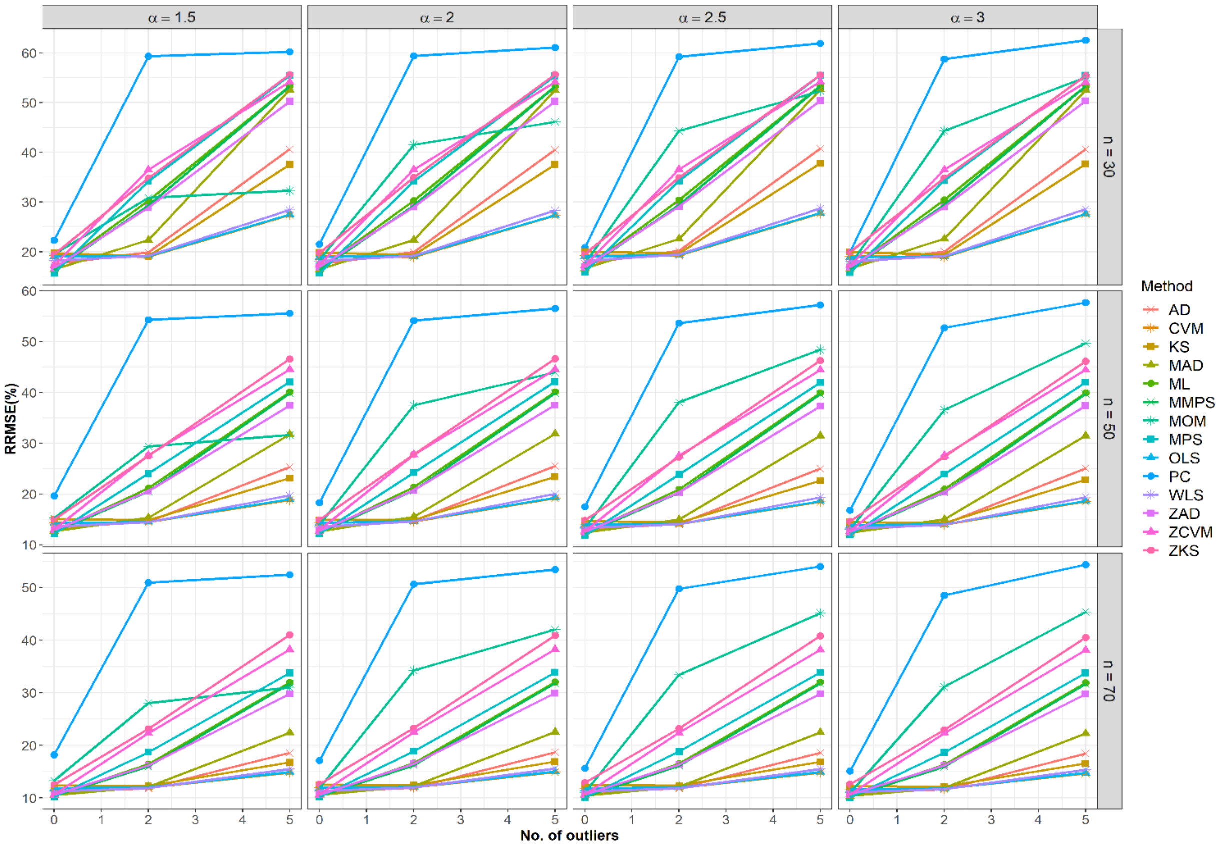

FIGURE 1. RRMSE results for estimations of the NP tail index θ = 1, 2, 3 with n = 30, 50, 70 and m = 0, 2, 5

