Bayesian correction for missing rich using a Pareto II tail with unknown 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 correction for missing rich using a Pareto II tail with unknown threshold: Combining EU-SILC and WID data

Mathias Silva
Michel Lubrano

WP 2023 Nr 20

Bayesian correction for missing rich using a Pareto II tail with unknown threshold: Combining EU-SILC and WID data*

Mathias Silva† and Michel Lubrano‡

August 2023

Abstract

Survey data are known for under-reporting rich households while providing large information on contextual variables. Tax data provide a better representation of top incomes at the expense of lacking any contextual variables. So the literature has developed several methods to combine the two sources of information. For Pareto imputation, the question is how to chose the Pareto model for the right tail of the income distribution. The Pareto I model has the advantage of simplicity. But Jenkins (2017) promoted the use of the Pareto II for its nicer properties, reviewing three different approaches to correct for missing top incomes. In this paper, we propose a Bayesian approach to combine tax and survey data, using a Pareto II tail. We build on the extreme value literature to develop a compound model where the lower part of the income distribution is approximated with a Bernstein polynomial truncated density estimate while the upper part is represented by a Pareto II. This provides a way to estimate the threshold where to start the Pareto II. Then WID tax data are used to build up a prior information for the Pareto coefficient in the form of a gamma prior density to be combined with the likelihood function. We apply the methodology to the EU-SILC data set to decompose the Gini index. We finally analyse the impact of top income correction on the Growth Incidence Curve between 2008 and 2018 for a group of 23 European countries.

Keywords: Bayesian inference, Pareto II, profile likelihood, Bernstein density estimation, top income correction, EU-SILC
JEL codes: C11, D31, D63, I31

*This work was supported by the French National Research Agency Grant ANR-17-EURE-0020, and by the Excellence Initiative of Aix-Marseille University - A*MIDEX. We have benefitted 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 prepared for the 2023 ESOBE conference in Glasgow, September 1-2, 2023.

†ENS Lyon, France and Aix Marseille Univ, CNRS, AMSE, Marseille, France
‡Aix Marseille Univ, CNRS, AMSE, Marseille, France

1 Introduction

Pareto models are convenient for modelling the right tail of an income distribution, for generating missing rich and for interpolating empirical quantiles when grouped data are used. The Pareto I model as been widely used in the literature (Atkinson 2005, Atkinson 2007, Cowell and Flachaire 2007, Bartels and Metzing 2019 to quote just a few) as it is easy to estimate when the threshold parameter is known and because Pareto interpolation is linear. However the Pareto I model is known to be rather restrictive as discussed for instance in Jenkins (2017). It leads to biased estimates if the truncation point is too low. The Generalised Pareto Distribution (GPD) introduced by Pickands (1975) or the Pareto II (the useful restriction of the GPD for income distributions) appears as an interesting alternative, obtained by adding just one parameter. In particular is leads to estimates that are less sensitive to the choice of the truncation point, but at the expense of a larger variance (see e.g. Jenkins 2017 or Charpentier and Flachaire 2022), at least in a classical framework. It covers the Pareto I as a particular case. Jenkins (2017) found that it was sometimes difficult to differentiate the two models for a given data set in term of fit, but let us note that the two models can be compared by testing for a parametric restriction. In this paper, we shall use the Pareto II for modelling the right tail of the income distribution.[1]

The next question is to know on which type of data to adjust a Pareto tail. Survey data, either raw or tabulated, are well-known for not reporting properly high incomes, either because of under-sampling which is a default of the sampling scheme itself (not enough rich households are included in the sampling design) or because of under-reporting when rich households are sampled but, either refuse to provide their true income, or report a lower value than their actual income. Tax data are most of the time tabulated and report more precisely rich households because it is difficult to escape the tax authorities even if tax avoidance and tax evasion are well-known phenomenons. However, if tax records are better equipped for reporting high incomes, they neglect low income holders who anyway do not pay taxes. This question is illustrated in Atkinson (2005) who makes use of the supertax data for the UK over the 20th century. One of the problems he had to face was to reconstruct the uncovered lower incomes as he wanted to compute the

[1]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-Lognormal or double Pareto-Lognormal distributions of Reed and Jorgensen (2004) used in Hajargasht and Griffiths (2013).

top 1% income share. So it is a problem to match exactly the two types of data sources.

Modelling the right tail of an income distribution using a Pareto model appears as one solution among others to the the question of under-sampling of rich households in sample surveys. This is what Jenkins (2017) call *method A* when his *methods B* and *method C* rely on the use of external information provided by administrative or tax data. With *method A*, when modelling the right tail of the income distribution, extreme observations are no longer considered as outliers, but as observations belonging to an under-sampled tail. However, Jenkins (2017) notes that using only one source of information is not enough and that tax return data are a precious external source of information that should not be neglected to correct for under-reporting. So, we have to find a way to combine these two sources of information with an appropriate matching device. The first aim of this paper is to propose a Bayesian approach to combine the two sources of information. We first use tax records to build up a prior information on a Pareto coefficient. We then model the right tail of survey data for the income distribution with a Pareto II model and use an informative prior on the Pareto coefficient.

An important question discussed amply in Jenkins (2017) is the determination of the threshold at which to start correcting for missing rich and use a Pareto tail. The economic literature is rather scarce and contradictory on the question. Bartels and Metzing (2019) opt for top 1% on the ground that it is the level at which survey and tax data start to diverge in their estimation of income shares in Germany. Jenkins (2017) performs some sensitivity analysis on the range 10% till 1%. On the contrary, but still confronting survey and administrative data, both Flachaire et al. (2022) (for Uruguay) and Angel et al. (2019) (for Austria) found that the correction should start at a much lower value, around the median, but this time independently of the question of estimating a Pareto tail. It is hard to find explanations to this puzzle, except that we have to be very careful on the data sources that we are comparing. A fact remains, 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, an approach that we shall adopt here. Jenkins (2017) quotes several graphical methods such as Pareto plots, mean excess plots or Zenga curves to which we can add Hill plots. These methods are not very precise according to the survey of Scarrott and MacDonald (2012). The extreme value literature provides more convincing methods with parametric and semi-parametric bulk mod-

els, where the lower part of the distribution follows for instance a parametric lognormal distribution when the upper part is modelled according to a Generalised Pareto. In this model, the threshold becomes a parameter to estimate and the uncertainty associated to its estimation can be measured. In order to limit the bias when estimating the threshold, Cabras and Castellanos (2011) propose to use a semi-parametric estimator for the lower part when the upper part is represented by a Generalised Pareto. A second aim of this paper is to use a semi-parametric bulk model, based on a Bernstein polynomial, to determine the value of h that leads to the best fit.

The third aim of the present paper is to combine the bulk model from the extreme value theory with the approach developed in Jenkins (2017) to correct inequality measures such as the Gini index in a Bayesian framework to finally explore the consequences of these corrections on the shape of the Growth incidence curves using EU-SILC data for a group of 23 European countries. We shall adopt one specific bulk model of the extreme value literature, more specifically a variant of one of the model detailed in Cabras and Castellanos (2011). In order to remove the constraint imposed by the lognormal for modelling the bulk of the income distribution, Cabras and Castellanos (2011) use a *profile likelihood* based on the method of Lindsey (1974) which proposes to approximate a truncated distribution (the bulk of the distribution below the threshold) by a polynomial. While Cabras and Castellanos (2011) use orthogonal polynomials, we propose in this paper to use a Bernstein polynomial approximation that we think is more adapted for future Bayesian developments and for which we have the experience acquired in Fourrier-Nicolai and Lubrano (2023). For the upper part, we adopt the the Pareto II. With this model, we will be able to make inference on the value of the threshold, using survey data. But the resulting value will correspond to the point where the survey data actually follow a Pareto tail, independently of the point where survey data need to be corrected because of under-sampling or under-reporting.

Adopting a Bayesian approach will be fruitful for three reasons. In a bulk model, the likelihood function is numerically intractable because it not differentiable and because imposing differentiability would result in a too much restricted model. It would be very difficult to provide an efficient method of maximum likelihood in this context. The second reason is that with Bayesian inference it is easy to obtain posterior confidence intervals. Finally and most importantly, Bayesian inference is the natural way of incorporating extra information when this information is scarce and uncertain. The main source of tax data is provided by the World Inequality Data base (WID). However, this information (the variable `fiinc` in the WID data base) is not available for every country and every year. Nevertheless, we can construct for each coun-

try a prior information on the Pareto coefficient computed from top income shares (the variable `scainc` in the WID data base) in order to determine the parameters of a gamma prior on the Pareto coefficient. This is a new way of combining two sources of information. From a bulk model estimated on EU-SILC data, incorporating prior information coming from the WID, we shall propose a Bayesian counterpart to the decomposition of the Gini index adopted in Jenkins (2017) in order to provide a new estimation of inequality in Europe. As a by-product we also obtain a corrected evaluation of a Growth Incidence Curve (GIC) to visualise which quantiles profited from growth between 2008 and 2018 and what is the impact of the missing rich on the shape of the GIC.

The paper is organised as follows. In section 2, we review the properties of the Pareto I and II models and classical inference including weights. We introduce the class of compound lognormal-Pareto II models and review other options for imbedding a Pareto tail in a general model. We finally compare the performance of the two models for estimating inequality in a small Monte Carlo experiment to justify the choice of a Pareto II tail. With section 3, we detail Bayesian inference for the Pareto II with an informative prior on the Pareto coefficient when h is known. We then introduce a profile likelihood based on Bernstein polynomials to make inference on h and present a general algorithm for inference. We finally detail how to decompose the Gini in a Bayesian framework when h is random. In section 4, we apply the method on the EU-SILC data. We show how to build a prior information, using the WID data. With section 5, we extend the method so as to consider the Growth Incidence Curve, based on the quantile function of the compound model. We present the GIC for a group of 23 European countries and analyse the impact of the correction on the general shape of this curve. Section 6 concludes.

2 Pareto models

Two main models are commonly used in the empirical literature for modelling high incomes: the simple Pareto I and the Pareto II which was progressively introduced (see e.g. Jenkins 2017) because of the limitations of the Pareto I. The Generalised Pareto model of Pickands (1975) appears now and then. It is slightly more general than the Pareto II as its exponent $\xi \in [-1, +\infty[$. The Pareto II corresponds to $\xi > 0$, the exponential to $\xi = 0$ and the uniform to $\xi = -1$. The Pareto II is recovered after a re-parameterisation (see Appendix B).

2.1 Pareto I

The well-known Pareto I is a good starting point to explain the need of a more general model for high incomes. For $0 < h \leq x$, its cdf and pdf are:

$$F(x) = 1 - (x/h)^{-\alpha}, \qquad f(x) = \alpha h^{\alpha} x^{-\alpha-1}.$$

Taking logs and rearranging terms of $F(x)$, we get the expression of a linear expression:

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

a fundamental characteristics of the Pareto I process. Plotting $\log(1 - \hat{F}(x))$ against $\log x$, we get the Pareto diagram as named by Cowell (2011). It should produce a straight line with a negative slope if the data follow a Pareto I distribution.

Example 1 *Let us check this using the French EU-SILC data for 2018. Figure 1 shows that the Pareto assumption can be assumed for high incomes, gross and disposable, starting over their respective medians. However, things are not so clear for very top incomes. Charpentier and Flachaire (2022) would interpret the last points as true outliers (see their Figure 7). However, we should note that the interpretation of the Pareto plot relies strictly on a Pareto I assumption. Taking the logs and rearranging the terms of $F(x)$ for the Pareto II does not lead to a linear relation. We shall detail this point later.*

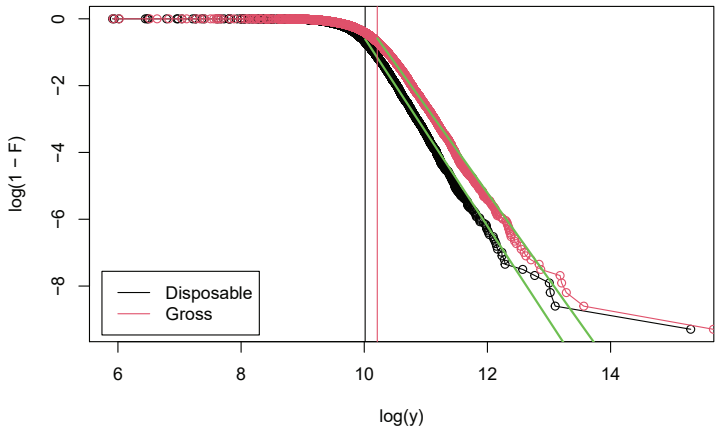

Figure 1: Pareto plot for French incomes in 2018

The Pareto I process is constrained because its conditional expectation is:

$$\mathrm{E}(x|X > h') = h' \frac{\alpha}{\alpha - 1}, \tag{1}$$

whatever the value of $h' > h$. This means that the average income over h' divided by h' is constant and equal to $b = \alpha/(\alpha - 1)$, a number called the inverted Pareto coefficient in the literature (Atkinson 2017). This means that inequality among the rich is constant, whatever the value of h'. The Gini coefficient depends solely on the value of α with:

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

The Generalised Pareto curve, detailed in Blanchet et al. (2022), shows that the assumption of a constant inequality over the quantiles is not empirically tenable, which justifies the need for a more complex model, the Pareto II being one of them.

