Bayesian inference for income inequality using a Pareto II tail with an uncertain threshold: Combining EU-SILC and WID data

Mathias Silva, Michel Lubrano

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Bayesian inference for income inequality using a Pareto II tail with an uncertain threshold:Combining EU-SILC and WID data

Mathias Silva
Michel Lubrano

WP 2024 - Nr 29

Bayesian inference for income inequality using a Pareto II tail with an uncertain threshold: Combining EU-SILC and WID data*

Mathias Silva† and Michel Lubrano‡

October 2024

Abstract

When estimated from survey data alone, the distribution of high incomes in a population may be misrepresented, as surveys typically provide detailed coverage of the lower part of the income distribution, but offer limited information on top incomes. Tax data, in contrast, better capture top incomes, but lack contextual information. To combine these data sources, Pareto models are often used to represent the upper tail of the income distribution. In this paper, we propose a Bayesian approach for this purpose, building on extreme value theory. Our method integrates a Pareto II tail with a semi-parametric model for the central part of the income distribution, and it selects the income threshold separating them endogenously. We incorporate external tax data through an informative prior on the Pareto II coefficient to complement survey micro-data. We find that Bayesian inference can yield a wide range of threshold estimates, which are sensitive to how the central part of the distribution is modelled. Applying our methodology to the EU-SILC micro-data set for 2008 and 2018, we find that using tax-data information from WID introduces no changes to inequality estimates for Nordic countries or The Netherlands, which rely on administrative registers for income data. However, tax data significantly revise survey-based inequality estimates in new EU member states.

Keywords: Top income correction, Pareto II, Bayesian inference, extreme value theory, EU-SILC
JEL Classification: C11, D31, D63, I31

* **Acknowledgements** We acknowledge financial support from the French government under the "France 2030" investment plan managed by the French National Research Agency Grant ANR-17-EURE-0020, and by the Excellence Initiative of Aix-Marseille University - A*MIDEX.
We have benefited from useful comments and remarks by Emmanuel Flachaire, Philippe van Kerm, Sylvia Kaufman and Duangkamon Chotikapanich made on a previous version of this paper. Remaining errors are solely ours. This paper was presented at the 2023 ESOBE conference in Glasgow, September 1-2, 2023.

† CERGIC, ENS Lyon, France and Aix Marseille Univ, CNRS, AMSE, Marseille, France. Email: mathias.silva_vazquez@ens-lyon.fr

‡ Aix Marseille Univ, CNRS, AMSE, Marseille, France. Orcid: 0000-0003-0448-0307. Email:michel.lubrano@univ-amu.fr, corresponding author.

1 Introduction

Tracking income distribution dynamics and inequality requires accurate insights into top incomes. This arises from the right-skewed nature of income distributions and the strong influence of high incomes on communicable conventional inequality measures like the Gini index or the top 1% income share.

For this purpose, Pareto distributions are convenient for representing the right tail of an income distribution above a given income threshold. Although their utility dates back to their introduction in Pareto (1896), they are nowadays vastly exploited for imputing observations, correcting data issues like high-income measurement or coverage errors (e.g., Bourguignon 2018, Bartels and Metzing 2019, Blanchet et al. 2022a), and for inter- or extrapolating empirical top quantiles when grouped data are used (Blanchet et al., 2022b).

The key issue in using Pareto models for high incomes (as we do in this paper), is to determine when and how additional available data sources on high incomes can provide most appropriate results for fitting the Pareto tail. Survey data is an essential source of information to study a population's incomes, along with detailed information on other contextual variables. However, survey data are evidenced to misrepresent high incomes, mainly due to under-sampling due to unreliable sampling schemes (e.g., not enough rich households are included in the survey's sampling design), or to non-sampling issues like high-income households being less likely to report information on their income when surveyed or more likely to under-declare their income in their responses (see e.g. Lustig 2020).

Tax data, though typically available only in tabulated form, provide information on top incomes more reliably due to tax authority scrutiny, although this relative improvement compared to survey data may also be nuanced by tax avoidance and tax evasion phenomena. While tax records are more accurate sources for information on high incomes, they lack coverage of low-income populations as these population groups are generally not concerned by tax filing regulations.[1] Their utility to study inequalities within entire populations using Pareto tails has therefore been approached through efforts to exploit them simultaneously to survey data.

This paper develops a new approach suited for the study of income distributions and inequality under Pareto tails, introducing information from tax data on high incomes to distributional estimates computed from household survey microdata. In doing so, it offers contributions to several strains of the recent literature on the topic.

Firstly, this paper contributes to the literature on empirical methods for reconciling information from survey microdata with Pareto tails from external data on high incomes for inequality analysis. A key obstacle is that conventional parametric distributions integrating Pareto tails, such as the Singh-Maddala or more general Generalized Beta distribution, lack an explicit setting for the threshold delimiting their asymptotic Pareto tail. This prevents these tails from being directly substituted or adjusted using tax data on high incomes. Recent approaches, such as those developed in Jenkins (2017) and Blanchet et al. (2022a), offer solutions to incorporate Pareto tails from tax data on incomes above a fixed threshold value to distributional estimates from other data sources for the rest of the distribution

[1] This question is illustrated in Atkinson (2005) who makes use of the super-tax data set for the UK over the 20th century to analyse top 1% income shares. A crucial challenge for this purpose is the computation of the income shares of low-income population groups excluded from the super-tax data.

below it. A challenge that has risen in consequence is that of devising the income threshold above which the Pareto distribution estimated on tax data is a better representation of the population's top incomes than that offered from survey data.

Specifying the threshold delimiting a Pareto distribution of high incomes is an exercise where several sources of uncertainty intersect. One source of uncertainty lies in the discrepancies in the populations and income concepts each source of data is concerned with, limiting the direct conversion of estimated quantities from tax data to survey data. Another important source of uncertainty is the sampling variability introduced by the sampling scheme yielding the survey sample, or by the possibility of undetected tax filing omissions in the available data. In light of the several uncertainties surrounding possible values for the threshold, a Bayesian treatment of the issue seems a feasible alternative. This paper offers a first proposal in this direction by treating this income threshold as an unknown parameter within a parametric income distribution model with a Pareto tail and by devising a Bayesian inference procedure to estimate posterior probability distributions for all model parameters. The proposed Bayesian strategy builds on the recent literature modelling extreme events using Pareto tails with an uncertain threshold (e.g., Cooray and Ananda 2005, Scarrott and MacDonald 2012, Majid and Ibrahim 2021b).

Secondly, this paper further contributes to the literature using Pareto distributions to correct survey microdata estimates for high-income under-reporting or non-response problems. An issue limiting this use of Pareto tails is that resulting inequality estimates are only often reasonable when the income threshold delimiting the tail is set to high values within the top decile of the distribution. This setting results restrictively too high, considering that significant high-income under-reporting issues are often evidenced to take place well below this segment of the income distribution (e.g., see matched survey and tax samples' analyses of Angel et al. 2019 for Austria and Flachaire et al. 2022 for Uruguay). Treating the income threshold as uncertain, our proposed approach can exploit external tax-data estimates if available by putting informative priors on the parameters ruling the Pareto tail.

We apply our methodology to the study of income distributions for most states covered by the European Union's Statistics on Income and Living Conditions (EU-SILC) household survey microdata in 2008 and 2018. To allow the possibility of revising high-income estimates through external tax-data estimates of top incomes available from the World Inequality Database (WID) we build informative prior distributions on the parameter ruling right-tail dispersion. This approach can introduce external information for inference if upper-tail inequality as estimable from EU-SILC microdata alone is significantly lower than that implied in the WID top income estimates. We find that using tax-data estimates from WID introduces no changes to inequality estimates for Nordic countries or The Netherlands, which rely on administrative registers for income data variables in the EU-SILC. We also find, however, that setting priors to be consistent with WID data significantly revises survey-based inequality estimates in new EU member states.

Finally, this paper also contributes to the recent literature seeking to extend the traditional Pareto I distribution to more elaborate Pareto distributions, allowing for more flexible high-income representations less sensitive to the issue of selecting the income threshold delimiting them. The original Pareto distribution is of particular popularity due to the simplicity of estimating its single parameter independent of data format when an appropriate income threshold delimiting the upper tail is known and because it serves as a practical linear interpolator between empirical top

quantiles. Two crucial restrictions of this Pareto I model are that it leads to biased estimates of income inequality if the threshold income is fixed to an inappropriately low value and that it assumes a common and constant level of inequality within any high-income population groups.

To address these limitations, more flexible models such as the Generalized Pareto Distribution (GPD) of Pickands (1975) have been proposed, with the Pareto II being a particularly useful variant for income distributions.[2] This latter distribution is obtained through the addition of one parameter to the conventional Pareto I model, effectively containing it as a particular case, and typically yield high-income estimates that are less sensitive to the choice of the income threshold, though possibly at a loss of precision (e.g., see Jenkins 2017 or Charpentier and Flachaire 2022). We illustrate in our applications how posterior distribution estimates obtained under our proposed approach can be used for probabilistically assessing the equivalence of inequality estimates obtained under a Pareto II distribution with those obtainable under a simpler Pareto I distribution, the former being evidenced as more appropriate in almost all our analyses.

The paper is organized as follows. In section 2, we present the EU-SILC data for 15 initial member states (EUR15) and 8 new member states (NMS). We analyse the limitation of this data set in terms of missing high-incomes information when it is confronted to the WID data set. We compare in a Bayesian framework the Pareto I tails of each data set. However, the Pareto model has limitations, compared to the Pareto II model. In section 3, we show how the compound model of the extreme value theory can be a basis for estimating the threshold. We detail Bayesian inference for the compound Pareto II model with an informative prior on the Pareto coefficient in section 4. We also detail in this section how to decompose the Gini coefficient in a Bayesian framework. In section 5, we apply the proposed method to analyse EU income distributions from EU-SILC microdata, exploiting external tax-data estimates available from WID. Finally, section 6 offers concluding discussions and several venues for future research stemming from these.