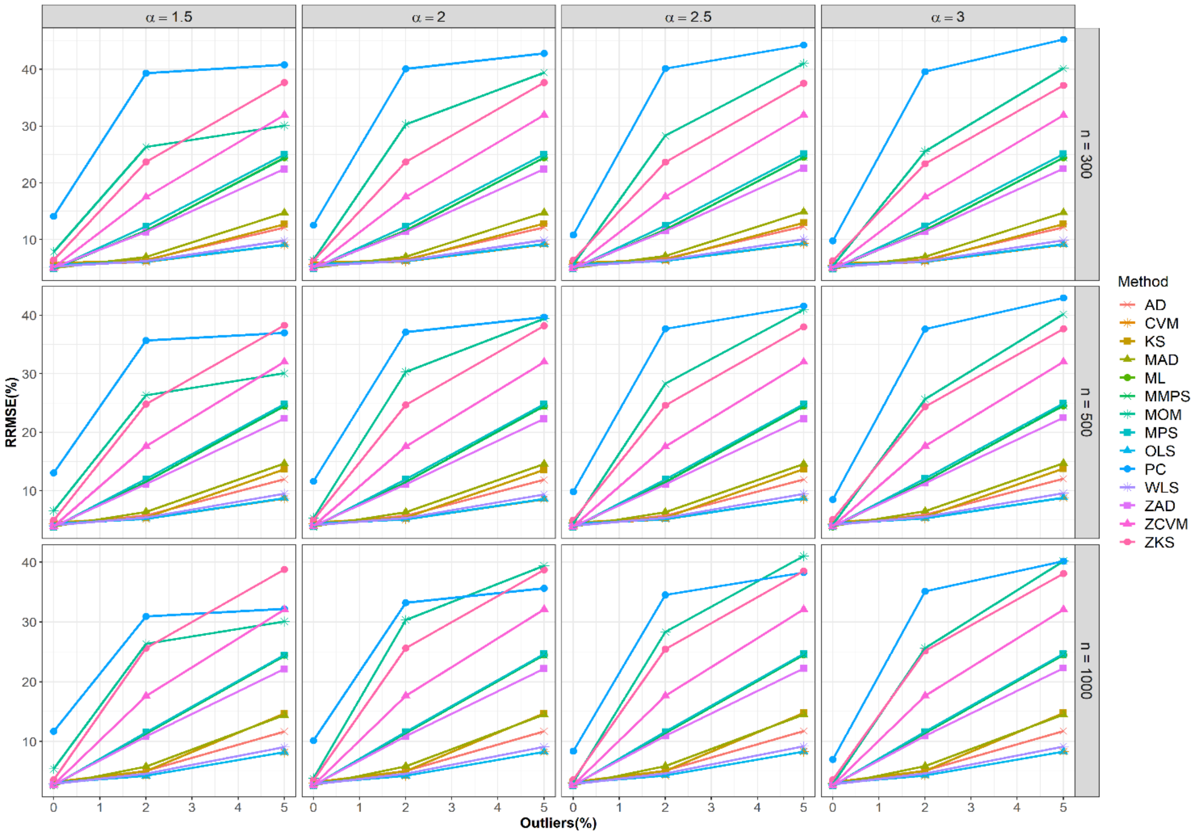

FIGURE 2. RRMSE results for estimations of the NP tail index θ = 1, 2, 3 with n = 300, 500, 1000 and ε = 0%, 2%, 5%

TABLE 1. The descriptive statistics of the Italian household income data

Year	Sample size (n)	Mean (€)	Median (€)	Min (€)	Max (€)	Variance	Skewness	Kurtosis
2014	8109	31433.95	25834.25	3.07	440199.10	503660705	3.3032	28.1831
2016	7366	30714.75	25200.92	1.75	541879.20	516663044	4.7770	65.4361

TABLE 2. The descriptive statistics of the Malaysian household income data

Year	Sample size (n)	Mean (RM)	Median (RM)	Min (RM)	Max (RM)	Variance	Skewness	Kurtosis
2014	24463	4982.73	3750.25	197.42	182311.40	23225628	8.0130	180.4427
2016	23536	5508.80	4163.08	269.58	274940.50	27747498	11.2428	393.0925

shown in Figure 4. Figures 3 and 4 show that the upper sections of the NP quantile plots almost follow a straight line, signifying that the high-income data from Italian and Malaysian households adhere to an NP distribution assumption. Moreover, certain data points that deviate from the fitted lines can be observed, highlighting the presence of extreme outliers – households in Italy and Malaysia with considerably higher incomes compared to others.

The NP quantile plot is also used as a visual tool to identify potential candidates for the optimal threshold. Practically, these candidates can be selected from the leftmost data point on the fitted line to a specified upper limit. In each NP quantile plot (Figures 3 & 4), the intersection between the red dashed horizontal line and the fitted straight line represents the leftmost data point, which is at the 45th percentile of Italian household income data. Moreover, the intersection point between the blue dashed horizontal line and the fitted straight line signifies the upper limit for the candidates of the optimal threshold, situated at the 95th percentile of Italian household income data. The optimal threshold for the NP model is then determined by pinpointing the threshold value that minimizes the KS statistic.

To gauge the tail index of the NP model, we employ 13 estimators, including ML, MoM, MPS, MMPS, OLS, WLS, KS, AD, MAD, CVM, ZKS, ZAD, and ZCVM. Due to its poor performance in the simulation study, the PC estimator is not included. To evaluate the efficacy of these methods in estimating the NP distribution's tail index, we employ the KS test, AD test, CVM test, and coefficient of determination (R^2) to assess the goodness-of-fit (GoF). The formulae for the KS, AD, and CVM statistics are given in Equations (17), (18), and (20), respectively. The coefficient of determination, R^2, is computed as follows:

$$R^2 = \frac{\sum_{i=1}^{n}\left[\hat{F}(x_i;\alpha,x_0)-\bar{F}(x;\alpha,x_0)\right]^2}{\sum_{i=1}^{n}\left[\hat{F}(x_i;\alpha,x_0)-\bar{F}(x;\alpha,x_0)\right]^2+\sum_{i=1}^{n}\left[F_n(x_i)-\hat{F}(x_i;\alpha,x_0)\right]^2}. \tag{26}$$

Here, F_n (x_i) represents the empirical cumulative probability for the ith observation above the threshold x_0, $\hat{F}(x_i;\ \alpha,\ x_0)$ is the estimated cumulative probability for the ith observation data above the threshold x_0 under the NP model, and $\hat{F}(x;\ \alpha,\ x_0)$ represents the mean of $\hat{F}(x_i;\ \alpha,\ x_0)$. Using all of these GoF measures, a global score (GS) criteria is computed to determine the best estimator. The process of computing the GS criteria is as follows: Step 1 Compute the KS statistic, AD statistic, CVM statistic, and 1 - R^2 value for each estimator. Step 2 Normalize each of the GoF measures from Step 1, converting them into standard normal random variables. This can be done using the formula:

$$z_{ij} = \frac{k_{ij}-\bar{k}_j}{s_j}, \quad for \quad i = 1,2,\ldots,13;\ j = 1,2,3,4. \tag{27}$$

Here, z_{ij} is the standardized score, k_{ij} is the ith value of the jth GoF measures, and $\bar{k}_j$ and s_j are the mean and standard deviation of the jth GoF measures, respectively. *Step 3* Convert the standardized score to a criteria value between 0 and 1, using the standard normal CDF, with the formula:

$$\Phi(z_{ij}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{ij}} e^{\frac{-t^2}{2}}\, dt. \qquad (28)$$

Step 4 Compute the GS by multiplying all the transformed standardized scores from Step 3. The formula for this is:

$$\text{GS} = \Phi(z_{i1}) \times \Phi(z_{i2}) \times \Phi(z_{i3}) \times \Phi(z_{i4}), \qquad (29)$$
$$for\ i = 1,2,\ldots,13.$$

The estimator which results in the smallest GS is considered the most optimal for estimating the NP tail index.