2.2 Pareto II

Following Arnold (2008), the Pareto II process is built from taking h as a location parameter and introducing a separate scale parameter β, leading to:

$$F(x) = 1 - \left(1 + \frac{x - h}{\beta}\right)^{-\alpha}, \quad f(x) = \frac{\alpha}{\beta}\left(1 + \frac{x - h}{\beta}\right)^{-\alpha - 1}, \quad 0 \leq h \leq x.$$

The Pareto I corresponds to the restriction $h = \beta$. Additionally, the Pareto I models the distribution of relative excesses x/h whereas the Pareto II models the distribution of absolute excesses $x - h$ (see e.g. Charpentier and Flachaire 2022). The scaled conditional expectation is no longer constant as:

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

It depends on both h and h' as documented e.g. in Charpentier and Flachaire (2022). They conclude that while the Pareto I model implies a constant inequality among the rich, the Pareto II allows for a varying inequality among the rich.

Let us now give some complementary results for this interesting model that can be found in Arnold (2008). The mean is:

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

The Gini has a quite complex expression in the general case. Arnold (2008, page 135) gives the Gini for the Pareto IV. The Gini for the Pareto II is found by imposing a restriction which leads to (3):

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

where $B(,)$ is the Beta function.[2] When the data are translated, which means that we subtract their minimum to the original data, the expression of the Gini is very much simplified as indicated in Arnold (2008, p. 135). In this case, we impose $h = 0$ and the above formula is simplified into:

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

Consequently, the Pareto I and Pareto II measure inequality in a quite different way for a given value of the Pareto coefficient α. This is well depicted in Figure 2.

Example 2 *For a given value of $\alpha = 1.7$ 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 $h = 0$ and is equal to (4) in this case. 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).*

2.3 Classical inference for Pareto II using weights

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:

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

and the log-likelihood:

$$\begin{aligned} l(x; \theta) &= \sum_{i=1}^{n} w_i \log f(x_i; h, \beta, \alpha) && (5) \\ &= -(\alpha + 1) \sum_{i=1}^{n} w_i \log\left(1 + \frac{x_i - h}{\beta}\right) - n \log\beta + n \log\alpha, && (6) \end{aligned}$$

[2] Jenkins (2017) provides a similar formula, but for a different parameterisation.

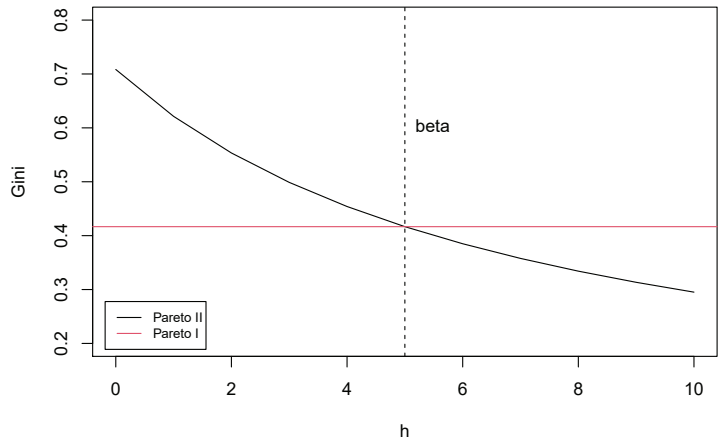

Figure 2: Gini constellation, varying h for a given $\beta = 5$ and $\alpha = 1.7$

leading to the normal equations:

$$\hat{\beta} = \frac{\hat{\alpha}+1}{n}\sum w_i(x_i - x_{[1]})\left[1 + \frac{x_i - x_{[1]}}{\hat{\beta}}\right]^{-1}, \quad (7)$$

$$\hat{\alpha} = \left[\frac{1}{n}\sum w_i \log\left(1 + \frac{x_i - x_{[1]}}{\hat{\beta}}\right)\right]^{-1}. \quad (8)$$

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:

$$\hat{\beta} = m_1 m_2/(m_2 - 2m_1^2), \quad (9)$$

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

$$\hat{\alpha} = n/\sum_{i=1}^{n} w_i \log(1 + (x_i - x_{[1]})/\hat{\beta}). \quad (10)$$

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

$$\hat{\alpha}_{PI} = n/\sum_{i=1}^{n} w_i \log(x_i/x_{[1]}). \quad (11)$$

2.4 The class of compound lognormal-Pareto II distributions

We have understood that deciding at which point should start the Pareto tail is a problem that we have skipped for the while, following the literature that considers h as fixed. So many authors had the idea of imbedding the Pareto model in a more general framework (see section 2.5). The framework of compound models provided by the extreme value theory is particularly useful for this purpose, due to its large flexibility, at least in the forms developed in the recent literature. The composite lognormal-Pareto model was introduced by Cooray and Ananda (2005) and Scollnik (2007). The lower part of a sample is modelled as a truncated lognormal while the upper part of the sample follows a Pareto I, a Pareto II or a Generalised Pareto distribution. This model was found useful to model extreme events in insurance claims, ecology and many other topics (see e.g. the references provided in Scollnik 2007, Cabras and Castellanos 2011 or Nadarajah and Bakar 2013). Some applications exist for modelling the income distribution with for instance Safari et al. (2018) or Abdul-Majid and Ibrahim (2021).

As the initial compound model of Cooray and Ananda (2005) and Scollnik (2007) was too restricted, it was later generalised to include more parameters so that Abdul-Majid and Ibrahim (2021) consider as a starting point the following model:

$$f(x|\theta) = \begin{cases} (1-\rho)\dfrac{f_\Lambda(x|\mu,\sigma^2)}{F_\Lambda(h|\mu,\sigma^2)}, & 0 < x < h, \\ \\ \rho f_{P2}(x|\alpha,\beta,h), & x \geq h. \end{cases} \tag{12}$$

In this writing, $f_\Lambda(x|\mu,\sigma^2)$ is the lognormal pdf with parameters μ and σ^2, $F_\Lambda(h|\mu,\sigma^2)$ the corresponding CDF, $f_{P2}(x|h,\alpha,\beta)$ is the pdf of the Pareto II with parameters α, β and h, the latter representing the cutting point between the two parts of the sample, while $0 < \rho < 1$ is the proportion of data coming from the Pareto II distribution. Without further restrictions, the overall pdf $f(x|\theta)$ presents an evident discontinuity at $x = h$. It can be made continuous by imposing a parametric restriction on ρ:

$$\rho(\theta) = \frac{\beta f_\Lambda(h|\mu,\sigma^2)}{\beta f_\Lambda(h|\mu,\sigma^2) + \alpha F_\Lambda(h|\mu,\sigma^2)}. \tag{13}$$

An alternative version of this model was first considered by Behrens et al. (2004). It corresponds to introducing the following simplifying assumption:

$$\rho = 1 - F_\Lambda(h|\mu,\sigma^2), \tag{14}$$

leading to the following data density:

$$f(x|\theta) = f_\Lambda(x|\mu,\sigma^2)\mathbb{1}(x<h)+(1-F_\Lambda(h|\mu,\sigma^2))f_{P2}(x|h,\alpha,\beta)\mathbb{1}(x\geq h), \quad (15)$$

where $\mathbb{1}(\cdot)$ is the indicator function. The CDF is given by (see e.g. Villa 2017):

$$\begin{aligned} F(x|\theta) &= F_\Lambda(x|\mu,\sigma^2)\mathbb{1}(x<h) + F_\Lambda(h|\mu,\sigma^2)\mathbb{1}(x\geq h) \\ &+ (1-F_\Lambda(h|\mu,\sigma^2)F_{P2}(x|h,\alpha,\beta)\mathbb{1}(x\geq h). \end{aligned} \quad (16)$$

This model is no longer exactly a mixture model as it is not the weighted sum of two densities, but a truncated lognormal distribution, the right tail of which has been replaced by a Pareto II. This model is called the bulk model. It is used in the extreme value literature with the aim of making inference for top quantiles and for h (see e.g. Cabras and Castellanos 2011, do Nascimento et al. 2012 or Villa 2017), using most of the time a Bayesian approach. Continuity is usually not imposed as this would mean a restriction on the Pareto II parameters. Nevertheless, we can note that the continuity restriction is, as indicated in Abdul-Majid and Ibrahim (2021):

$$\beta = \alpha\frac{1-F_\Lambda(h|\mu,\sigma^2)}{f_\Lambda(h|\mu,\sigma^2)}. \quad (17)$$

So we have five parameters without the continuity restriction and a Pareto II tail (this number is reduced by one if we impose a continuity restriction). To summarise, with the compound Pareto model, we have a framework where we are free to choose the shape of the bulk of the distribution (here the lognormal for a simplified exposition) and the shape of the tail which can be Pareto I, Pareto II or Generalised Pareto. The threshold parameter h can be estimated in this framework.

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 lognormal 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* $\beta = 3$*. The Pareto II right tail is well above the lognormal tail. But imposing continuity would mean* $\beta = 1.605324$*, a restriction that lowers the position of the Pareto II tail.*

2.5 Other imbedding models

Several other models are said to have a Pareto-like tail. Limiting ourselves to three parameter distributions, the first candidate is of course the famous

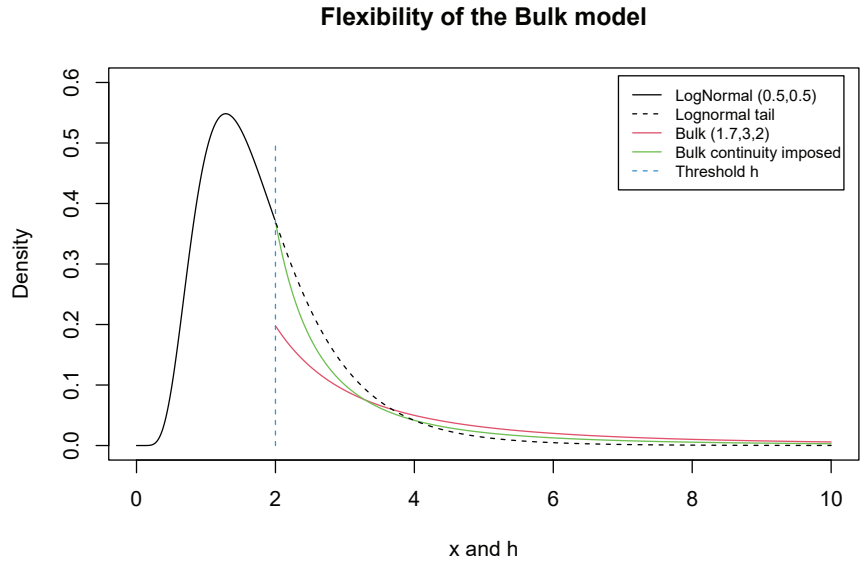

Figure 3: The Compound Pareto II family

Singh and Maddala (1976) distribution with:

$$F(x) = 1 - \frac{1}{(1+(x/b)^a)^q}, \qquad f(x) = \frac{aq}{b}\frac{(x/b)^{a-1}}{(1+(x/b)^a)^{q+1}}.$$

For $a = 1$, we have the Pareto II distribution. The Gini coefficient is given by:

$$G = 1 - \frac{B(q, 2q - 1/a)}{B(q - 1/a, 2q)},$$

where $B(a, b)$ is the incomplete Beta function. The success of the Singh-Maddala distribution is explained by the possibility of a Pareto II tail. But the latter is obtained by imposing a parametric restriction. So there is a strict dependence between the shape of the bulk of the distribution and the shape of the tail. With the compound Pareto II distribution, we do not have such a restriction.