2 The EU-SILC data set and its missing rich

The European Community Statistics on Income and Living Conditions (EU-SILC) aims at collecting comparable data on income, poverty and living conditions at the European level. Income data can have alternative sources depending on the way they are collected. The first source comes from surveys, which means that the respondent provides directly her income. The second source comes from administrative data, covering various sources such as social security or fiscal declarations, which are supposed to be of a better quality, not suffering from under-reporting. It means that when a respondent is surveyed, her income is taken from the administrative source, under the condition that this respondent accepts to be surveyed. So these data can nevertheless suffer from under-sampling, similarly to usual survey data, but less from under-reporting. The source can be also mixed, which means that the

[2]The Pareto family also includes the Pareto III and Pareto IV which are all particular cases of the Feller-Pareto distribution detailed in Arnold (2008). Note also that other models corresponding to a Pareto tail have been proposed in the literature, such as the extended Pareto distribution of Beirlant et al. (2009) used in Charpentier and Flachaire (2022) or the Pareto-Log-normal or double Pareto-Log-normal distributions of Reed and Jorgensen (2004) used in Hajargasht and Griffiths (2013).

source, survey or administrative, depends on the year of collection, most of the time without further precision. The variable `HX090` corresponds to household income, normalized by the OECD equivalence scale.

The main source for external tax data is provided by the World Inequality Data base (WID). However, this information (the variable `fiinc`) is not available for every country and for every year. Nevertheless, this database provides valuable information when using an alternative income variable (`scainc` or its variant `sdiinc`) that we shall use instead to construct our prior information. It represents post tax income. This variable is available in the form of grouped data when the EU-SILC data are individual data. These data were the object of several corrections for data issues and in particular for top incomes as detailed in Alvaredo et al. (2016).

2.1 The quantity of information provided by WID data

Assuming a Pareto I for top 5% incomes, we analyse, in Table 1 for two years (2008 and 2018) how the extra source of information provided by WID data, manages to modify the estimation of the Pareto coefficient α when using EU-SILC data. For doing this, we first make inference on α, using a non-informative prior. This is a textbook exercise (see e.g. Arnold 2008). For a known threshold, this posterior density is a Gamma. We then do proceed to inference, using a gamma prior based on the WID data, derived from the top 5% income shares, which is typically interpreted as an inverse measure of the concentration of top incomes (Atkinson 2017).[3] We finally measure the Kullback-Leibler divergence between these two posterior densities. In theory, this distance should be minimum when the source of the data is mixed or administrative, and maximum when the source is survey.[4]

The main message of Table 1 is that the WID data provide on average a much more dispersed right tail than the EU-SILC, justifying the need of extra information for correcting for top incomes in surveys. In other terms, α_0 is always lower than α, except for Portugal in 2008. The maximum KL distance is for Estonia (EE) which uses survey data and minimum for the Netherlands (NL) which makes use of register data. Top incomes are particularly under-sampled in the New Member States. Otherwise, there does not seem to be any relation between the source of the income data and the KL distance, showing the need of a correction, whatever the source. This is a puzzling fact for Nordic countries, which all have register data that have the reputation of being quite accurate. This is the main message brought by the simple Pareto I model, with a fixed threshold. However, this model has

[3]For EU-SILC micro data, we use the income variable `HX090` with weights `DB090`. When h is known, the posterior density of α is a gamma density with parameters n and $\sum \log x_i/h$, where n is the number of observations in the tail above h.

For the WID tabulated data, we use the `scainc` income variable. The Pareto coefficient α is estimated using the two top income shares and the formula given for instance in Atkinson (2007, page 24). The gamma prior is indexed by ν_0 and $\nu_0\alpha_{WID}$, where ν_0 is the prior precision.

The WID data are available at `https://wid.world/fr/donnees/` and can be downloaded using the package `wid` available at `devtools::install_github("WIDworld/wid-r-tool")`. Income data are available in the form of income shares s_π with for instance $\pi = p95p100$, meaning in this case the top 5% share. The Pareto I distribution implies that the relative share of two groups is given by $s_1/s_2 = (\pi_1/\pi_2)^{(\alpha-1)/\alpha}$ leading for instance to the following formula $\hat{\alpha} = 1/(1 - (\log(s_{0.05}/s_{0.01})/\log(0.05/0.01)))$.

[4]The Kullback-Leibler divergence between two gamma densities $\mathrm{G}(\nu_1, s_1)$ and $\mathrm{G}(\nu_2, s_2)$ is given by $D_{KL}(G_1||G_2) = (\nu_1 - \nu_2)dG(\nu_1) - \log(\Gamma(\nu_1)/\Gamma(\nu_2)) + \nu_2 \log(s_1/s_2) + \nu_1(s_2 - s_1)/s_1$, where $\mathrm{dG}(\cdot)$ is the digamma function.

Table 1: The quantity of extra information contained in WID data assuming that the 5% tail is a Pareto I

		2008				2018			
Country	Source	α	α_0	α_*	KL	α	α_0	α_*	KL
DK	Register	3.440	1.735	3.125	2.426	2.788	1.665	2.532	1.988
FI	Register	3.291	1.896	3.148	0.954	3.162	2.007	3.044	0.700
SE	Register	3.867	1.855	3.500	2.649	3.861	2.214	3.543	1.868
IE	Register	2.901	2.301	2.741	0.620	3.038	2.104	2.734	1.696
UK	Survey	2.647	1.839	2.497	0.909	2.975	2.028	2.869	0.587
AT	Mixed	3.445	1.738	3.006	3.477	3.648	1.921	3.251	2.800
BE	Mixed	3.183	2.791	3.093	0.197	3.701	2.953	3.511	0.561
DE	Survey	2.948	1.958	2.817	0.775	3.08	1.888	2.905	1.147
FR	Mixed	2.844	2.045	2.73	0.586	2.766	2.125	2.667	0.433
LU	Mixed	3.314	1.676	2.749	4.81	3.485	2.422	3.123	1.751
NL	Register	3.044	2.761	3.004	0.068	3.069	2.708	3.022	0.096
EL	Survey	2.940	2.514	2.832	0.298	3.049	2.137	2.969	0.404
ES	Mixed	3.469	2.231	3.305	0.877	3.838	1.853	3.591	1.745
IT	Mixed	3.276	2.455	3.208	0.268	2.888	1.746	2.797	0.643
PT	Survey	2.858	3.017	2.908	0.087	3.393	2.389	3.257	0.617
EE	Survey	4.036	1.896	3.24	6.100	10.846	1.767	8.274	11.871
LT	Mixed	2.731	2.05	2.517	1.071	2.819	2.024	2.572	1.349
LV	Mixed	3.022	2.365	2.799	0.892	3.011	2.416	2.821	0.693
CZ	Survey	3.410	1.597	3.063	2.921	4.141	1.816	3.642	3.698
HU	Survey	3.388	2.106	3.15	1.408	2.907	1.991	2.637	1.597
PL	Survey	2.948	1.759	2.762	1.332	3.998	1.854	3.687	2.200
SI	Register	4.591	2.841	4.288	1.326	4.213	2.559	3.904	1.528
SK	Survey	3.762	2.143	3.311	2.837	6.284	2.353	5.247	5.839

α is the posterior expectation of the Pareto coefficient under a non-informative prior, α_0 is the prior expectation of α derived from WID data and α_* the posterior expectation of α under the informative prior. KL is the Kullback-Leibler divergence between the posterior under a non-informative and the same posterior under an informative prior, centred on the WID data with $\nu_0 = 100$. For EU-SILC data, means and quantiles were computed using the `R` package `DescTools`. We selected $\rho = 0.95$ (top 5% incomes) for both sources as in Atkinson (2017).

many limitations that can be removed by using the more elaborate Pareto II model promoted by Jenkins (2017).

2.2 The Pareto II distribution

Pareto I and Pareto II are intimately related (see e.g. Arnold 2008). For $0 < h \leq x$, the cdf and pdf of the Pareto I are:

$$F_{P1}(x) = 1 - (x/h)^{-\alpha}, \qquad f_{P1}(x) = \alpha h^{\alpha} x^{-\alpha-1}. \tag{1}$$

The Pareto II process is built from these expressions when taking h as a location parameter and introducing a separate scale parameter β, leading to:

$$F_{P2}(x) = 1 - \left(1 + \frac{x-h}{\beta}\right)^{-\alpha}, \quad f_{P2}(x) = \frac{\alpha}{\beta}\left(1 + \frac{x-h}{\beta}\right)^{-\alpha-1}. \tag{2}$$

A Pareto I corresponds to the testable restriction $h = \beta$. A Pareto I models the distribution of relative excesses, x/h, whereas a Pareto II models the distribution of

absolute excesses $x - h$ (see e.g. Charpentier and Flachaire 2022). The mean is:

$$\mathrm{E}(x) = h + \frac{\beta}{\alpha - 1}.$$

The Gini can be deduced from Arnold (2008, page 135) with:

$$G(x) = 1 - \frac{h + 2\alpha\beta B(2\alpha - 1, 2)}{h + \alpha\beta B(\alpha - 1, 2)}, \tag{3}$$

where $B(,)$ is the Beta function. Pareto I and Pareto II measure inequality in a quite different way for a given value of the Pareto coefficient α, as the Gini for the Pareto I is:

$$G = \frac{1}{2\alpha - 1}. \tag{4}$$

This opposition in measuring inequality is well depicted in Figure 1 of example 1.

Remark 1. *A prior on α can be translated directly into a prior on the value of the Gini in the Pareto I case. For a Pareto II, we have no longer this one-to-one correspondence, as the value of the Gini depends both on α and on the value of the difference $h - \beta$. If the sign of $h - \beta$ is negative, the Pareto II Gini will be greater, and lower in the reverse case.*

Example 1. *For a given value of $\alpha = 1.75$ and $\beta = 5$, we let h vary between 0 and 10. The Gini of the Pareto I corresponds to the particular case $h = \beta$. Depending on the value of h, compared to that of β, the Pareto II process can display either more or less inequality than the Pareto I process. Maximum inequality is obtained for $h = 0$. This behaviour is related to the fact that the Gini is invariant by scaling (i.e. change in a monetary unit), but not invariant by translation (when the same sum is given to or taken from everybody). In Figure 1, we have plotted the value of the Gini against $(h - \beta)/\beta$, which provides a scale free graph. The Pareto I Gini is equal to 0.4. For the Pareto II Gini to vary between reasonable bounds (0.3 to 0.5), $(h - \beta)/\beta$ has to be limited to the range (-0.466 to 0.778).*

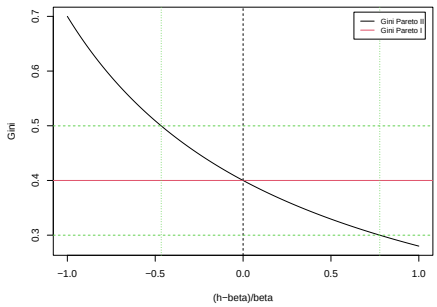


Figure 1: Gini constellation, varying h for a given $\beta = 5$ and $\alpha = 1.7$

2.3 Comparing Pareto I and Pareto II tails

Jenkins (2017) discusses the difficulty of empirically distinguishing between these two Pareto processes, leading eventually to the question: Is a Pareto II model really needed to fit well the data? To answer briefly this question, we have run a simulation exercise and exploited the message given by the Pareto plot. This example shows that depending on the sign of the difference $h - \beta$, the Pareto plot presents quite different configurations that identify the deviance of extreme observations from a simple Pareto I process. The Pareto diagram, as named by Cowell (2011), takes logs in the expression of the Pareto CDF and leads to the linear expression:

$$\log(1 - F(x)) = -\alpha \log x + \alpha \log h.$$

So a plot of $\log(1-\hat{F}(x))$ against $\log x$ provides a straight line with a negative slope if the data follow a Pareto I distribution. Data generated according to a Pareto II leads to particular results depending on the sign of $h - \beta$.