Tables 3 and 4 detail the parameter calculations and GoF for the NP distribution's upper tail concerning Italian and Malaysian household data, respectively. As evidenced in Table 3, the OLS and WLS methods exhibit the lowest GS in 2014 and 2016, respectively, indicating their superior efficacy in estimating the NP tail index for the corresponding years in the context of Italian household income. Moving to the Malaysian context, as delineated in Table 4, the MAD and OLS emerge as the most successful performers for estimating the NP tail index for 2014 and 2016, respectively. Using these optimal estimators, the NP tail index (α) is observed to hit its lowest point in 2016 for Italy and in 2014 for Malaysia, suggesting the heaviest NP tail concentration in those years.

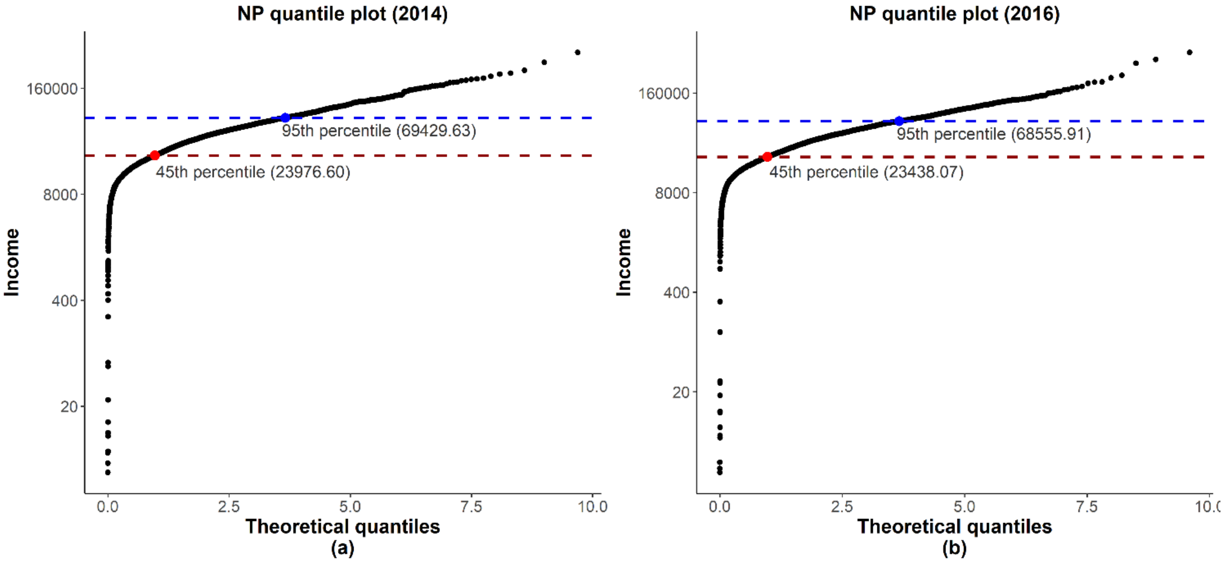

FIGURE 3. NP quantile plots of Italian household income data for (a) 2014 and (b) 2016

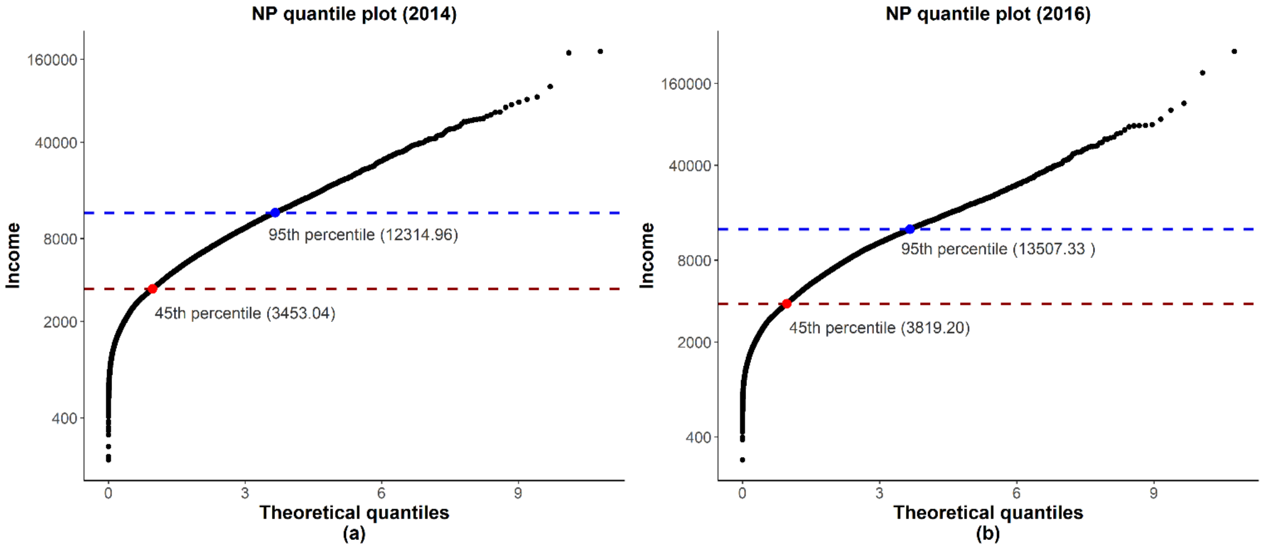

FIGURE 4. NP quantile plots of Malaysian household income data for (a) 2014 and (b) 2016

TABLE 3. Parameter estimates and GoFs of the NP distributions for the upper-tail data of Italian household income using several different estimators