The properties of the Pareto-lognormal distribution of Reed and Jorgensen (2004) were amply discussed in Hajargasht and Griffiths (2013) as well as estimation procedures for grouped data. Let us define the transformed variable $z = (\log x - m)/\sigma$ and let Φ and ϕ be the CDF and pdf of the Gaussian distribution. Let us also define the function $R(t) = [1 - \Phi(t)]/\phi(t)$. Then the Pareto-lognormal distribution is:

$$F(x) = \Phi(z) - \phi(z)R(\alpha\sigma - z), \qquad f(x) = \frac{\alpha}{x}\phi(z)R(\alpha\sigma - z).$$

It has a Pareto I tail of the form $x^{-\alpha-1}$ for $x \to \infty$ (see e.g. Reed 2003). It was found to provide a very good fit by Hajargasht and Griffiths (2013) and totaly comparable to those of the double Pareto-lognormal or of the GB2. The Gini coefficient is:

$$G = \frac{2\exp(\alpha(\alpha-1)\sigma^2)}{2\alpha-1}\Phi\left(\frac{(1-2\alpha)\sigma}{\sqrt{2}}\right) + 2\Phi(\sigma/\sqrt{2}) - 1.$$

The right tail of the Pareto-lognormal distribution is directly a Pareto I, without a parametric restriction, but only for very large values of x. So we do not know where the Pareto tail starts exactly.

Recently, the Kaniadakis distribution, or κ-Generalised Distribution (see e.g. Clementi and Gallegati 2016 for details and complementary references) has gained the reputation of fitting the income data better than the Singh-Maddala distribution while keeping the same advantages of analytical results, a Pareto-like tail and the possibility of a zero-mode (see the empirical results in Clementi et al. 2012 using the GSOEP, BHPS and PSID data sets). Let us first define generalised exponential and logarithmic functions as $\exp_\kappa(x) = (\sqrt{1+\kappa^2x^2} + \kappa x)^{1/\kappa}$ and $\log_\kappa(x) = (x^\kappa - x^{-\kappa})/(2\kappa)$ with $\kappa > 0$. Then, the CDF and the pdf are:

$$F(x) = 1 - \exp_\kappa -(x/\beta)^\alpha, \qquad f(x) = \frac{\xi}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\frac{\exp_\kappa(-(x/\beta)^\alpha)}{\sqrt{1+\kappa^2(x/\beta)^{2\alpha}}}.$$

There is one scale parameter ($\beta > 0$) and two shape parameters, $\alpha > 0$ and $\kappa > 0$. The shape of the lower tail is governed by α with a zero-mode obtained when $0 < \alpha \leq 1$. The shape of the upper tail is governed by both α and κ. The Gini coefficient, as given in Clementi and Gallegati (2016, page 29) is:

$$G = 1 - \frac{2\alpha + 2\kappa\Gamma(\frac{1}{\kappa} - \frac{1}{2\alpha})\Gamma(\frac{1}{2\kappa} + \frac{1}{2\alpha})}{2\alpha + \kappa\Gamma(\frac{1}{\kappa} + \frac{1}{2\alpha})\Gamma(\frac{1}{2\kappa} - \frac{1}{2\alpha})}.$$

We can add for further use the quantile function:

$$Q(p) = \beta\left[\log_\kappa\left(\frac{1}{1-p}\right)\right]^{1/\alpha}.$$

We compare these three distributions in Figure 4.

Example 4 *We first draw a Pareto-lognormal, using the estimated parameters obtained for urban India in Hajargasht and Griffiths (2013). We then obtain the parameters for the Sing-Maddala and the Kaniadakis distributions by minimising their distance to the Pareto-lognormal. The estimated Gini coefficients are 0.388 for the Pareto-lognormal, 0.458 for the Singh-Maddala and 0.370 for the Kaniadakis distributions. The Kaniadakis distribution has the fatter tail and the Pareto-lognormal the thinnest tail.*

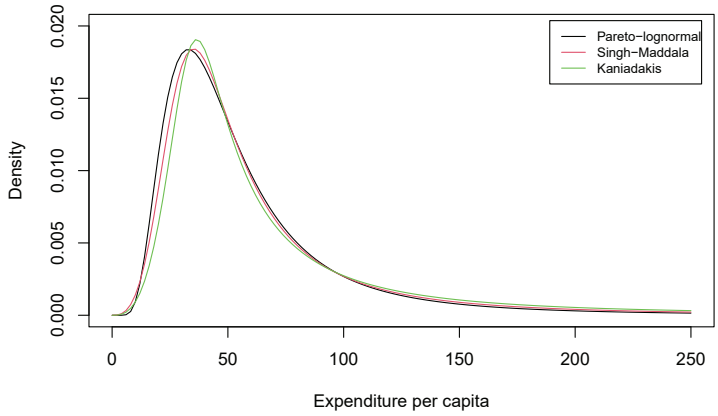

Figure 4: Comparing three distributions with Pareto tails

2.6 Comparing Pareto I and Pareto II tails

Jenkins (2017) discuss the difficulty of empirically distinguishing between the two Pareto processes, leading eventually to the question: Is a Pareto II model really needed to fit the data? To answer briefly this question, we have run a small Monte Carlo experiment.

Example 5 *The Kaniadakis distribution, as described above, seems to be the best candidate for simulating random numbers as it has the fattest tail and an analytical quantile function. We have generated* $m = 1,000$ *samples of size* $n = 1,000$ *with parameters* $\alpha = 1$ *to obtain a zero mode compatible with a Pareto shape and* $\kappa = 0.6$*. We have then estimated the sample Gini, the MLE estimate of the Pareto I coefficient with (11) and the same parameter for the Pareto II, using the moment estimators (9)-(10). From these estimates, we deduce the corresponding Gini with (2) and (4). The results are impressive as displayed in Figure 5. The data being generated with a zero-mode distribution, the parameters of the two Pareto processes could in theory be estimated using the whole sample. With the Pareto II, the estimated inequality with on average 0.621 is slightly lower than the sample inequality which is 0.642 on average. To get sensible results with the Pareto I, we had to estimate the Pareto parameter with the top 40% of the sample (the quantile 0.60 is on average the point at which a Pareto plot produces a straight line). We get an average Gini of 0.697. Otherwise, the obtained estimates of* α *are much lower than 0.5, which is the lower bound for computing the corresponding Gini. So the Pareto I model can greatly over-estimate inequality while the Pareto II*

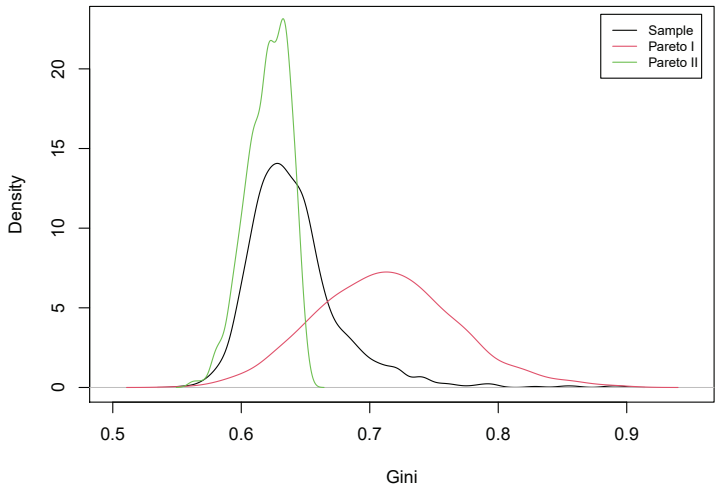

Figure 5: Measuring inequality with two different Pareto models, data generated with Kaniadakis distribution (1.0,1.0,0.6)

slightly under-estimate it, at least in this example. The second result is that the empirical distribution of the Gini is very concentrated for the Pareto II and very dispersed for the Pareto I model. These results militate in favour of using the Pareto II model for adjusting the right tail of income distributions.

Let us now come back to the interpretation of the Pareto plot in Figure 1 obtained for French incomes in 2018, especially the meaning of the extreme points which are over the straight line. Again, a small Monte Carlo experiment provides a nice answer.

Example 6 *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 in Figure 6. With* $h > \beta$*, extreme points are located above the Pareto line, a configuration that was qualified of outliers in Charpentier and Flachaire (2022, Figure 7). With* $h < \beta$*, we have the reverse situation. These two plots depict average situations, which can be repeated with the same final configuration, by running the same experiment several times.*

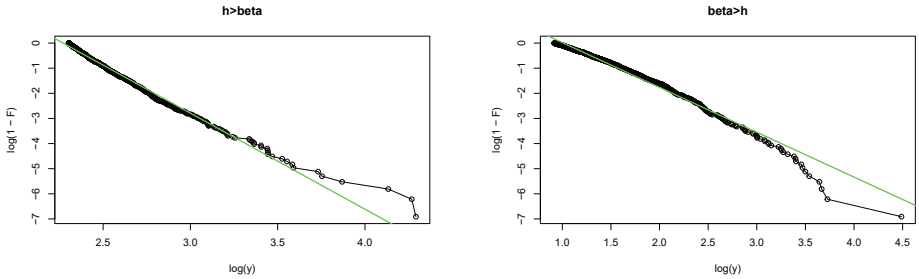

Figure 6: Comparing Pareto tails obtained from the Pareto II

3 Bayesian inference with a Pareto II tail

In order to cope with the missing rich, the economic literature has focussed on imputing Pareto top quantiles, but has given very little clues about how to select the truncation point. For instance Bartels and Metzing (2019) decided for all EU-SILC countries to replace the top 1% income by a Pareto I imputed income using the external information provided in the WID data set. Their choice for 1% was based on the fact that for Germany survey and tax income shares start to differ for the 1% income shares, but are quite similar below that threshold. Jenkins (2017) performed a sensitivity analysis, using different values for the threshold, between top 10% and top 1%. In this section, we will show how to correct for the missing rich in a Bayesian framework, when assuming that the top of the income distribution has a Pareto II tail. The correction will be operated by introducing an informative prior on the Pareto coefficient, building our prior information using the WID data base. We shall see how to make inference on h. The optimal h will correspond to a notion of best fit under the assumption of a Pareto II tail. We finally detail how to operate the decomposition of the Gini index in a Bayesian framework in order to measure the impact of our procedure on the measurement of inequality.

3.1 A Gibbs sampler when the threshold is known

We first present Bayesian inference for the Pareto II model when the threshold is known. The complete likelihood of the Pareto II using exogenous weights can be expressed as:

$$L(x; h, \beta, \alpha) = \beta^{-n}\alpha^n \exp -(\alpha+1)\sum w_i \log\left(1+\frac{x_i-h}{\beta}\right)\mathbb{1}(x_i \geq h).$$

Bayesian inference for the Pareto II model started with Arnold and Press (1983) who proposed an empirical Bayes approach, assuming that $h = x_{[1]}$, then translating the data defining $\tilde{x} = x - x_{[1]}$ and finally considering a model for $\tilde{x}$, assuming as a consequence that $h = 0$ in the original model. All papers in the following literature (Castellanos and Cabras 2007, Mokrani et al. 2016, Hu and Gui 2018, to quote a few) assume that h is known and fixed, reducing inference to a two parameter problem on the process for $\tilde{x}$. With this simplifying assumption, the posterior density of (α, β) does not belong to a known family, but the conditional posterior density of α is a gamma density. This suggests a Gibbs sampler with a Metropolis step and the possibility of a natural conjugate gamma informative prior on α. Let us note our informative prior as:

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

with prior expectation $\mathrm{E}(\alpha) = \nu_0/s_0$. It is with this informative prior density that we shall introduce our prior information coming from the WID data set. A non-informative prior would correspond to $\nu = 0$ and $s_0 = 0$ leading to $\varphi(\alpha) \propto 1/\alpha$. We have chosen a non-informative prior for β with:

$$\varphi(\beta) \propto \frac{1}{\beta}.$$

With our partially informative prior, the joint posterior density is proportional to:

$$\begin{aligned} \varphi(\alpha, \beta|x, w) &\propto \alpha^{n+\nu_0-1}\beta^{-n-1} \\ &\quad \exp\left(-\alpha[s_0 + \sum w_i \log(1 + \tilde{x}/\beta)] + \sum w_i \log(1 + \tilde{x}/\beta)\right). \end{aligned}$$

The two conditional distributions are found by discarding alternatively the proportional terms on which we condition. We get:

$$\begin{aligned} \varphi(\alpha|\beta, x, w) &\propto \alpha^{n+\nu_0-1} \exp\left(-\alpha[s_0 + \sum w_i \log(1 + \tilde{x}_i/\beta)]\right) \\ &= f_G\left(\alpha|\nu_0 + n, s_0 + \sum w_i \log(1 + \tilde{x}_i/\beta)\right), \qquad (18) \\ \varphi(\beta|\alpha, x, w) &\propto \beta^{-n-1} \exp\left(-(\alpha + 1) \sum w_i \log(1 + \tilde{x}_i/\beta)\right). \qquad (19) \end{aligned}$$

We can draw directly from the conditional $\varphi(\alpha|\beta, x)$. The question is to decide how to draw from $\varphi(\beta|\alpha, x)$. The literature is opting for a Metropolis step (Gilks et al. 1995) with different choices for the proposal. In fact, a Metropolis step is interesting when the density to simulate is multivariate.

