**Daniel Kosiorowski**[1][2]

# Income Distribution Models and Income Inequality Measures from the Robust Statistics Perspective Revisited

## 1. Introduction

Considerations related to income distribution and income inequalities in populations of economic agents belong to the core of the modern economic theory. They appear also in a public debate concerning postulates as to taxation or pension politics, in theories of a human capital creation or searching for regional development factors. Correct estimation of parameters of the income distribution its derivative measures of income inequality like *Gini* coefficient or *Theil* Index is important from several reasons – it is source of the knowledge about a structure of income in society and also could be a base for further economic issues such as changing taxation system or government aid programs in order to redistribute some part of wealth. Underestimation of the parameters of income distribution could lead to a conclusion that inequalities are too high and trigger some corrective action like rising taxes in high income group. If there is too much severity in changing tax bracket it may have influence on productivity and investment activities among well-paid citizens. Overestimation of the parameters could have opposite but also harmful effect for health of the economy, because when taxation is too liberal then it will be a huge probability that low-paid people get insufficient public transfers. Moreover income distribution affects economic growth, market demand, and is important factor in determining the amount of savings in a society (Kleiber and Kotz, 2003).

In real economic data sets, it often happens that some observations are different from the majority. Such outlying observations cause problems because they may strongly influence the result of an economic analysis. Robust statistics aims at detecting the outliers by searching for the model fitted to the majority of the data. All classical statistical methods (e.g. discriminant analysis, factor analysis, regression analysis, estimation of time series models parameters) can be severely distorted by outliers. It should be stressed that statistical inferences (an important part of each economic analysis) are based only in part upon the observations. An equally important base is formed by prior assumptions about the underlying situation. Even in the simplest cases, there are explicit or implicit assumptions about randomness and independence, about distributional models, perhaps prior distributions for some unknown parameters, and so on.

In this paper we show selected aspects of robust estimation of the income distribution. We focus our attention on two well-known models for the income distribution namely on the Pareto and log-normal distributions and on popular income inequality measures namely on the Lorentz curve and the Gini coefficient. The presented arguments however are applicable to a wide class of over 100 models used for income distributions modelling which are by default estimated by means of maximal likelihood methodology.

The rest of the paper is organized as follows. In Section 2, selected income distribution models are presented. In Section 3, selected robust estimators of income distribution are briefly presented. In Section 4, popular income distribution inequality measures are recalled. In Section 5, results of simulation as well as empirical studies of statistical properties of considered estimators are presented. The paper ends with conclusions and references.

---

[1] Dr hab. Daniel Kosiorowski, Department of Statistics, Cracow University of Economics, Department of Management, daniel.kosiorowski@uek.krakow.pl

[2] Author thanks for the polish NCS financial support DEC-011/03/B/HS4/01138

## 2. Selected Income Distribution Models

A modern concern about the income distribution started with Pareto's research related with his discussion with the French and Italian Socialists who were insisting on institutional reforms to reduce inequality in the distribution of income. Pareto studied the economic agent's income distribution for tax purposes. The distribution was truncated to the left at a point $x_m$, the maximum non-taxable income $x_m > 0$. He found a regularity of observed income distribution obtained from tax records - a stable linear relation of the form $\log N(x) = A - \alpha \log x$ , $x \geq x_m > 0$ , $\alpha > 1$ , where $N(x)$ is the number of economic units with income $X > x$ , $X$ being the income variable with range $[x_m, \infty)$. The Pareto type I model is the solution of that linear relationship. In the same context in 1898, March proposed the gamma probability density function (pdf) and fitted it to the distribution of wages in France, Germany, and the United States. Nowadays, there are over 100 models used for the income distribution modelling (see Kleiber and Kotz, 2002). The Pareto distribution for modelling high income groups and to deal with positive asymmetric distribution having heavy weight tails with either finite or infinite variance – still stands in a center of income distributions considerations however. It is mainly due to its elegance, interpretation possibilities and its relation to the popular inequality measures. The Pareto distribution as well as others skewed size distributions appear also in a context of economic data stream analysis i.e., e.g. for modelling sizes of data packages in the Internet (see Kosiorowski, 2012).

For purposes of this paper it is enough to consider a broad classification of the income distribution according to tail behaviour: Pareto type distributions (polynomially decreasing tails), lognormal distribution (intermediate case) and gamma – type distribution (exponentially decreasing tails).We focus our attention on two estimation difficulties which are good illustration for the robust analysis of income distribution. We start with the Pareto model $P(x_m, \alpha)$ which is suitable to model relatively high probability in the upper tail (right-skewed tail), where lower α shape parameter determine the lower probability mass at $x_m$ point. Thanks to that property of the model it is useful and relatively effective to apply in actuarial applications, risk management and Economy of Welfare.

A simple Pareto distribution $P(x_m, \alpha)$ is given by its cumulative distribution function (cdf)

$$F(x) = 1 - \left( \frac{x_m}{x} \right)^{\alpha} , \tag{1}$$

for $x > x_m$ , where $\alpha$ is the shape parameter that characterizes the tail of the distribution and $x_m > 0$ is the scale parameter.

The Pareto distribution has pdf $\frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$ for $x > x_m$ and following formulas for the expected value

$$E(X) = \begin{cases} \infty & \alpha \leq 1 \\ \frac{\alpha x_m}{\alpha - 1} & \alpha > 1 \end{cases}$$ , and the variance $$D^2(X) = \begin{cases} \infty & \alpha \in (1, 2] \\ \frac{x_m^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \alpha > 2 \end{cases}$$ ,

the median $x_m \sqrt[\alpha]{2}$ , and the mode $x_m$ .