Year	Estimator	$\hat{x}_0$	$\hat{\alpha}$	KS statistic (*p*-value)	AD statistic (*p*-value)	CVM statistic (*p*-value)	R^2	GS (Rank)
2014	ML	37765.10	3.8004	0.0087 (0.9953)	0.3058 (0.9338)	0.0338 (0.9623)	0.9998197	0.0161 (7)
	MOM	37726.64	3.7838	0.0090 (0.9928)	0.3226 (0.9203)	0.0415 (0.9252)	0.9997740	0.4104 (12)
	MPS	37765.10	3.7952	0.0084 (0.9969)	0.3157 (0.9259)	0.0362 (0.9519)	0.9998047	0.0204 (9)
	MMPS	37765.10	3.8000	0.0086 (0.9955)	0.3063 (0.9334)	0.0339 (0.9617)	0.9998188	0.0128 (5)
	OLS	38836.58	3.9156	0.0086 (0.9955)	0.2980 (0.9396)	0.0308 (0.9734)	0.9998333	0.0018 (1)
	WLS	38836.58	3.9026	0.0089 (0.9950)	0.2835 (0.9500)	0.0334 (0.9636)	0.9998131	0.0042 (3)
	KS	37765.10	3.7941	0.0083 (0.9972)	0.3184 (0.9237)	0.0368 (0.9491)	0.9998010	0.0158 (6)
	AD	38836.58	3.9022	0.0090 (0.9947)	0.2835 (0.9500)	0.0336 (0.9629)	0.9998119	0.0052 (4)
	MAD	38773.81	3.8905	0.0091 (0.9934)	0.3291 (0.9148)	0.0370 (0.9481)	0.9997944	0.2505 (11)
	CVM	38836.58	3.9159	0.0089 (0.9954)	0.2987 (0.9391)	0.0308 (0.9734)	0.9998334	0.0041 (2)
	ZKS	37461.85	3.7488	0.0094 (0.9863)	0.3414 (0.9041)	0.0523 (0.8625)	0.9997194	0.9411 (13)
	ZAD	37765.10	3.7939	0.0084 (0.9971)	0.3189 (0.9233)	0.0369 (0.9486)	0.9998004	0.0280 (10)
	ZCVM	37765.10	3.7977	0.0085 (0.9962)	0.3102 (0.9303)	0.0349 (0.9576)	0.9998126	0.0162 (8)
2016	ML	36024.62	3.6888	0.0135 (0.8239)	0.3870 (0.8615)	0.0525 (0.8607)	0.9997145	0.2376 (11)
	MOM	36024.62	3.6773	0.0122 (0.9023)	0.3733 (0.8747)	0.0485 (0.8854)	0.9997315	0.0013 (3)
	MPS	36024.62	3.6831	0.0129 (0.8650)	0.3773 (0.8710)	0.0500 (0.8763)	0.9997259	0.0246 (7)
	MMPS	36024.62	3.6885	0.0135 (0.8262)	0.3864 (0.8621)	0.0524 (0.8617)	0.9997152	0.2220 (9)
	OLS	35604.93	3.6317	0.0122 (0.8994)	0.3903 (0.8583)	0.0489 (0.8831)	0.9997320	0.0038 (5)
	WLS	36024.62	3.6767	0.0121 (0.9044)	0.3733 (0.8748)	0.0484 (0.8860)	0.9997318	0.00119 (1)
	KS	35604.93	3.6307	0.0121 (0.9046)	0.3917 (0.8569)	0.0489 (0.8829)	0.9997313	0.0034 (4)
	AD	36024.62	3.6766	0.0122 (0.9038)	0.3733 (0.8748)	0.0484 (0.8861)	0.9997319	0.00121 (2)
	MAD	36024.62	3.6837	0.0130 (0.8612)	0.3780 (0.8703)	0.0502 (0.8751)	0.9997251	0.0329 (8)
	CVM	35754.88	3.6450	0.0121 (0.9035)	0.3855 (0.8630)	0.0491 (0.8815)	0.9997288	0.0041 (6)
	ZKS	35709.15	3.6470	0.0134 (0.8285)	0.4252 (0.8234)	0.0547 (0.8471)	0.9997035	0.7943 (13)
	ZAD	36033.60	3.6727	0.0124 (0.8961)	0.4346 (0.8138)	0.0523 (0.8621)	0.9997100	0.2307 (10)
	ZCVM	36033.60	3.6738	0.0125 (0.8902)	0.4351 (0.8133)	0.0526 (0.8606)	0.9997092	0.2954 (12)

TABLE 4. Parameter estimates and GoFs of the NP distributions for the upper-tail data of Malaysian household income using several different estimators