Here, β is one-dimensional. So a either a rejection technique or the numerical inverse transformation method are well adapted. The former requires determining an envelope, the later requires determining an exploration grid.

We propose to implement an enriched version of the Griddy-Gibbs of Bauwens and Lubrano (1998). We use moment estimators to determine an initial value $\beta^{(0)}$ and a grid bp of k points on the range $[\beta^{(0)}/2, 2\beta^{(0)}]$. The moment estimator (9) serves to determine a plausible starting value $\beta^{(0)}$. Conditionally on $\beta^{(0)}$, we draw a first value $\alpha^{(1)}$ from the conditional posterior gamma density. Conditionally on $\alpha^{(1)}$, we evaluate $\varphi(\beta|\alpha, x)$ on the predetermined grid and derive a normalised empirical CDF. Then a sample from this empirical CDF is obtained in the logic of the inverse transformation method. For this purpose, we sample a random value from a uniform distribution over $[0, 1]$, determine its position in the CDF and then proceed by linear interpolation to determine the corresponding value $\beta^{(1)}$ on the predefined grid of k points. We regroup these results in Algorithm 1.

Algorithm 1 A Griddy-Gibbs algorithm for Bayesian inference on Pareto II models with fixed threshold h

1: Translate the data, subtracting the minimum and set $h = 0$
2: Initialise $\beta^{(0)}$ with a moment estimator
3: Determine an initial grid bp of np points $b_1, \cdots, b_{np}$ for β with unit interval dt
4: **for** $j = 1, \ldots, m$ **do**
5: Sample $\alpha^{(j)}$ from $f_G(\alpha|n + \nu_0, s_0 + \sum w_i \log(1 + \tilde{x}_i/\beta^{(j-1)})$
6: Evaluate the conditional posterior density

$$\varphi(\beta|\alpha^{(j)}, x) \propto \beta^{-n-1} \exp(-(\alpha^{(j)} + 1) \sum w_i \log(1 + \tilde{x}_i/\beta))$$

for β on the k points of the grid
7: Cumulate $\varphi(\beta|\alpha^{(j)}, x)$ using the Trapezoidal rule to obtain the cdf
8: Normalise the cdf by its integral
9: Draw $u \sim U(0, 1)$
10: Find the rank i of the first occurrence of $cdf > u$
11: Using linear interpolation, deliver $\beta^{(j)} = bp[i-1] + (u - cdf[i-1]) \times dt/(cdf[i] - cdf[i-1])$
12: **end for**
13: Discard the initial draws for computing posterior moments and posterior densities

As in every integration problem, scaling is of primary importance. The log of the conditional posterior of β is evaluated over the grid, scaled by sub-

tracting its maximum value on the grid, before taking the exponential. To be operational, the starting values for the grid have to be carefully chosen. The algorithm can be enriched by introducing a variable grid for β. After a period of initial *burn-in* draws, the mean and standard deviation of sampled $\beta^{(j)}$ values can be computed to adjust the grid. The grid is updated as an interval around the current mean of the already sampled $\beta^{(j)}$ plus or minus three times its standard deviation. This updating has to be done a limited number of times (say twice) and the draws used for the updating have to be discarded. Convergence can be checked using the normalised version of the CUMSUM plot of Yu and Mykland (1998) suggested in Bauwens and Lubrano (1998).

Remark 1 *We can note that for a given and fixed h, the posterior density of α in the Pareto I process would be also a gamma density with:*

$$\begin{aligned} \varphi(\alpha|x,h) &\propto \alpha^{\nu_0+n-1}\exp -\alpha(s_0+\sum w_i \log x_i/h) \\ &= f_G(\alpha|\nu_0+n, s_0+\sum w_i \log x_i/h), \end{aligned} \tag{20}$$

to be compared to (18).

3.2 A profile likelihood for making inference on h

Let us now turn to the case where we want to make inference on h, instead of working conditionally on it. For this purpose, we have introduced in section 2.4 the class of compound lognormal-Pareto II model, which consists in imbedding the Pareto II model in a more general model that also copes with the lower part of the sample. A simplified version of this general model is the bulk model (15) for which we shall now detail Bayesian inference, building on some of the solutions proposed in Cabras and Castellanos (2011). The likelihood function, when adding weights, is:

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

However, because the assumption of a truncated lognormal might be too restrictive, Cabras and Castellanos (2011) consider estimating the lower part of the density in a semi-parametric way, conditionally on h, leading to a *profile likelihood* function. This means that the parameter space is reduced to (α,β,h) and the profile likelihood is:

$$L_p(x;h,\alpha,\beta) = \prod_{i,x_i\leq h} \hat{f}_h(x_i)^{w_i} \hat{F}(h) \prod_{i,x>h} (1-\hat{F}(h)) f_{P2}(x|\alpha,\beta,h)^{w_i}, \tag{21}$$

where $\hat{f}_h(x_i)$ is a semi-parametric estimate of the truncated distribution, say $f(\cdot)\mathbb{1}(x \leq h)/\hat{F}(h)$ and $\hat{F}(h)$ the value of the estimated CDF at h, using the lower data points $\{x_i : x_i \leq h\}$.

For making inference on h, we have to specify a prior density $\varphi(h)$. We have chosen for simplicity a uniform prior between bounds:

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

This prior has been amply discussed in Abdul-Majid and Ibrahim (2021). In the Bulk model, there is a one-to-one relation between h and ρ with:

$$\rho = 1 - \hat{F}(h).$$

So, a prior on h can be translated into a prior on ρ, with the advantage that the latter is scaled-free and consequently is much easier to interpret. However, as noted in Abdul-Majid and Ibrahim (2021), a uniform prior on h does not mean a uniform prior on ρ.

We have to estimate $f(x)$ and $F(x)$, conditionally on the maximum a priori range, that is $[0, \overline{h}]$, $\overline{h}$ being the upper bound of the prior. There are several ways of estimating a truncated density. Cabras and Castellanos (2011) have chosen orthogonal polynomials. We have preferred to rely on our experience with Bernstein polynomials which provide a simple solution in the case of further developments (see a use of Bernstein polynomials in Fourrier-Nicolai and Lubrano 2023). A kernel estimate using a lognormal kernel could be an alternative. But remember that kernel estimators (even with a lognormal kernel) are not very precise for estimating the right tail of a distribution. And also they are not designed to estimate a density on a truncated range. We present in Appendix C an implementation of the method. The estimated density is noted $\hat{f}_k(x)$ and the corresponding CDF $\hat{F}_k(x)$, k being the degree of the Bernstein polynomial.

The conditional posterior distribution of α is just an extension of (18) to include h and thus corresponds again to a gamma density with:

$$\varphi(\alpha|h, \beta, x, w) = f_G(\alpha|n + \nu_0, s_0 + \sum_{i, x_i \geq h} w_i \log(1 + (x_i - h)/\beta)). \tag{22}$$

The conditional posterior distribution of β is an extension of (19) with:

$$\varphi(\beta|h, \alpha, x, w) \propto \beta^{-n-1} \exp(-(\alpha + 1) \sum w_i \log(1 + (x_i - h)/\beta)), \tag{23}$$

and can be simulated using the same procedure as in Algorithm 1. These two densities are evaluated on a sub-sample of x, conditionally on the drawn

value of h. The conditional distribution of h, on the contrary, depends on the whole sample with:

$$\varphi(h|\alpha,\beta,x,w) \propto \prod_{i,x_i\leq h} \hat{F}(h)\hat{f}_h(x_i)^{w_i} \times \prod_{i,x>h} (1-\hat{F}(h)) \left[\frac{\alpha}{\beta}(1+(x_i-h)/\beta)^{-\alpha-1}\right]^{w_i} \times \varphi(h). \quad (24)$$

We use a Griddy Gibbs to draw from this distribution, in the same spirit as the one used for drawing from the conditional posterior density of β.

Let us now regroup these results into Algorithm 2 to propose a Gibbs sampler for making inference on α, β and h. The update of the initial grids for β has to be done in the same cautious way as before.

Algorithm 2 Bayesian inference for α, β and h using a profile likelihood

1: Choose a prior range for h as a function of a prior range for ρ
2: Choose an initial h, e.g. $h^{(0)} = \text{quantile}(x, 0.90)$
3: Build the initial grid hp of np points $h_1, \cdots, h_{np}$ for h
4: Compute an initial estimate of α and β, conditionally on $h^{(0)}$, using a method of moments
5: Determine an initial grid bp of np points $b_1, \cdots, b_{np}$ for β
6: Select a value k for the degree of the Bernstein polynomial
7: Estimate the CDF of x, $\hat{F}_k(x)$, limited to the range corresponding to $\rho \in [0, 0.995]$.
8: Estimate the pdf of x, $\hat{f}_k(x)$, limited to the same range.
9: **for** $j = 1, \ldots, m$ **do**
10: Select $y = x[x > h^{(j-1)}] - h^{(j-1)}$
11: Sample $\alpha^{(j)}$ from $f_G(\alpha|n+\nu_0, s_0 + \sum w_i \log(1+y_i/\beta^{(j-1)})$
12: Draw $\beta^{(j)} \sim \varphi(\beta|h^{(j-1)}, y)$, using a Griddy Gibbs
13: Draw $h^{(j)} \sim \varphi(h|\alpha^{(j)}, \beta^{(j)}, x)$, using a Griddy Gibbs
14: Update the grid of β
15: **end for**
16: Discard the initial draws for computing posterior moments and posterior densities

3.3 Gini decomposition using the Pareto II

The aim of modelling the right tail of the income distribution by a Pareto II is to obtain a better representation of the rich and the missing rich, treating

outliers as regular observations and treating under-reporting by an informative prior on the Pareto coefficient.

Jenkins (2017) has proposed to decompose the Gini index in a classical framework between observed low incomes, corresponding to the lower p 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 $s_u = \pi_u \bar{x}_u/\bar{x}$, $s_l = 1 - s_u$. Alvaredo (2011) has shown that the between groups inequality can be simplified to $s_u - \pi_u$ leading to the decomposition formula:

$$G = \pi_l \times s_l \times G_l + \pi_u \times s_u \times G_u + s_u - \pi_u, \tag{25}$$

where the Gini of the two groups are G_u and G_l. G_l is the empirical Gini for the lower group while G_u is the parametric Gini given by the Pareto tail.

We have now to investigate how this approach is modified when the information coming from the upper group is provided by draws from the parameters of the Pareto II model, namely $\theta^{(j)} = (\alpha^{(j)}, \beta^{(j)}, h^{(j)})$, that weights are introduced and that the bulk of the distribution is smoothed using a Bernstein polynomial of degree k, noted $\hat{F}_k(x|h^{(j)})$. Conditionally on a value of $h^{(j)}$, let us call $x_l(h)$ the lower part of the complete observed sample x, and $x_u(h)$ its upper part. Changes are as follows. *First*, because h is random, the population shares become random and have to be computed for each draw of $\theta^{(j)}$. *Second*, the value of the Gini coefficient for the lower part of the sample becomes also random and has also to be computed for each value of $h^{(j)}$. *Third*, Bernstein smoothing provides a continuity between the sparse sample points of $F(x)$ so that the Gini coefficient can be advantageously computed as $G = \bar{x}_l(h)^{-1} \int_0^h \hat{F}_k(x|h^{(j)})[1 - \hat{F}_k(x|h^{(j)})]\, dx$.

These remarks being made, we arrive finally at the following algorithm (Algorithm 3), assuming that the weights sum to n and the we have stored the draws $\theta^{(j)}$.