If sample observations follow the postulated model $P(x_m, \alpha)$, then it is well known that for large data sets, maximum likelihood estimator (MLE) attains the minimum possible variance among a large class of competing estimators,

$$\hat{\alpha}_{ML} = \frac{n}{\sum_{i=1}^{n} \log(X_i / x_m)} . \tag{3}$$

It can be easily found that $2n\alpha / \hat{\alpha}_{ML}$ has cdf $\chi^2_{2n}$ (see Brazauskas and Serfling, 2000). Although $\hat{\alpha}_{ML}$ is biased, it is easy to find its unbiased version (MLE)

$$MLE = \frac{n-1}{\sum_{i=1}^{n} \log(X_i / x_m)}, \tag{4}$$

For large sample size $n$ , $MLE$ is approximately $N\left(\alpha, \frac{\alpha^2}{n}\right)$ . In case of the scale estimator we have following maximal likelihood formula

$$MLE(x_m) = \min_i \{X_i\}. \tag{5}$$

The Pareto distribution is widely used in the Economics due to its elegance and clear relations with popular measures of income inequality called the **Gini coefficient** $GINI = 1/(2\alpha - 1)$ for $\alpha \geq 1$ or popular risk measures like value at risk. It should be stressed however that even small relative error in estimation of $\alpha$ in $P(x_m, \alpha)$ may lead to a large relative error in estimated quantiles or tail probabilities based on $\alpha$. For the quantile $q_\varepsilon$ corresponding to upper tail probability $\varepsilon$ , it follows that $q_\varepsilon = x_m \varepsilon^{-1/\alpha}$. For $\varepsilon = 0.001$ underestimation of $\alpha = 1$ by only 5% leads to overestimation of $q_{0.001}$ by 58%. Errors in estimation of $\alpha$ may result in errors in estimation of basic measure of social inequity and lead to incorrect social politics.

Next important distribution for modelling incomes is the lognormal distribution discovered for economic purposes by Gibrat in 1931. A random variable $Y$ has **a lognormal distribution** $L(\mu, \sigma)$ if $X = \log Y$ has the normal distribution $N(\mu, \sigma^2)$ .

Three parameter form $L(\mu, \sigma, \tau)$ is the distribution of $Y = \tau + e^X$ , where $\tau$ represents a threshold value and $X$ is a random variable with mean $\mu$ and stdandard deviation $\sigma$.

In many applications a problem of efficient and robust estimation of the expected value of this distribution $E(Y) = e^{\mu + \sigma^2/2}$ appears (we assume the threshold $\tau$ is known). The problem leads to a nontrivial issue of robust joint estimation of $\mu$ and $\sigma$ in the context of thecorresponding model $N(\mu, \sigma)$. For a sample $Y^n = \{Y_1, ..., Y_n\}$ from the model $L(\mu, \sigma)$, transformation to the equivalent model $N(\mu, \sigma)$ yields the well-known ML estimators of the location $\mu$ and $\sigma$ scale parameter and depending on them expected value estimator:

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} \log Y_i, \tag{6}$$

$$\hat{\sigma}_{ML} = \left( \frac{1}{n} \sum_{i=1}^{n} (log Y_i - \hat{\mu}_{ML})^2 \right)^{1/2}, \tag{7}$$

and to the estimator of the expected value

$$E(Y) = e^{\hat{\mu}_{ML} + \hat{\sigma}^2_{ML}/2}. \tag{8}$$

Estimators (6), (7), (8) have good properties, minimal asymptotic variance but they fail to be robust, i.e., their breakdown point BP=0, and their influence function IF is unbounded.

As a last landmark distribution for income consider the generalized gamma distribution with pdf:

$$f(x) = \frac{a}{\beta^{ap} \Gamma(p)} x^{ap-1} e^{-(x/\beta)^a}, \tag{9}$$

where $x > 0$ , $\beta = b^{1/a}$ scale parameters, $a, p$ shape parameters.

The model (9) is usually estimated via maximal likelihoods methodology which lead to estimators which are not robust.

Each of the above distributions and in particular their parameters have interesting economic interpretations expressed in terms of elasticity of survival function, maximization of entropy, probability to increase an income of an agent under some condition etc. A discrimination between these three landmark distributions in case of a presence of inliers or outliers within a data by means of classical model selection may be a very difficult task. Empirical justification of a theoretical concepts explaining a form of the income distribution may be doubtful. Take for instance Mandelbrot's (1960) who argued that incomes follow what he calls a Pareto – Levy distribution – maximally skewed stable distribution with a characteristic exponent $\alpha$ between 1 and 2.

## 3. Robust estimators of the income distributions

Kalecki (1945) found that increments of the income are proportional to the excess in ability of given members of the distribution over the lowest (or median) member. He considered log – normal distribution for United Kingdom personal incomes for 1938 – 1939 and found that lognormal distribution fits well only when certain part of the data is omitted. He introduced therefore three parameter lognormal distribution. Kalecki can be treated as pioneer of robust approach to income distribution analysis.

Robust estimation in terms of bounded influence function of income distribution parameters was extensively studied by Victoria-Fezer (2000) basing on M-estimation approach (see Marona et all, 2006). We focus our attention on a less known approach but in our opinion very interesting related to works Brazauskas and Serfling.

We understand robustness of the estimator in terms of the **influence function** (IF) and in terms of the **finite sample breakdown point** (BP) – for further details see Maronna et. al. (2006).