Year	Estimator	$\hat{x}_0$	$\hat{\alpha}$	KS statistic (*p*-value)	AD statistic (*p*-value)	CVM statistic (*p*-value)	R^2	GS (Rank)
2014	ML	4750.85	2.7408	0.0063 (0.8701)	0.5907 (0.6569)	0.0924 (0.6232)	0.9998764	0.0190 (9)
	MOM	4812.92	2.7762	0.0079 (0.6393)	1.1199 (0.3001)	0.1811 (0.3073)	0.9997567	0.9874 (13)
	MPS	4734.92	2.7361	0.0062 (0.8761)	0.5654 (0.6813)	0.0925 (0.6230)	0.9998769	0.0130 (6)
	MMPS	4750.85	2.7408	0.0063 (0.8712)	0.5906 (0.6569)	0.0924 (0.6233)	0.9998764	0.0190 (8)
	OLS	4709.63	2.7274	0.0060 (0.8966)	0.5615 (0.6851)	0.0899 (0.6367)	0.9998811	0.0069 (4)
	WLS	4474.31	2.6562	0.0061 (0.8673)	0.6898 (0.5674)	0.0808 (0.6875)	0.9999012	0.0073 (5)
	KS	4694.39	2.7233	0.0060 (0.9028)	0.6233 (0.6262)	0.0953 (0.6085)	0.9998745	0.0160 (7)
	AD	4474.31	2.6559	0.0060 (0.8734)	0.6897 (0.5675)	0.0802 (0.6904)	0.9999018	0.0057 (2)
	MAD	4474.31	2.6542	0.0060 (0.8723)	0.6919 (0.5656)	0.0778 (0.7046)	0.9999045	0.0046 (1)
	CVM	4709.63	2.7275	0.0060 (0.8956)	0.5614 (0.6852)	0.0899 (0.6367)	0.9998811	0.0069 (3)
	ZKS	4727.10	2.7354	0.0064 (0.8510)	0.5778 (0.6693)	0.0951 (0.6094)	0.9998739	0.0231 (10)
	ZAD	4785.76	2.7584	0.0071 (0.7706)	0.7649 (0.5071)	0.1198 (0.4962)	0.9998390	0.3647 (11)
	ZCVM	4785.76	2.7592	0.0072 (0.7521)	0.7685 (0.5043)	0.1205 (0.4935)	0.9998383	0.3890 (12)
2016	ML	7278.60	3.3076	0.0084 (0.8587)	0.4820 (0.7652)	0.0569 (0.8336)	0.9998651	0.0338 (7)
	MOM	6772.12	3.1590	0.0092 (0.7065)	1.0770 (0.3193)	0.1290 (0.4604)	0.9997348	0.9334 (13)
	MPS	7087.99	3.2584	0.0082 (0.8560)	0.5448 (0.7017)	0.0733 (0.7318)	0.9998366	0.1164 (9)
	MMPS	7278.60	3.3076	0.0084 (0.8584)	0.4824 (0.7648)	0.0569 (0.8333)	0.9998649	0.0341 (8)
	OLS	7739.05	3.4727	0.0068 (0.9822)	0.5139 (0.7327)	0.03902 (0.9381)	0.9998991	0.0015 (1)
	WLS	7739.05	3.4679	0.0069 (0.9790)	0.5071 (0.7395)	0.0400 (0.9331)	0.9998954	0.0020 (3)
	KS	7739.05	3.4866	0.0065 (0.9895)	0.5901 (0.6574)	0.0466 (0.8967)	0.9998829	0.0030 (5)
	AD	7739.05	3.4671	0.0070 (0.9785)	0.5070 (0.7397)	0.0403 (0.9314)	0.9998944	0.0025 (4)
	MAD	7739.05	3.4578	0.0074 (0.9601)	0.5261 (0.7203)	0.0481 (0.8876)	0.9998723	0.0126 (6)
	CVM	7739.05	3.4729	0.0068 (0.9823)	0.5142 (0.7323)	0.0390 (0.9381)	0.9998991	0.0015 (2)
	ZKS	6891.51	3.1896	0.0088 (0.7668)	0.7203 (0.5421)	0.0951 (0.6096)	0.9997972	0.5504 (12)
	ZAD	7087.99	3.2567	0.0083 (0.8447)	0.5564 (0.6901)	0.0739 (0.7279)	0.9998348	0.1360 (10)
	ZCVM	7044.47	3.2427	0.0082 (0.8499)	0.5781 (0.6690)	0.0778 (0.7048)	0.9998281	0.1680 (11)

Figures 5 and 6 illustrate the best-fitted PDF plot of the NP distribution to the high-income data of Italian and Malaysian households, respectively. It is evident from these figures that the NP model fits well to the high-income data, suggesting that this model is well-suited to describe the income distribution of the wealthiest households in both countries.

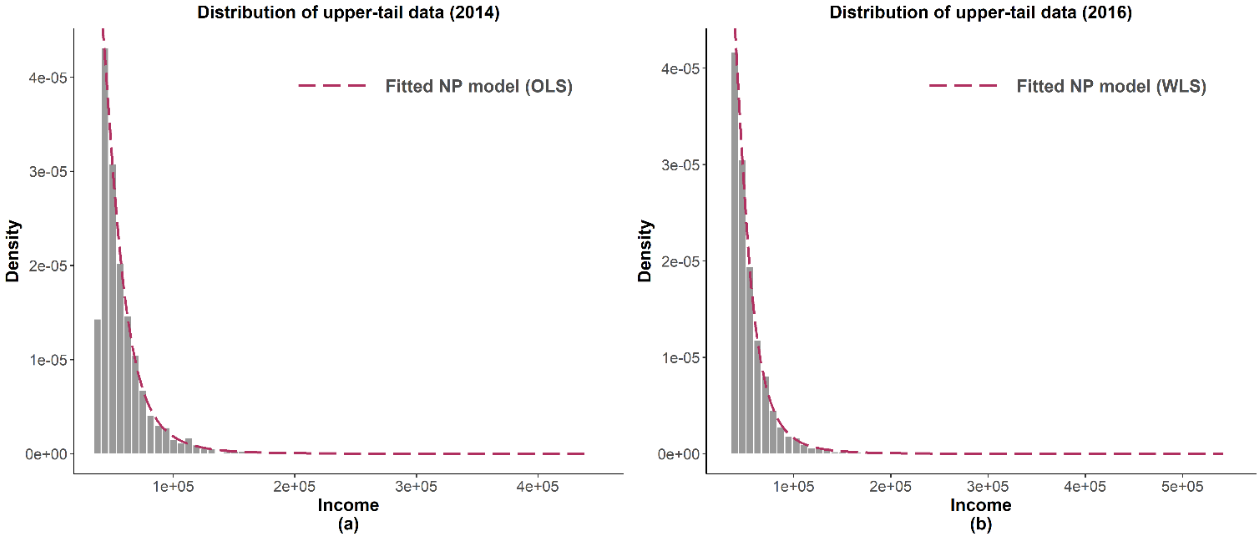

FIGURE 5. Best fitted NP density on a histogram for the upper-tail data of Italian household income for (a) 2014 and (b) 2016