Remark 2 *The decomposition of the Gini relies on the computation of income shares and consequently on the existence of the means. This implies the supplementary condition that $\alpha^{(j)} > 1$, a condition that is not formally needed to compute the Gini coefficients (2) or (4).*

Remark 3

The above decomposition can be applied to the Pareto I model. We just have to change the following expressions:

$$\begin{aligned} E(x_u|\alpha^{(j)}) &= h^{(j)} \frac{\alpha^{(j)}}{\alpha^{(j)} - 1}, \\ G_u(\alpha^{(j)}) &= \frac{1}{2\alpha^{(j)} - 1}. \end{aligned}$$

Algorithm 3 Bayesian decomposition of the Gini

From the stored values of $\theta^{(j)} = (\alpha^{(j)}, \beta^{(j)}, h^{(j)})$
for $j = 1, m$ **do**
  $n(h^{(j)}) = \sum_{i, x_i \leq h^{(j)}} w_i$
  $\pi_u(h^{(j)}) = 1 - n(h^{(j)})/n$
  $\pi_l(h^{(j)}) = n(h^{(j)})/n$
  $\bar{x}_l(h^{(j)}) = \sum_{i, x \leq h^{(j)}} w_i x_i / n(h^{(j)})$
  $\mathrm{E}(x_u|\theta^{(j)}) = h^{(j)} + \beta^{(j)}/(\alpha^{(j)} - 1)$
  $\mathrm{E}(x|\theta^{(j)}) = \pi_l(h^{(j)})\, \bar{x}_l(h^{(j)}) + \pi_u(h^{(j)})\, \mathrm{E}(x_u|\theta^{(j)})$
  $s_u(\theta^{(j)}) = \pi_u(h^{(j)})\, \mathrm{E}(x_u|\theta^{(j)}) / \mathrm{E}(x|\theta^{(j)})$
  $s_l(\theta^{(j)}) = 1 - s_u(\theta^{(j)})$
  $G_u(\theta^{(j)}) = 1 - \frac{h^{(j)} + 2\alpha^{(j)}\beta^{(j)} B(2\alpha^{(j)} - 1, 2)}{h^{(j)} + \alpha^{(j)}\beta^{(j)} B(\alpha^{(j)} - 1, 2)}$
  $G_l(h^{(j)}, k) = \frac{1}{\bar{x}_l(h^{(j)})} \int_0^{h^{(j)}} \hat{F}_k(x|h^{(j)})[1 - \hat{F}_k(x|h^{(j)})]\, dx$
  $G(\theta^{(j)}) = \pi_l(h^{(j)})\, s_l(\theta^{(j)})\, G_l(h^{(j)}, k) + \pi_u(h^{(j)})\, s_u(\theta^{(j)})\, G_u(\theta^{(j)}) + s_u(\theta^{(j)}) - \pi_u(h^{(j)})$.
end for

4 Top income correction for EU-SILC data

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 survey, which means that the respondent provides 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 suffer from under-sampling in a similar way as usual survey data, but less from under-reporting. The source can be also mixed, which means that the source, survey or administrative, depends on the year of collection, most of the time without further precision.

4.1 The need for correction

We use the income variable `HX090` with weights `DB090`. We report in Tables 1 and 2 the Pareto coefficient, using the top 5% income EU-SILC data for the EU-15 and the NMS for 2008 and 2018, indicating the source of the data

(survey, mixed or administrative). In the second part of Tables 1 and 2, we report the same Pareto coefficients, but this time using the tax data of the WID, when available (`fiinc`, in WID coding). Available countries are: Denmark, France, Germany, Ireland, Italy, the Netherlands, Spain, Sweden, and the United Kingdom, while the periods covered after the year 2000 vary greatly. This means that external fiscal data information is available for 9 members only of the EU-15 and for none of the NMS. For countries without available tax data, we use available disposable income shares from the WID (`scainc`, in WID coding) instead. This latter income distribution might not necessarily be of better representativeness of high incomes than that on EU-SILC samples but are still object to several corrections for data issues (see Alvaredo et al. 2016 for details).[3] In the last columns of Tables 1 and 2, we provide the Gini decomposition (25) for these two sets of α estimates. The WID estimates always provide a larger estimate of inequality (lower

Table 1: The impact of a classical correction on inequality measurement: EU-15

	EU-SILC α		WID α		Gini SILC		Gini WID	
Country	2008	2018	2008	2018	2008	2018	2008	2018
DK (Register)	3.43	2.78	1.51	1.49	0.24	0.27	0.33	0.36
FI (Register)	3.29	3.16	1.77	1.84	0.27	0.26	0.33	0.31
SE (Register)	3.86	3.86	1.73	1.79	0.25	0.27	0.31	0.32
IE (Register)	2.90	3.03	1.98	1.46	0.31	0.30	0.34	0.40
UK (Survey)	2.64	2.97	1.61	1.68	0.34	0.33	0.40	0.39
AT (Mixed)	3.44	3.64	2.23	1.84	0.28	0.27	0.31	0.32
BE (Mixed)	3.18	3.70	2.54	2.67	0.27	0.26	0.28	0.27
DE (Survey)	2.94	3.07	1.54	1.62	0.31	0.30	0.39	0.37
FR (Mixed)	2.84	2.76	1.87	2.02	0.29	0.28	0.33	0.31
LU (Mixed)	3.31	3.48	1.53	2.27	0.28	0.31	0.37	0.33
NL (Register)	3.04	3.06	2.50	2.45	0.26	0.28	0.28	0.29
EL (Survey)	2.94	3.04	2.34	1.94	0.32	0.31	0.34	0.34
ES (Mixed)	3.46	3.83	1.87	1.70	0.32	0.32	0.37	0.39
IT (Mixed)	3.27	2.88	2.11	1.57	0.31	0.33	0.34	0.40
PT (Survey)	2.85	3.39	3.54	2.39	0.36	0.33	0.35	0.35

Means, quantiles and Gini were computed using the `R` package `DescTools`. We selected $p = 0.95$ in order to have enough observations for every country. Total Gini are corrected with a Pareto I assumption for the tail, using either SILC or WID estimates for α together with (25).

Pareto coefficient), except for Portugal in 2008. The trend in inequality

[3]When micro-data are available, the Pareto coefficient can be estimated as the inverse of the mean of $\log x/h$, where h is the 5% quantile and x the part of the sample corresponding to the 5% upper tail. When tabulated data are available, α is estimated using the two top income shares, using the formula given for instance in Atkinson (2007).

is not changed for Nordic countries which all have register incomes when passing from SILC to WID data. This is not the case for the other groups of countries, and that independently of the sources used for income.

Table 2: The impact of a classical correction on inequality measurement: NMS

	EU-SILC α		WID α		Gini SILC		Gini WID	
Country	2008	2018	2008	2018	2008	2018	2008	2018
EE (Survey)	4.03	10.84	1.76	1.58	0.32	0.32	0.39	0.43
LT (Mixed)	2.73	2.81	2.04	2.02	0.35	0.39	0.38	0.42
LV (Mixed)	3.02	3.01	2.43	2.41	0.40	0.38	0.41	0.39
CZ (Survey)	3.41	4.14	1.44	1.64	0.25	0.25	0.36	0.33
HU (Survey)	3.38	2.90	2.06	1.92	0.25	0.29	0.28	0.32
PL (Survey)	2.94	3.99	1.54	1.66	0.32	0.29	0.40	0.37
SI (Register)	4.59	4.21	2.67	2.36	0.25	0.25	0.27	0.28
SK (Survey)	3.76	6.28	1.98	2.09	0.24	0.20	0.28	0.25

The income distribution of Estonia (EE) in 2018 displays some unexpected features when compared to 2008, indicating a possible unadapted type of correction for the higher quantiles. Means, quantiles and Gini were computed using the `R` package `DescTools`. We selected $p = 0.95$ in order to have enough observations for every country. Total Gini are corrected with a Pareto I assumption for the tail, using either SILC or WID estimates for α together with (25).

Corrections brought by the WID are mild for Baltic countries. But they are quite important for the Czech republic, Hungary, Poland that use survey incomes. They are mild for Slovenia (register) and also for Slovakia (survey).

The main message of these two tables is that the WID data provide on average a much important evaluation of inequality than the EU-SILC, justifying the need of extra information for correcting for top incomes, even for countries that are using administrative data for reporting incomes.

4.2 Building prior information from the WID

Let us now see how to build prior information for our Pareto II process. We shall propose for each country a prior mean, using Tables 1-2 which report the estimated values for α using the WID data. A gamma prior has two parameters, ν_0 and s_0 with $\mathrm{E}(\alpha) = \nu_0/s_0$. The estimated values of α provide information on the ratio ν_0/s_0. We have now to decide for a value for ν_0 which represents the degrees of freedom or in a natural conjugate framework the size of the hypothetical sample on which the prior information is built. In $\varphi(\alpha|\beta, x)$, ν_0 is added to the sample size, more precisely the size of the sub-sample representing the top quantile determined by h. We can decide for a rule saying that the size hypothetical sample represents 10% of the size

of the top 0.10 quantile or 1% of the total sample size. Let us summarise this prior information in Table 3.

Table 3: Prior information derived from the WID data

	2008		2018			2008		2018	
Cty	E(α)	ν_0	E(α)	ν_0	Cty	E(α)	ν_0	E(α)	ν_0
DK	1.51	60	1.49	60	EE	1.76	50	1.58	60
FI	1.77	100	1.84	100	LT	2.05	50	2.03	50
SE	1.73	75	1.79	60	LV	2.43	50	2.41	60
AT	2.23	60	1.84	60	CZ	1.44	110	1.64	90
BE	2.55	60	2.67	60	HU	2.06	90	1.92	75
DE	1.54	120	1.62	120	PL	1.54	140	1.66	150
FR	1.87	100	2.02	100	SI	2.67	90	2.36	90
LU	1.53	40	2.28	40	SK	1.98	55	2.09	55
NL	2.50	100	2.45	120					
IE	1.98	50	1.46	50					
UK	1.61	90	1.68	170					
EL	2.34	60	1.94	240					
ES	1.87	130	1.70	130					
IT	2.11	200	1.57	200					
PT	3.54	50	2.39	140					

Source: Own calculation from WID.
Note: ν_0 represents the degrees of freedom for the gamma prior and s_0 is the scale parameter. The prior expectation of α was computed using WID data of Tables 1-2 with a Pareto I assumption. The prior value for ν_0 was chosen so as to represent 1% of the total sample size. Once ν_0 is chosen, the value of s_0 is given by $s_0 = \nu_0/\mathrm{E}(\alpha)$.

4.3 A Pareto II for Gini correction with a random h

We can now propose a Bayesian correction for the measurement of inequality, using a Pareto II tail, an estimated value for h and the prior information given by the WID data base. We use for h a uniform prior corresponding to the range $\rho \in [0.650, 0.995]$. The statistical model is the bulk model (15) with a profile likelihood, based on the Bernstein approximation, with $k = 6$ for the CDF and the pdf and using weights. The CDF and the pdf are estimated on the range corresponding to $\rho \leq 0.995$, the upper bound of the uniform prior. As a lower bound for the prior, we took 0.650, so allowing for a quite large prior range for h. We use $m = 5,000$ draws plus 500 draws to warm up the chain. The grid for β was adjusted twice on the next 1,500 draws, which were then discarded to report posterior moments. For ease of presentation, we report the results without standard deviations in Table 4 for EU-15 and

in Table 5 for the New Member States. Later down, we shall provide plots of the posterior density of the corrected Gini. Several salient facts appear from

Table 4: Gini correction using Pareto II tail, prior information from the WID and Bernstein smoothing: EU-15

	2008					2018				
	Gini y	Gini	Δ %	Pr($\Delta > 0$)	p	Gini y	Gini	Δ %	Pr($\Delta > 0$)	p
DK	0.234	0.231	-1.32	0.108	0.888	0.314	0.272	-13.27	0.000	0.893
FI	0.269	0.274	1.77	0.983	0.908	0.255	0.258	1.43	0.954	0.919
SE	0.273	0.262	-4.21	0.001	0.872	0.287	0.267	-7.01	0.000	0.862
IE	0.301	0.353	17.01	1.000	0.744	0.312	0.339	8.58	1.000	0.734
UK	0.345	0.384	11.10	1.000	0.883	0.339	0.365	7.67	1.000	0.918
AT	0.293	0.310	6.01	1.000	0.828	0.279	0.293	4.71	1.000	0.820
BE	0.285	0.292	2.44	0.957	0.833	0.263	0.285	8.15	1.000	0.819
DE	0.333	0.341	2.54	0.986	0.909	0.318	0.338	6.09	1.000	0.902
FR	0.286	0.320	11.68	1.000	0.903	0.282	0.304	7.78	1.000	0.904
LU	0.270	0.313	16.11	1.000	0.687	0.321	0.329	2.53	0.959	0.734
NL	0.273	0.272	-0.43	0.364	0.906	0.295	0.296	0.32	0.611	0.917
EL	0.320	0.365	14.04	1.000	0.845	0.324	0.357	10.04	1.000	0.943
ES	0.328	0.367	11.77	1.000	0.883	0.332	0.362	9.01	1.000	0.896
IT	0.314	0.341	8.55	1.000	0.939	0.338	0.360	6.52	1.000	0.930
PT	0.355	0.390	9.83	1.000	0.683	0.330	0.382	15.67	1.000	0.879

Simulation with 5,000 draws+500 for warming up the chain. Extra 1,500 draws were discarded due to the updating of β. Convergence was checked, using CUMSUM plots. The value of k for the Bernstein polynomials was equal to 6. Sensitivity analysis with $k = 3$ and $k = 12$ shows that the results are not very much changed.