Let us recall that for a given distribution $F$ in $\mathbb{R}$ and an $\varepsilon > 0$, the version of $F$ contaminated by an $\varepsilon$ amount of an arbitrary distribution $G$ in $\mathbb{R}$ is denoted by $F(\varepsilon, G) = (1-\varepsilon)F + \varepsilon G$. The **influence function** (IF) of an estimator $T$ at a given $x \in \mathbb{R}$ for a given $F$ is defined

$$IF(x;T,F) = \lim_{\varepsilon \to 0^+} \left(T(F(\varepsilon, \delta_x)) - T(F)\right)/\varepsilon \ , \tag{10}$$

where $\delta_x$ is the point-mass probability measure at $x \in \mathbb{R}$.

The $IF(x,T,F)$ describes the relative effect (influence) on $T$ of an infinitesimal point-mass contamination at $x$, and measures the local robustness of $T$. An estimator with bounded IF (with respect to a given norm) is therefore robust (locally, as well as globally) and very desirable.

Let $X^n = \{X_1, ..., X_n\}$ be a sample of size $n$ from $X$ in $\mathbb{R}$. The **replacement breakdown point** (BP) of an estimator $T$ at $X^n$ is defined as

$$BP(T, X^n) = \left\{ \frac{m}{n} : \left\| T(X_m^n) - T(X^n) \right\| > \delta \right\} \ , \tag{11}$$

where $X_m^n$ is a contaminated sample resulting from replacing $m$ points of $X^n$ with arbitrary values, $\|\cdot\|$ denotes a norm, $\delta$ is certain content-related threshold, i.e., for the Gini coefficient we can take $\delta = 0.3$ if for that value we face with different social politics basing on the Gini coefficient.

The BP point serves as a measure of global robustness, while the IF function captures the local robustness of estimators. In the context of the simple Pareto, lognormal or gamma distribution estimation it is useful to discriminate between sample contamination with lower values (**LBP**) and sample contamination with upper values (**UBP**).

It is beyond a scope of this paper to introduce the reader into the formal details of the robust statistics. An excellent introduction into the matter could be found for example in Huber and Ronchetti (2009) or Marona et all. (2006). For our purposes it is enough to intuitively understand a following simple

example. Suppose we have five measurements of five monthly salaries (in PLN) in Poland in 2011 year: 3225; 3103; 2944; 3100; 1123, and our aim is to estimate a true value of the "center salary" in Poland in 2011 year. Calculating *the mean* we obtain 2699 but calculating *the median* we obtain 3100. The median is the middle value and in contrary to the mean is not affected by outlying salary 1123. We say that the median is more robust against the outlier than the mean. Similarly calculating a typical measure of dispersion the standard deviation (SD) we obtain 886.63 but calculating robust measure of dispersion – the median of absolute deviations from the median (MAD) we obtain 185.23. We can say that MAD shows the differences in the salaries in a robust way in contrary to the SD. The mean and the SD have unbounded influence functions and their BP are equal to zero. The median and the MAD have bounded IF and maximal BP values.

### 3.1 Robust estimators of Pareto and lognormal distribution.

Let us recall that for specified $\beta_1$ and $\beta_2$ satisfying $0 \le \beta_1, \beta_2 < 1/2$ , a **trimmed mean** is formed by discarding the population $\beta_1$ lowest observations and the proportion of $\beta_2$ uppermost observations and averaging the remaining ones in some sense. In particular, for estimating $\alpha$ , with **known** $x_m$

Brasauskas and Serfling (2000) proposed **the trimmed mean estimator**

$$\hat{\alpha}_{TM} = \left( \sum_{i=1}^{n} c_{ni} \log\left( X_{(i)} / x_m \right) \right)^{-1}, \tag{12}$$

with $c_{ni} = 0$ for $1 \le i \le [n\beta_i]$, $c_{ni} = 0$ for $n - [n\beta_2] + 1 \le i \le n$ and $c_{ni} = 1/d(\beta_1, \beta_2, n)$ for $[n\beta_1] + 1 \le i \le n - [n\beta_1]$, where $[\cdot]$ denotes "*greatest integer part*" and $d\left(\beta_1, \beta_2, n\right) = \sum_{j=[n\beta_1]+1}^{n-[n\beta_2]} \sum_{i=0}^{j-1} (n-i)^{-1}$ .

Next robust estimator appeals to idea of the **generalized median (GM) statistic.** The GM statistics are defined by taking median of the $\binom{n}{k}$ evaluations of a given kernel $h(x_1, ..., x_k)$ over all $k-$ subsets of the data. Brazauskas and Serfling (2002) proposed estimator for the parameter $\alpha$ in Pareto model in case of $x_m$ :

$$\hat{\alpha}_{GM} = MED\left\{ h\left( X_{i1}, ..., X_{ik} \right) \right\}, \tag{13}$$

with a particular kernel $h(x_1, ..., x_k)$ :

$$h(x_1, ..., x_k; x_m) = \frac{1}{C_k} \frac{k}{\sum_{j=1}^{k} \log\left( x_j / x_m \right)}, \tag{14}$$

where $C_k$ is a multiplicative, the median – unibasing factor i.e. chosen so that the distribution of $h(x_1, ..., x_k; x_m)$ has median $\alpha$ - values of $C_k$ for $k = 2$ , $C_2 = 1.1916$ , $k = 3$ $C_3 = 1.1219$ .