Example 2. *We have generated two series of Pareto II random numbers with* $\alpha = 2.5$ *and* $\beta = 5$*. One is obtained with* $h = 2.5$*, so with* $h < \beta$ *and the other with* $h = 10$*, so with* $h > \beta$*. For each of these two samples, we draw the corresponding Pareto plot and compare them in Figure 2. With* $h-\beta > 0$*, extreme points are located above the Pareto line, a configuration that was qualified of outliers in Charpentier and Flachaire (2022, Figure 7). With* $h - \beta < 0$*, we have the reverse situation. We can conclude that when* $h - \beta < 0$*, rich households are under-sampled, and when* $h - \beta > 0$ *rich households are better represented.*

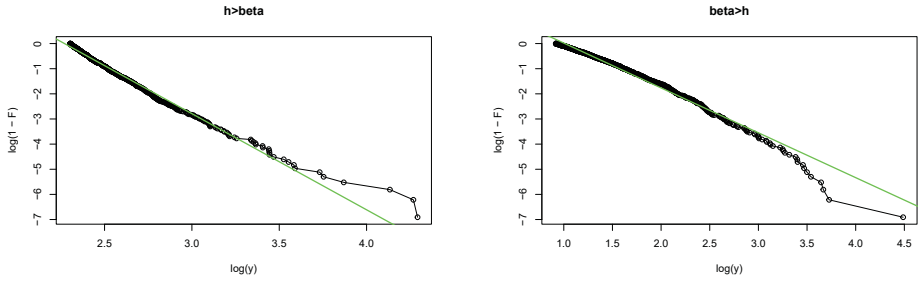


Figure 2: Comparing Pareto tails obtained from the Pareto II

Example 2 shows how Pareto II processes can generate quite different tail behaviours depending on the sign of $h - \beta$. In the empirical application of section 5, the EUR15 states and typically the Nordic countries will in general correspond to $h - \beta > 0$, while the NMS will correspond to $h - \beta < 0$.

3 Compound log-normal-Pareto II models

Deciding at which point should start the Pareto tail is an important problem that cannot be solved easily, as discussed in Scarrott and MacDonald (2012).[5] An elegant

[5]For instance, a graphical solution such as the Hill plot has been referred to as the Hill horror plot in the literature, due to its very poor performance, see Scarrott and MacDonald (2012).

solution consists in considering a complete model mixing a truncated distribution (the central model) for the observations below the threshold h and a tail model belonging to the Pareto family above the threshold. The threshold is treated as a parameter to be estimated. The extreme value literature has proposed many ways to combine these two models. The composite log-normal-Pareto model was introduced by Cooray and Ananda (2005) and Scollnik (2007). This model was found useful to model extreme events in insurance claims, ecology and many other topics including modelling the income distribution (see e.g. the references provided in Scollnik 2007, Cabras and Castellanos 2011 or Nadarajah and Bakar 2013).

3.1 The bulk model as a useful restriction

The initial compound model can be written as a two component mixture:

$$f(x|\theta) = \rho f_1(x|\theta) + (1-\rho) f_2(x|\theta), \tag{5}$$

where ρ is the proportion of observations below the threshold h. The truncated log-normal distribution with parameters μ and σ^2 is usually chosen for $f_1(x|\theta)$:

$$f_1(x|\theta_1) = \frac{f_\Lambda(x|\mu,\sigma^2)}{F_\Lambda(h|\mu,\sigma^2)} \mathbb{1}(x \leq h), \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function. We have chosen the Pareto II for $f_2(x|\theta)$. The literature (see e.g. Behrens et al. 2004) has proposed to impose the restriction:

$$\rho = 1 - F_\Lambda(h|\mu,\sigma^2), \tag{7}$$

leading to what is called the bulk model. Its PDF is:

$$f(x|\theta) = f_\Lambda(x|\mu,\sigma^2)\mathbb{1}(x<h) + (1-F_\Lambda(h|\mu,\sigma^2)) f_{P2}(x|h,\alpha,\beta)\mathbb{1}(x \geq h), \tag{8}$$

and its corresponding CDF:

$$\begin{aligned} F(x|\theta) &= F_\Lambda(x|\mu,\sigma^2)\mathbb{1}(x<h) + F_\Lambda(h|\mu,\sigma^2)\mathbb{1}(x \geq h) \\ &+ (1-F_\Lambda(h|\mu,\sigma^2) F_{P2}(x|h,\alpha,\beta)\mathbb{1}(x \geq h). \end{aligned} \tag{9}$$

A model with a Pareto I tail is obtained by replacing $f_{P2}(x|h,\alpha,\beta)$ by $f_{P1}(x|h,\alpha)$ and $F_{P2}(x|h,\alpha,\beta)$ by $F_{P1}(x|h,\alpha)$ in these expressions.

The bulk model presents a discontinuity at h as shown in Figure 3 of Example 3. Continuity is usually not imposed, as this would mean a restriction on the Pareto parameters. For the Pareto II, the continuity restriction would mean a restriction on the value of β with (see Majid and Ibrahim 2021b):

$$\beta = \alpha \frac{1-F_\Lambda(h|\mu,\sigma^2)}{f_\Lambda(h|\mu,\sigma^2)}. \tag{10}$$

For the Pareto I, the continuity restriction is even more severe, as it fully determine the value of α:

$$\alpha = h \frac{f_\Lambda(h|\mu,\sigma^2)}{1-F_\Lambda(h|\mu,\sigma^2)}. \tag{11}$$

Example 3. *Let us give now an idea of the shape of the bulk model with Figure 3. In this example, the parameters of the log-normal component are* $\mu = 0.5$ *and* $\sigma = 0.5$*. We have then added the Pareto II component with* $h = 2.0$*,* $\alpha = 1.7$ *and* $\beta = 3$*, so* $(h - \beta) < 0$*. The Pareto II right tail is well above the log-normal tail. Imposing continuity would mean* $\beta = 1.605324$*, a restriction that lowers the position of the Pareto II tail and implies* $(h - \beta) > 0$*. The Pareto I tail corresponds to another restriction with* $(h - \beta) = 0$*.*

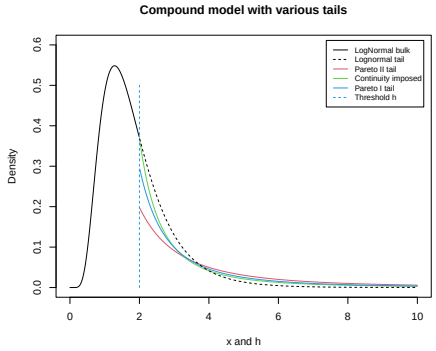


Figure 3: The Compound lognormal-Pareto II family

To summarize, with the compound Pareto model, we have a framework where we are free to choose the central model (here the log-normal density) and the shape of the tail which can be Pareto I, Pareto II or Generalized Pareto. The threshold parameter h can be estimated in this framework. Due to the discontinuity, the Bayesian approach is most of the time privileged.

Several other models are said to have a Pareto-like tail. They all include a continuity restriction, limiting thus the shape of the right tail. Limiting ourselves to three parameter distributions, the first candidate is of course the famous Singh and Maddala (1976) distribution which embed directly a Pareto II tail as can be seen from its CDF $F(x) = 1 - (1 + (x/\beta)^q)^{-\alpha}$, to be compared to (2). Our second candidate is the Pareto-log-normal distribution of Reed and Jorgensen (2004). Its properties were amply discussed in Hajargasht and Griffiths (2013). Further comparisons can be found in Majid and Ibrahim (2021a). Finally, we want to mention the Kaniadakis distribution (see Kaniadakis 2013 or Clementi and Gallegati 2016 for complementary details). Clementi et al. (2012) provide an interesting exploration of this density for income distributions, using the GSOEP, BHPS and PSID data sets.

3.2 Alternative likelihood functions

The likelihood function is the central ingredient for Bayesian inference. Using Bayes' theorem, it serves to revise the prior density $\varphi(\theta)$, leading to the posterior density $\varphi(\theta|x)$. In most cases, this posterior density has to be explored using simulation methods, one of them being Monte Carlo Markov Chain (MCMC). In our context, the likelihood function of model (8), when adding weights, is:

$$L(x;\theta) = \prod_{i,x_i \leq h} f_\Lambda(x_i|\mu,\sigma^2)^{w_i} \prod_{i,x>h} \left(1 - F_\Lambda(h|\mu,\sigma^2)\right) f_{P2}(x|\alpha,\beta,h)^{w_i}, \qquad (12)$$

with $\theta = (\mu, \sigma, \alpha, \beta, h)$. Majid and Ibrahim (2021a) have compared several composite models for modelling the income distribution in Malaysia, using either a Pareto I or a Pareto II for the tail and various uni-modal parametric distributions for the central model, such as log-normal, gamma, Weibull with two parameters or Dagum and Singh-Maddala with three parameters. It appeared that the log-normal-Pareto II composite model provided the most satisfactory fit. However, they did not measure the impact of these choices on the estimation of h, and consequently on their measurement of inequality.

On the contrary, the main objective of Cabras and Castellanos (2011) was to find the best way to estimate extreme values for insurance claims. They propose to treat the parameters of the central model as nuisance parameters, which leads them to consider a profile likelihood function where the central model is replaced by a MLE estimator of a semi-parametric data density procedure based on orthogonal polynomials. The alternative likelihood function of their model is:

$$L_{Prof}(x;\theta) = \prod_{i,x_i \leq h} \hat{F}(h)\hat{f}_h(x_i) \prod_{i,x>h} (1-\hat{F}(h)) f_{P2}(x|\alpha,\beta,h)^{w_i}, \tag{13}$$

where $\hat{F}(h)$ is the proportion of observations below h and $\hat{f}_h(x_i)$ a polynomial approximation of the truncated distribution of the observations below h.[6] In this approach, the number of parameters to estimate is reduced to three with $\theta = (\alpha, \beta, h)$. We shall see later that choosing between $f_\Lambda(x_i|\mu,\sigma^2)$ and $\hat{f}(x_i)$ can have a tremendous effect on the estimation of the threshold h, the value of $h-\beta$ and consequently on the decomposition of the corresponding Gini coefficient, as detailed below.