Table 4.

1. First, the Pareto tail is estimated on a quite large range, from the last 32% for Luxembourg and Portugal to the last 6% for Greece and Italy. There is thus much variation, either over time or between countries. This variety motivates the interest in estimating h instead of fixing it at the same value for each country. This variety is due to the capabilities of the Pareto II, compared to the Pareto I, as shown in section 2.6.

2. Second the correction is either zero or very small (Denmark, Finland, Sweden, the Netherlands) for some countries using register data for income.

3. Third, the correction can be quite important (Ireland, Luxembourg, Greece in 2008, 16% for Portugal in 2018), but can vary a lot for the same country, depending on the year (Luxembourg, Portugal).

4. Finally, using register data is not a guaranty for a small correction,

witness the case Ireland which experience a correction of 17% in 2008, despite the use of register data.

Table 5: Gini correction using Pareto II tails and prior information from the WID and Bernstein smoothing: NMS

	2008					2018				
	Gini y	Gini	Δ %	Pr($\Delta > 0$)	p	Gini y	Gini	Δ %	Pr($\Delta > 0$)	p
EE	0.341	0.373	9.54	1.000	0.714	0.335	0.368	9.68	1.000	0.661
LT	0.367	0.389	6.17	1.000	0.801	0.408	0.417	2.19	0.950	0.750
LV	0.392	0.454	15.76	1.000	0.723	0.376	0.431	14.70	1.000	0.735
CZ	0.276	0.308	11.74	1.000	0.889	0.251	0.295	17.46	1.000	0.828
HU	0.253	0.277	9.45	1.000	0.871	0.312	0.342	9.79	1.000	0.848
PL	0.333	0.384	15.32	1.000	0.895	0.294	0.353	20.09	1.000	0.882
SI	0.266	0.269	0.99	0.887	0.849	0.275	0.268	-2.27	0.011	0.835
SK	0.261	0.262	0.47	0.589	0.797	0.212	0.227	7.31	1.000	0.784

Simulation with 5,000 draws+500 for warming up the chain. Extra 1,500 draws were discarded due to the updating of β. Convergence was checked, using CUMSUM plots. The value of k for the Bernstein polynomials was equal to 6. Sensitivity analysis with $k = 3$ and $k = 12$ shows that the results are not very much changed.

Table 5 depict a less contrasted situation for New Member States. The estimated value of p is still varying among the countries, corresponding to a range between 0.66 and 0.90. The correction for the Gini is significant for all countries and more important than for the EUR-15. The exception is Slovenia with a correction that can be zero, perhaps due to the fact that this country is using registered data.

4.4 Posterior densities for the corrected Gini

In this section, we provide the posterior density of the corrected Gini indices. We have regrouped the countries into categories. The plots are given, using the same scale for ease of comparison. On each plot, we have indicated with a vertical dashed line the value of the sample Gini without correction, in order to visualise the importance of the correction brought by our calculations.

Nordic countries (Figure 7) are typical of the use of registered data. For Denmark, the vertical line is at the very end of each posterior density. For Sweden, the vertical line is within the range of the posterior density of the Gini. For these two countries, we can accept the fact that there is no need for correction, a configuration that has to be contrasted with Finland, where a correction is needed, however small. The trend in inequality can be clearly identified and is significant for the Nordic countries, increasing for Denmark and Sweden, decreasing for Finland. On average the range of the posterior corrected Gini is from 0.23 (Denmark) to 0.31 (Finland).

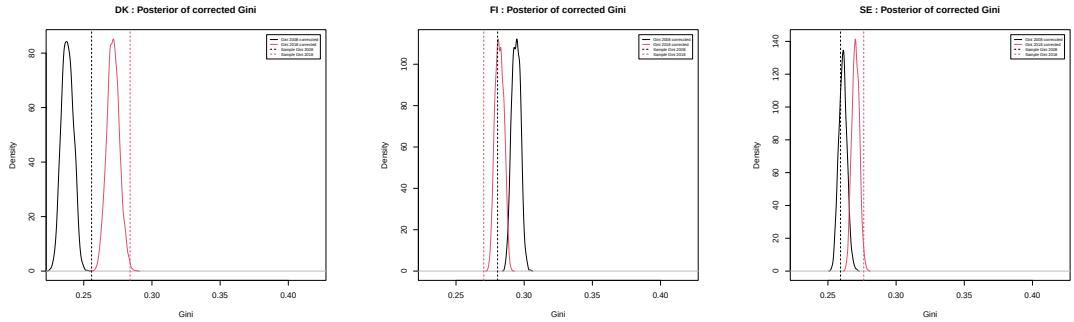

Figure 7: Posterior densities of corrected Gini for Nordic countries

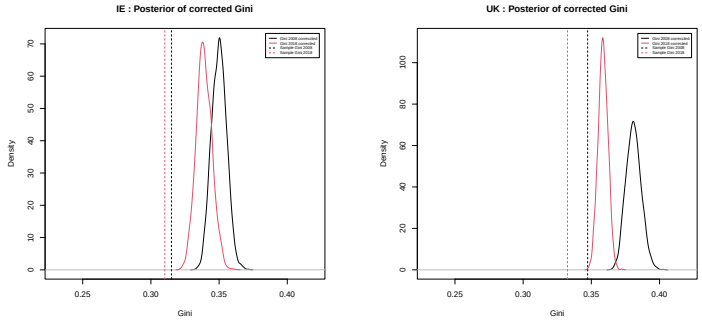

Figure 8: Posterior densities of corrected Gini for Anglo-Saxon countries

For *Anglo-Saxon countries* (Figure 8), the Gini has been significantly corrected, irrespective of the origin of the data, register for Ireland and survey for the UK. There is a decreasing trend in inequality, significant for the UK, but more hazardous for Ireland as its two posterior densities overlap. Inequality is much more important (0.35, Ireland to 0.38, the UK) than in the Nordic countries.

For *Northern European countries* (Figure 9), the correction can lead some paradoxical results. For instance it is important for Luxembourg in 2008, not significant in 2018, so that finally the two posterior densities overlap. Finally the corrected measure has not changed over time. In the Netherlands, there is no correction for 2008, a mild correction for 2018, but it amplifies the increase of inequality over time. For the other countries, the correction can be important (France) and inequality is decreasing over time. On average, the range of the Gini is between 0.26 (the Netherlands) and 0.35 (Germany), a range even greater than that of Anglo-Saxon countries.

For *Southern European countries* (Figure 10), the case of Spain is interesting. The correction is important, but the posterior densities of the two

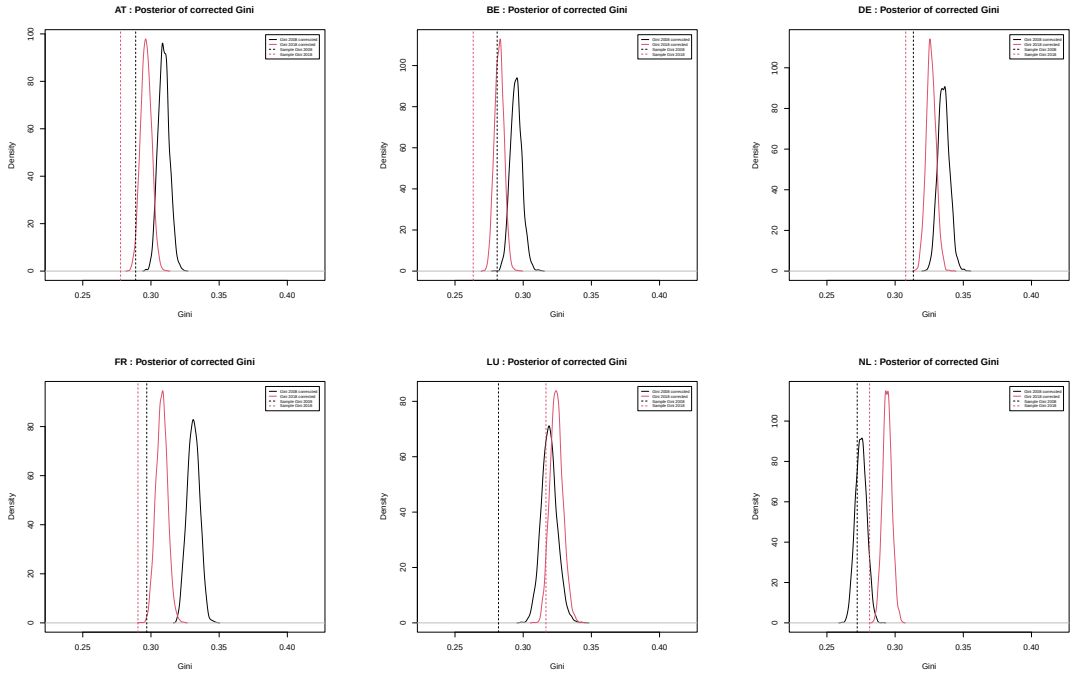

Figure 9: Posterior densities of corrected Gini for Northern European countries

Gini for 2008 and 2018 strictly coincide. For Greece the correction strongly amplifies the decrease in inequality. In Italy, on the contrary, the correction reduces the increase in inequality over time. In Portugal, the strong correction preserves the trend for decreasing inequality. The range of the corrected Gini is even greater than in Northern European countries with 0.33 (Greece) to 0.43 (Portugal).

The case of *Baltic and Eastern countries* (Figures 11 and 12) is exemplary for the need of corrections. The corrections are quite important except for Slovenia (register). They however introduce no major reversal in the evolution of inequality: increasing for Lithuania and Hungary, stationary for Estonia, Czech Republic and Slovenia, decreasing for Latvia, Poland and Slovakia. The range of the posterior corrected is between 0.22 (Slovakia) and 0.46 (Latvia), so the much larger of all the group of European countries, mainly due to the Baltic countries. Without the correction, the Czech Republic, Hungary and Slovakia have a level of inequality lower than that of the Nordic countries. With the correction, they reach a level of inequality comparable to that of Northern European countries.

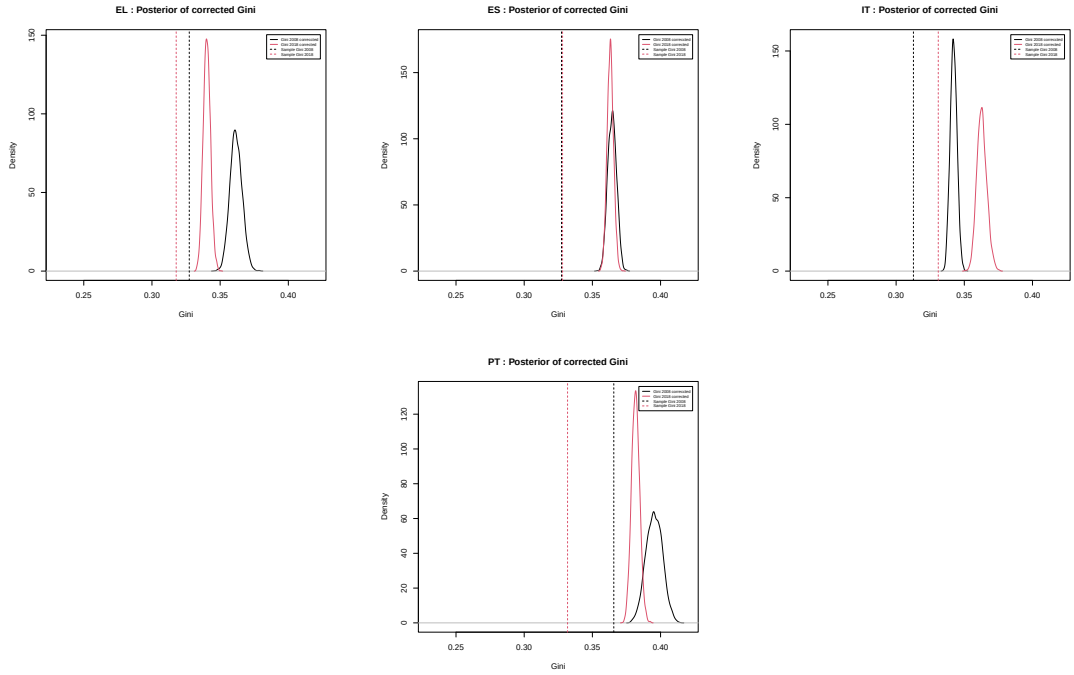

Figure 10: Posterior densities of corrected Gini for Southern European countries

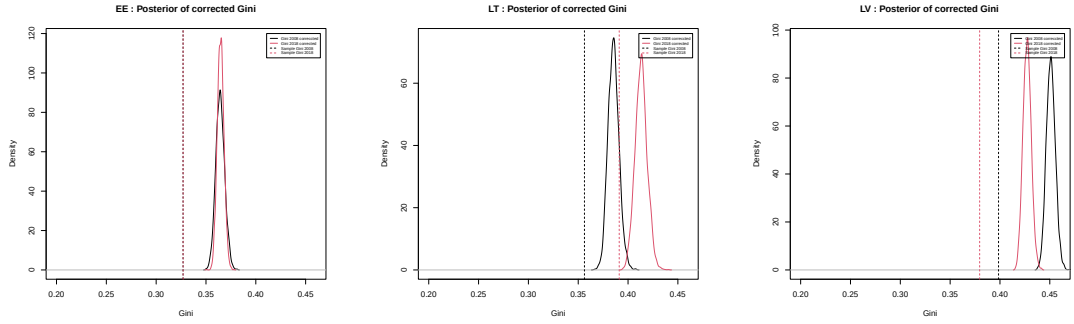

Figure 11: Posterior densities of corrected Gini for Baltic countries

4.5 The impact of adopting a Bayesian compound model

What is brought in by adopting our Bayesian modelling of the income distribution compared to the simple classical calculations based on a Pareto I assumption (instead of a Pareto II), a fixed value for h as reported in Table 1? Most of the time, the Bayesian correction is slightly lower than that reported in Table 1. But in few cases, the Bayesian correction can be either the