For the **lognormal** distribution $L(\mu, \sigma)$ Serfling (2004) introduced GM estimators and obtained their properties. A kernel for the GM **location** estimator takes a form

$$h_1(x_1, ..., x_k) = \frac{1}{k} \sum_{i=1}^{k} \log x_i, \tag{15}$$

$$\hat{\mu}_{GM}(k) = median\left\{ h_1\left( X_1, ..., X_k \right) \right\}. \tag{16}$$

This estimator has the $BP(\hat{\mu}_{GM}(k)) = 1 - (1/2)^{1/k}$ and smooth and bounded IF.
In case of a **scale** estimator, Serfling (2004) proposes using a following kernel

$$h_2(x_1, ..., x_m) = \frac{1}{mM_{m-1}} \sum_{1 \le i < j \le m} (x_i - x_j)^2, \tag{17}$$

which leads to a following robust estimator of scale in lognormal model

$$\hat{\sigma}^2_{GM}(m) = median\left\{h_2\left(X_1,...,X_m\right)\right\}, \tag{18}$$

This estimator has the $BP(\hat{\sigma}^2_{GM}(m)) = 1-(1/2)^{1/m}$ and smooth and bounded IF.

## 4. Measures of income inequality

A measurement of income inequality within a population of economic agents is very closely related to estimation of a probability distribution of income. Incorrect estimates of the distribution may led to incorrect evaluations of the inequalities and incorrect social politics. It should be stressed that we can evaluate a degree of income inequality assuming certain model (i.e., e.g., Pareto model), estimate it and then use known relation between parameters of this model and a measure of the inequality for evaluation of a degree of inequality in the population. From other point of view, it is possible to estimate degree of inequality nonparametrically – i.e., without assumptions on the probability distribution generating the data. The first method is commonly said to be more elegant and easier for economic interpretations. The second method however is in general "closer to a reality" of the observed data.

**Fig. 1**: Pereto densities and corresponding Gini inequality coefficiens.

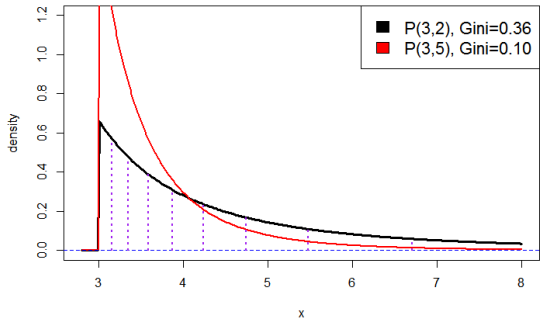
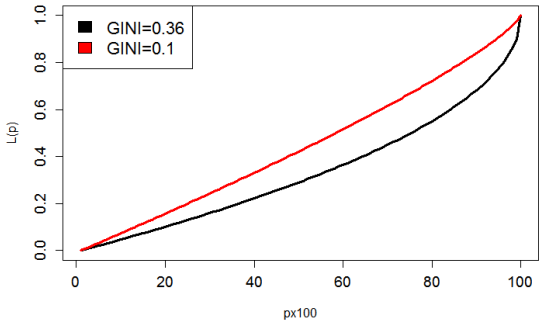



**Fig. 2**: Lorenz corves for Pereto densities and corresponding Gini coefficiens.




Although there are at least twenty popular measures of income inequalities used, a benchmark measure is the Lorentz Curve, which is a graphical representation of the CDF of the empirical probability of wealth. For a discrete probability function $f(y)$ , let $y_i$ , $i=1$ to $n$ be the points with non-zero probabilities indexed in increasing order $y_i < y_{i+1}$ . The **Lorentz Curve** is the continuous piecewise linear function connecting the points $(F_i, L_i)$ , $i=1,...,n$ , where $F_0 = 0$ , $L_0 = 0$ and $F_i = \sum_{j=1}^{i} f(x_i)$, $S_i = \sum_{j=1}^{i} f(x_j)x_j$ , $L_i = S_i / S_n$ . For a pdf function $f(x)$ with the cdf $F(x)$ , the Lorentz curve $L(F(x))$ is given by

$$L(F(x)) = \frac{\int_{-\infty}^{x} tf(t)dt}{\int_{-\infty}^{\infty} tf(t)dt} = \frac{\int_{-\infty}^{x} tf(t)dt}{\mu} . \tag{19}$$

The next popular measure of the income inequality is **the Gini coefficient** which is half the relative mean difference and is usually defined basing on the Lorentz Curve. For random nonzero variable $X$ with cdf $F$ and expected value $\mu$ the Gini coefficient is defined as

$$G = 1 - \frac{1}{\mu}\int_0^{\infty} \left(1 - F(x)\right)^2 dx = \int_0^{\infty} F(x)\left(1 - F(x)\right) dx \ . \tag{20}$$

The **mean difference** is defined as the expected value of the absolute difference of two random variables $X$ and $Y$ independently and identically distributed with the same unknown distribution $MD = E\left[\left|X - Y\right|\right]$ . For a sample $X^n = \{x_1, ..., x_n\}$ it means

$$MD = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|, \tag{21}$$

and the **relative mean difference** is defined as

$$RMD = \frac{MD}{\overline{x}} = 2 \cdot GINI \tag{22}$$

Other popular measures involve the Pietra coefficient, variance of logarithms, Zenga curve, Atchinson or Generalized entropy measure.

Loking into (19), (20), (21) and (22) it is easy to notice that robustness of the sample Lorentz Curve is related to robustness of the sample mean and robustness of a probability density estimator. The Gini coefficient may be calculated on several ways which may give different results in case of an existence of outliers or inliers within the data. Popular method of "robustifying" an estimator involving for example trimming of the data are applicable for (21). We should notice however that the Gini coefficient takes a value from a bounded interval and its breakdown should be understand in a spirit of a certain decision process basing on the Gini estimates. Theory for inequality measures may be obtained within theory of empirical processes, where for example the Gini coefficient is treated as functional of the empirical Lorenz process or within theory sample quantiles and in the same way theory for their robustness may be obtained.