3.3 Gini decomposition using the Pareto II

Jenkins (2017) has proposed to decompose the Gini index in a classical framework between observed low incomes, corresponding to the ρ lower quantiles (those corresponding to $x < h$) and upper incomes modelled with a Pareto II. The population shares are π_u for the upper quantiles and $\pi_l = 1 - \pi_u$ for the lower quantiles. Corresponding income shares are $s_u = \pi_u \bar{x}_u / \bar{x}$, $s_l = 1 - s_u$, where $\bar{x}_u$ is the average income for upper quantiles and $\bar{x}$ the average income. Alvaredo (2011) has shown that the between groups inequality can be simplified to $s_u - \pi_u$ leading to the Gini decomposition formula:

$$G = \pi_l \times s_l \times G_l + \pi_u \times s_u \times G_u + s_u - \pi_u, \tag{14}$$

where G_l is the empirical Gini for the lower group, while G_u is the parametric Gini given by the Pareto tail. There are two important parameters in this decomposition: the population share π_u which is a direct function of h and the value of G_u which depends on both α and $h - \beta$.

Using a bulk model brings in new features in this decomposition, due to the fact that h is now random and that we have stored a full MCMC output $\theta^{(j)} = (\alpha^{(j)}, \beta^{(j)}, h^{(j)})$ (a $m \times 3$ matrix of draws from the posterior density $\varphi(\theta|x)$) if we adopt a profile likelihood. Because h is now a random variable, π_l, s_l and G_l become random variables. First, $\pi_l(h^{(j)}) = \hat{F}(h^{(j)})$. Then, for each draw $h^{(j)}$, we have a new sample separation on which we can compute the empirical values $\bar{x}_l(h^{(j)})$, $s_l(h^{(j)})$

[6] In Appendix B, we develop an approach based on Bernstein polynomials for $\hat{f}(x)$ and $\hat{F}(x)$.

and $G_l(h^{(j)})$. Collecting these results, we arrived at Algorithm 1, assuming that the weights were taken into account when estimating θ and that the MCMC output $[\theta^{(j)}]$ has been stored.

Algorithm 1 Bayesian decomposition of the Gini index

From the m stored values of $\theta^{(j)} = (\alpha^{(j)}, \beta^{(j)}, h^{(j)})$
for $j = 1, m$ **do**
  $n_j = \sum \mathbb{1}(x < h^{(j)})$
  $\pi_l(h^{(j)}) = \hat{F}_\Lambda(h^{(j)})$
  $x_l(h^{(j)}) = x[x < h^{(j)}]$
  $\bar{x}_l(h^{(j)}) = \sum x_l(h^{(j)})/n_j$
  $G_l(h^{(j)}) = \text{Gini}(x_l(h^{(j)}))$
  $\pi_u(h^{(j)}) = 1 - \pi_l(h^{(j)})$
  $\text{E}(x_u|\theta^{(j)}) = h^{(j)} + \beta^{(j)}/(\alpha^{(j)} - 1)$
  $G_u(\theta^{(j)}) = 1 - \dfrac{h^{(j)} + 2\alpha^{(j)}\beta^{(j)}B(2\alpha^{(j)} - 1, 2)}{h^{(j)} + \alpha^{(j)}\beta^{(j)}B(\alpha^{(j)} - 1, 2)}$
  $\text{E}(x|\theta^{(j)}) = \pi_l(h^{(j)})\,\bar{x}_l(h^{(j)}) + \pi_u(\theta^{(j)})\,\text{E}(x_u|\theta^{(j)})$
  $s_u(\theta^{(j)}) = \pi_u(\theta^{(j)})\,\text{E}(x_u|\theta^{(j)})/\text{E}(x|\theta^{(j)})$
  $s_l\theta^{(j)}) = 1 - s_u(\theta^{(j)})$
  $G(\theta^{(j)}) = \pi_l(\theta^{(j)})\,s_l(\theta^{(j)})\,G_l(\theta^{(j)}) + \pi_u(\theta^{(j)})\,s_u(\theta^{(j)})\,G_u(\theta^{(j)}) + s_u(\theta^{(j)}) - \pi_u(\theta^{(j)})$.
end for

Using Algorithm 1, we get draws from the posterior density of the Gini coefficient.

4 Bayesian inference for a Pareto II tail

For decomposing the Gini index in a Bayesian way, we need a MCMC output $[\theta^{(j)}]$. We now detail how to obtain it for the bulk model (8).

4.1 Priors for making inference on h

Let us decompose the prior density related to the likelihood function (12) of model (8) as:

$$\varphi(\theta) = \varphi(\mu|\sigma^2)\varphi(\sigma^2)\varphi(\alpha)\varphi(\beta)\varphi(h).$$

We have a strong interest in being informative on the Pareto parameter α, in order to introduce our prior information coming from the WID data set. A gamma informative prior on α is natural conjugate and corresponds to:

$$\varphi(\alpha|\nu_0, s_0) \propto \alpha^{\nu_0 - 1} \exp(-\alpha s_0),$$

with prior expectation $\text{E}(\alpha) = \nu_0/s_0$. The WID estimated values α_0 of Table 1 provide a value for the ratio ν_0/s_0, leaving aside the question of prior precision. A non-informative prior corresponds to $\nu_0 = 0$ and $\varphi(\alpha) \propto 1/\alpha$. A non-informative prior is equivalent to *method A* of Jenkins (2017). The prior precision increases with the value of ν_0 and thus for scientific reporting it is essential to provide results for $\nu_0 = 0$ and for say $\nu_0 = 100$. Note that an informative prior on α directly provides a specific prior information on the concentration of high incomes, but not on the value of the implied Gini coefficient, which, in a Pareto II context, depends also on the sign of $h - \beta$. We can have a much higher Gini if $h - \beta < 0$, or a much lower Gini if $h - \beta > 0$ as was illustrated in Figure 1.

For $\varphi(h)$, Cabras and Castellanos (2011) have chosen a uniform prior between bounds:

$$\varphi(h) \propto 1, \quad h \in [\underline{h}, \overline{h}].$$

This prior has been amply discussed in Majid and Ibrahim (2021b). In the Bulk model, there is a one-to-one relation between h and ρ, which is the proportion of observations in the tail, as given in (7). So, a prior on h can be translated into a prior on ρ, with the advantage that the latter is scale-free and consequently much easier to elicit.[7]

Jenkins (2017) conducted a sensitivity analysis for different values of ρ, the starting point of the Pareto tail. He has chosen $\rho = 0.90, 0.95, 0.99, 0.995$. Amongst other recent studies relating ρ to the upper sample proportion of incomes for which data issues should be corrected, Bartels and Metzing (2019) opted for $\rho = 0.99$, while Angel et al. (2019) or Flachaire et al. (2022) found that the corrections could start as early as the median. With a Bayesian approach and a bulk model, we integrate over a possible range, letting the sample choose the most plausible range. So, we decided for $\underline{\rho} = 0.650$ and $\overline{\rho} = 0.995$. The corresponding prior range for h corresponds to the empirical quantiles $[\underline{h} = Q_x(0.650), \overline{h} = Q_x(0.995)]$.

We can choose to be non-informative on the scale parameter β, so that:

$$\varphi(\beta) \propto 1/\beta.$$

Because the central model is not our prime interest, we decide to be non-informative on (μ, σ^2), the parameters of the log-normal, so that:

$$\varphi(\mu, \sigma^2) \propto 1/\sigma^2.$$

With the profile likelihood (13), this prior is no longer relevant.

4.2 Conditional posterior distributions

The posterior distribution is proportional to the product of the likelihood function and the prior, using either (12) or (13):

$$\varphi(\theta|x) \propto \varphi(\theta) L(x;\theta), \quad \text{or} \quad \varphi(\theta|x) \propto \varphi(\theta) L_{Prof}(x;\theta). \tag{15}$$

As this posterior density has no closed form, we propose in a next section a Gibbs sampler to produce draws from an approximation to it, draws that will serve to decompose the Gini index, as explained above.

Conditionally on h, we can separate inference on the central part and on the tail. The two conditional posterior distributions of the Pareto II are found by discarding alternatively the proportional terms on which we condition in one of

[7] As noted in Majid and Ibrahim (2021b), a uniform prior on h does not mean a uniform prior on ρ. We must keep in mind that the threshold h has two meanings. An optimal h can first correspond to the level at which a correction has to be done. It can be rather low or high depending on the quality of the survey (for instance if the source used for income is administrative or results from a simple interview). A second meaning is at which level a Pareto model best fits the data. A statistical approach defines h according to an optimal fit. The posterior density of h will correspond to this second meaning. We thank Emmanuel Flachaire for pointing out this distinction.

the two complete posterior densities in (15). The conditional posterior distribution of α corresponds to a gamma density with:

$$\varphi(\alpha|h,\beta,x,w) = f_G(\alpha|\nu_0+n, s_0+\sum_{i,x_i\geq h} w_i \log(1+(x_i-h)/\beta)). \tag{16}$$

Conditionally on a previous draw of β and h, it is easy to draw random numbers from this density.

The conditional posterior distribution of β does not correspond to a known density with:

$$\varphi(\beta|h,\alpha,x,w) \propto \beta^{-n-1}\exp(-(\alpha+1)\sum w_i \log(1+(x_i-h)/\beta)). \tag{17}$$

We propose to implement an enriched version of the Griddy-Gibbs of Bauwens and Lubrano (1998) to draw from this conditional posterior density. We use a moment estimator (see 25 in Appendix A) to determine an initial value $\beta^{(0)}$ and a grid *bp* of k points on the range $[\beta^{(0)}/3, 3\beta^{(0)}]$. Conditionally on a draw of α, we evaluate $\varphi(\beta|\alpha,x)$ on the predetermined grid and derive its normalized empirical CDF. Equipped with this empirical CDF, we use the logic of the inverse transformation method, sampling a random value from a uniform distribution over $[0,1]$, determining its position in the CDF and then proceeding by linear interpolation to determine the corresponding value $\beta^{(1)}$ on the predefined grid of k points. The method can be enriched by considering an update of the initial exploration range $[\beta^{(0)}/3, 3\beta^{(0)}]$, adjusting the lower and upper bounds of the grid, using the minimum and maximum draws obtained during the warming up of the chain.