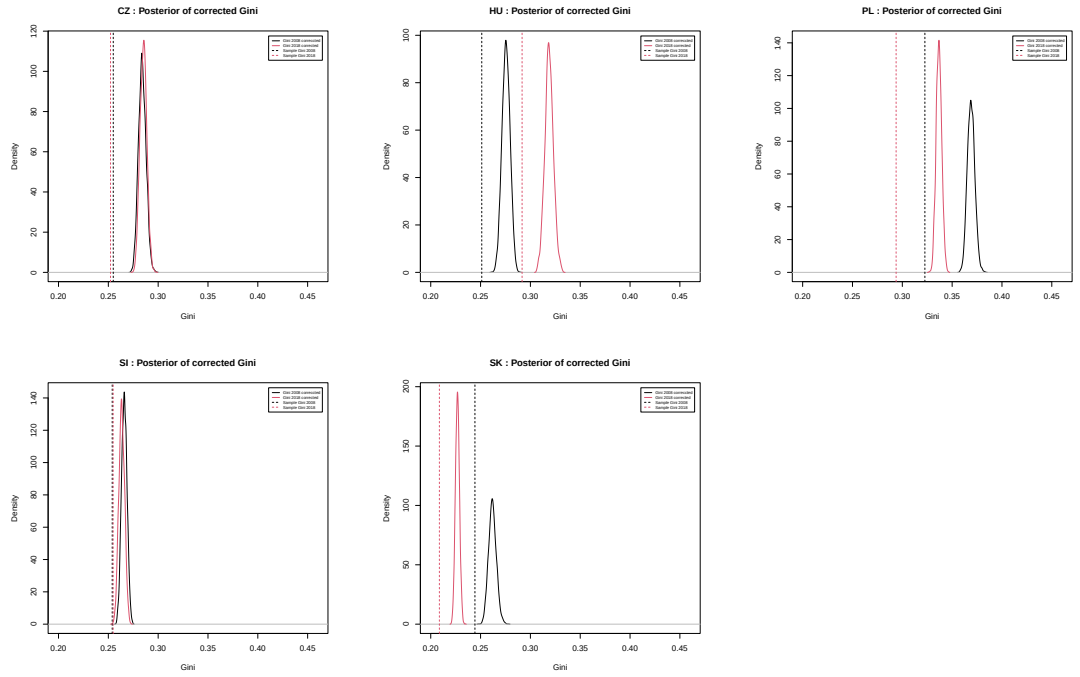

Figure 12: Posterior densities of corrected Gini for other eastern countries

same or greater (see Tables 6 and 7). In fact the individual survey data do not bring exactly the same message as the grouped WID data. For instance, we have used weights when these of course are not available in grouped data. And a sample Gini using weights is usual smaller than a Gini without weights, at least in the EU-SILC data sets. On the other side, computing the Gini for the lower sample as:

$$G_l(h,k) = \frac{1}{\bar{x}_l(h)} \int_0^h \hat{F}_k(x)[1 - \hat{F}_k(x)]\, dx,$$

provides usually a slightly larger value that that obtained using directly the truncated sample. Finally, h is estimated and might thus be different both over time and between countries. And as underlined in Jenkins (2017), the lower h, the higher the correction. So the differences comes from a bunch of different reasons. The corrections provided by the Bayesian approach seem in total fairly reasonable.

We present in Appendix D the need for a Pareto II by testing the restriction $\beta = h$. Most of the time this restriction is rejected by the sample, confirming the need for a Pareto II model.

Table 6: Bayesian versus classical Gini using the WID: EUR-15

	2008		2018		2008	2018
Cty	Bayes	WID	Bayes	WID		
DK	0.231	0.33	0.272	0.36	$B < W$	$B < W$
FI	0.274	0.33	0.258	0.31	$B < W$	$B < W$
SE	0.262	0.31	0.267	0.32	$B < W$	$B < W$
IE	0.353	0.34	0.339	0.40	$B > W$	$B < W$
UK	0.384	0.40	0.365	0.39	$B < W$	$B < W$
AT	0.310	0.31	0.293	0.32	$B = W$	$B < W$
BE	0.292	0.28	0.285	0.27	$B > W$	$B > W$
DE	0.341	0.39	0.338	0.37	$B < W$	$B < W$
FR	0.320	0.33	0.304	0.31	$B = W$	$B = W$
LU	0.313	0.37	0.329	0.33	$B < W$	$B = W$
NL	0.272	0.28	0.296	0.29	$B = W$	$B = W$
EL	0.365	0.34	0.357	0.34	$B > W$	$B > W$
ES	0.367	0.37	0.362	0.39	$B = W$	$B < W$
IT	0.341	0.34	0.360	0.40	$B = W$	$B < W$
PT	0.390	0.35	0.382	0.35	$B > W$	$B > W$

The columns Bayes correspond to the values of the corrected Gini obtained in Table 4, using a Pareto II tail and a prior information coming from the WID data base. The columns WID corresponds to the values reported in Table 1, using a Pareto I tail.

Table 7: Bayesian versus classical Gini using the WID: NMS

	2008		2018		2008	2018
Cty	Bayes	WID	Bayes	WID		
EE	0.373	0.39	0.368	0.43	$B < W$	$B < W$
LT	0.389	0.38	0.417	0.42	$B > W$	$B = W$
LV	0.454	0.41	0.431	0.39	$B > W$	$B > W$
CZ	0.308	0.36	0.295	0.33	$B < W$	$B < W$
HU	0.277	0.28	0.342	0.32	$B = W$	$B > W$
PL	0.384	0.40	0.353	0.37	$B < W$	$B < W$
SI	0.269	0.27	0.268	0.28	$B = W$	$B < W$
SK	0.262	0.28	0.227	0.25	$B < W$	$B < W$

The columns Bayes correspond to the values of the corrected Gini obtained in Table 4, using a Pareto II tail and a prior information coming from the WID data base. The columns WID corresponds to the values reported in Table 1, using a Pareto I tail.

5 The impact of the correction on inequality dynamics

The *Growth Incidence Curve* (GIC) of Ravallion and Chen (2003) provides a tool to measure how growth has benefited to the different quantiles of an income distribution. Its level indicates the average growth rate. If it is

downward slopping, inequality has decreased over the period, which means that lower quantiles have benefited more from economic growth than higher quantiles. And the reverse if it is upward slopping. The GIC is obtained simply by computing the difference between the logs of two quantiles functions computed at two different points in time. By considering a compound model and prior information on the shape of the Pareto II tail, we have introduced a correction for potentially missing rich, which means that the original data gave a biased representation of the very top quantiles, and thus potentially introducing a deformation of the GIC. What is the impact of this correction on the shape of the GIC? We try to answer to this question in this last section of the paper.

5.1 Quantile function for the compound model

We have first to derive the quantile function for our compound model. Let us consider the log-normal-Pareto II compound model as a simplification. The CDF was in given in (16). Due to the very particular structure of the bulk model, the quantile function can be easily derived. There is a one-to-one correspondence between ρ (the proportion of observations below h) and the value of h with $\rho = 1 - F_\Lambda(h|\mu, \sigma^2)$. Let us define the cutting point $\rho_h = F_\Lambda(h|\mu, \sigma^2)$. We can express the overall quantile function for $p \in [0, 1]$ as:

$$\begin{aligned} Q(p|\theta) &= Q_\Lambda(p|\mu, \sigma^2)\mathbb{1}(p \leq \rho_h) + Q_\Lambda(\rho_h|\mu, \sigma^2)\mathbb{1}(p > \rho_h) \\ &+ (1 - \rho_h)Q_{P2}(p|h, \alpha, \beta)\mathbb{1}(p > \rho_h), \end{aligned} \tag{26}$$

where the quantile function for the lognormal and the Pareto II are:

$$Q_\Lambda(p|\mu, \sigma^2) = \exp(\mu + \sigma\Phi^{-1}(p)) \tag{27}$$

$$Q_{P2}(p|h, \alpha, \beta) = h - \beta + \beta(1 - p)^{-1/\alpha}. \tag{28}$$

The GIC is obtained as the difference of the log of two quantiles functions, each computed at two points in time. Given a draw $\theta^{(j)}$ from the posterior density of the parameters, a draw from the posterior density of the GIC is given by:

$$g_t(p|\theta^{(j)}) = \log Q_t(p|\theta_t^{(j)}) - \log Q_{t-1}(p|\theta_{t-1}^{(j)}). \tag{29}$$

Remark 4 *In section 3.2, we have replaced the lognormal distribution by a Bernstein polynomial approximation. A direct approximation exist also for the quantile function of the lower part of the distribution as was used for instance in Fourrier-Nicolai and Lubrano (2023). The case is quite simple as we have simply to smooth the empirical quantile function which is naturally*

defined on [0,1]. Let us partition the vector p into $p = [p_l, p_u]$ *where* $p_l \in [0, \tau[$ *and* $p_u \in [\tau, 1]$ *and* τ *is the limit between the two components. We then build the* $n \times (k+1)$ *matrix Z:*

$$Z_k(p) = [B_k(p, 0), \cdots, B_k(p, k)],$$

An estimate of the lower quantile function is obtained from the regression

$$Qs = Z_k(p)\delta + u, \tag{30}$$

where Qs is the vector of empirical quantiles of x. The smoothed lower quantile function is given by $Z_k(p_l)\hat{\delta}$.

Let us now suppose that we have obtained m draws from the posterior density of δ *in regression (30), using a non-informative prior and m draws for the parameters* α, β *and h of the Pareto II model. A posterior draw from the quantile function of the compound model is given by:*

$$Q(p|\theta^{(j)}) = [Z_k(p_l)\delta^{(j)}, (1 - \rho(h^{(j)}))(h^{(j)} - \beta^{(j)} + \beta^{(j)}(1 - p_u)^{-1/\alpha^{(j)}})], \tag{31}$$

where we have chosen a fixed grid for p for ease of presentation. This extension will be used in a future version of the paper. However, the empirical results should not be too much altered.

