Pareto Models for Top Incomes

Arthur Charpentier, Emmanuel Flachaire

Pareto Models for Top Incomes

by

Arthur Charpentier

Université du Québec à Montréal (UQAM)
201, avenue du Président-Kennedy,
Montréal (Québec), Canada H2X 3Y7
arthur.charpentier@uqam.ca

and

Emmanuel Flachaire

Aix-Marseille Université
AMSE, CNRS and EHESS,
5 bd Maurice Bourdet,
13001, Marseille, France,
emmanuel.flachaire@univ-amu.fr

May 2019

Abstract

Top incomes are often related to Pareto distribution. To date, economists have mostly used Pareto Type I distribution to model the upper tail of income and wealth distribution. It is a parametric distribution, with an attractive property, that can be easily linked to economic theory. In this paper, we first show that modelling top incomes with Pareto Type I distribution can lead to severe over-estimation of inequality, even with millions of observations. Then, we show that the Generalized Pareto distribution and, even more, the Extended Pareto distribution, are much less sensitive to the choice of the threshold. Thus, they provide more reliable results. We discuss different types of bias that could be encountered in empirical studies and, we provide some guidance for practice. To illustrate, two applications are investigated, on the distribution of income in South Africa in 2012 and on the distribution of wealth in the United States in 2013.

JEL: C46, D31
Keywords: Pareto distribution, top incomes, inequality measures

1 Introduction

Income and wealth distributions are skewed to the right, with thick upper tails. They are often related to Pareto distribution. In his initial work, Vilfredo Pareto suggested several distributions (Pareto 1895, 1896). The term "Pareto distribution" refers to both Pareto I and Generalized Pareto distributions. Rytgaard (1990) wrote that the Pareto I distribution is the common used definition of the Pareto distribution in Europe, and the Generalized Pareto distribution in America. To some extent, this distinction could also be made between economists and statisticians.

To date, economists have mainly used the Pareto I distribution to fit the upper tail of income and wealth distribution (Atkinson 2007, 2017). It is a simple power function, with a single parameter, which can be related to theoretical models that can explain the generation of his thick upper tail (Jones 2015, Benhabib and Bisin 2018).

Recently, Jenkins (2017) addressed the question: *What model should be fitted to top incomes?* In an empirical application, based on tax data in U.K. over 1995-2011, he finds that the preferred specification is the Generalized Pareto distribution (GPD). In this paper, we provide theoretical foundation to this result, and we go further.

Our contribution is manifold. We first show that the Pareto I distribution is very sensitive to the choice of the threshold. In particular, a threshold too low can lead to over-estimate the heaviness of the distribution and, thus, to over-estimate inequality. This bias comes from a misspecification, it does not disappear as the sample size increases. Next, we show that the GPD is less sensitive to the threshold, but its estimation is less accurate. We also show that the Pareto I behaves like the GPD only at (much) higher threshold. Then, we introduce the Extended Pareto distribution (EPD), which is even less sensitive to the threshold and which provides more reliable results. Finally, we discuss different types of bias that could be encountered in practice, and we illustrate our findings through two applications, on income distribution in South-Africa in 2012 and on wealth distribution in the U.S. in 2013. A R package (`TopIncomes`) can be used to reproduced this analysis[1].

In section 2, we present Pareto I and Generalized Pareto distributions. We discuss how sensitive they are to the choice of the threshold. In section 3, we introduce the Extended Pareto distribution. In section 4, we conduct several simulation experiments. In section 5, we discuss different type of

[1] https://github.com/freakonometrics/TopIncomes.

biases. In section 6, we derive formulas for Lorenz curves and top share indices. In section 7, two applications are investigated. Section 8 concludes.

2 Strict Pareto models

Pareto models have been often used for modelling the upper tail of distributions in economic inequality and economic losses, in finance and insurance (Albrecher et al. 2017, Beirlant et al. 2004, Kleiber and Kotz 2003). In this section, we present Pareto Type I and Generalized Pareto distributions and we discuss the close relationship between them. In the following, we consider distributions with finite mean, that is, with Pareto tail parameter greater than one ($\alpha > 1$). For more details on properties of Pareto distributions, see Arnold (2015).

2.1 Pareto I distribution

For Pareto Type I distribution bounded from below by $u > 0$, with tail parameter α, the probability density function and the cumulative density function (CDF) are, respectively, equal to

$$f(x) = \frac{\alpha u^\alpha}{x^{\alpha+1}} \qquad \text{and} \qquad F(x) = 1 - \left(\frac{x}{u}\right)^{-\alpha}, \qquad \text{for } x \geq u \tag{1}$$

If a random variable X has (1) as CDF, we will write $X \sim \mathcal{P}_1(u, \alpha)$.

In the economic inequality literature, Pareto Type I distribution has been always used to fit the upper tail of income and wealth distributions.This distribution has an attractive property: the average above a threshold is proportional to the threshold, it does not depend on the scale parameter u,

$$\mathbb{E}(X|X > u') = \frac{\alpha u'}{\alpha - 1}, \qquad \alpha > 1 \tag{2}$$

where $u' \geq u$. For instance, if the inverted Pareto coefficient $\alpha/(\alpha - 1) = 2$, the average income of individuals with income above \$100,000 is \$200,000 and, the average income of individuals with income above \$1 million is \$2 million (Piketty 2007, Atkinson et al. 2011). This nice property makes easy calculation of the mean and leads to simple formulas of inequality measures (see section 6). Moreover, simple theoretical economic models can be used to explain the basic mechanism that gives rise to Pareto distributions (Jones 2015).

2.2 Generalized Pareto distribution

For Generalized Pareto Distribution (GPD), also known as Pareto Type II distribution, bounded from below by $u \geq 0$, with scale parameter σ and tail parameter α, the cumulative density function (CDF) is

$$F(x) = 1 - \left[1 + \left(\frac{x-u}{\sigma}\right)\right]^{-\alpha} \qquad \text{for } x \geq u \tag{3}$$

where $\sigma > 0$ and $\alpha \in (0, \infty]$.[2] If a random variable X has (3) has its CDF, we will write $X \sim \mathcal{GPD}(u, \sigma, \alpha)$. The $\mathcal{GPD}(0, \sigma, \alpha)$ distribution is also called "Lomax" in the literature (Lomax 1954).

From (1) and (3), we can see that Pareto I is used to model the distribution of relative excesses X/u given $X > u$, while GPD is used to model the distribution of absolute excesses $X - u$ given $X > u$. And because of the additional parameter, there is an affine transformation that links these relative and absolute excesses. The GPD is "generalized" in the sense that Pareto I distribution is a special case, when $\sigma = u$:

$$\mathcal{GPD}(u, u, \alpha) = \mathcal{P}_1(u, \alpha) \tag{4}$$

GPD lets the model decide if the upper tail is better modeled as a power law from relative excesses ($\sigma = u$) or from absolute excesses ($\sigma \neq u$). Overall, GPD is more flexible than Pareto I distribution.

The average above a higher threshold, $u' \geq u$, depends on all parameters of the distribution,

$$\mathbb{E}(X|X > u') = \frac{\sigma - u}{\alpha - 1} + \frac{\alpha}{\alpha - 1}u', \tag{5}$$

The linearity of this function characterizes the GPD class (see Guess and Proschan 1988 and Ghosh and Resnick 2010).

In the statistical literature, GPD is often used to fit the upper tail of heavy-tailed distributions (Beirlant et al. 2004). Indeed, one of the most important result in the extreme value theory states that, for most heavy-tailed distributions, the conditional excess distribution function, above a threshold u, converges towards a GPD distribution as u goes to infinity (Pickands 1975, Balkema and de Haan 1974), for some parameters α and σ,

$$F_u(x) \longrightarrow \text{ GPD (or Pareto II)} \quad \text{ as } u \to +\infty \tag{6}$$

[2] or $\alpha \in (1, \infty]$ if we want to compute averages that have a probabilistic interpretation. The limiting case $\alpha \to \infty$ corresponds to an exponential distribution.

where $F_u(x) = P(X - u \leq x | X > u)$. This result is known as the Pickands-Balkema-de Haan theorem, also called the second theorem in extreme value theory (as discussed in footnote 6, page 7). It provides strong theoretical support for modelling the upper tail of heavy-tailed distributions with GPD, or Pareto Type II distribution. In a very general setting, it means that there are α and σ such that F_u can be approximated by the CDF of a $\mathcal{GPD}(u, \sigma, \alpha)$, see Embrechts et al. (1997, Theorem 3.4.13, p.165).

2.3 Threshold selection

Estimation of the upper tail of a distribution with a Pareto distribution proceeds with a preliminary choice of the threshold u by the researcher. In statistics, the choice of the threshold is notoriously difficult, and the results are often quite sensitive to this choice in empirical studies.

The choice of the threshold means that only the k largest observations are considered, and the distribution of those k observations is supposed to be Pareto, with tail index α. There is a bias-variance trade-off: if the threshold is too high, k is small and estimators can be very volatile, while if the threshold is too small, there is less variability, but the bias can be large. In practice, an *optimal* threshold would then be the lowest threshold that does not generate significant bias in parameter estimates.

Whether to use a Pareto I or GPD model to fit the upper tail is related to the choice of the threshold. From (1) and (3), we have seen that Pareto I is a special case of GPD, when $\sigma = u$. They differ by an affine transformation when $\sigma \neq u$. A key property of Pareto distributions is that, if a distribution is Pareto for a fixed threshold u, it is also Pareto with the same tail parameter α for a higher threshold $u' \geq u$. For GPD, we have

$$\overline{F}_u = \mathcal{GPD}(u, \sigma, \alpha) \quad \Rightarrow \quad \overline{F}_{u'} = \mathcal{GPD}(u', \sigma + u' - u, \alpha). \tag{7}$$

where $\overline{F}_u$ is the survival excess function above u.[3] Thus, Pareto I and GPD are the same for all $u' \geq u$, if $\sigma = u$. Otherwise, we have $\sigma + u' - u \approx u'$ for very large values of u' only. It follows that a GPD above a threshold will behave approximately as a Pareto I above a *higher* threshold, much higher as σ differs from u.

[3] $\overline{F}_{u'}$ is a truncated Pareto distribution, with density equals to $f(x)/(1 - F(u'))$. Note that this property is quite intuitive, since the GPD distribution appears as a limit for exceeding distributions, and limit in asymptotic results are always fixed points: the Gaussian family is stable by addition (and appears in the Central Limit Theorem) while Fréchet distribution is max-stable (and appears in the first theorem in extreme value theory).

To illustrate, Figure 1 shows boxplots of the maximum likelihood estimator (MLE) of the tail index α, from Pareto I (left) and GPD (right) models, as the threshold increases.[4] The values are obtained from 1,000 samples of 1,000 observations drawn from a GPD where σ differs from u, $\mathcal{GPD}(0.5, 1.5, 2)$.

For Pareto I model (left), the MLE of the tail index is the Hill estimator,

$$\widehat{\alpha}_k = \left[\frac{1}{k} \sum_{i=n-k+1}^{n} \log x_{(i)} - \log x_{(n-k+1)} \right]^{-1} \tag{8}$$

from a sample $\{x_1, \cdots, x_n\}$, with the ordered version $x_{(1)} \leq x_{(2)} \leq \cdots x_{(n)}$, where only the k largest values are considered. We can see that the boxes are always below the true value (dashed line), and the dispersion increases as the threshold increases. It shows that, if the threshold is not very high, Pareto I model provides severe bias in the estimation of the tail parameter. For instance, with a threshold $u = 0.5$, we have $\hat{\alpha} \approx 1$, that is, the estimated tail parameter is half as small as the true value, $\alpha = 2$.