Let us only briefly recall that the Lorenz curve may be generalized to a multivariate case within a data depth concept. The generalization was proposed by Mosler (see Mosler 2013). **Data depth concept** was originally introduced as a way to generalize the concepts of median and quantiles to the multivariate framework. A depth function $D(\cdot, F)$ associates with any $\mathbf{x} \in \mathbb{R}^d$ a measure $D(\mathbf{x}, F) \in [0,1]$ of its centrality w.r.t. a probability measure $F \in \mathcal{P}$ over $\mathbb{R}^d$ or w.r.t. an empirical measure $F_n \in \mathcal{P}$ calculated from a sample $\mathbf{X}^n = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. The larger the depth of $\mathbf{x}$, the more central $\mathbf{x}$ is w.r.t. to $F$ or $F_n$. As an example of depth let us recall the weighted $L^p$ depth from sample $\mathbf{X}^n = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ which is computed as follows:

$$L^p D(\mathbf{x}, \mathbf{X}^n) = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} w\left(\left\|\mathbf{x} - \mathbf{X}_i\right\|_p\right)}, \tag{23}$$

where $w$ is a suitable is non-decreasing and continuous on $[0, \infty)$ weight function, and $\left\|\cdot\right\|_p$ stands for the $L^p$ norm (when $p = 2$ we have the usual Euclidean norm and so called spatial depth).
The set of points for which depth takes value not smaller than $\alpha \in [0,1]$ is multivariate analogue of the quantile and is called the $\alpha-$ central region,

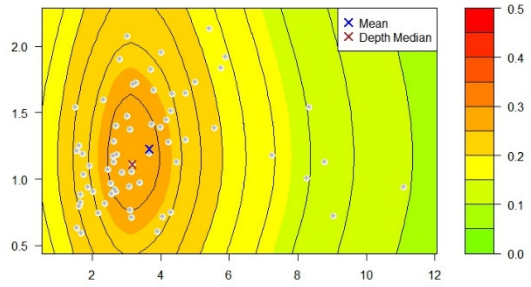
$$D_\alpha(\mathbf{X}) = \{\mathbf{x} \in \mathbb{R}^d : D(\mathbf{x}, \mathbf{X}) \geq \alpha\}. \tag{24}$$

**Multivariate Lorentz curve** is defined proportion of the mean confined to the central region $D_\alpha(\mathbf{X})$ to the overall mean. Let $f(\mathbf{x})$ denote wealth of a point $\mathbf{x}=(x_1,...,x_d)$, i.e., the coordinates of points may represent amounts of *d* goods in an agent disposal. We can define **multivatiate Lorenz Curve** as

$$L(\alpha)=\alpha\times\frac{E\left(f(\mathbf{x})\,|\,\mathbf{x}\in D_\alpha(\mathbf{X})\right)}{E(f(\mathbf{x}))}. \tag{25}$$

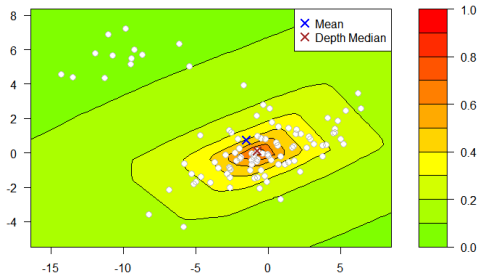
Please note that parameter $\alpha\in(0,1)$ expresses outlyingness of a point w.r.t. a center i.e. multivariate median induced by a depth function. It is however possible to use depth regions consisting probability mass not smaller than $\alpha\in(0,1)$ and hence order them by consisting probability.

**Fig. 3**: A contour plot for sample $L^2$ depth.




**Source:** DepthProc R package

**Fig. 4**: A contour plot for sample projection.




**Source:** DepthProc R package

Fig. 3 presents a contour plot for $L^2$ sample depth and Fig. 4 presents a contour plot for projection sample depth. It is easy to notice that (25) shows an allocation of wealth with respect to a departure from a central object (a multivariate median) – what for several socio-economic reasons may be more interesting than a relation of the object to group of very rich or very poor objects.

## 5. Properties of the robust estimators of income distribution

In order to critically study a performance of known robust estimators of income distributions and income inequalities we conducted intensive simulation as well as empirical studies. Below we present only small part of the results[3]. In a context of the Pareto model estimation we considered MLE, TM and GM estimators which were compared with Victoria-Faser bounded IF proposals as well as with constrained local polynomial estimator proposed by Hyndeman and Yao (2002). We performed similar analysis for the lognormal distribution estimators, Dagum distribution estimators and generalized gamma distribution.

In case of the Pareto distribution we performed intensive simulation studies involving simulated datasets of size 500 observations from the following mixtures of distributions

1. Mixture of $P(1,5)\times10\%$ and $P(10,5)\times90\%$.
2. Mixture of lognormal distribution $LN(2.14,1)\times10\%$ and $P(7,2)\times90\%$.
3. Mixture of normal distribution $N(3300,500)\times10\%$ and $P(2500,4)\times90\%$.
4. Mixture of uniform $U[0,0.1]\times10\%$ distribution and $P(2500,4)\times90\%$ distribution.

Fig. 5–8 present the estimated log densities for the mixtures and $x_m$ taken as minimum. It is easy to notice, that the estimator of the $x_m$ has a crucial issue for the performance of the estimators. With classical MLE estimator for the $x_m$, all the shape parameter estimators perform relatively poorly.

[3] The rest of the results, R codes for calculating the robust estimators are available by request.