The conditional distribution of h depends on the whole sample, and thus of course on the central model. In the case of a profile likelihood, we have:

$$\begin{aligned}\varphi(h|\alpha,\beta,x,w) \quad \propto \quad & \varphi(h) \times \prod_{i,x_i\leq h} \hat{F}(h)\hat{f}(x_i) \times \\ & \prod_{i,x>h} \left(1-\hat{F}(h)\right)\left[\frac{\alpha}{\beta}(1+(x_i-h)/\beta)^{-\alpha-1}\right]^{w_i}.\end{aligned} \tag{18}$$

Weights were already used for estimating $\hat{f}(x_i)$ and $\hat{F}(h)$. The Griddy-Gibbs of Bauwens and Lubrano (1998) is here again a good solution to draw from this conditional posterior density. The exploration grid is calibrated from the prior range $[\underline{h}, \overline{h}]$. It can be adjusted in the same way as the grid used for β.

If we decide to opt for a truncated log-normal central model, the conditional posterior distribution of h has the following form:

$$\begin{aligned}\varphi(h|\alpha,\beta,\mu,\sigma,x,w) \quad \propto \quad & \varphi(h) \times \prod_{i,x_i\leq h} f_\Lambda(x_i|\mu,\sigma^2)^{w_i} \times \\ & \prod_{i,x>h} \left(1-F_\Lambda(h|\mu,\sigma^2)\right)\left[\frac{\alpha}{\beta}(1+(x_i-h)/\beta)^{-\alpha-1}\right]^{w_i}.\end{aligned} \tag{19}$$

It depends both on draws of (α,β), but also on draws for (μ,σ^2). So we have to find the conditional posterior density of (μ,σ^2) for a given h. Due to the truncation, this conditional distribution does not belong to a known family. It is obtained from the

likelihood function (12), neglecting the factors that do not depend on (μ, σ^2):

$$\varphi(\mu, \sigma^2|h, x) \propto \frac{1}{\sigma^2} \prod_{i, x_i \leq h} f_\Lambda(x_i; \mu, \sigma^2)^{w_i} \prod_{i, x > h} \left(1 - F_\Lambda(h; \mu, \sigma^2)\right). \tag{20}$$

We propose an independent Metropolis step within Gibbs algorithm. We first estimate $(\hat{\mu}, \hat{\sigma}^2)$ by maximizing numerically a conditional likelihood function built on (6), for a fixed h, corresponding to the top 5% tail. The proposal is a truncated bi-variate normal indexed by the MLE:

$$q(\mu, \sigma^2) = f_{TN}(\mu, \sigma|(\hat{\mu}, \hat{\sigma}), \hat{H}^{-1}),$$

where $\hat{H}$ is the Hessian matrix from the MLE. A draw $\zeta = (\mu^{(j+1)}, \sigma^{2(j+1)})$ from this proposal is accepted with probability p given by:

$$p = \min \left[\frac{\varphi(\zeta|h, x) q(\theta^{(j)})}{\varphi(\theta^{(j)}|h, x) q(\zeta)}, 1 \right],$$

where $\varphi(\zeta|h, x)$ is given in (20). Otherwise the previous draw $\theta^{(j)} = (\mu^{(j)}, \sigma^{2(j)})$ is kept.

4.3 A Gibbs sampler

Let us now regroup all these results into Algorithm 2 to propose a Gibbs sampler for making inference on all the parameters in the case of a profile likelihood. For a maximum range $\rho = 0.995$, we first estimate $\hat{f}(x|k)$ and $\hat{F}(x|k)$ where k is the degree of the Bernstein polynomial (see Appendix B for more details on density estimation using Bernstein polynomials). Then, for an initial guess for h, a sample separation is found, $\hat{F}(h|k)$ and $\hat{f}(x|x < h, k)$ are evaluated. The algorithm simulates draws for α and β. Given these draws, a new draw for h is proposed. The obtained MCMC output will be used to propose a new measure of inequality, thanks to a decomposition of the Gini detailed in section 3.3.

5 Top income correction for EU-SILC data

First, we show that the truncated log-normal is a poor choice for the central model, compared to the Bernstein density estimator. Equipped with a profile likelihood of the bulk model, we then detail how the correction for missing information on high incomes very much depends on the sign and amplitude of $(h-\beta)/\bar{\beta}$. We finally show that the high incomes correction is not effective for the Nordic countries, is mild for the remaining EUR15 countries and important for the New Member States.

5.1 The pitfalls of a truncated log-normal central model

We provide in Figures 4 and 5 the histograms of the 2008 income data for the EU15 and the NMS. On the same plot, we provide the Bernstein density estimates of the truncated distribution in red and of the truncated log-normal in green for h corresponding to the 95% lower tail. If a truncated log-normal central model provides a reasonable fit for some countries (Luxembourg, Netherlands, Portugal, Hungary), for all the other countries the error committed for estimating the 0.95

Algorithm 2 Bayesian inference for the bulk model

Select m the number of draws and *mdrop* the size of the warming chain
Choose a prior range for ρ, $[\underline{\rho}, \overline{\rho}]$ and determine the corresponding values of $[\underline{h}, \overline{h}]$ using the empirical CDF of the sample
Estimate $\hat{F}(x|k)$ and $\hat{f}(x|k)$ on the restricted sample $x \leq \overline{h}$ using Bernstein polynomials of degree k
Chose a starting value for ρ, e.g. $\rho^{(0)} = 0.95$ (inside the prior range) and the corresponding $h^{(0)}$
Build the initial grid hp of np points $(\underline{h}, h_2, \cdots, h_{np-1}, \overline{h})$ for h
Compute an initial estimate of α and β, conditionally on $h^{(0)}$, using a method of moments
Determine an initial grid bp of np points $(b_1, \cdots, b_{np})$ for β
for $j = 1, \ldots, m + mdrop$ **do**
    Select the sample vector $xl = x[x \leq h^{(j-1)}]$
    Compute $\rho^{(j)} = 1 - \hat{F}(h^{(j-1)}|k)$
    Compute $\hat{f}(x_l|k)$
    Select the sample vector $xu = x[x > h^{(j-1)}] - h^{(j-1)}$
    Sample $\alpha^{(j)}$ from $f_G(\alpha|n + \nu_0, s_0 + \sum w_i \log(1 + xu_i/\beta^{(j-1)})$
    Draw $\beta^{(j)} \sim \varphi(\beta|h^{(j-1)}, \alpha^{(j)}, xu)$, using a Griddy-Gibbs
    Draw $h^{(j)} \sim \varphi(h|\alpha^{(j)}, \beta^{(j)}, x)$, using a Griddy-Gibbs
    if $j = mdrop$ **then**
        update the grid for h using the min and max of the previous draws of h
        update the grid for β using the min and max of the previous draws of β
    end if
end for
Discard the initial *mdrop* draws for computing posterior moments and densities

truncation point is too important. This will have crippling consequences for inference on the Gini coefficient. On the contrary, the error committed by the Bernstein estimator is always lower than half a percent (see Table 4 of Appendix B). So, the profile likelihood based on a Bernstein polynomial (with $k = 15$) provides a much better approach and we shall stick to it.

5.2 Bayesian Gini correction

Let us first compare the results obtained for the UK by Jenkins (2017) to those provided by our approach and highlight the contrast. Jenkins (2017) recommends that the truncation point should be 95% or over, a finding based on data accuracy. His corrected Gini was estimated at 0.49 for 2007 with a Pareto I or II for a value of h corresponding to 0.99, results based on the data of the UK Survey of Personal Income. Using SILC data, we found the much lower values of 0.350 for 2008 and 0.339 for 2018 (the WID data alone would have meant a Gini of 0.373 for 2008 and 0.327 for 2018, under a Pareto I assumption).

Concerning the posterior expectations for ρ, we found for the UK in Table 2 0.926 and 0.954 in 2008 and 2018. These are values slightly lower than those recommended by Jenkins (2017). Moreover, we have found for other countries that $\mathrm{E}(\rho|x)$ could be as low as 0.82 (Estonia, 2008) and were never higher than 0.95 (Tables 2 and 3). The Pareto II model is found to provide a better fit than Pareto I, when inspecting the scaled posterior distribution of $h - \beta$. A Pareto I corresponds to the restriction $h = \beta$. Zero never belongs to a posterior 95% credible interval, except for Greece in 2008. For the UK, the value of the information given by the prior is rather weak, with a very low KL distance between the posterior of the Gini with and without an informative prior on α. This could question the adequacy of WID data for the

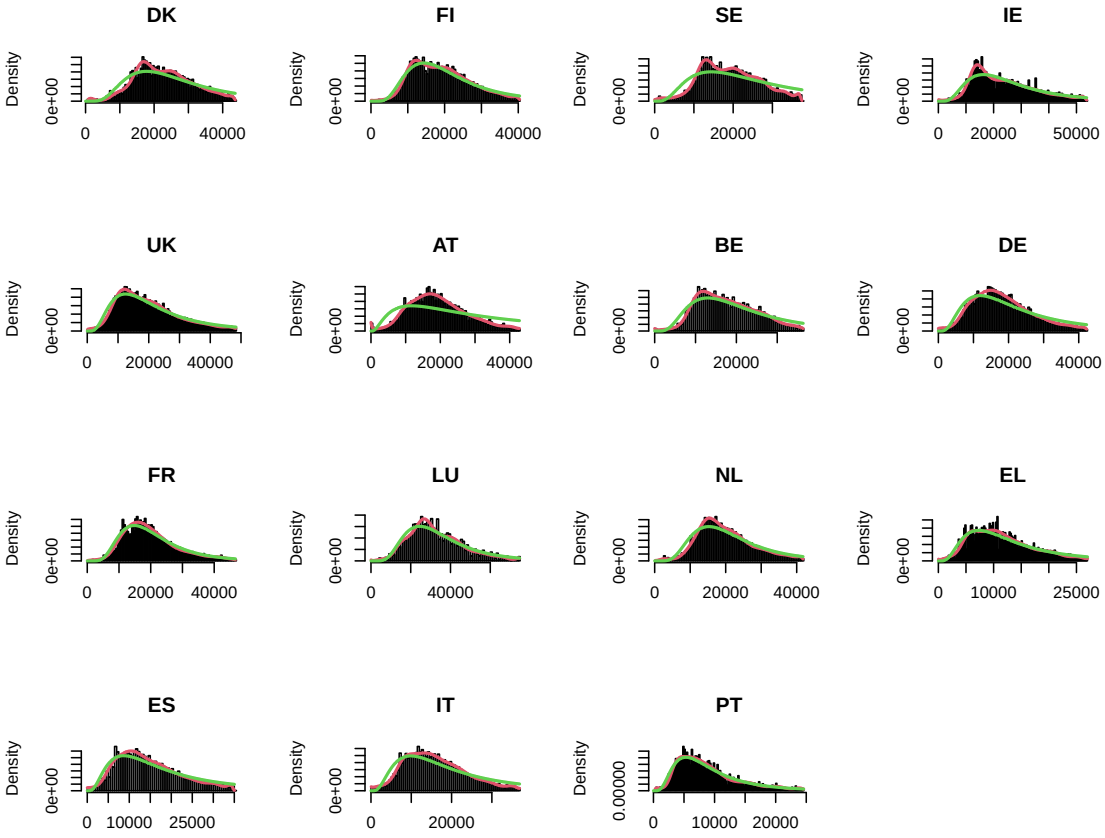


Figure 4: Truncated density estimates with Bernstein or Log-normal, EU15 2008

UK. The posterior expectation of the Gini has 0.687 chances of being greater than the sample Gini in 2008 and 0.997 in 2018, implying that a correction was effective with a Pareto II tail for 2018, but less for 2008. The use of survey versus register data does not make a difference when comparing Ireland and the UK, as the KL distances reported in the last column of Table 2 are comparable.