For GPD model (right), we can see that there is no bias, but the dispersion is much greater. Moreover, several values are extremes, it can be more than twice the true value (circles in the plot, see footnote 4). Here, the MLE of the tail index has no analytical solution, and numerical methods are required for the estimation. It is known that the MLE of the GPD can be very unstable and, in some cases, has no solution (Castillo and Hadi 1997, Hosking and Wallis 1987, del Castillo and Daoudi 2009).

[Figure 1 about here.]

Overall, this Figure illustrates that Pareto Type I distribution behaves like a GPD only at much higher threshold, when σ differs from u. From his analysis on UK income data, Jenkins (2017) concludes that "the Pareto I model is as good as GPD only at extremely high incomes, beyond a range of thresholds usually considered". It is consistent with the values of $\sigma - u$ that he found, always very different from zero.[5]

[4] Boxplots provide information on the median, skewness, dispersion and outliers. The median is the band inside the box. The first and third quartiles (q1,q3) are the bottom and the top of the box. The outlier detection is based on the interval $[\underline{b}; \overline{b}]$, where $\underline{b} = q_1 - 1.5IQR$, $\overline{b} = q_3 + 1.5IQR$ and $IQR = q_3 - q_1$ is the interquartile range. Any values that fall outside the interval [b;b] are detected as outliers, they are plotted as individual circles. The horizontal lines at the top and bottom of each boxplot correspond to the highest and smallest values that fall within the interval $[\underline{b}; \overline{b}]$, see Pearson (2005).

[5] The plots of $u - \sigma$ can be found in bottom Figures in Appendix H, in the online supplementary material of Jenkins' paper, pp.102-116.

3 Pareto-type models

In the previous section, we have seen that lower threshold can be used with GPD model, compared to Pareto I. Nevertheless, a (very) high threshold may still be required to have (nearly) unbiased estimation of the tail parameter with GPD model. To select not too high threshold, we need to consider Pareto-type models, rather than strictly Pareto models. It leads us to consider higher order of regular variation, from which an Extended Pareto distribution (EPD) can be derived. The EPD provides better fit of the upper tail when the distribution becomes strictly Pareto in the very top of the distribution only.

3.1 First- and second-order regular variation

The tail index α is related to the max-domain of attraction of the underlying distribution, while parameter σ is simply a scaling parameter.[6] The shape of the conditional excess cumulative distribution function is a power function (the Pareto distribution) if the threshold is large enough. Tails are then said to be Pareto-type, and can be described using so called *regularly varying* functions (see Bingham et al. 2013).

First and second order regular variation were originally used in extreme value theory, to study respectively the tail behavior of a distribution and the speed of convergence of the extreme value condition (see Bingham et al. 2013, de Haan and Stadtmüller 1996, Peng and Qi 2004, or section 2 in de Haan and Ferreira 2006 for a complete survey). A function H is said to be regularly varying (at infinity) with index $\gamma \in \mathbb{R}$ if

$$\lim_{t\to\infty} \frac{H(tx)}{H(t)} = x^{\gamma} \quad \text{or} \quad \lim_{t\to\infty} x^{-\gamma}\frac{H(tx)}{H(t)} = 1. \tag{9}$$

A function regularly varying with index $\gamma = 0$ is said to be slowly varying.

[6] Historically, extremes were studied through block-maximum - yearly maximum, or maximum of a subgroup of observations. Following Fisher and Tippett (1928), up to some affine transformation, the limiting distribution of the maximum over n i.i.d observations is either Weibull (observations with a bounded support), Gumbel (infinite support, but light tails, like the exponential distribution) or Fréchet (unbounded, with heavy tails, like Pareto distribution). Pickands (1975) and Balkema and de Haan (1974) obtained further that not only the only possible limiting conditional excess distribution is GPD, but also that the distribution of the maximum on subsamples (of same size) should be Fréchet distributed, with the same tail index γ, if $\gamma > 0$. For instance in the U.S., if the distribution of maximum income per county is Fréchet with parameter γ (and if county had identical sizes), then the conditional excess distribution function of incomes above a high threshold is a GPD distribution with the same tail index γ.

Observe that any regularly varying function of index $-\gamma$ can be written $H(x) = x^{-\gamma}\ell(x)$ where ℓ is some slowly varying function.

Consider a random variable X, its distribution is regularly-varying with index $-\gamma$ if, up-to some affine transformation, its survival function is regularly varying. Hence,

$$\lim_{t\to\infty} x^{-\gamma}\frac{\overline{F}(tx)}{\overline{F}(t)} = 1 \tag{10}$$

or

$$\overline{F}(x) = x^{-\gamma}\ell(x), \tag{11}$$

where $\overline{F}(x) = 1 - F(x)$. A regularly varying survival function is then a function that behaves like a power law function near infinity. Distributions with survival function as defined in (11) are called *Pareto-type distributions*. It means that the survival function tends to zero at polynomial (or power) speed as $x \to \infty$, that is, as $x^{-\gamma}$. For instance, a Pareto I distribution, with survival function $\overline{F}(x) = x^{-\alpha}u^{\alpha}$, is regularly varying with index $-\alpha$, and the associated slowly varying function is the constant u^{α}. And a GPD or Pareto II distribution, with survival function $\overline{F}(x) = \left(1 + \sigma^{-1}x\right)^{-\alpha}$, is regularly varying with index $-\alpha$, for some slowly varying function. But in a general setting, if the distribution is not *strictly* Pareto, ℓ will not be constant, and it will impact the speed of convergence.

In de Haan and Stadtmüller (1996), a concept of second-order regular variation function is introduced, that can be used to derive a probabilistic property using the quantile function,[7] as in Beirlant et al. (2004). Following Beirlant et al. (2009), we will consider distributions such that an extended version of equation (10) is satisfied,

$$\lim_{t\to\infty} x^{-\gamma}\frac{\overline{F}(tx)}{\overline{F}(t)} = 1 + \frac{x^{\rho} - 1}{\rho}, \text{ for some } \rho \leq 0, \tag{12}$$

that we can write, up to some affine transformation,

$$\overline{F}(x) = cx^{-\gamma}[1 - x^{\rho}\ell(x)], \tag{13}$$

for some slowly varying function ℓ and some second-order tail coefficient $\rho \leq 0$. The corresponding class of Pareto-type distributions defined in (13) is often named the Hall class of distributions, referring to Hall (1982). It includes the Singh-Maddala (Burr), Student, Fréchet and Cauchy distributions. A mixture of two strict Pareto-I distributions will also belong to this class.

[7] The quantile function U is defined as $U(x) = F^{-1}(1 - 1/x)$.

Since $\rho \leq 0$, $x \mapsto 1 - x^{\rho}\ell(x)$ is slowly varying, and therefore, a distribution $\overline{F}$ that satisfies (13) also satisfies (11). More specifically, in (13), the parameter ρ captures the rate of convergence of the survival function to a strict Pareto distribution. Smaller is ρ, faster the upper tail behaves like a Pareto, as x increases. Overall, we can see that

- γ is the first-order of the regular variation, it measures the tail parameter of the Pareto distribution,
- ρ is the second-order of the regular variation, it measures how much the upper tail deviates from a strictly Pareto distribution.

In the following, we will write $\text{RV}(-\gamma, \rho)$. There are connexions between tail properties of the survival function $\overline{F}$ and the density f (see also Karamata theory for first order regular variation). More specifically, if $\overline{F}$ is $\text{RV}(-\gamma, \rho)$, with $\gamma > 1$, then f is $\text{RV}(-\gamma - 1, \rho)$.

For instance, consider a Singh-Maddala (Burr) distribution, with survival distribution $\overline{F}(x) = [1 + x^a]^{-q}$, then a second order expansion yields

$$\overline{F}(x) = x^{-aq}[1 - qx^{-a} + o(x^{-a})] \text{ as } x \to \infty \tag{14}$$

which is regularly varying of order $-aq$ and with second order regular variation $-a$, that is $\text{RV}(-aq, -a)$.

Schluter (2018) shows that the bias, induced by higher order regular variation in the Burr case (14), for the estimator of the inverted tail parameter $\xi = 1/\alpha$ in Pareto I distribution, obtained from an OLS regression of log sizes on log ranks, is equal to

$$b_{k,n} = \frac{1}{2}\xi \frac{2-\rho}{(1-\rho)^2}\left(\frac{n}{k}\right)^{\rho}, \tag{15}$$

where the k largest observations are considered to fit a Pareto I and n is the sample size. Since $\rho \leq 0$, the bias is positive, it increases as ρ and k increases. Thus, the distribution will be estimated heavier than it is really, as ρ tends to zero, and the bias can be large if the threshold is not high enough. For values of ρ close to zero, the bias of the tail parameter of Pareto I model is then expected to be negligible for extremely high threshold only.

3.2 Extended Pareto distribution

Beirlant et al. (2009) show that (13) can be approximated by

$$\overline{F}(x) = \left[x\left(1 + \delta - \delta x^{\tau}\right)\right]^{-\alpha} \qquad \text{for } x \geq 1 \tag{16}$$

where $\tau \leq 0$ and $\delta > \max(-1, 1/\tau)$.[8] The main feature of this function is that it captures the second-order regular variation of the Hall class of distributions, that is, deviation to a strictly Pareto tail. For more details, see also Albrecher et al. (2017, section 4.2.1).

From (16), we can define the Extended Pareto Distribution (EPD), proposed by Beirlant et al. (2009), as follows:

$$F(x) = 1 - \left[\frac{x}{u}\left(1 + \delta - \delta\left(\frac{x}{u}\right)^{\tau}\right)\right]^{-\alpha} \qquad \text{for } x \geq u \tag{17}$$

where $\tau \leq 0$ and $\delta > \max(-1, 1/\tau)$. If a random variable X has (17) has its CDF, we will write $X \sim \mathcal{EPD}(u, \delta, \tau, \alpha)$.

Pareto I is a special case when $\delta = 0$ and GPD is a special case when $\tau = -1$:

$$\mathcal{EPD}(u, 0, \tau, \alpha) = \mathcal{P}_1(u, \alpha) \tag{18}$$

$$\mathcal{EPD}(u, \delta, -1, \alpha) = \mathcal{GPD}(1, u/(1+\delta), \alpha) \tag{19}$$

The mean over a threshold for the EPD distribution has no closed form expression.[9] Numerical methods can be used to calculate it. Since $u' \geq u > 0$, X given $X > u'$ is a positive random variable and

$$\mathbb{E}[X|X > u'] = \int_0^{\infty} \overline{F}_{u'}(x)dx \tag{20}$$

where $\overline{F}_{u'}(x) = \mathbb{P}[X > x|X > u']$ for $x > u$, i.e.

$$\overline{F}_{u'}(x) = \frac{\overline{F}(x)}{\overline{F}(u')} \text{ where } \overline{F} \text{ is the s.d.f. of } X \tag{21}$$

Thus

$$\mathbb{E}[X|X > u'] = u' + \frac{1}{\overline{F}(u')}\int_{u'}^{\infty} \overline{F}(x)dx \tag{22}$$

The integral in (22) can be computed numerically. Since numerical integration over a finite segment could be easier, we can make a change of variable $(1/x)$ to obtain an integral over a finite interval:

$$E_{u'} = \int_{u'}^{\infty} \overline{F}(x)dx = \int_0^{1/u'} \frac{1}{x^2}\,\overline{F}\left(\frac{1}{x}\right)dx \tag{23}$$

[8] Using the expansion $(1 + y^a)^b \approx 1 + by^a$, for small y^a, in (16) yields (13).

[9] Albrecher et al. (2017, section 4.6) give an approximation, based on $(1+\delta-\delta y^{\tau})^{-\alpha} \approx 1 - \alpha\delta + \alpha\delta y^{\tau}$, which can be very poor. Thus, we do not recommend to use it.