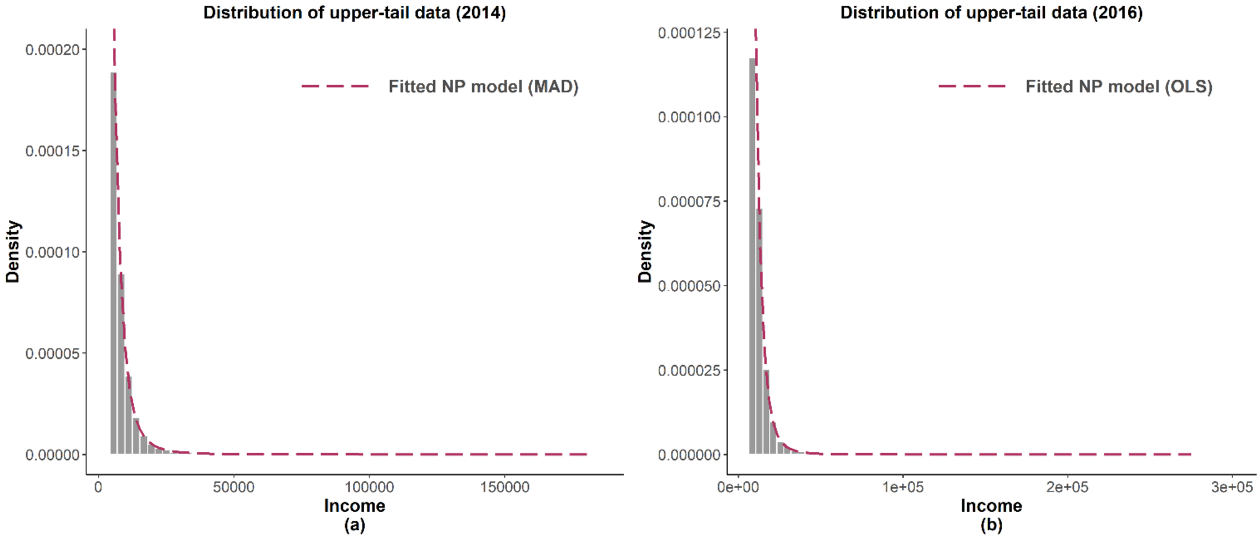

FIGURE 6. Best fitted NP density on a histogram for the upper-tail data of Malaysian household income for (a) 2014 and (b) 2016

Utilizing the Lorenz curve (LC) and the Gini coefficient from the NP model (Abd Raof et al., 2022; Sarabia, Jordá & Prieto 2019), we examine income disparity amongst the highest earners in Italy and Malaysia. Table 5 presents an overview of the Gini coefficients, and Figure 7 graphically represents the Lorenz Curves (LCs) for the upper-income segments of both Italy and Malaysia, specifically for the years 2014 and 2016. A Gini coefficient less than 0.3, as shown in Table 5, points to low levels of income inequality among the top-earning households in both nations. This observation of minimal income inequality is further corroborated by the LC (Figure 7), showing a close alignment with the line of perfect equality. It is important to note that insights into income inequality can also be gleaned from the estimated NP tail index alone; smaller values of the NP tail index correlate with higher Gini values, indicating a more skewed income distribution. The LC allows us to discern the distribution of income among different tiers within the top earning households. If we segment the upper-class households in both countries into two divisions, the lower 80% and the upper 20%, as illustrated in Figure 7, the lower 80% group commands between 60.67% and 68.17% of the total top earners' income. On the other hand, the upper 20% possesses approximately 31.83% to 39.33% of the top earners' total income. From these observations, it is clear that the conventional 80/20 Pareto principle does not hold true in this instance.

TABLE 5. Estimated Gini coefficients based on NP model for the upper-class earners in Italy and Malaysia

Country	Year	Estimated NP tail index	Gini
Italy	2014	3.9156	0.1764
	2016	3.6767	0.1893
Malaysia	2014	2.6542	0.2759
	2016	3.4727	0.2021

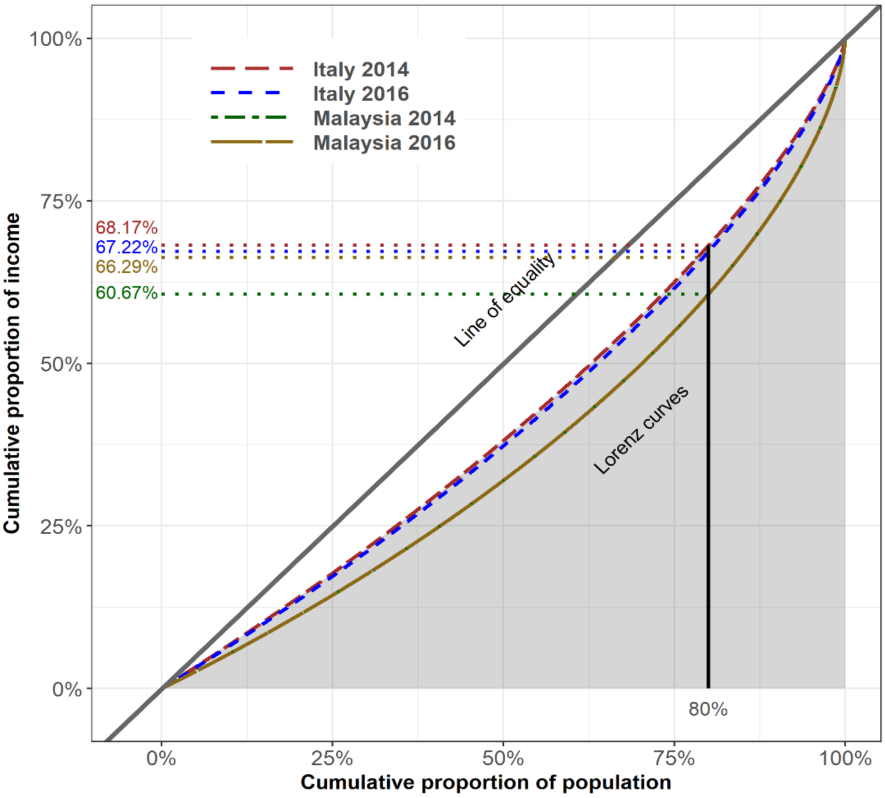

FIGURE 7. The fitted LCs based on NP model for upper-class earners in Italy and Malaysia