Because we are using the EU-SILC data, we have access to a much diverse number of cases, examining a total of 23 countries. Using this diversity, we shall see that the differences between survey and register data are significant for some groups of countries, but not for all (Nordic countries versus the others). And adopting a Pareto I can lead to biased results for the Gini, when the posterior expectation of $(h-\beta)/\bar{\beta}$ is important as it is in Nordic countries.

For the Nordic countries that are using register data, our approach tells that there is no need for a correction: adopting a Pareto tail does not manage to increase the value of the sample Gini. This is also the case for France (mixed) and the Netherlands (register). For those countries, the posterior expectation of the normalized difference $h-\beta$ is positive and can be quite large. This means that we have a quite large number of observations which are above the Pareto plot, a situation qualified as outliers in Charpentier and Flachaire (2022). So high incomes are well

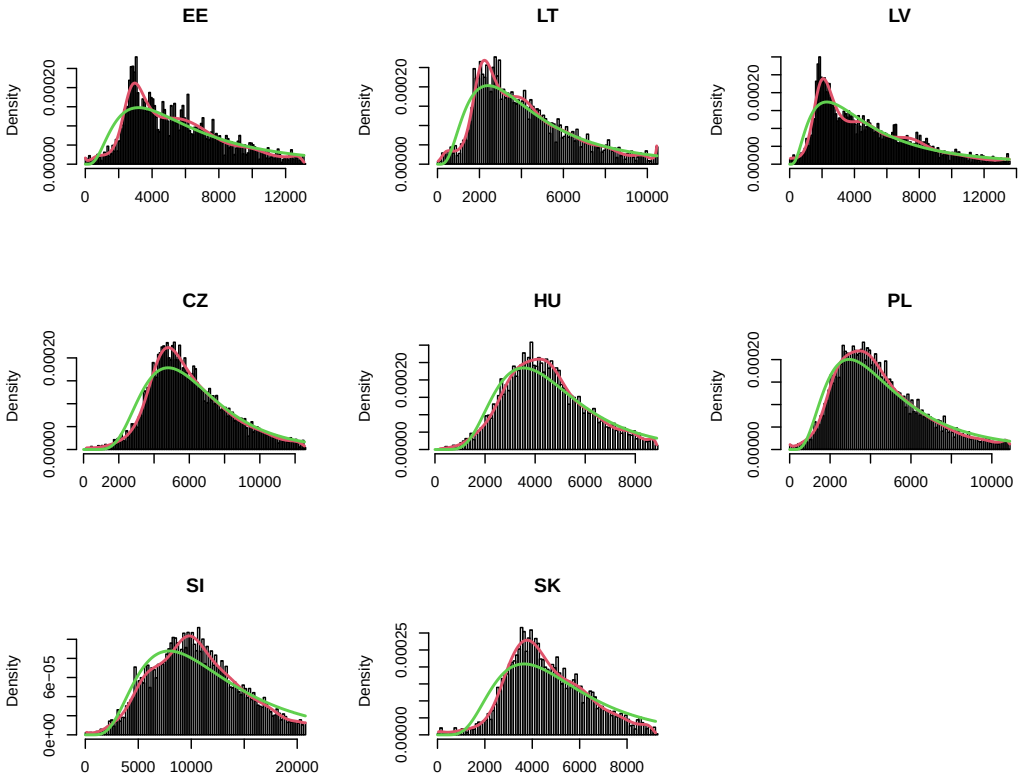
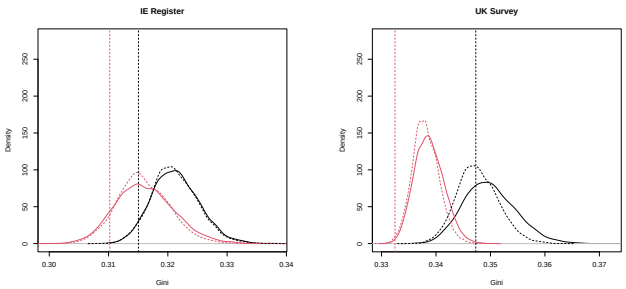


Figure 5: Truncated density estimates with Bernstein or Log-normal, NMS 2008

represented in the original data.



Black lines correspond to 2008 and red lines to 2018. Plain lines are for the Pareto II and dotted lines for the Pareto I. Vertical dotted lines represent the sample Gini computed with weights.

Figure 6: Posterior density of the Gini index: Ireland and the UK

In Figures 6, 7, 8, 9, we provide the graph of the posterior density of the corrected Gini, using either a Pareto II or a Pareto I right tail. As already said, the difference is mild for Ireland and the UK. This is also the case for most of the New Member

Table 2: Bayesian correction for the Gini index: EU15

Ctry	Data	E(Gini)	E(ρ)	E($h-\beta$)	Pr($G_B > G_y$)	KL
DK	Register	0.237	0.937	1.812	0.000	1.575
		0.268	0.928	0.740	0.000	0.370
FI	Register	0.261	0.945	1.531	0.000	0.935
		0.254	0.950	0.653	0.000	3.241
SE	Register	0.252	0.926	1.705	0.000	0.424
		0.261	0.920	1.846	0.000	0.391
IE	Register	0.321	0.895	0.224	0.951	0.202
		0.316	0.854	0.412	0.892	0.369
UK	Survey	0.350	0.926	0.638	0.687	0.038
		0.339	0.954	0.314	0.997	0.242
AT	Mixed	0.297	0.896	0.550	0.992	1.169
		0.278	0.903	0.396	0.524	2.132
BE	Mixed	0.270	0.901	1.049	0.001	0.420
		0.264	0.908	0.628	0.562	0.313
DE	Survey	0.316	0.947	0.722	0.824	0.009
		0.312	0.942	0.346	0.933	0.902
FR	Mixed	0.292	0.937	0.457	0.033	0.208
		0.285	0.934	0.381	0.058	0.024
LU	Mixed	0.301	0.833	0.162	1.000	3.016
		0.319	0.837	0.361	0.689	0.108
NL	Register	0.262	0.943	0.969	0.000	0.222
		0.278	0.941	0.697	0.047	0.172
EL	Survey	0.334	0.908	0.033	0.971	0.077
		0.323	0.962	0.129	0.982	0.013
ES	Mixed	0.330	0.942	0.405	0.928	0.370
		0.330	0.941	0.493	0.908	1.775
IT	Mixed	0.312	0.956	0.747	0.428	0.070
		0.333	0.955	0.784	0.783	0.244
PT	Survey	0.378	0.851	-0.465	0.996	0.472
		0.343	0.936	-0.545	1.000	7.580

Obtained with 5000 draws plus 500 for warming the chain. E(Gini) is the posterior expectation of the Bayesian corrected Gini, using the informative prior on α. KL is the Kullback-Leibler distance between the Gini estimated under an informative prior and without this prior information. E(ρ) is the posterior expectation (under the informative prior) of the proportion of observations concerned by the Pareto II tail. E($(h-\beta)$) indicates the posterior expectation of the normalized difference $(h-\beta)$. Pr($(G_B > G_y)$) indicates the posterior probability that the Bayesian corrected Gini is greater than the measured data Gini without Pareto II correction. For each country, the first line is for 2008 and the second line for 2018.

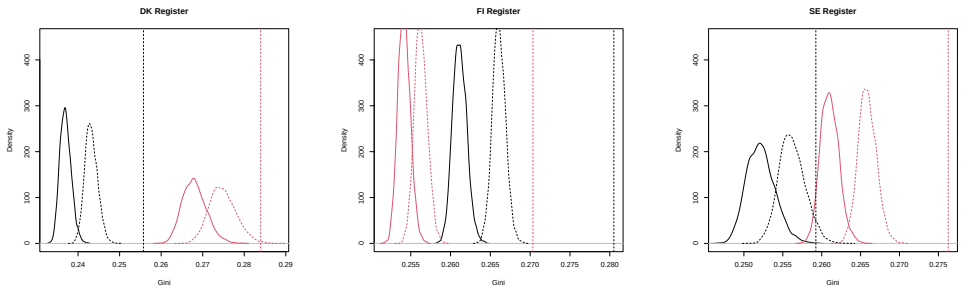
States. However, the difference is important for the Nordic countries, where the posterior expectation of $(h-\beta)/\bar{\beta}$ is quite large.

When turning to the new member states, we find two main differences. First, the value of the prior information is much higher, on average, as measured by the KL distance. Second, the posterior density of $(h-\beta)/\bar{\beta}$ is now either very low or negative. This means that the income data are lacking extreme values, or in other words, that the rich are under-represented. We are in the second case of Example 2.