The Extended Pareto distribution has a stable tail property: if a distribution is EPD for a fixed threshold u, it is also EPD for a higher threshold $u' \geq u$, with the same tail parameter α. Indeed, deriving a truncated EPD distribution, we find

$$\overline{F}_u = \mathcal{EPD}(u, \delta, \tau, \alpha) \quad \Rightarrow \quad \overline{F}_{u'} = \mathcal{EPD}(u', \delta', \tau, \alpha). \tag{24}$$

where $\delta' = \delta(u'/u)^\tau/[1+\delta-\delta(u'/u)^\tau]$. A plot of estimates of the tail index α for several thresholds would then be useful. If the distribution is Extended Pareto, a stable horizontal straight line should be plotted. This plot is similar to Hill plot for Hill estimates of α from Pareto I distribution. It is expected to be more stable if the distribution is not strictly Pareto. Indeed, Embrechts et al. (1997, p.194) and Resnick (2007, p.87) illustrate that the Hill estimator can perform very poorly if the slowly varying function is not constant in (11). It can lead to very volatile Hill plots, also known as Hill *horror* plots.

4 Simulations

In this section, we use Monte Carlo experiments to study the sensitivity of Pareto models to the choice of the threshold and to the degree of deviation from a strict-Pareto distribution in the upper tail.

4.1 Sensitivity to the threshold

Figure 2 shows boxplots of values of the tail parameter α, estimated by maximum likelihood from Pareto I (left), GPD (middle) and EPD (right) models, as the threshold increases. The x-axis is the percentage of the largest observations used to fit Pareto distributions. The values are obtained from 1,000 samples of 50,000 observations drawn from a Singh-Maddala distribution that closely mimics the US 2013 income distribution, $SM(2.07, 1.14, 1.75)$.[10] This distribution is Pareto-type, with RV(-3.63,-2.07), see (14).

[Figure 2 about here.]

For Pareto I model (left), we can see that the income distribution is estimated too heavy, whatever the threshold is (blue boxes are always below

[10] These parameters are obtained from the estimation of a Singh-Maddala distribution with the US incomes in 2013 provided by LIS cross-national center in Luxemburg ($a = 2.07$ and $q = 1.75$ are shape parameters, $b = 1.14$ is a scale parameter)

the true value, given by the dashed line). This results is expected from Schluter's result in (15). Even when the Pareto I distribution is estimated with the top 1% observations only (highest threshold, $k = 500$), the box is clearly below the true value. In other words, the Hill estimator is biased downward: the lower the threshold, the more biased it is.

For GPD model (middle), we can see that the income distribution is estimated less heavy than it is in reality (green boxes are always above the true value). It is only for very high thresholds that the estimates/boxes are around the true value. Moreover, we can see that the dispersion is much higher than for Pareto I model.

For Extended Pareto model (right), we can see that the estimation of the tail parameter outperforms Pareto I and GPD models. Indeed, red boxes are always much closer to the true value, compared to blue boxes (Pareto I) and green boxes (GPD), for given threshold. The dispersion is slightly larger than Pareto I, and much smaller than GPD. [11]

[Figure 3 about here.]

It is important to note that the bias doesn't disappear as the sample size increases, since it is a misspecification problem. To illustrate, Figure 3 shows similar results from huge samples, that is, from samples of 1 million of observations. Compared to Figure 2, Figure 3 shows clearly that the dispersion decreases as the sample size increases, not the bias. Thus, having several millions of observations does not reduce the bias.

4.2 Sensitivity to deviations from a strict-Pareto

Using a Singh-Maddala (Burr) distribution allows us to control the first- and second- order of the regular variation, see (14). Moreover, this distribution is often used to estimate density of income distribution in parametric approaches, see Cowell and Flachaire (2015).

In our experiments, we generate 10,000 samples of 50,000 observations from Singh-Maddala distributions, $SM(-\rho, 0.5, -\alpha/\rho)$, with an upper tail that deviates more and more from a strict-Pareto, that is, with an increasing 2nd-order RV parameter $\rho = -2, -1, -0.75, -0.5$. For each sample, we estimate the tail index α by maximum likelihood based on Pareto I, GPD

[11] We have also made similar experiments for two extensions of the GPD model, proposed in the literature: Pareto IV (Arnold 2015) and Extended Generalized Pareto, or EGPD3 (Papastathopoulos and Tawn 2013). We do not report the results here, since the dispersion is always much larger and they are outperformed by Extended Pareto model.

and EPD models, fitted on, respectively, the top 10%, 5% and 1% largest observations:

- Pareto 1 models fitted on, respectively, the top 10%, 5%, 1% largest observations are, respectively, denoted `P1`, `P1`', `P1`",
- GPD models fitted on, respectively, the top 10%, 5%, 1% largest observations are, respectively, denoted `GPD`, `GPD`', `GPD`",
- EPD models fitted on, respectively, the top 10%, 5%, 1% largest observations are, respectively, denoted `EPD`, `EPD`', `EPD`".

Atkinson et al. (2011, p.15) observe that, in practice, the tail index α of income distributions typically vary between 1.5 and 3. When $\alpha < 2$, the variance is infinite. In the following, we first consider the case of finite variance ($\alpha = 3$) and the case of infinite variance ($\alpha = 1.5$).

Finite variance

[Figure 4 about here.]

Figure 4 shows boxplots of the tail index estimates $\hat{\alpha}$, when the true value is $\alpha = 3$. The case $\rho = -2$ (top-left) is quite similar to the case used in the previous subsection and the results have similar patterns as those in Figure 2. The other cases, where ρ increases, are used to analyze what happens when the deviation from a strict Pareto in the upper tail increases. We can see that:

- Pareto I model (Hill estimator) shows severe under-estimation of the tail index α, when ρ increases (boxes are increasingly below the horizontal dashed line). As expected by Schluter (2018), the distribution is estimated too heavy with a Pareto I model, see (15).
- GPD model exhibits less bias, but much more dispersion, when ρ increases (boxes are closer to the dashed line, but wider). Several outliers are far from the true value. It suggests that GPD can perform better than Pareto I, but it can also provide very poor results.[12]

[12] See the discussion in section 2.3. Note that GPD seems to have no bias when $\rho = -1$ (boxes are around the dashed line). It is not surprising, since the Singh-Maddala distribution is a GPD when the shape parameter $a = 1$ in (14), which corresponds to $\rho = -1$.

- EPD model outperforms Pareto I model. The dispersion is slightly increased, but the bias is largely reduced, when ρ increases. However, the EPD model does not correct completely the bias when ρ is large (boxes are still below the horizontal dashed line).[13]

Infinite variance

[Figure 5 about here.]

Figure 5 shows boxplots of the tail index estimates $\hat{\alpha}$, when the true value is $\alpha = 1.5$. The results are quite similar to the case of finite variance. It provides poor results with Pareto I model. Fitted on the top 10% of the sample (`P1`), with a large deviation to a strict-Pareto ($\rho = -0.5$), the tail index is estimated smaller than 1, which suggests that the mean doesn't exist and Pareto models cannot be used. Note that similar results have been obtained in empirical studies on wealth distribution. For instance, Cowell (2013) finds $\hat{\alpha} = 0.48, 0.52, 0.73$ from a Pareto I model fitted on, respectively, top 10%, 5% and 1% observations of the U.S. wealth distribution.

Overall, the simulation results show that Pareto I model can lead to severe over-estimation of the heaviness of the distribution, when the threshold is not large enough and when the distribution deviates from a strict-Pareto. At given threshold, the EPD model can provide significant improvement over Pareto I model.

5 From theory to practice

In empirical studies, the analysis of Pareto-type upper tail is often inferred from some plots. For Pareto Type I distribution, with tail parameter α, rearranging (1) and applying a logarithmic transformation, we have

$$\log(1 - F(x)) = c - \alpha \log x \tag{25}$$

where $c = \alpha \log u$. Given a sample $\{x_1, \cdots, x_n\}$, let us consider the plot of the log of incomes in the x-axis and the log of the survival function in the y-axis:

$$\{\log x,\, \log\left(1 - F(x)\right)\} \tag{26}$$

[13] This model is based on the approximation $(1 + \delta - \delta y^{\rho})^{-\alpha} \approx 1 - \alpha\delta + \alpha\delta y^{\rho}$, which is better as ρ is more and more negative when $y > 1$.

or its empirical version

$$\left\{\log x_i,\, \log\left(1-\widehat{F}(x_i)\right)\right\} \tag{27}$$

This plot, known as the Pareto diagram (Cowell 2011) or Zipf plot (Cirillo 2013), shows the proportion of the population with x or more against x itself on a double-logarithmic diagram. If the distribution is strictly Pareto, it should exhibit a linear function, where the slope coefficient (taken positively) is equal to the tail index α. If the distribution is not strictly Pareto, the curve obtained from a Pareto diagram is not linear.[14] This interpretation can be carried over to the general case of Pareto-type distributions since then ultimately for $x \to \infty$ the Pareto diagram is still linear with slope α at some set of largest values (Albrecher et al. 2017, p.70).

In practice, tail index estimation is subject to three potential sources of bias: misspecification bias, estimation bias and sampling bias.

5.1 Misspecification bias

Let us assume that the CDF is known, $F(x)$.

[Figure 6 about here.]

Figure 6 shows a Pareto diagram, based on a Singh-Maddala distribution that closely mimics the US 2013 income distribution, $SM(2.07, 1.14, 1.75)$.[10] This plot exhibits a concave curve, which becomes linear in the right part, when $\log x > 1$. As expected, it suggests that the distribution behaves like a Pareto distribution in the upper tail. It also exhibits two linear functions, as defined in (25), obtained when we consider two different threshods, $\log u = -2, 1$. Since α is both the slope of the linear function (taken positively) and the tail index of the Pareto distribution, we can see that:

- α increases as the threshold u increases, until it is constant,
- the distribution is then estimated too heavy if u is not large enough.

The simulation results in Figures 2 and 3 are obtained with samples drawn from the Singh-Maddala distribution used to draw the Pareto diagram in

[14] Another popular plot is the Pareto QQ-plot, $\{-\log(1-F(x_i)),\, \log x_i\}$. It is obtained from the tail quantile function of Pareto I distribution, $Q(p) = (1-p)^{-1/\alpha}$. Hence, $\log Q(p) = -\frac{1}{\alpha}\log(1-p)$. If the distribution is Pareto I, we expect to see a linear function with a slope coefficient equal to $1/\alpha$. Note that these plots may not be helpful to distinguish between a Lognormal and a Pareto distributions (Cirillo 2013).

Figure 6. The concavity of the Pareto diagram explains why the Pareto I model (Hill estimator) under-estimates the tail index and, thus, over-estimates the heaviness of the distribution.

Most Pareto diagrams based on the CDF exhibit typically a concave-like curvature, at least for large x, when the distribution is heavy-tailed.[15] Thus, we may expect that distributions are often estimated too heavy with Pareto I model, if the threshold is too low.[16] This problem occurs because the linear approximation is not appropriate. It is important to note that it is a misspecification bias, which does not disappear as the sample size increases to infinity (see section 4.1).

5.2 Estimation bias

In empirical studies, the CDF is in general unknown. We replace it by a consistent estimate, $\hat{F}(x)$. When we consider the estimation of α, the replacement of the CDF by an estimate introduce another type of bias: an estimation bias.