Conclusion

This research explored the estimation of the NP tail index through the lens of 14 diverse estimators. These included ML, MoM, MPS, MMPS, OLS, WLS, PC, KS, AD, MAD, CVM, ZKS, ZAD, and ZCVM. The efficiency of each estimator was scrutinized in the contexts of both the absence and presence of outliers, assessed through a comprehensive Monte Carlo simulation. The findings pointed towards MPS, MMPS, and ML as the three most robust estimators for the NP tail index in data devoid of outliers. In contrast, in the presence of outliers, CVM, OLS, and WLS emerged as the top performers. It is important to note that the PC estimator delivered the weakest results for estimating the NP tail index.

A new graphical instrument, named the NP quantile plot, was introduced to verify the assumption of an NP distribution in upper-tail data. When data points on this plot align to form an almost straight line, it suggests that the upper-tail data are compliant with the NP model. This plot also proves beneficial in pinpointing outliers within the upper-tail data, as it highlights data points that deviate from the fitted line. Additionally, a straightforward procedure was developed to discern the threshold of the NP distribution, where the optimal threshold is selected to minimize the KS statistic.

Applying these methodologies to actual datasets, the study modeled the upper-tail data of household income for Italy and Malaysia in 2014 and 2016. The NP quantile plot substantiated the applicability of the NP distribution assumption to these datasets and identified the existence of outliers. Excluding the PC, all 13 estimators were employed to estimate the NP tail index. An amalgamated approach, utilizing the KS statistic, AD statistic, CVM statistic, and $1 - R^2$, was used to pinpoint the best estimator. The lowest GS value, resulting from this integrated approach, determined the top estimator for the NP tail index. The investigation showed that OLS (2014) and WLS (2016) emerged as the superior estimators for Italy, while MAD (2014) and OLS (2016) excelled for Malaysia. This analysis reaffirmed that the NP model was a good fit for the upper-tail data of Italian and Malaysian household income, suggesting its effectiveness in explaining the income dynamics of the top earners in these countries.

Acknowledgments

The authors extend their thanks to the Department of Statistics Malaysia for supplying the HIS data. We appreciate the support from Universiti Putra Malaysia under the Geran Putra-Inisiatif Putra Muda (project code GP-IPM/2022/9714900) and Universiti Malaya (grant number GPF088B-2020).

References

Abd Raof, A.S., Haron, M.A., Safari, M.A.M. & Siri, Z. 2022. Modeling the incomes of the upper-class group in Malaysia using new Pareto-type distribution. *Sains Malaysiana* 51(10): 3437-3448.

Alfons, A., Templ, M. & Filzmoser, P. 2013. Robust estimation of economic indicators from survey samples based on Pareto tail modelling. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 62(2): 271-286.

Amoroso, L. 1938. Vilfredo Pareto. *Econometrica: Journal of the Econometric Society* 6(1): 1-21.

Banca d'Italia. 2008. *Survey of Household Income and Wealth (SHIW) of the Bank of Italy*. https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/index.html

Bee, M., Riccaboni, M. & Schiavo, S. 2019. Distribution of city size: Gibrat, Pareto, Zipf. In *The Mathematics of Urban Morphology*, edited by D'Acci L. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser, Cham. pp. 77-91. https://doi.org/10.1007/978-3-030-12381-9_4

Beirlant, J., Vynckier, P. & Teugels, J.L. 1996. Tail index estimation, pareto quantile plots regression diagnostics. *Journal of the American Statistical Association* 91(436): 1659-1667.

Bourguignon, M., Saulo, H. & Fernandez, R.N. 2016. A new Pareto-type distribution with applications in reliability and income data. *Physica A: Statistical Mechanics and Its Applications* 457: 166-175.

Cheng, R.C.H. & Amin, N.A.K. 1983. Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society: Series B (Methodological)* 45(3): 394-403.

Cheng, R.C.H. & Stephens, M.A. 1989. A goodness-of-fit test using Moran's statistic with estimated parameters. *Biometrika* 76(2): 385-392.

Cirillo, P. 2013. Are your data really Pareto distributed? *Physica A: Statistical Mechanics and Its Applications* 392(23): 5947-5962.

Cirillo, P. & Hüsler, J. 2009. On the upper tail of Italian firms' size distribution. *Physica A: Statistical Mechanics and Its Applications* 388(8): 1546-1554.

Coronel-Brizio, H.F. & Hernandez-Montoya, A.R. 2005. On fitting the Pareto–Levy distribution to stock market index data: Selecting a suitable cutoff value. *Physica A: Statistical Mechanics and Its Applications* 354: 437-449.

Department of Statistics Malaysia. 2017. *Household Income and Basic Amenities Survey Report 2016*.

Díaz, J.D., Cubillos, P.G. & Griñen, P.T. 2021. The exponential Pareto model with hidden income processes: Evidence from Chile. *Physica A: Statistical Mechanics and Its Applications* 561: 125196.

Dunford, R., Su, Q. & Tamang, E. 2014. The pareto principle. *The Plymouth Student Scientist* 7(1): 140-148.

Filimonov, V. & Sornette, D. 2015. Power law scaling and 'Dragon-Kings' in distributions of intraday financial drawdowns. *Chaos, Solitons & Fractals* 74: 27-45.