Table 3: Bayesian correction for the Gini index: NMS

Country	Data	E(Gini)	E(ρ)	E($h-\beta$)	Pr($G_B > G_y$)	KL
EE	Survey	0.364	0.820	-0.194	1.000	7.332
		0.355	0.822	-0.041	1.000	3.773
LT	Mixed	0.371	0.888	-0.298	1.000	3.038
		0.404	0.879	0.014	0.994	0.648
LV	Mixed	0.429	0.865	-0.386	1.000	2.802
		0.409	0.882	-0.452	1.000	2.515
CZ	Survey	0.280	0.923	0.033	1.000	5.339
		0.268	0.910	0.275	1.000	4.415
HU	Survey	0.256	0.929	0.215	0.981	1.022
		0.328	0.894	-0.327	1.000	3.164
PL	Survey	0.345	0.935	-0.191	1.000	4.792
		0.312	0.932	-0.188	1.000	7.160
SI	Register	0.254	0.924	0.140	0.625	1.393
		0.258	0.923	-0.097	0.983	3.429
SK	Survey	0.248	0.898	0.862	0.897	0.216
		0.222	0.899	-0.379	1.000	25.134

Obtained with 5000 draws plus 500 for warming the chain. E(Gini) is the posterior expectation of the Bayesian corrected Gini, using the informative prior on α. KL is the Kullback-Leibler distance between the Gini estimated under an informative prior and without this prior information. E(ρ) is the posterior expectation (under the informative prior) of the proportion of observations concerned by the Pareto II tail. E($(h-\beta)$) indicates the posterior expectation of the normalized difference $(h-\beta)$. Pr($(G_B > G_y)$) indicates the posterior probability that the Bayesian corrected Gini is greater than the measured data Gini without Pareto II correction. For each country, the first line is for 2008 and the second line for 2018.

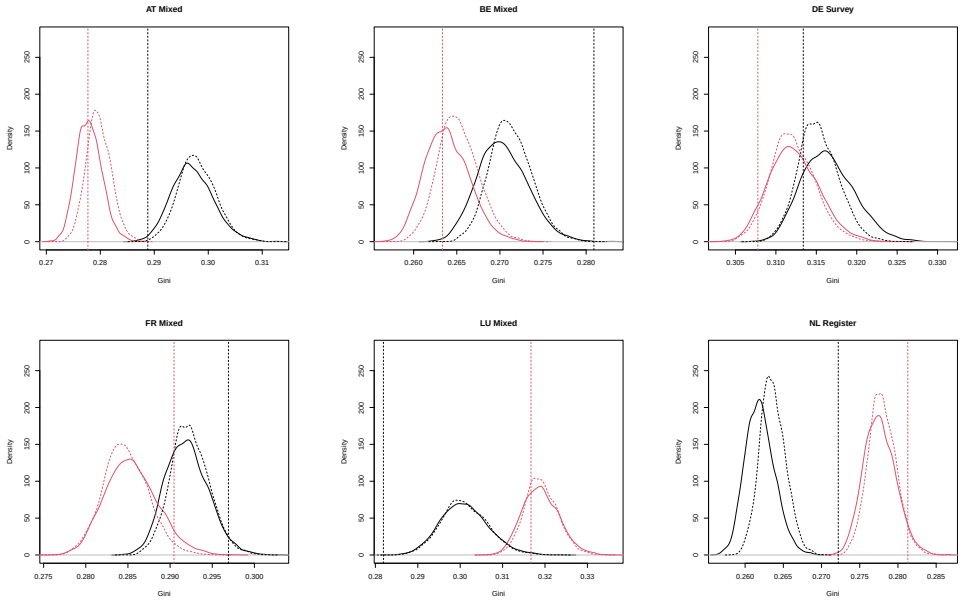


Black lines correspond to 2008 and red lines to 2018. Plain lines are for the Pareto II and dotted lines for the Pareto I. Vertical dotted lines represent the sample Gini computed with weights.

Figure 7: Posterior density of the Gini index: Nordic countries

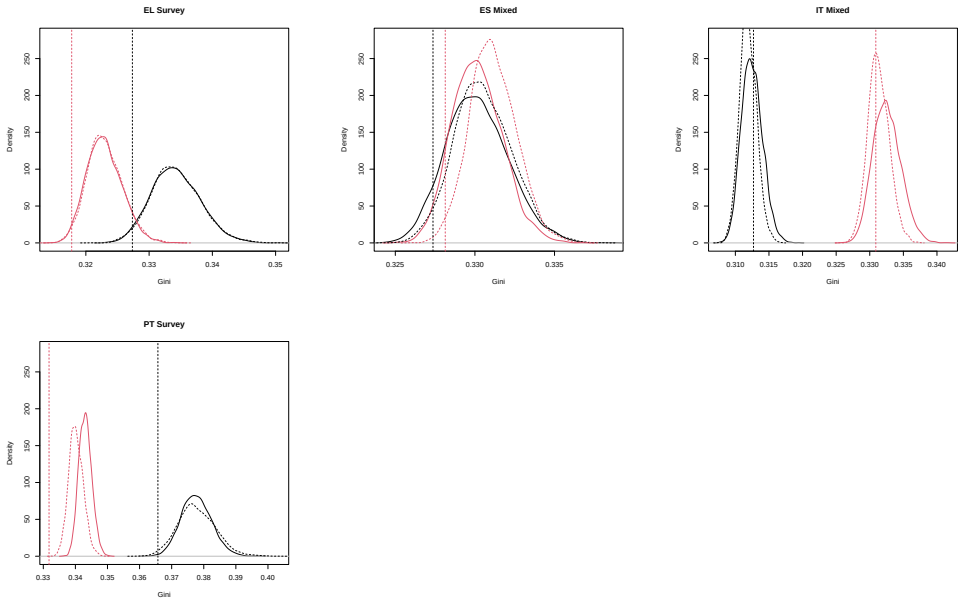
6 Conclusion

By embedding the Pareto tail into a larger model, we could estimate the threshold in a sound and logical way. However, for doing this, we had to use a semi-parametric approach for the central model, as a simple log-normal would have led to biased results. We have shown that the threshold posterior expectations can be quite diverse, even for the same country, as we gave estimates for two years. These estimates gave a value at which the Pareto II tail adjusted the best. They could range between 0.882 and 0.95, on average. We did not report standard deviations, they are in general not very large. For instance, if the posterior expectation is 0.95,



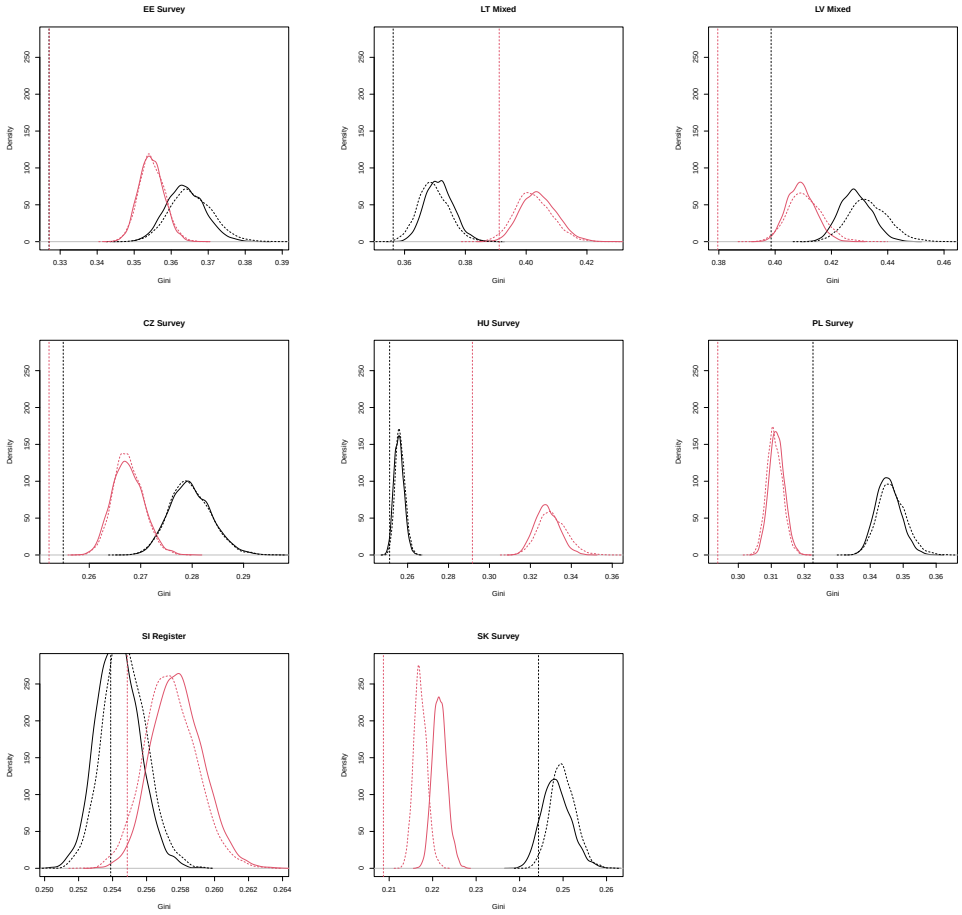
Black lines correspond to 2008 and red lines to 2018. Plain lines are for the Pareto II and dotted lines for the Pareto I. Vertical dotted lines represent the sample Gini computed with weights.

Figure 8: Posterior density of the Gini index: Northern Europe



Black lines correspond to 2008 and red lines to 2018. Plain lines are for the Pareto II and dotted lines for the Pareto I. Vertical dotted lines represent the sample Gini computed with weights.

Figure 9: Posterior density of the Gini index: Southern Europe



Black lines correspond to 2008 and red lines to 2018. Plain lines are for the Pareto II and dotted lines for the Pareto I. Vertical dotted lines represent the sample Gini computed with weights.

Figure 10: Posterior density of the Gini index: New Member States

a 95% credible interval would be around [0.94, 0.96]. But, we have provided the plots of the posterior density of the corrected Gini, posterior densities that were quite concentrated.

Using a Pareto II model provides interesting information on the shape of the tail of the income distribution. In particular, the posterior expectation of $h - \beta$ indicates if there are important outliers or if very rich people are really absent from the sample. Most of the posterior expectations of this normalized quantity were negative for the new member states and also for Portugal, indicating the real need of a correction for missing information on high incomes.

Our attempt to correct for missing information on high incomes implies the modelling of an underlying income distribution. We had results for two years, which means that we can follow how quantiles have evolved over time. Moreover, we have these results for a group of countries, belonging to the same economic space. So it is tempting to derive the Growth Incidence Curve for these countries and what was the impact of integrating the NMS on the European Income Distribution, when

taking into account the missing information on high incomes. This is planned for future work.

References

Alvaredo, F. (2011). A note on the relationship between top income shares and the Gini coefficient. *Economics Letters*, 110:274–277.

Alvaredo, F., Atkinson, A., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2016). Distributional national accounts guidelines: Methods and concepts used in the world inequality database. *WID.world Working Paper*, 2016(2).

Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2019). What did you really earn last year? Explaining measurement error in survey income data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 182(4):1411–1437.