A standard choice is the EDF, which is a consistent estimator of the CDF. If we sort a sample in ascending order, $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, the Pareto diagram is then the log of the ordered observations in the x-axis and the log of the *empirical* survival function in the y-axis:[17]

$$\texttt{Pareto diagram:} \quad \left\{ \log x_{(i)},\, \log\left(1 - \frac{i}{n+1}\right) \right\} \quad i = 1, \ldots, n \tag{28}$$

The Pareto diagram often exhibits an erratic behavior in the extreme right part. It is because the EDF provides a poor estimation of the upper tail of the CDF, especially when the distribution is heavy-tailed.[18] Indeed, sample quantiles poorly estimate higher population quantiles when the distribution is heavy-tailed.[19]

[15] Schluter (2019, appendix A) shows formal derivations on the concavity of Pareto QQ-plots, with the Hall class of distributions. Using many different parameters and parametric distributions, we always found concave Pareto diagrams, based on the CDF.

[16] Beirlant et al. (2004, p.113) argue that, in many cases, the Hill estimator overestimates the population value of $1/\alpha$.

[17] With R, we can plot a Pareto diagram with: `plot(log(sort(x)),log((n:1)/(n+1)))`, and a Pareto QQ-plot with `plot(-log((n:1)/(n+1)),log(sort(x)))`.

[18] This is why Davidson and Flachaire (2007) and Cowell and Flachaire (2007) proposed to fit the upper tail with a parametric Pareto distribution to improve inference.

[19] The (asymptotic) variance of sample quantiles is equal to $n^{-1}F(x)(1-F(x))/f(x)^2$. For Pareto I distribution, $F(x) = 1-x^{-\alpha}$, the variance is then equal to $n^{-1}(x^{\alpha+2}-x^2)/\alpha^2$, which explodes as $x \to +\infty$.

To illustrate the consequences of this erratic behavior, let us consider the value of the maximum income in the sample, $x_{(n)}$. The Pareto diagram tells us that the proportion of population with the maximum income or more is equal to $1/(n+1)$. It follows that,

- if the maximum income is too large (Bill Gate's income belongs to the sample),[20] the proportion of the population with income more or equal to $x_{(n)}$ is over-estimated: $1/(n+1) > 1-F(x_{(n)})$. Thus, the heaviness of the upper tail is over-estimated.

- if the maximum income is too small (topcoding, censoring, truncation or underreporting),[21] the proportion of the population with income more or equal to $x_{(n)}$ is under-estimated: $1/(n+1) < 1-F(x_{(n)})$. Thus, the heaviness of the upper tail is under-estimated.

Figure 7 illustrates these features. A Pareto diagram is obtained from a sample of 20,000 observations drawn from the Singh-Maddala distribution $SM(2.07, 1.14, 1.75)$.Firstly, we gradually increase the largest incomes, that is, those such that $\log(x) > 2$. The Figure shows that the presence of outliers pushes the extreme part of the curve to the right, which becomes convex before becoming concave again. Secondly, we topcode the largest incomes, that is, every log-income greater than 2 is set equal to 2. The Figure shows that the topcoding pushes the extreme part of the curve to the left, with a more pronounced concavity.

[Figure 7 about here.]

Overall, the estimation bias can go in opposite directions. It will depend on the exact nature of the problem. The presence of outliers in the sample is known to biased downward classical estimators of the tail index (Hubert et al. 2013). Thus, it tends to estimate the upper tail too heavy. To the opposite, top-coding or underreporting from the rich[22] tend to biased upward estimators of the tail index. Thus, it tends to under-estimate the heaviness of the upper tail.

In applications, the main question is which type of bias is the most pronounced and to what extent? Examining the Pareto diagrams can help to answer this question, as we will show in the application section.

[20] More precisely, if $x_{(n)}$ is greater than the $n/(n+1)$-percentile of the population.

[21] More precisely, if $x_{(n)}$ is smaller than the $n/(n+1)$-percentile of the population.

[22] Linking a restrictive subsample of Uruguay's official household survey to tax data, Higgins et al. (2018) show that the rich tend to underreport their income.

5.3 Sampling bias

It is often argued that surveys do not capture well the top of income distributions, because the rich may be harder to reach or more likely to refuse to participate (Atkinson 2007, Atkinson et al. 2011). If some members of the population are more or less likely to be included than others, a *sampling* bias is introduced in the estimation. To correct this bias, surveys often include weights to make the sample representative of the overall population.[23] It is particularly important to use the weights when provided by data producers, otherwise the estimated distribution maybe quite different from the population distribution.

To illustrate, recall that a sample $\{x_1, \cdots, x_n\}$ is based on individual observations x_i of individuals who agreed to respond. Let X denote the (true) variable of interest (individual income, or wealth), and D denote the response (to the survey) variable (1 if the individual responds). Let $\boldsymbol{Z}$ denote some possible covariates (age, gender, etc) and X the income. Since all computations should be conditional on $\boldsymbol{Z}$, we will skip it to avoid too heavy notations.[24]

Following Horvitz and Thompson (1952), assume that the variable of interest X has distribution F_θ. The Horvitz-Tompson estimator of θ should be based on weights $W = \mathbb{P}[D = 1|X]^{-1}$. Here our sample of x_i's are be seen as realizations of variables $X|D = 1$. Using Bayes formula, and the weights, it will be possible to link the true distribution of income, and the one of our sample. Hence,

$$\mathbb{P}[X|D = 1] = \frac{\mathbb{P}[X] \cdot \mathbb{P}[D = 1|X]}{\mathbb{P}[D = 1]} \propto \underbrace{\mathbb{P}[X]}_{=F_\theta} \cdot \underbrace{\mathbb{P}[D = 1|X]}_{=W^{-1}} \tag{29}$$

Consider here some sort of proportional hazard model for the weight function, with either

$$\mathbb{P}[D = 1|X = x] = S^x \tag{30}$$

where S is some baseline with value in $[0, 1]$, or

$$\mathbb{P}[D = 1|X = x] = x^{-a} \tag{31}$$

If X is Pareto distributed with tail index α, with density proportional to $x^{-(1+\alpha)}$, then, with model (30), density of $X|D = 1$ should be proportional

[23] To correct the sampling weights for unit nonresponse, see Korinek et al. (2006, 2007)

[24] For instance, $\mathbb{P}[\mathrm{X}|D = 1, \boldsymbol{Z}]$ could be written $\mathbb{P}_{\boldsymbol{Z}}[\mathrm{X}|D = 1]$, but instead of using $\mathbb{P}_{\boldsymbol{Z}}$ in all computations, we will simply use $\mathbb{P}$

to

$$x \mapsto x^{-(1+\alpha)} \cdot S^x \tag{32}$$

and the survey distribution has a Weibull distribution. With model (31), density of $X|D=1$ should be proportional to

$$x \mapsto x^{-(1+\alpha)} \cdot x^{-a} \tag{33}$$

which is another Pareto distribution.

In practical applications, even if the true distribution of X is Pareto I, if the response variable is (strongly) correlated with the income, either the sample is also Pareto I, with possibly another tail index, or there might be some second order effect. Thus, to prevent this possible bias, weights can be used. With a sample $\{(x_1, w_1), \cdots, (x_n, w_n)\}$, where w_i's are given by statistical institutes providing the datasets, our estimators should take them into account. With maximum likelihood estimation, we need to rewrite the likelihood function as the product of each individual contribution multiplied by its weight. For instance, the *weighted* version of the Hill estimator would be

$$\widetilde{\alpha}_{k,\boldsymbol{w}} = \left[\sum_{i=n-k+1}^{n} \frac{w_{(i)}}{\bar{w}_k} \log y_{(i)} - \log y_{(n-k+1)} \right]^{-1} \tag{34}$$

where $\bar{w}_k = \sum_{i=n-k+1}^{n} w_{(i)}$. The special case $w_i = c$, where c is a constant value, corresponds to the standard Hill estimator, as defined in (8). The estimation of GPD and EPD distributions with weights, as well as plotting Pareto diagram with weights, are not implemented in standard softwares. We develop functions in R to do this, which we make available on GitHub.[25]

6 Lorenz curve and top shares

In this section, we derive inequality measures based on distributions being Pareto in the upper tail (with finite mean, $\alpha > 1$).

The top $p100\%$ income share can be defined as follows:

$$\mathrm{TS}_p = \frac{p\,\mathbb{E}(X|X > Q(1-p))}{\mathbb{E}(X)} \tag{35}$$

[25] All R programs developed in this paper are available at the following website https://github.com/freakonometrics/TopIncomes

where $Q(t)$ is the quantile function, $Q(t) = \inf\{x \in \mathbb{R} : t \leq F(x)\}$. A Lorenz curve is defined by $(p, L(p))$, where $L(p) = 1 - \mathrm{TS}_{1-p}$ and $p \in (0, 1)$. A top income share is then related to a single value from the Lorenz curve.

From a sample of n incomes and weights, $\{(x_1, w_1), \cdots, (x_n, w_n)\}$, the sample top income share, based on the Empirical Distribution Function (EDF), is computed as:

$$\mathrm{TS}_p^{(\mathrm{edf})} = \frac{\sum_{i=1}^n w_i \, x_i \, \mathbb{1}(x_i > \widehat{Q}(1-p))}{\sum_{i=1}^n w_i \, x_i} \tag{36}$$

where $\mathbb{1}(\cdot)$ is an indicator function and $\widehat{Q}(1-p)$ is a sample weighted quantile.[26] Without weights, the sample top income share can be computed from (36), with $w_i = 1$.

Let us consider a distribution where the upper tail, above a threshold or cut-point $Q(1-q) = u$, is modelled by a Pareto distribution $\mathcal{P}$, and the remaining distribution is modelled by another distribution $\mathcal{F}$. For a two-component sliced distribution, with a Pareto distribution $\mathcal{P}$ for the top $q100\%$ and a distribution $\mathcal{F}$ for the bottom $(1-q)100\%$, we need to consider two cases. When the top $p100\%$ share is located in the upper tail modelled by the Pareto distribution, we have

$$\mathrm{TS}_{p \leq q} = \frac{p \, \mathbb{E}_{\mathcal{P}}(X | X > Q_{\mathcal{P}}(1 - p/q))}{(1-q)\mathbb{E}_{\mathcal{F}}(X) + q\mathbb{E}_{\mathcal{P}}(X)} \qquad \text{if } p \leq q \tag{37}$$

When the top $p100\%$ share is located below the upper tail modelled by the Pareto distribution, we have

$$\mathrm{TS}_{p \geq q} = 1 - \frac{(1-p) \, \mathbb{E}_{\mathcal{F}}(X | X < Q_{\mathcal{F}}(1-p))}{(1-q)\mathbb{E}_{\mathcal{F}}(X) + q\mathbb{E}_{\mathcal{P}}(X)} \qquad \text{if } p > q \tag{38}$$

If $\mathcal{F}$ is the empirical distribution function (EDF) and if the sample is given with weights, the threshold u is a weighted quantile.[27]

Pareto I and GPD models: we consider that the upper tail, above a threshold $Q(1-q) = u$, is modelled by a GPD distribution, and the remaining distribution is modelled by the empirical distribution function (EDF). The quantile function of a GPD is equal to

$$Q^{\mathcal{GPD}}(t) = \sigma(1-t)^{-1/\alpha} - \sigma + u \tag{39}$$

[26] With $\sum_{i=1}^n w_i = 1$, the weighted quantile $\widehat{Q}(1-p)$ is the ordered observation $x_{(k)}$ satisfying $\sum_{i=1}^{k-1} w_{(i)} \leq 1-p$ and $\sum_{i=k+1}^n w_{(i)} \leq p$.

[27] With $\sum_{i=1}^n w_i = 1$, the cut-point u is the ordered observation $x_{(l)}$ satisfying $\sum_{i=1}^{l-1} w_{(i)} \leq 1-q$ and $\sum_{i=l+1}^n w_{(i)} \leq q$.