**Fig. 5** The estimated densities for the first mixture and $x_m$ taken as as quantile 12%.

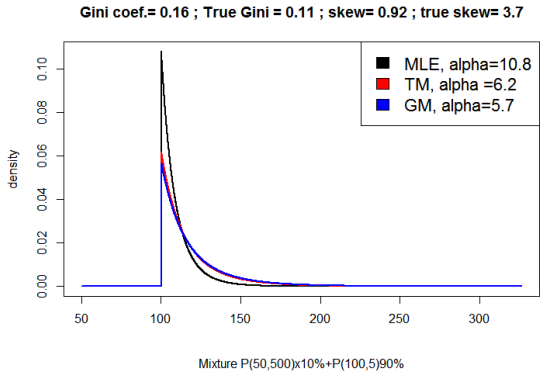



**Fig. 6** The estimated densities for the second mixture and $x_m$ taken as as quantile 12%.

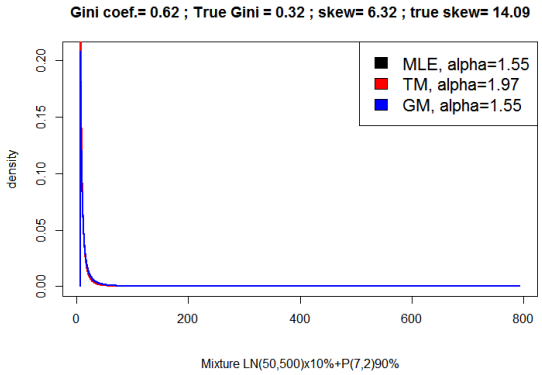
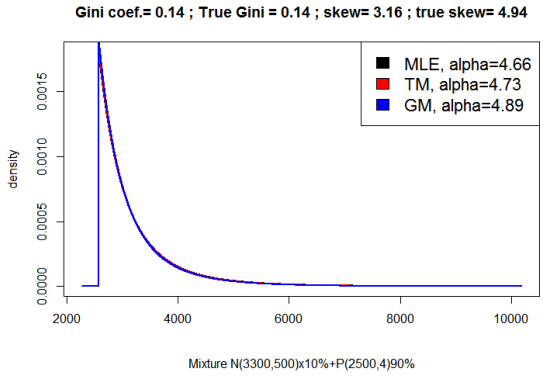



**Fig. 7** The estimated densities for the third mixture and $x_m$ taken as quantile 12%.

**Fig. 8** The estimated densities for the fourth mixture and $x_m$ taken as quantile 12%.



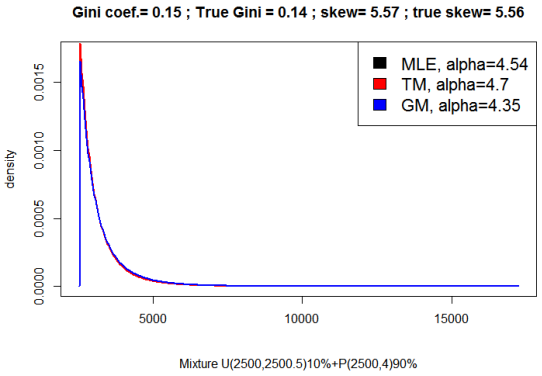
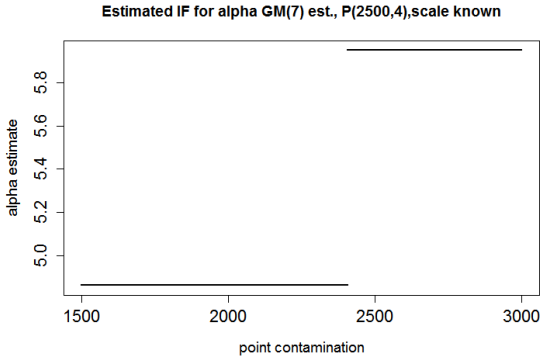




**Fig. 9** The estimated IF for the MLE estimator and stylized sample of 100 obs.

**Fig. 10** The estimated IF for the MLE estimator and stylized sample of 100 obs.



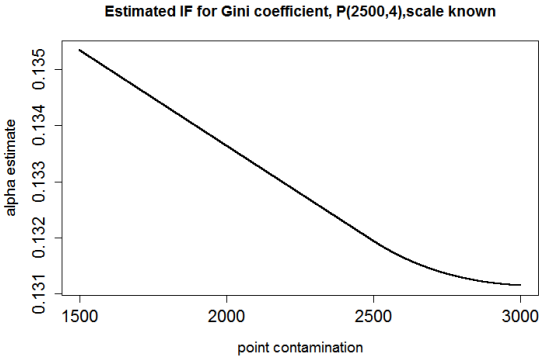




Fig. 9 presents stylized empirical influence function for the GM estimator in case of subsamples consisting of 7 points, Pareto P(2500,4) model and scale estimator taken as quantile of order 0.12. In this case the GM estimator can be treated as robust. Fig. 10 presents stylized empirical influence function for the Gini coefficient. It is easy to notice that this measure of inequality is not robust. Results of simulation lead to the similar conclusions which are similar for other well known income distribution models estimators and popular inequality measures. The conclusions may be summarized as follows:

1. The GM estimators with scale (threshold in three parameter lognormal model) estimated as quantile of order $\beta \in (0,0.3)$, where $\beta$ is optimized using Kolmogorov - Smirnoff goodness of fit

statistics outperforms classical MLE as well as TM estimator. The estimators are computationally intensive however. We recommend using GM type estimator for the scale estimation.

2. It is worth considering to estimate the income distribution nonparametrically - we recommend the constrained local polynomial estimator of Hyndeman and Yao (2002) which provides also estimates of the density derivatives - at least on the explanatory step of a research.