Arnold, B. C. (2008). Pareto and generalized Pareto distributions. In Chotikapanich, D., editor, *Modeling Income Distribuions and Lorenz Curves*, volume 5 of *Economic Studies in Equality, Social Exclusion and Well-Being*, chapter 7, pages 119–145. Springer, New-York.

Arnold, B. C. (2015). *Pareto Distributions*. Chapman and Hall, New York.

Atkinson, A. (2005). Top incomes in the UK over the 20th century. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):325–343.

Atkinson, A. (2007). Measuring top incomes: Methodological issues. In Atkinson, A. and Piketty, T., editors, *Incomes over the Twentieth Century: a Contrast Between Continental European and English-Speaking Countries*, chapter 2, pages 18–42. Oxford University Press, Oxford.

Atkinson, A. B. (2017). Pareto and the upper tail of the income distribution in the UK: 1799 to the present. *Economica*, 84(334):129–156.

Babu, G. J., Canty, A. J., and Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, 105(2):377–392.

Bartels, C. and Metzing, M. (2019). An integrated approach for a top-corrected income distribution. *Journal of Economic Inequality*, 17:125–143.

Bauwens, L. and Lubrano, M. (1998). Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal*, 1:C23–C46.

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4:227–244.

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

Blanchet, T., Flores, I., and Morgan, M. (2022a). The weight of the rich: improving surveys using tax data. *The Journal of Economic Inequality*, 20(1):119–150.

Blanchet, T., Fournier, J., and Piketty, T. (2022b). Generalized Pareto curves: theory and applications. *Review of Income and Wealth*, 68(1):263–288.

Bourguignon, F. (2018). Simple adjustments of observed distributions for missing income and missing people. *The Journal of Economic Inequality*, 16:171–188.

Cabras, S. and Castellanos, M. E. (2011). A Bayesian approach for estimating extreme quantiles under a semiparametric mixture model. *ASTIN Bulletin: The Journal of the IAA*, 41(1):87–106.

Charpentier, A. and Flachaire, E. (2022). Pareto models for top incomes and wealth. *The Journal of Economic Inequality*, 20(1):1–25.

Clementi, F. and Gallegati, M. (2016). *The Distributionof Income and Wealth: Parametric Modeling with the k-Generalized Family*. Springer International Publishing Cham, Switzerland.

Clementi, F., Gallegati, M., and Kaniadakis, G. (2012). A new model of income distribution: the κ-generalized distribution. *Journal of Economics*, 105(1):63–91.

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

Cowell, F. A. (2011). *Measuring Inequality*. Oxford University Press, Oxford, third edition.

Flachaire, E., Lustig, N., and Vigorito, A. (2022). Underreporting of top incomes and inequality: A comparison of correction methods using simulations and linked survey and tax data. *Review of Income and Wealth*, 69(4):1033–1059.

Hajargasht, G. and Griffiths, W. (2013). Pareto-lognormal distributions: Inequality, poverty, and estimation from grouped income data. *Economic Modelling*, 33:593–604.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica*, 84(334):261–289.

Kaniadakis, G. (2013). Theoretical foundations and mathematical formalism of the power-law tailed statistical distributions. *Entropy*, 15(10):3983–4010.

Lustig, N. (2020). The missing rich in household surveys: Causes and correction approaches. *ECINEQ Working Paper No. 2020-520*.

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

Majid, M. H. A. and Ibrahim, K. (2021b). On Bayesian approach to composite Pareto models. *PLoS ONE*, 16(9).

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

Pareto, V. (1896). *Cours d'économie politique: professé à l'Université de Lausanne*, volume 1. F. Rouge, Lausanne.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3:119–131.

Reed, W. J. and Jorgensen, M. (2004). The double Pareto-Lognormal distribution: A new parametric model for size distributions. *Communications in Statistics - Theory and Methods*, 33(8):1733–1753.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10:33–60.

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

Singh, S. K. and Maddala, G. S. (1976). A function for size distribution of incomes. *Econometrica*, 44(5):963–970.

Vitale, R. A. (1975). A Bernstein polynomial approach to density function estimation. In Puri, M. L., editor, *Statistical Inference and Related Topics*, pages 87–99. Academic Press, New-York.

Appendix A Classical inference for the Pareto II

For classical inference, Arnold (2008) considers estimating the threshold by $\hat{h} = x_{[1]}$ and then solving numerically the normal equations of the likelihood function. Using exogenous weights w_i summing to n, the full likelihood is:

$$L(x;\theta) = \prod_{i=1}^{n} f(x_i;h,\beta,\alpha)^{w_i}.$$

The log-likelihood:

$$\begin{aligned} l(x;\theta) &= \sum_{i=1}^{n} w_i \log f(x_i;h,\beta,\alpha) && (21)\\ &= -(\alpha+1)\sum_{i=1}^{n} w_i \log\left(1+\frac{x_i-h}{\beta}\right) - n\log\beta + n\log\alpha, && (22)\end{aligned}$$

leads to the normal equations:

$$\begin{aligned} \hat{\beta} &= \frac{\hat{\alpha}+1}{n}\sum w_i(x_i-x_{[1]})\left[1+\frac{x_i-x_{[1]}}{\hat{\beta}}\right]^{-1}, && (23)\\ \hat{\alpha} &= \left[\frac{1}{n}\sum w_i \log\left(1+\frac{x_i-x_{[1]}}{\hat{\beta}}\right)\right]^{-1}. && (24)\end{aligned}$$

A method of moments can also be implemented, using the translated raw moments m_1 and m_2 and equating them to their theoretical counterparts. We start from the definition of the weighted sampling moments:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} w_i (X_i - X_{[1]})^r \ , \ r = 1,2,$$

Arnold (2015, page 255) proposes the following estimator for β:

$$\hat{\beta} = m_1 m_2/(m_2 - 2m_1^2), \quad (25)$$

from which we can deduce an estimator for α, using the normal equation:

$$\hat{\alpha} = n/\sum_{i=1}^{n} w_i \log(1+(x_i-x_{[1]})/\hat{\beta}). \quad (26)$$

Finally, let us recall that the MLE estimate of the Pareto coefficient in the Pareto I process is simply given by:

$$\hat{\alpha}_{PI} = n/\sum_{i=1}^{n} w_i \log(x_i/x_{[1]}). \quad (27)$$

Appendix B Bernstein polynomials for density estimation

Vitale (1975) was the first to propose a density estimator based on Bernstein polynomials. Let us suppose that we have n observations with distribution $f(x)$ from

which we form histogram values of $k+1$ bins. Let x_j be the centre of each class and n_j the corresponding frequencies. A semi-parametric estimator of the density is then formed by a polynomial approximation of the empirical function described by the $k+1$ couples (x_j, n_j).

We propose here another Bernstein density estimator, where the coefficients of the polynomial approximation are obtained by a regression. Let us first recall the expression of a Bernstein polynomial defined for $x \in [0,1]$:

$$B_k(x,j) = C_k^j x^j (1-x)^{k-j}, \tag{28}$$

where C_k^j is the binomial coefficient. This polynomial has, among many, the properties that $\sum_j B_k(x,j) = 1$ and $B_k(x,j) \geq 0$. If the range of x is $[a,b]$, then we can always use the transformation $y = (x-a)/(b-a)$ and use $B_k(y,j)$ instead of $B_k(x,j)$. The estimator proposed by Vitale (1975) corresponds to:

$$\hat{f}_{n,k}(x) = (k+1)\sum_{j=0}^{k} \frac{n_j}{n} B_k(x,j). \tag{29}$$

We propose to approximate the coefficients in (29) by using a regression of the log of the vector of the histogram frequencies nj over $B_k(xj,j)$ where xj is the vector of the cell midst x_j with the advantage of choosing the degree k of the Bernstein polynomial independently of the number of cells of the histogram:

$$\log(nj) = B_k(xj,0)\delta_0 + \cdots + B_k(xj,k)\delta_k + \epsilon.$$

Calling $\hat{\delta}_j$ the estimated regression coefficients, the new density estimator is:

$$\hat{f}_{n,k}(x) = \exp(\sum_{j=0}^{k} B_k(x,j)\,\hat{\delta}_j). \tag{30}$$

It has to be normalized to one by numerical integration. Using a regression on the logs and then predicting the exponential is a way to impose the positivity of the density estimate.

The same approach can be used for estimating the CDF. Let us assume that the vector of the n values of x has been sorted and let $Fn = (1, \cdots, n)/(n+1)$.[8] The estimator proposed by Babu et al. (2002):

$$\hat{F}_{n,k}(x) = \sum_{j=0}^{k} F_n(j/k) B_k(x,j),$$

has the same problem as before, that we solve by introducing a new regression. Because an estimated cumulative is not only a positive, but also an increasing function of x, we have to impose this supplementary restriction inside the new regression. For this, we use a logistic regression, obtained by regressing the log of $(1-F_n)/F_n$ over the Bernstein basis $Z_k(x) = [B_k(x,j)]$ with:

$$\log[(1-F_n)/F_n] = Z_k(x)\delta + \epsilon.$$

[8] If we have weights w summing to n, then $Fn = \texttt{cumsum}(w)/(n+1)$ where `cumsum` in the operator giving the cumulative sum.

The estimated CDF is then obtained by the inverse transformation with:

$$\hat{F}_{n,k}(x) = \frac{1}{1 + \exp(Z_k(x)\hat{\delta})}. \tag{31}$$

Remains the question of the range of x which is not [0,1] in empirical applications. This time, we use the following logistic transformation of the x, $y = 1/(1+\exp(x/\bar{x}))$, the initial transformation $y = (x-a)/(b-a)$ producing unsatisfactory results at the top of the distribution.

In Table 4, we provide the error committed when estimating the 95^{th} percentile with the EU-SILC data.

Table 4: Error in percentage for estimating a 0.95 truncating point

Country	Bernstein	Log-normal	Country	Bernstein	Log-normal
DK	0.04	9.35	ES	0.48	9.31
FI	0.23	2.06	IT	0.21	9.47
SE	0.16	20.73	PT	0.46	1.49
IE	0.16	2.76			
UK	0.22	2.08	EE	0.17	6.93
AT	0.19	32.64	LT	0.49	2.93
BE	0.14	8.85	LV	0.78	2.25
DE	0.23	7.43	CZ	0.09	0.92
FR	0.16	2.24	HU	0.34	1.41
LU	0.27	1.04	PL	0.24	0.68
NL	0.20	1.16	SI	0.09	7.55
EL	0.20	3.97	SK	0.13	4.69