Using (5) and (39) in (37) and (38), we obtain

$$\mathrm{TS}_{p,q}^{(\mathcal{GPD})} = \begin{cases} p\dfrac{[\alpha/(\alpha-1)]\sigma(p/q)^{-1/\alpha} + u - \sigma}{(1-q)\bar{x}_q + q\sigma/(\alpha-1) + qu} & \text{if } p \leq q \\ 1 - \dfrac{(1-p)\,\bar{x}_p}{(1-q)\bar{x}_q + q\sigma/(\alpha-1) + qu} & \text{if } p > q \end{cases} \tag{40}$$

where $\bar{x}_q$ ($\bar{x}_p$) is the weighted mean of the $(1-q)100\%$ ($(1-p)100\%$) smallest ordered observations.

Top shares from Pareto I model for top incomes are obtained from (40), with $\sigma = u$.

EPD model: we consider that the upper tail, above a threshold $Q(1-q) = u$, is modelled by an Extended Pareto Distribution (EPD), and the remaining distribution is modelled by the empirical distribution function (EDF). There are no closed form expressions for the quantile function and for the mean of EPD distribution. However, we can use numerical methods. The quantile function is defined as

$$Q^{\mathcal{EPD}}(t) = \{x \geq u : (x/u)[1 - \delta + \delta(x/u)^{\tau}] = (1-t)^{-1/\alpha}\} \tag{41}$$

The equation in parenthesis can be solved numerically, by findings root for the difference between the two terms, or by minimizing the square of the difference.[28]

From (22), (23), (37) and (38), top $p100\%$ shares from EPD model are defined as follows:[29]

$$\mathrm{TS}_{p,q}^{(\mathcal{EPD})} = \begin{cases} \dfrac{pu' + qE_{u'}}{(1-q)\bar{x}_q + q(u + E_u)} & \text{if } p \leq q \\ 1 - \dfrac{(1-p)\,\bar{x}_p}{(1-q)\bar{x}_q + q(u + E_u)} & \text{if } p > q \end{cases} \tag{42}$$

where E_u and $E_{u'}$ are obtained by numerical integration from (23), and u' is obtained from (41) with $t = 1 - p/q$.

7 Applications

In this section, we consider two applications: the income distribution in South-Africa in 2012 and, the wealth distribution in the United-States in 2013. A R package (`TopIncomes`) can be used to reproduced this analysis[30].

[28] The function `qepd` of the `ReIns` R package minimizes the square of the difference, with $u = 1$.

[29] Since the distribution is Pareto in the upper tail only, we have $\overline{F}(u') = p/q$.

[30] https://github.com/freakonometrics/TopIncomes.

7.1 Income distribution in South-Africa

The income distribution in South-Africa in 2012 is obtained from the Luxembourg Income Study (LIS) Cross-National Data Center in Luxembourg. We use similar incomes as those used to generate the *Key Figures*, except that we do not bottom- and top-code incomes.[31] Income is equal to disposable household income divided by the square root of the number of household members ($x = \texttt{DHI}/\sqrt{\texttt{NHHMEM}}$). All households where disposable income is missing or exactly equal to zero are excluded, and we use person-level adjusted weights ($w = \texttt{HWGT} * \texttt{NHHMEM}$). The number of observations is equal to $n = 7{,}990$.

Figure 8 shows values of the tail index estimated by (weighted) maximum likelihood, $\hat{\alpha}$, as k the number of the largest observations used for Pareto estimation increases. The vertical dashed lines marked q_{90}, q_{95} and q_{99} correspond to, respectively, the 90%, 95% and 99% weighted quantiles.[32] The Pareto I curve shows $\hat{\alpha}$ obtained from Pareto I model, also known as Hill plot. The two other curves show $\hat{\alpha}$ obtained from GPD and EPD models. We can see that the Pareto I curve decreases as k increases. It is never stable and horizontal, the Hill plot is then not very revealing. The GPD curve is unstable for $k < 200$, otherwise it increases as k increases. To the opposite, the EPD curve looks stable when more than 200 observations are used to fit the Pareto distribution, suggesting that the tail index is slightly greater than 2. These results are consistent with those obtained by simulations, in Figure 2 and 3.

[Figure 8 about here.]

Figure 9 shows values of the top 1% share, as k increases. The horizontal line is the value of the sample top 1% share, given in (36). We can see that the top 1% share obtained from Pareto I model increases as k increases. Since more inequality is expected when the distribution is heavier (α is smaller), these results are consistent with those obtained in Figure 8. To the opposite, the top 1% obtained from GPD and EPD models are much more stable. The values given by the EPD model are slightly higher than those given by the GPD model.

[Figure 9 about here.]

[31] see http://www.lisdatacenter.org/data-access/key-figures/methods/

[32] Note that the 200 largest observations corresponds to the top 5% incomes ($x \geq q_{95}$), but not to 5% of observations in the sample ($n = 7{,}990$) because of the weights.

The following Table shows values from Figures 8 and 9 for three thresholds, q_{90}, q_{95} and q_{99}, that is, when the Pareto distribution is fitted on, respectively, the top 10%, 5% and 1% observations.

	tail index			top 1%		
threshold	q_{90}	q_{95}	q_{99}	q_{90}	q_{95}	q_{99}
Pareto I	1.742	1.881	2.492	0.192	0.171	0.146
GPD	2.689	2.935	19.249	0.142	0.141	0.139
EPD	2.236	2.255	4.198	0.148	0.149	0.139

We can see that the GPD and EPD models provides very stable values for the top 1% shares, while the Pareto I is quite sensitive to the choice of the threshold. The tail index values for q_{99} may appear odd. They are not reliable, since they use $k = 26$ observations only. Finally, this Table is not as informative as Figure 8: it does not capture decreases of the tail index from GPD model, as the threshold increases, from q_{90} to q_{94}.[33]

Figure 10 shows a Pareto diagram, with the Pareto I, GPD and EPD distributions fitted on the top 10% observations. We can see that the straight lines of the Pareto I model decreases more slowly than the curves of the GPD and EPD models, above the top 1% income (when $\log(x) \geq q_{99}$). The EPD model captures deviations from a strict Pareto distribution, it seems to better fit the Pareto diagram. The extreme right part exhibits an erratic behavior, but we should not pay attention to this part (see section 5.2).

[Figure 10 about here.]

Overall, we can see that a threshold too low leads to under-estimate the tail index and to over-estimate the top share, with a Pareto I model. It illustrates the misspecification bias, as explained in section 5.1, which is not expected to disappear as the sample size increases. Figures (8), (9) and (10) suggest that the EPD model provides more reliable results.[34] The top 1% is equal to 14.8%, with an EPD distribution fitted on the top 10% observations (q_{90}), which is slightly more than the sample top 1% share equals to 13.3%.

7.2 Wealth distribution in the United-States

The wealth distribution in the United-States in 2013 is obtained from the Luxembourg Wealth Study (LWS) database. We use disposable household

[33] That is, the GPD curve that increases as k increases, for $k > 200$, in Figure 8.

[34] Figure 8 suggests that similar results could be obtained with Pareto I and GPD models, with a higher threshold. However, the choice of the threshold is difficult based on Pareto I and GPD models only, since the curves are never stable and horizontal.

net worth divided by the square root of the number of household members ($x = \texttt{DNW}/\sqrt{\texttt{NHHMEM}}$). All households where disposable net worth is missing are excluded. We do not bottom- and top-code wealth, and we use person-level adjusted weights ($w = \texttt{HWGT} * \texttt{NHHMEM}$). The number of observations is equal to $n = 6{,}015$.[35]

Figure 11 shows values of the tail index estimated by (weighted) maximum likelihood, $\hat{\alpha}$, as k the number of the largest observations used for Pareto estimation increases. The vertical dashed lines marked q_{90}, q_{95} and q_{99} correspond to, respectively, the 90%, 95% and 99% weighted quantiles. The Pareto I curve shows $\hat{\alpha}$ obtained from Pareto I model, also known as Hill plot. The two other curves show $\hat{\alpha}$ obtained from GPD and EPD models. We can see that the Pareto I curve decreases as k increases. It becomes quite stable when the threshold is below q_{99} (above the top 1% observations), with a tail index slightly greater than 1.5. The EPD curve also decreases as k increases. It is not very stable when the threshold is between q_{95} and q_{99}, but it isn't still decreasing, suggesting a tail index slightly greater than 1.5. The GPD curve is not very different from the EPD curve, but it is slightly less stable.

[Figure 11 about here.]

Figure 12 shows values of the top 1% share, as k increases. We can see that the top 1% shares obtained from Pareto I model with a threshold greater than q_{99}, and from GPD and EPD models with a threshold greater than q_{95}, are very stable and very similar to the sample top 1% share (given by the horizontal line). Otherwise, the top 1% share increases as k increases, faster with Pareto I model.

[Figure 12 about here.]

The following Table shows values from Figures 11 and 12 for three thresholds, q_{90}, q_{95} and q_{99}, that is, when the Pareto distribution is fitted on, respectively, the top 10%, 5% and 1% observations.

[35] The data provider conducted a multiple imputation procedure to impute missing values. The number of observations in the original file, equals to 30,075, is five times the actual number of households, because the imputations are stored as five successive implicates of each record (see the LWS database user guide). In our sample, we use the first imputation of each record.

	tail index			top 1%		
threshold	q_{90}	q_{95}	q_{99}	q_{90}	q_{95}	q_{99}
Pareto I	0.931	1.088	1.637	-	0.767	0.421
GPD	1.265	1.540	1.490	0.571	0.455	0.437
EPD	1.317	1.486	1.517	0.540	0.466	0.436

We can see that the Pareto I model is very sensitive to the choice of the threshold, compared to the others. For instance, the Pareto I model with q_{95} and q_{99} provides very different estimates of the top 1% share (76.7% and 42.1%), while EPD model gives quite similar values (46.6% and 43.6%). Moreover, Pareto I model with q_{90} gives inconsistent results: $\hat{\alpha} < 1$, meaning that the mean doesn't exist and, thus, the top share is undefined. If we assume that the top 1% of the population has less than 100% of income or wealth, the tail index has to be greater than one, $\alpha > 1$.

Figure 13 shows a Pareto diagram, with the Pareto I, GPD and EPD models fitted on the top 3% observations (q_{97}). We can see that the GPD and EPD models provide a much better fit of the Pareto diagram, compared to the Pareto I model.

[Figure 13 about here.]

Overall, we can see that the EPD model is much less sensitive to the choice of the threshold.

Finally, it is important to notice that the measurement of inequality of wealth is somewhat more challenging than the analysis of income or consumption, because most sample data include a substantial fraction of negative net worth.[36] It makes the interpretation of standard measures of inequality not so easy, see Cowell and Van Kerm (2015). For instance, with negative observations, the top share is not defined over the interval $[0, 1]$, but over $]-\infty, +\infty[$.[37] It can take any negative or positive value, even greater than one.

8 Conclusion

Since income and wealth distributions are often assumed to be heavy-tailed, economists have mainly used the Pareto Type I distribution, or Pareto I model, to estimate the upper tail in empirical studies.

[36] In our sample, 747 observations are negative (12.4%)

[37] From (35), top share can take large positive (negative) values when the overall mean is close to zero and positive (negative). It is equal to infinity when the overall mean is 0.

In this paper, we first show that the Pareto I model (Hill estimator) is very sensitive to the choice of the threshold. When the threshold is too low, it can lead to severe under-estimation of the tail index and, thus, to severe over-estimation of inequality.

Then, we provide evidence that the GPD model, based on the Generalized Pareto Distribution, is less sensitive to the choice of the threshold. We show that Pareto I model is as good as the GPD model but only for higher threshold, much higher as the scale parameter differs from the threshold ($\sigma \neq u$). However, the estimation of the GPD distribution turns out to be quite inaccurate in our simulation results.