Gabaix, X. 2009. Power laws in economics and finance. *Annu. Rev. Econ.* 1(1): 255-294.

García, I.G. & Caballero, A.M. 2021. Models of wealth and inequality using fiscal microdata: Distribution in Spain from 2015 to 2020. *Mathematics* 9(4): 377.

Giesen, K., Zimmermann, A. & Suedekum, J. 2010. The size distribution across all cities - Double Pareto lognormal strikes. *Journal of Urban Economics* 68(2): 129-137.

Giorgi, G.M. & Gigliarano, C. 2017. The Gini concentration index: A review of the inference literature. *Journal of Economic Surveys* 31(4): 1130-1148.

Hlasny, V. & Verme, P. 2018. Top incomes and the measurement of inequality in Egypt. *The World Bank Economic Review* 32(2): 428-455.

Jiang, R. 2013. A modified MPS method for fitting the 3-parameter Weibull distribution. *2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE)*. pp. 983-985.

Kao, J.H.K. 1958. Computer methods for estimating Weibull parameters in reliability studies. *IRE Transactions on Reliability and Quality Control.* pp. 15-22.

Luceño, A. 2008. Maximum likelihood vs. maximum goodness of fit estimation of the three-parameter Weibull distribution. *Journal of Statistical Computation and Simulation* 78(10): 941-949.

Luceño, A. 2006. Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics & Data Analysis* 51(2): 904-917.

Lux, T. & Alfarano, S. 2016. Financial power laws: Empirical evidence, models, and mechanisms. *Chaos, Solitons & Fractals* 88: 3-18. https://doi.org/https://doi.org/10.1016/j.chaos.2016.01.020

Majid, M.H.A. & Ibrahim, K. 2021. Composite Pareto distributions for modelling household income distribution in Malaysia. *Sains Malaysiana* 50(7): 2047-2058.

Majid, M.H.A., Ibrahim, K. & Masseran, N. 2023. Three-part composite Pareto modelling for income distribution in Malaysia. *Mathematics* 11(13): 2899.

Meyer, S. & Held, L. 2014. Power-law models for infectious disease spread. *Annals of Applied Statistics* 8(3): 1612-1639.

Newman, M.E.J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46(5): 323-351.

Oancea, B., Andrei, T. & Pirjol, D. 2017. Income inequality in Romania: The exponential-Pareto distribution. *Physica A: Statistical Mechanics and Its Applications* 469: 486-498.

Pinto, C.M.A., Lopes, A.M. & Machado, J.A.T. 2012. A review of power laws in real life phenomena. *Communications in Nonlinear Science and Numerical Simulation* 17(9): 3558-3578.

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

Safari, M.A.M., Masseran, N. & Ibrahim, K. 2018b. 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. & Hussain, S.I. 2021. Measuring income inequality: A robust semi-parametric approach. *Physica A: Statistical Mechanics and Its Applications* 562: 125359.

Safari, M.A.M., Masseran, N., Ibrahim, K. & AL-Dhurafi, N.A. 2020. The power-law distribution for the income of poor households. *Physica A: Statistical Mechanics and Its Applications* 557: 124893.

Safari, M.A.M., Masseran, N., Ibrahim, K. & Hussain, S.I. 2019. A robust and efficient estimator for the tail index of inverse Pareto distribution. *Physica A: Statistical Mechanics and Its Applications* 517: 431-439.

Sarabia, J.M., Jordá, V. & Prieto, F. 2019. On a new Pareto-type distribution with applications in the study of income inequality and risk analysis. *Physica A: Statistical Mechanics and Its Applications* 527: 121277.

Soriano-Hernández, P., del Castillo-Mussot, M., Córdoba-Rodríguez, O. & Mansilla-Corona, R. 2017. Non-stationary individual and household income of poor, rich and middle classes in Mexico. *Physica A: Statistical Mechanics and Its Applications* 465: 403-413.

Xu, Y., Wang, Y., Tao, X. & Ližbetinová, L. 2017. Evidence of Chinese income dynamics and its effects on income scaling law. *Physica A: Statistical Mechanics and Its Applications* 487: 143-152.

*Corresponding author; email: aslam.safari@upm.edu.my

APPENDIX

Theorem 1 The natural logarithms of an NP(α, x_0 = 1) random variable adopt an exponential-type distribution.

Proof of Theorem 1 Let's assume that the random variable *X* obeys NP(α, x_0) with a PDF defined by Equation (1). Let's introduce a new random variable $Y = log\ (X)$, and consider the transformation $y = log\ (x)$, its inverse $x = e^y$, and the corresponding Jacobian,

$$\frac{\partial x}{\partial y} = e^y.$$

Employing the transformation technique, we can derive the PDF of *Y* as:

$$\begin{aligned} f(y) &= f(g^{-1}(y)) \left|\frac{\partial x}{\partial y}\right| \\ &= \frac{2\alpha x_0^\alpha (e^y)^{\alpha-1}}{((e^y)^\alpha + x_0^\alpha)^2} |e^y| \\ &= \frac{2\alpha x_0^\alpha e^{\alpha y}}{(e^{\alpha y} + x_0^\alpha)^2}, \quad y > log(x_0). \end{aligned}$$

When we set $x_0 = 1$, we arrive at:

$$f(y) = \frac{2\alpha e^{\alpha y}}{(e^{\alpha y} + 1)^2}, \quad y > 0.$$

This establishes that $f(y)$ is the PDF of an exponential-type distribution. The CDF and quantile function of this exponential-type distribution are:

$$F(y) = \frac{e^{\alpha y} - 1}{e^{\alpha y} + 1}, \quad y > 0,$$

$$Q(z) = \frac{log\left(\frac{z+1}{1-z}\right)}{\alpha}, \quad 0 < z < 1.$$