3. We recommend calculating the Gini coefficient "nonparametrically" i.e., without using an assumption of Pareto, log-normal, gamma distributed data. For popular scalar measures of inequality involving the Gini coefficient or Pierta coefficient it is possible to apply the generalized median approach (see Kosiorowski and Tracz 2014b).

For evaluation of the considered robust estimators in case of real data we focused our attention on data considered in Kosiorowski et all (2014) – census data from MINESSOTA POPULATION CENTER (https://international.ipums.org/international/)

We considered data on TOTAL INCOME from the following countries:

**PANAMA** 1960, 1970, 1980, 1990, 2000, 2010
**MEXICO** 1960, 1970, 1990, 1995, 2000, 2005, 2010
**PUERTO RICO** 1970, 1980, 1990, 2000, 2005
**CANADA** 1971, 1981, 1991, 2001
**BRAZIL** 1960, 1970, 1980, 1991, 2000, 2010
**USA** 1960, 1970, 1980, 1990, 2000, 2005, 2010

Each time we estimated the density by means of the GM, TM and M-type estimators (parametrically) after selection of the models by means of information criterion and value of Kolmogoroff goodness of fit statistics. Fig. 11 - 16 present obtained densities by means of constrained local polynomial method which in our opinion is the best counterpart to both classical as well robust estimators. The empirical data showed us a rich set of difficulties related with robust model selection issue. These difficulties are automatically omitted in case of the considered nonparametric method application. It is worth noticing that a kernel used within this method locally protects us against outliers. It is possible using k- nearest neighbours type kernel for a protection against inliers as well. For each case the density was estimated by means of the local linear polynomial estimator in equally spaced grid of 500 points.