Next, we introduce the EPD model, based on the Extended Pareto Distribution proposed by Beirlant et al. (2009), which allows to capture deviations from a strict Pareto distribution. Our simulation results and two applications show that the EPD model is much less sensitive to the choice of the threshold, and its estimation is quite accurate.

Finally, we discuss the fact that the tail index estimation is sensitive to several biases, which can go in opposite directions. To summarize, the tail index estimation of a Pareto model is mostly: (1) downward biased, when the threshold is not high enough and/or in the presence of outliers ; (2) upward biased with topcoding, censored data and when the rich underreport their income. It leads us to provide new highlights on several common beliefs.

It is widely believed that the tail index estimation of Pareto I model based on surveys is biased upward, because these data are subject to topcoding, censoring and underreporting of the rich. In a seminal paper, Atkinson et al. (2011, footnote 8, p.11) write that: "The Pareto parameter is estimated using the ratio of the top 5 percent income share to the top decile income share (...). Because those top income shares are often based on survey data (and not tax data), they likely underestimate the magnitude of the changes at the very top."[38] It is true, under the (strong) assumption that the distribution is strictly Pareto I above the top decile, and if there is no outlier in the sample. Otherwise, the tail index can be biased downward, and the upper tail might just as easily be over-estimated.

It is also widely believed that the tail index estimation is much more reliable with tax data, because there is no topcoding or censoring and, above all, these data are much less sensitive to misreporting. Indeed, tax data are much less sensitive, if not at all, to estimation bias. However, they are still sensitive to misspecification bias. Thus, a threshold too low may lead to substantial downward bias of the tail index and to severe over-estimation of

[38] For more details on the Pareto parameter calculation, see Atkinson (2007, p.24).

inequality, even with millions of observations. From tax data in the U.K., for several years 1995-2010, Jenkins (2017, p.279) finds that the optimal thresholds for Pareto I model are at around the 99.5-percentile or higher, that is, well above the thresholds commonly used. This suggests that fitting Pareto models to tax data should be done with more caution than has been done so far.

References

Albrecher, H., J. Beirlant, and J. L. Teugels (2017). *Reinsurance: Actuarial and Statistical Aspects*. Wiley Series in Probability and Statistics.

Arnold, B. C. (2015). *Pareto Distributions*. 2nd edition, CRC Press, Taylor and Francis Group.

Atkinson, A. B. (2007). Measuring top incomes: Methodological issues. In A. B. Atkinson and T. Piketty (Eds.), *Top Incomes over the 20th Century: A contrast between continental European and English-speaking countries*, Chapter 2, pp. 18–42. Oxford University Press.

Atkinson, A. B. (2017). Pareto and the upper tail of the income distribution in the UK: 1799 to the present. *Economica 84*, 129–156.

Atkinson, A. B., T. Piketty, and E. Saez (2011). Top incomes in the long run of history. *Journal of Economic Literature 49*(1), 3–71.

Balkema, A. and L. de Haan (1974). Residual life time at great age. *Annals of Probability 2*, 792–804.

Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics.

Beirlant, J., E. Joossens, and J. Segers (2009). Second-order refined peaks-over-threshold modelling for heavy-tailed distributions. *Journal of Statistical Planning and Inference 139*, 2800–2815.

Benhabib, J. and A. Bisin (2018). Skewed wealth distributions: Theory and empirics. *Journal of Economic Literature 56*, 1261–1291.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (2013). *Regular Variation*. Cambridge University Press.

Castillo, E. and A. Hadi (1997). Fitting the Generalized Pareto Distribution to data. *Journal of the American Statistical Association 92*, 1609–1620.

Cirillo, P. (2013). Are your data really Pareto distributed? *Physica A 392*, 5947–5962.

Cowell, F. A. (2011). *Measuring Inequality* (Third ed.). Oxford: Oxford University Press.

Cowell, F. A. (2013). UK wealth inequality in international context. In J. R. Hills (Ed.), *Wealth in the UK*, Chapter 3. Oxford: Oxford University Press.

Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics 141*, 1044–1072.

Cowell, F. A. and E. Flachaire (2015). Statistical methods for distributional analysis. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Volume 2. New York: Elsevier.

Cowell, F. A. and P. Van Kerm (2015). Wealth inequality: A survey. *Journal of Economic Surveys 29*, 671–710.

Davidson, R. and E. Flachaire (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics 141*, 141 – 166.

de Haan, L. and A. Ferreira (2006). *Extreme Value Theory: An introduction*. Springer Series in Operations Research and Financial Engineering.

de Haan, L. and U. Stadtmüller (1996). Generalized regular variation of second order. *Journal of the Australian Mathematical Society 61*, 381–395.

del Castillo, J. and J. Daoudi (2009). Estimation of the generalized Pareto distribution. *Statistics and Probability Letters 79*, 684–688.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modelling Extremal Events*. Applications of Mathematics: Stochastic Modelling and Applied Probability. Berlin, Heidelberg: Springer-Verlag.

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Proceedings of the Cambridge Philosophical Society*, Volume 24, pp. 180–290.

Ghosh, S. and S. Resnick (2010). A discussion on mean excess plots. *Stochastic Processes and their Applications 120*(8), 1492 – 1517.

Guess, F. and F. Proschan (1988). 12 mean residual life: Theory and applications. In *Quality Control and Reliability*, Volume 7 of *Handbook of Statistics*, pp. 215 – 224. Elsevier.

Hall, P. (1982). On some simple estimate of an exponent of regular variation. *Journal of the Royal Statistical Society: Series B 44*, 37–42.

Higgins, S., N. Lustig, and A. Vigorito (2018). The rich underreport their income: Assessing bias in inequality estimates and correction methods using linked survey and tax data. Ecineq WP 475.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Hosking, J. R. M. and J. R. Wallis (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics 29*, 339–3349.

Hubert, M., G. Dierckx, and D. Vanpaemel (2013). Detecting influential data points for the Hill estimator in Pareto-type distributions. *Computational Statistics & Data Analysis 65*, 13–28.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica 84*, 261–289.

Jones, C. I. (2015). Pareto and Piketty: The macroeconomics of top income and wealth inequality. *Journal of Economic Perspectives 29*, 29–46.

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

Korinek, A., J. A. Mistiaen, and M. Ravallion (2006). Survey nonresponse and the distribution of income. *Journal of Economic Inequality 4*, 33–55.

Korinek, A., J. A. Mistiaen, and M. Ravallion (2007). An econometric method of correcting for unit nonresponse bias in surveys. *Journal of Econometrics 136*, 213235.

Lomax, K. S. (1954). Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association 49*(268), 847–852.

Papastathopoulos, I. and J. A. Tawn (2013). Extended generalised pareto models for tail estimation. *Journal of Statistical Planning and Inference 143*, 131–142.

Pareto, V. (1895). La legge della domanda. *Giornale degli Economisti* (10), 59–68.

Pareto, V. (1896). La courbe de la répartition de la richesse. In C. Viret-Genton (Ed.), *Recueil publié par la Faculté de Droit à l'occasion de l'exposition nationale suisse, Geneva 1896*, pp. 373–387. Lausanne: Université de Lausanne.

Pearson, R. K. (2005). *Mining Imperfect Data: Dealing with Contamination and Incomplete Records*. Society for Industrial and Applied Mathematics.

Peng, L. and Y. Qi (2004). Estimating the first- and second-order parameters of a heavy-tailed distribution. *Australian & New Zealand Journal of Statistics 46*, 305–312.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics 23*, 119–131.

Piketty, T. (2007). Top incomes over the twentieth century: A summary of the main findings. In A. B. Atkinson and T. Piketty (Eds.), *Top Incomes Over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries*. Oxford University Press.

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering.

Rytgaard, M. (1990). Estimation in the Pareto distribution. *ASTIN Bulletin 20*, 201–216.

Schluter, C. (2018). Top incomes, heavy tails, and rank-size regressions. *Econometrics - MDPI 6*(1), 1–16.

Schluter, C. (2019). On Zipf's law and the bias of Zipf regression.

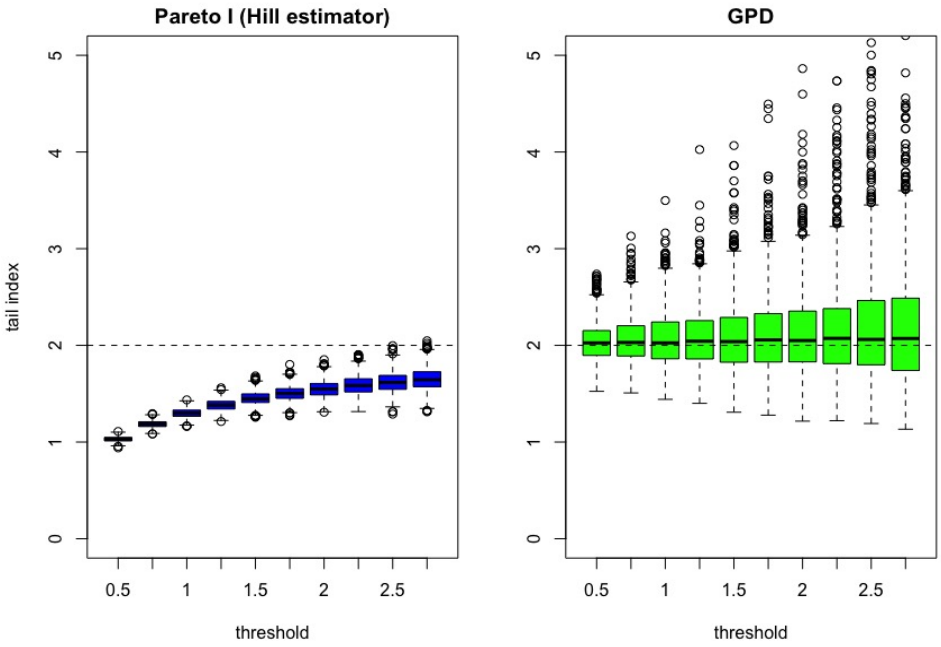

Figure 1: Boxplots of maximum likelihood estimators of the tail index from Pareto I (left) and GPD (right) models, as the threshold increases: 1,000 samples of 1,000 observations drawn from a GPD distribution, $\mathcal{GPD}(0.5, 1.5, 2)$.

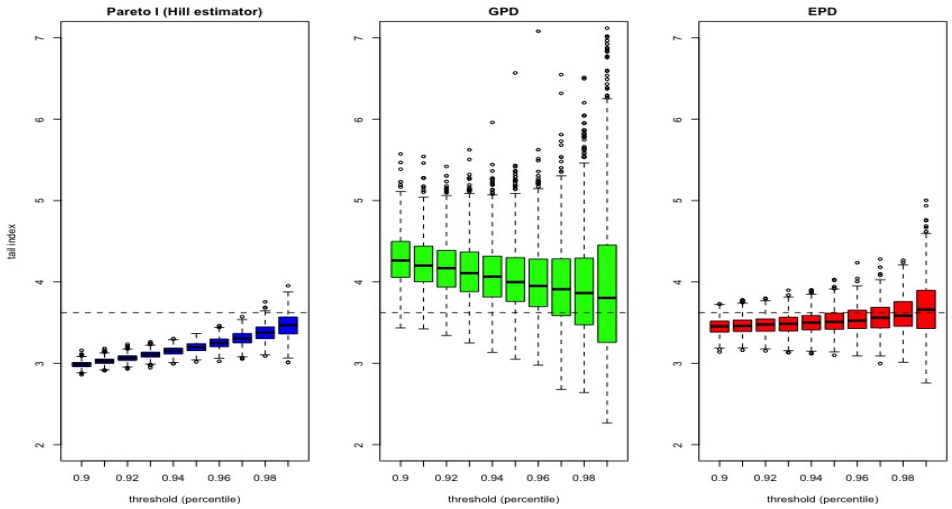

Figure 2: Boxplots of maximum likelihood estimators of the tail index: 1,000 samples of 50,000 observations drawn from a Singh-Maddala distribution, $SM(2.07, 1.14, 1.75)$. From the left to the right, the x-axis is the threshold (percentile) used to fit Pareto I (blue), GPD (green) and EPD (red) models.

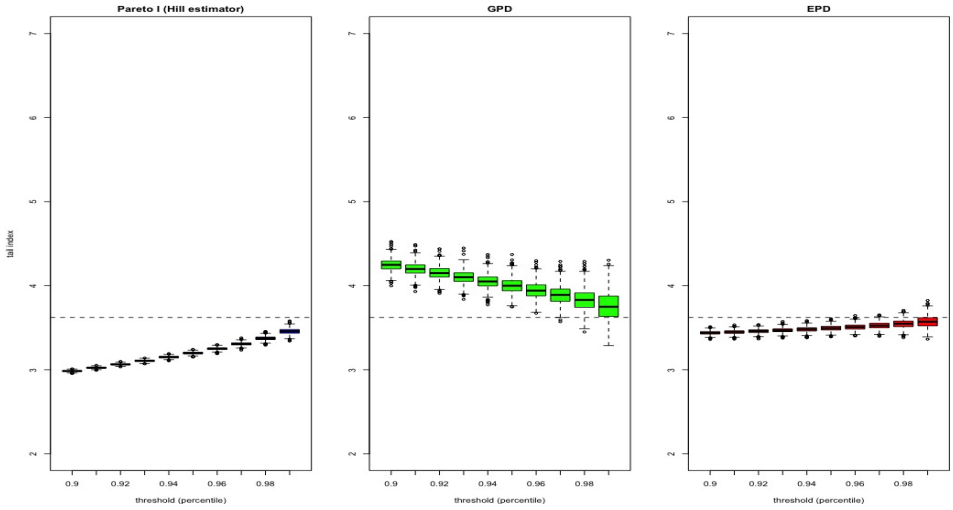

Figure 3: Boxplots of maximum likelihood estimators of the tail index: 1,000 samples of 1,000,000 observations drawn from a Singh-Maddala distribution, $SM(2.07, 1.14, 1.75)$. From the left to the right, the x-axis is the threshold (percentile) used to fit Pareto I (blue), GPD (green) and EPD (red) models.

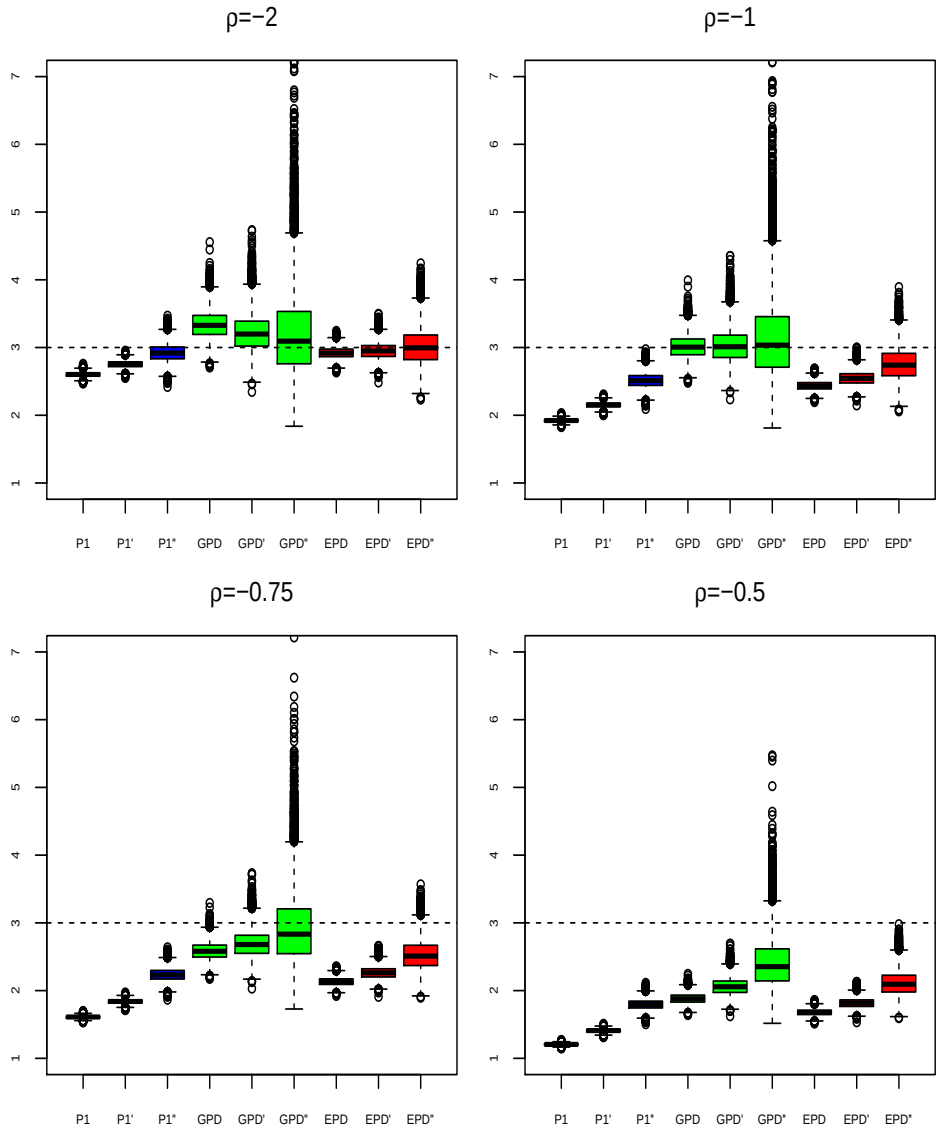

Figure 4: **Finite variance** ($\alpha = 3$) - Boxplots of **tail index** estimates $\hat{\alpha}$ based on 10,000 samples of 50,000 observations drawn from Singh-Maddala distributions with a upper tail that deviates more and more from a strict-Pareto (as ρ increases). Pareto I (blue), GPD (green) and EPD (red) models, fitted on, respectively, the top 10%, 5% and 1% of the sample.

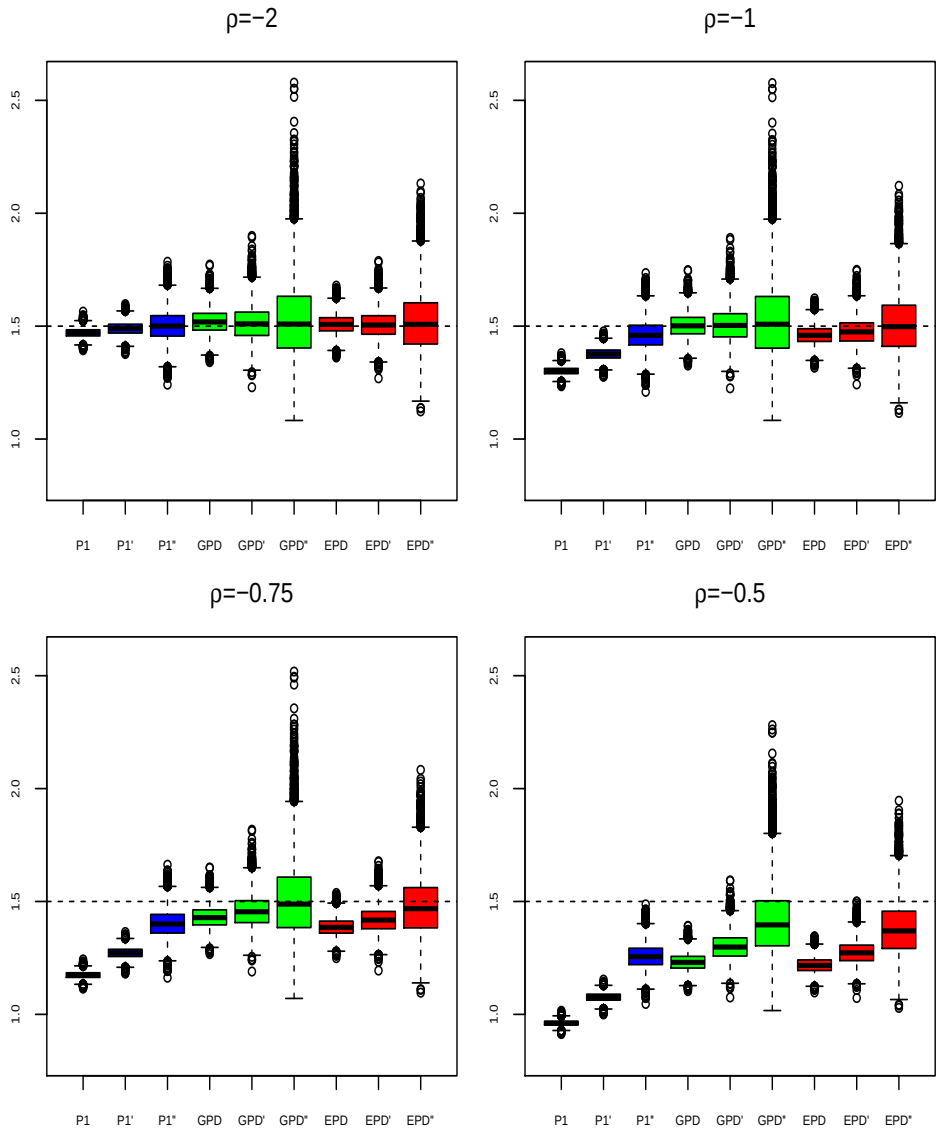

Figure 5: **Infinite variance** ($\alpha = 1.5$) - Boxplots of **tail index** estimates $\hat{\alpha}$ based on 10,000 samples of 50,000 observations drawn from Singh-Maddala distributions with a upper tail that deviates more and more from a strict-Pareto (as ρ increases). Pareto I model (blue), GPD model (green) and EPD (red), fitted on, respectively, the top 10%, 5% and 1% of the sample.