**Fig. 11** The estimated income densities in CANADA 1971, 1981, 1991, 2001.

**Fig. 12** The estimated income densities in PUERTO RICO 1970, 1980, 1990, 2000.

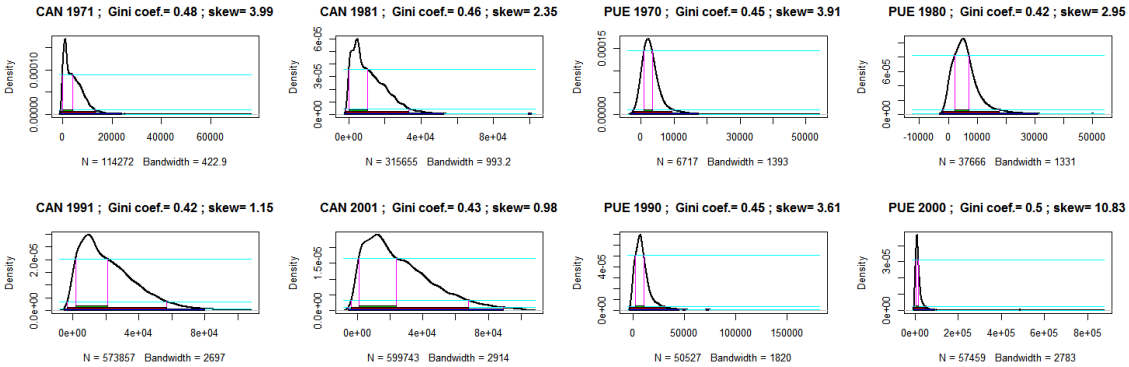


**Fig. 13** The estimated income densities in MEXICO 1960, 1990, 2000, 2010.

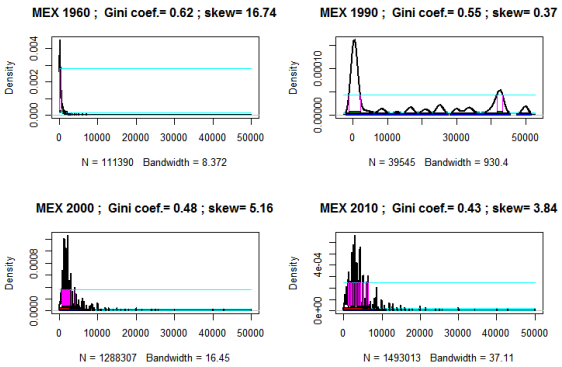



**Fig. 14** The estimated income densities in USA 1960, 1990, 2000, 2010.

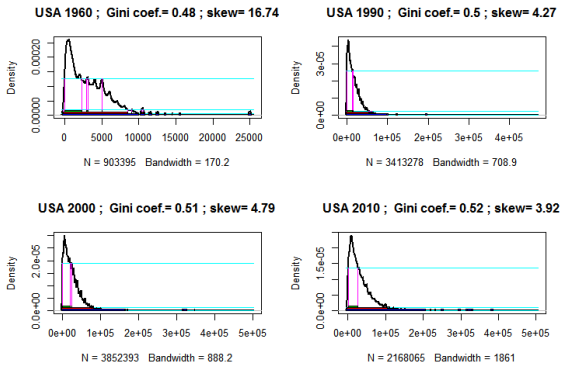



**Fig. 15** The estimated income/ median (income) densities in MEXICO 1960, 1990, 2000, 2010.

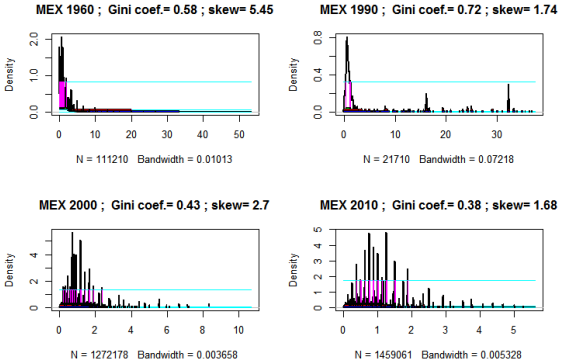



**Fig. 16** The estimated income/ median (income) densities in CANADA 1971, 1981, 1991, 2001.

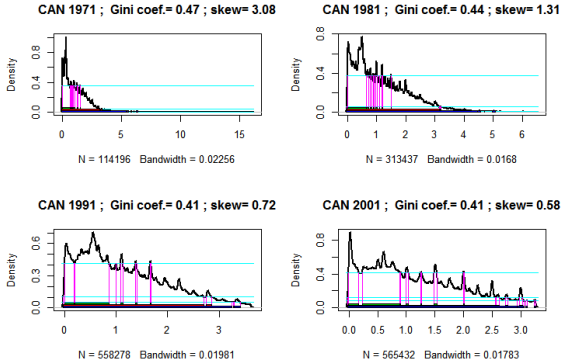



Fig. 15 – 16 presents estimation results for data divided by median incomes. The nonparametric estimator better underlies heterogeneity of the incomes and should be considered at least at a preliminary research step.

## 6. Conclusions

Considerations related to a nature of allocation of wealth within a populations have a central position in the Economics and public debate related to social justice and social solidarity. Arguments used within these debates strongly depend on properties of statistical procedures used for estimation of income distributions and income distribution measures. Classical maximal likelihood estimators of the income distribution parameters are not robust to outliers as well as inliers within the data. There are good robust and/or nonparametric alternatives for them however. We recommend using general generalized median approach of Brazauskas and Serfling in case of an existence of some knowledge on the considered phenomena and the constrained local polynomial estimator in case of lack of knowledge on a subject of study.

## References


1. Brazauskas V., Serfling R. (2000), Robust and Efficient Estimation of the Tail Index of a Single-Parameter Pareto Distribution, North American Actuarial Journal, 4, 12-27.
2. Brazauskas V., Serfling R. (2001), Robust Estimation of Tail Parameters for Two-Parameter Pareto and Exponential Models via Generalized Quantile Statistics, Extremes, 3, 231-249
3. Brazauskas V., Serfling R. (2004), Favorable Estimators for Fitting Pareto Models: A Study Using Goodness-of-Fit Measures with Actual Data" ASTIN Bulletin, 2, 365-381.

4. Dagum, C. (2001) A systemic approach to the generation of income distribution models. In *Income Distribution*, vol. I, M. Sattinger, ed. E. Elgar, Northampton, 32-53.
5. Hyndman, J. R., Yao, Q., 2002. Nonparametric estimation and symmetry tests for conditional density functions. Journal of Nonparametric Statistics 14 (3), 259 278.
6. Kalecki, M. (1945). On the Gibrat distribution. Econometrica, 13, 161–170.
7. Kleiber, C. and Kotz, S. (2002): A characterization of income distributions in terms of generalized Gini coefficients. *Social Choice and Welfare*, 19, 789-794.
8. Kleiber C., Kotz S. (2003), Statistical Size Distributions in Economics and Actuarial Sciences, Wiley, New Jersey
9. Kosiorowski, D., Zawadzki, Z., (2014). DepthProc: An R package for robust exploration of multidimensional economic phenomena. Submitted
10. Kosiorowski, D., Tracz, D. (2014a) On robust estimation of Pareto models and its consequences for government aid programs evaluation, CFM 2014, submitted
11. Kosiorowski, D., Tracz, D. (2014b) A generalized median approach in size distribution modelling and social inequalities evaluation, submitted
12. Kosiorowski, D., Mielczarek, D., Rydlewski, J., Snarska, M. (2014), Applications of the functional data analysis for extracting meaningful information from families of yield curves and income distribution densities, CFM 2014, submitted
13. Maronna, R. A., Martin, R., D. and Yohai, V., J. (2006), Robust Statistics - Theory and Methods, Chichester: Wiley.
14. Mosler, K., 2013. Depth statistics. In: Becker, C., Fried, R., S., K. (Eds.), Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather. Springer, 17-34.
15. Pawlak W., Sztaudynger J.J. (2008), Wzrost gospodarczy a optymalne zróżnicowanie dochodów w USA i Szwecji, Annales – Etyka w życiu gospodarczym, 1, 259-271
16. Serfling, R. (2002), Efficient and Robust Fitting of Lognormal Distributions
17. Victoria – Feser, M. – P. (2000), Robust Methods for the Analysis of Income Distribution, Inequality and Poverty. International Statistical Review, 68, 277 – 293.

**Summary**

Considerations related to income distribution and income inequalities in populations of economic agents belong to the core of the modern economic theory. They appear also in a public debate concerning postulates as to taxation or pension politics, in theories of a human capital creation or searching for regional development factors.
Results of statistical inference conducted for giving arguments pro or against particular hypotheses, strongly depend on properties of statistical procedures used within this process. We mean here for example: a quality of probability density estimator in case of missing data, a quality of skewness measure in multivariate case departing from normality, or a quality of dimension reduction algorithm in case of existence of outliers.
In this paper from the robust statistics point of view, we analyze difficulties related to statistical inference on income distribution models and income inequalities measures. Theoretical considerations are illustrated using real data obtained from Eurostat and Minessota Population Center (IMPUS).

**Streszczenie**

**Wybrane zagadnienia modelowania rozkładu dochodu oraz pomiaru nierówności dochodowych rozpatrywane z punktu widzenia statystyki odpornej**

W pracy prezentowane są wybrane zagadnienia związane z odporną estymacją popularnych rozkładów dochodów oraz z odporną estymacją popularnych miar nierówności dochodowych. Rozważania teoretyczne ilustrowane są przekładami empirycznymi jak również za pomocą symulacji komputerowej.