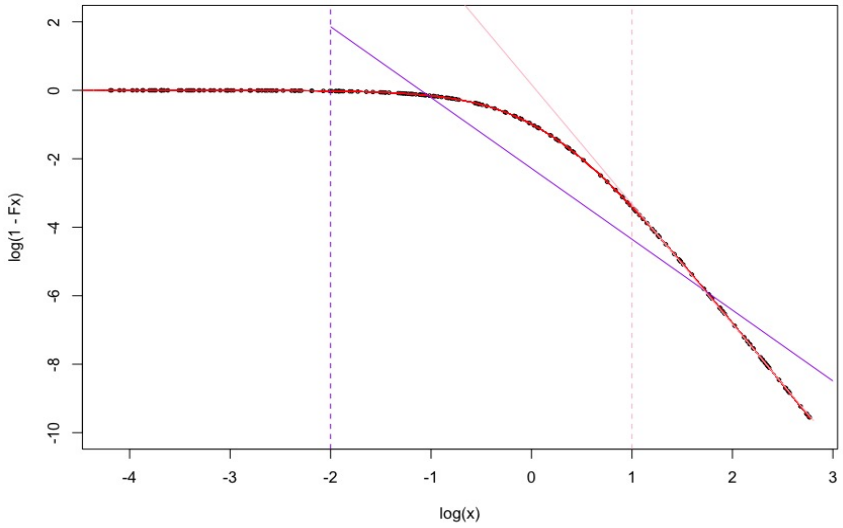

Figure 6: Pareto diagram based on the true CDF, with two linear approximations based on $\log x \geq -2$ and $\log x \geq 1$

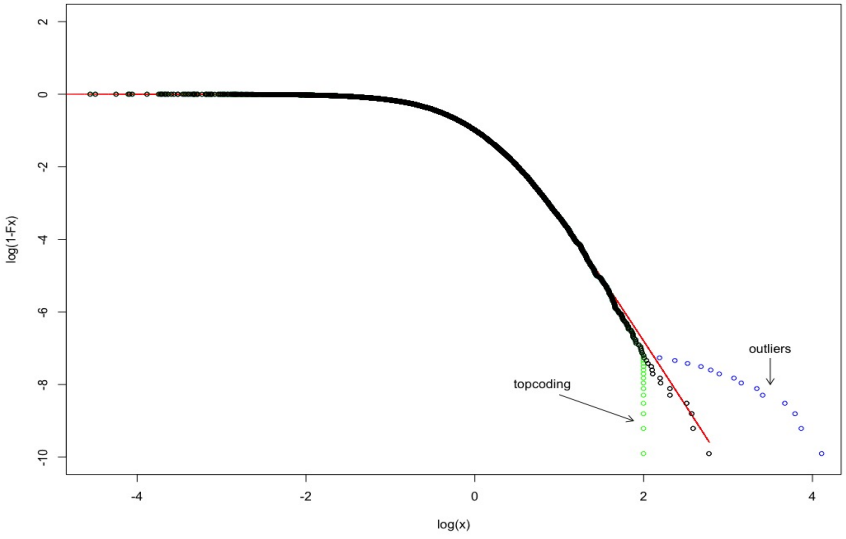

Figure 7: Pareto diagram based on a sample (black), with an artificial increase of the largest observations (blue) and with topcoding (green).

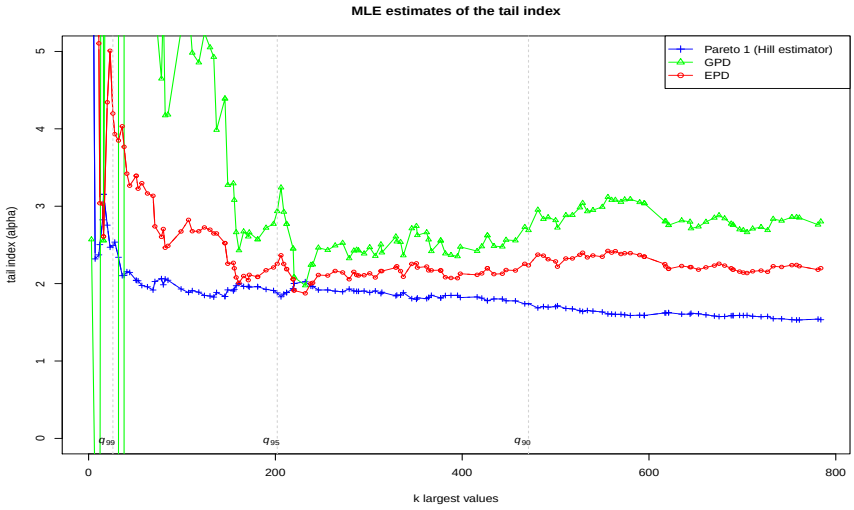

Figure 8: Incomes in South-Africa 2012: MLE estimates of the tail index, $\hat{\alpha}$, from Pareto I (blue), GPD (green) and EPD (red) models, as the number of the k-largest observations used for Pareto estimation increases.

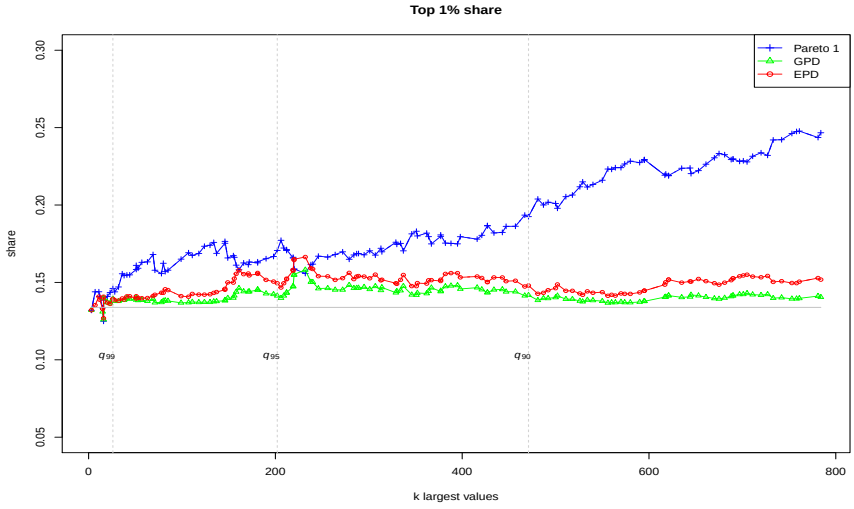

Figure 9: Incomes in South-Africa 2012: Top 1% income shares obtained from Pareto I (blue), GPD (green) and EPD (red) models, as the number of the k-largest observations used for Pareto estimation increases.

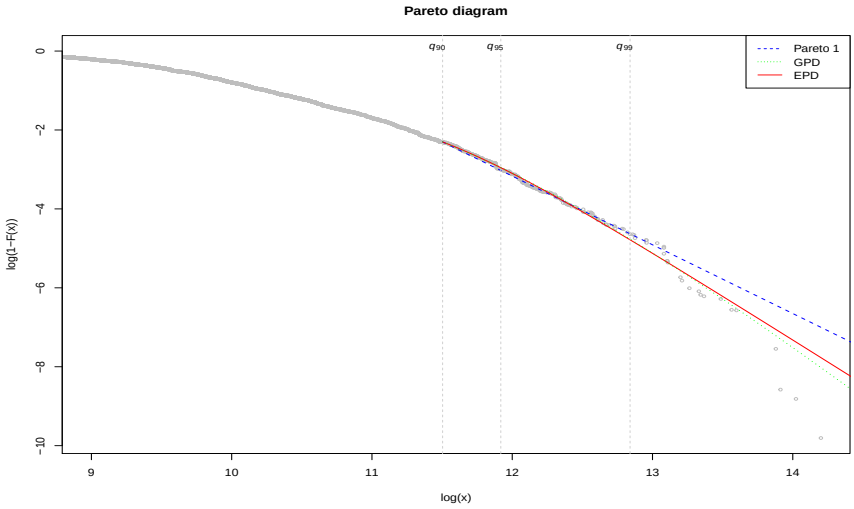

Figure 10: Incomes in South-Africa 2012: Pareto diagram, with Pareto I (blue), GPD (green) and EPD (red) models fitted on the top 10% incomes.

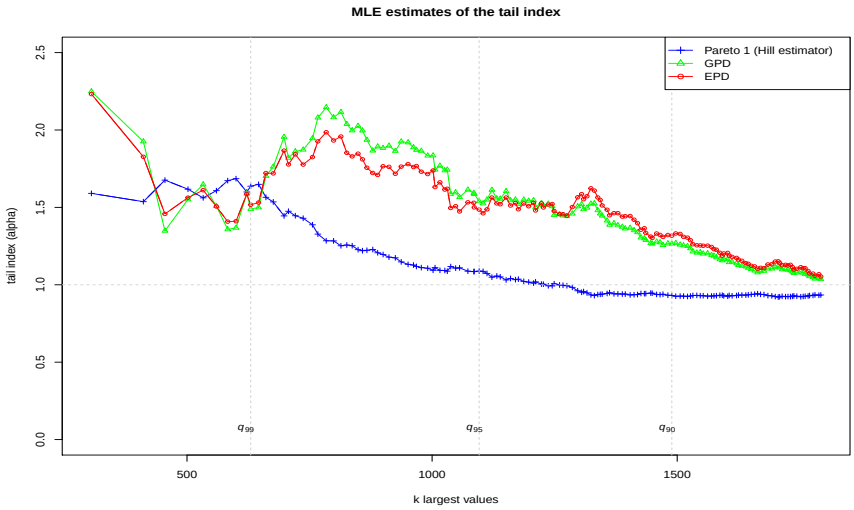

Figure 11: Wealth in the United States 2013: MLE estimates of the tail index, $\hat{\alpha}$, from Pareto I (blue), GPD (green) and EPD (red) models, as the number of the k-largest observations used for Pareto estimation increases.

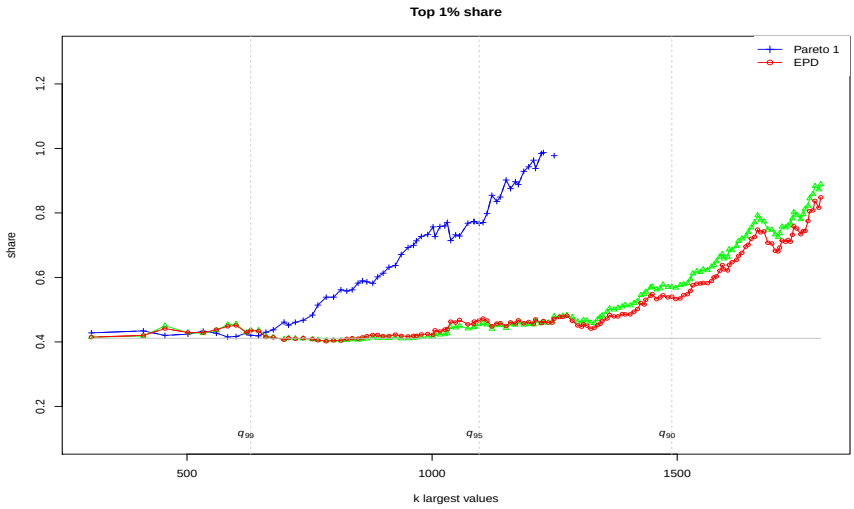

Figure 12: Wealth in the United States 2013: Pareto diagram, with Pareto I (blue) and EPD (red) models fitted on the top 5% incomes.

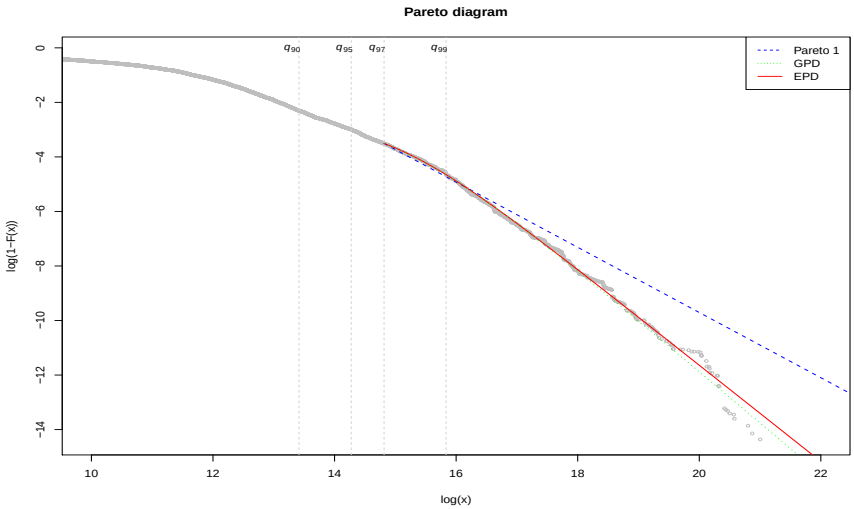

Figure 13: Wealth in the United States 2013: Pareto diagram, with Pareto I (blue) and EPD (red) models fitted on the top 5% incomes.