5.2 GIC for a group of countries

Finding the GIC for a group of countries is a specific problem. We have to estimate the income distribution for this group as a mixture of country income distributions. It is usual very easy to find the corresponding CDF, as the CDF of a mixture is just the mixture of the member CDFs. The same property however does not hold for the quantile function which has to be derived numerically, by solving for each draw a one variable equation. The method is detailed in Fourrier-Nicolaï and Lubrano (2021). Let Pop_i be the population of country i in percentage of the total population of the group and $F_i(x|\theta)$ the CDF of country i. The CDF of a group of countries is given by:

$$F(x|\theta^{(j)}) = \sum_i Pop_i F_i(x|\theta_i^{(j)}).$$

For each draw of the parameters, $F_i(x|\theta_i^{(j)})$ can be deduced from (16) to give:

$$\begin{aligned} F_i(x|\theta_i^{(j)}) &= F_\Lambda(x|\hat{\mu}_i, \hat{\sigma}_i^2)\mathbb{1}(x \leq h_i^{(j)}) \\ &+ \mathbb{1}(x > h_i^{(j)})[F_\Lambda(h_i^{(j)}|\hat{\mu}_i, \hat{\sigma}_i^2) + (1 - F_\Lambda(h_i^{(j)}|\hat{\mu}_i, \hat{\sigma}_i^2)F_{P2}(x|h_i^{(j)}, \alpha_i^{(j)}, \beta_i^{(j)})], \end{aligned}$$

with:

$$F_{P2}(x|h_i^{(j)}, \alpha_i^{(j)}, \beta_i^{(j)}) = 1 - \left(1 + \frac{x - h^{(j)}}{\beta^{(j)}}\right)^{-\alpha^{(j)}}.$$

The corresponding quantile function for each draw of the parameters has to be found by solving in x the equation:

$$F(x|\theta^{(j)}) - p = 0.$$

This equation can be solved, using Brent (1973) algorithm, programmed in `R` with the `uniroot` function. As this algorithm does not use derivatives, there is in general no problem of convergence. The execution time is rather quick.

5.3 Is there an Elephant curve in Europe?

Lakner and Milanovic (2016) have shown that when plotting the GIC of the world income distribution, the latter had the shape of an elephant. This means that, economic growth has benefited a lot to the world middle class, not at all to the world upper middle class and a lot to the world very rich households (assuming that the elephant is raising its trump). The last part of the plot is especially amplified when correcting for missing rich. At the world level, the middle class corresponds to the emergence of China, which is both one of the largest populated country and has benefited of a huge economic growth. Do we find the same type of phenomenon at the European level? The 8 new member states have experienced between 2008 and 2018 of a large economic growth, much more important than that of the initial 15 member states, as we can see by comparing the GIC plots in Figure 13 corresponding to these two groups of countries. Average growth is around 36% over the

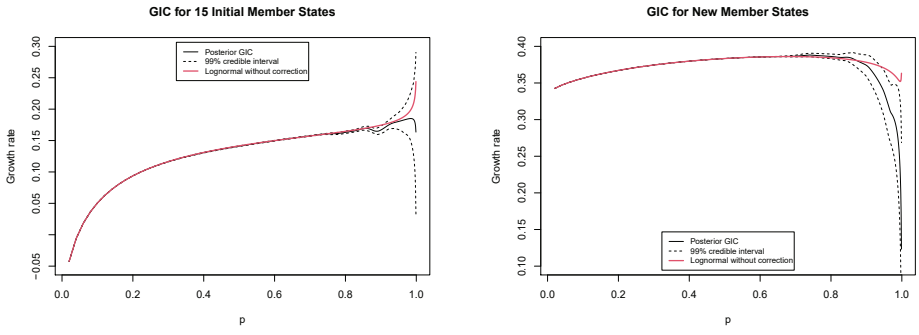

Figure 13: Comparing GICs for EUR-15 and NMS based on a lognormal assumption and a Pareto II tail

period for the NMS against a mere 10% for the EUR-15. However, the NMS are very small countries with a total population of 71.8 millions against the 410.6 millions of the EUR-15. The distribution of inequality over the quantiles is also very different. For the NMS, all quantiles seem to have benefited from growth and our Pareto tail correction indicates a sharp decline in the growth rate for top quantiles. On the contrary, for the EUR-15, the extreme low quantiles have lost, while the top quantiles have gained. However, the Pareto tail correction indicates a large uncertainty.

Mixing those two groups of countries produces the GIC displayed in Figure 14. The mixing of different groups of countries changes completely the

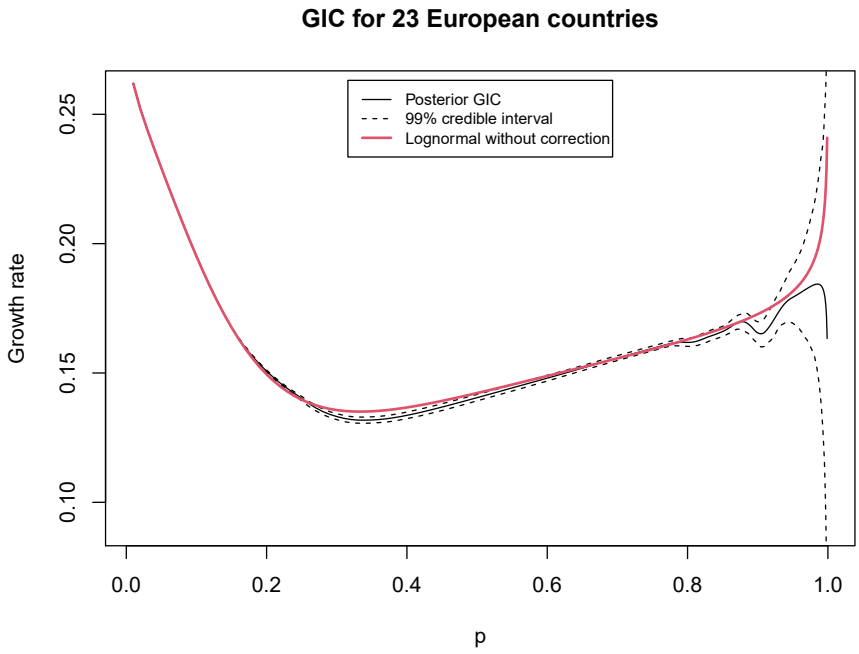

Figure 14: GIC for 23 European states based on a lognormal assumption

lower part of the global Gic. When lower quantiles of the EUR-15 countries were the losers of economic growth, at the complete group of countries, the lower quantiles were the net beneficiaries of growth up to the 0.15 quantile. Then, the Pareto tail correction has an impact starting very early around the 0.20 quantile. The GIC issued from a simple lognormal assumption is lowered between quantiles 0.30 and 0.60. Finally the uncertainty at the level of top quantiles remains the same as that depicted in Figure 13 for EUR-15.[4]

[4]We have computed the Gic with a higher precision for top quantiles. For the last

We do not have the usual elephant curve of Lakner and Milanovic (2016). How can we explain this difference with the World GIC? The shape of the latter was mainly due to the impact of the huge economic growth of China and partially of India. In the GIC depicted in Figure 14, the position of China would correspond to the 0.05 quantile in 2008 and 0.15 quantile in 2018. So we observe only the trump of the world elephant curve and not its body.[5]

6 Conclusion

We have assumed that the lower part of the sample correctly reported incomes, while some incomes were either missing or under-reported in the upper part of the survey sample. The main question was how to empirically determine the point of junction between the two parts of the sample. We have proposed a method coming from the extreme-value literature. We have corroborated one of the findings of Jenkins (2017) that the amount of correction for measuring inequality depended very much on the threshold. However, we did not corroborate the second finding of Jenkins (2017) that the direction of correction depended little on the choice of threshold. The reason is that with our enlarged model, we estimate the value of h in every case and this value can change a lot, both over time and between countries. This has to be contrasted with the empirical literature (Jenkins 2017, Bartels and Metzing 2019 and others) that opted for quite high and fixed values of the threshold, between the top 0.90 quantile and the top 0.99 quantile.

Our Bayesian approach allows the sample to react in its own way to the prior information based on WID data. It can lead to lower corrections and in a few cases to higher corrections. We found only a weak relation between the source of the data (register versus survey) and the amount of correction, which is in a way surprising.

This varying h has also important consequences on the shape of the GIC which tries to depict which quantiles benefited from growth in income. When measured for a group of countries, here the 15+8 European countries, the

0.01%, the corrected Gic is decreasing while the uncorrected one is increasing. But there is so much uncertainty that the difference is not significant.

[5]China GDP per capita was $7,413 in 2008 and $15,134 in 2018 (computed in 2017 $PPP). The same indicators for Germany were respectively $47,643 and $53,431. To translate these figures in average incomes in €, we can confront them to the figures of average income obtained for Germany with the SILC data. They are 20,697 €in 2008 and 25,144 €in 2018. So the ratio between GDP per capita in $ PPP and mean income in €is 20697/47643=0.43 in 2008 and 25144/53431=0.47 in 2018. The corresponding mean income for China is thus 7,413*0.43=3,188 €in 2008 and 15134.47=7,113 €in 2018.

Pareto correction modifies its shape at the level of intermediate quantiles, because the EU-15 and the NMS have intrinsic different characteristics.

This work is not exempt of limitations. The main limitation is due to the interpretation of h which is here the threshold where a Pareto II tail best fit the sample. This threshold can be rather low, due to the nice properties of the Pareto II, compared to those of the Pareto I. There are plenty of other reasons, besides under-sampling, for explaining the possible bias in survey incomes. They have been investigated by Angel et al. (2019) and they might be different according to the nature of income (wages, pension, unemployment allowances). And this analysis requires linked samples, which are not very frequent. This is beyond the scope of the present paper.

References

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

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

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.

Arnold, B. C. and Press, S. J. (1983). Bayesian inference for Pareto populations. *Journal of Econometrics*, 21:287–306.

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

Brandolini, A. and Rosolia, A. (2021). The distribution of well-being among Europeans. In *Europe's Income, Wealth, Consumption, and Inequality*. Oxford University Press, Oxford.

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

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.

Castellanos, M. E. and Cabras, S. (2007). A default Bayesian procedure for the generalized Pareto distribution. *Journal of Statistical Planning and Inference*, 137:473–483.

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.

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

do Nascimento, F. F., Gamerman, D., and Lopes, H. F. (2012). A semiparametric Bayesian approach to extreme value estimation. *Statistics and Computing*, 22:661–675.

Ederer, S., Ćetković, P., Humer, S., Jestl, S., and List, E. (2020). Distributional national accounts (DINA) with household survey data: Methodology and results for European countries. *Review of Income and Wealth*, 68(3):667–688.

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*, To appear.

Fourrier-Nicolaï, E. and Lubrano, M. (2021). Bayesian inference for parametric growth incidence curves. In Bandyopadhyay, S., editor, *Research on Economic Inequality: Poverty, Inequality and Shocks (Research on Economic Inequality)*, volume 29, pages 31–55. Emerald Publishing Limited, Bingley.

Fourrier-Nicolai, E. and Lubrano, M. (2023). Bayesian inference for non-anonymous growth incidence curves using Bernstein polynomials: an application to academic wage dynamics. *Studies in Nonlinear Dynamics and Econometrics*, Forthcoming.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics),*, 44(4):455–472.

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

Hlasny, V. and Verme, P. (2018). Top incomes and inequality measurement: A comparative analysis of correction methods using the EU-SILC data. *Econometrics*, 6(2):1–21.

Hu, X. and Gui, W. (2018). Bayesian and non-Bayesian inference for the generalized Pareto distribution based on progressive type II censored sample. *Mathematics*, 6(319):0.

Jantti, M., Tormalehto, V.-M., and Marlier, E. (2013). The use of registers in the context of EU–SILC: challenges and opportunities. Statistical working papers, Eurostat.

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

Kranzinger, S. (2020). The decomposition of income inequality in the EU-28. *Empirica*, 47:643–668.

Lakner, C. and Milanovic, B. (2016). Global income distribution: From the fall of the Berlin wall to the great recession. *World Bank Economic Review*, 30(2):203–232.

Lindsey, J. K. (1974). Comparison of probability distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):38–47.

Mokrani, F., Fellag, H., and Necir, A. (2016). Robust Bayesian inference of generalized Pareto distribution. *Afrika Statistika*, 11(2):1061–1074.

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

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

Ravallion, M. and Chen, S. (2003). Measuring pro-poor growth. *Economics Letters*, 78(1):93–99.

Reed, W. J. (2003). The Pareto law of incomes: An explanation and an extension. *Physica A*, 319:469–486.

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.

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

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.

Van Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. Working Paper 2007-01, EPS-Instead IRISS.

Villa, C. (2017). Bayesian estimation of the threshold of a generalised Pareto distribution for heavy-tailed observations. *Test*, 26(1):95–118.

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.

Yu, B. and Mykland, P. (1998). Looking at Markov samplers through CUM-SUM path plots: a simple diagnostic idea. *Statistics and Computing*, 8(3):275–286.

Appendix

A The EU-SILC data set

The EU-SILC data set is both a cross section and a panel survey. But the panel aspect is limited as it is a rotational household panel with a quarter of respondents being exchanged each year. So it is possible to follow the same household only during four consecutive years. Data are available starting in 2004 for the initial 15 EU members and slightly later for the new member states.

A.1 Coverage

We provide in Tables 8 and 9 the coverage of the survey for the initial 15 members and for the new member states that joined the EU later. We indicate the origin of the income variables (survey, administrative or mixed). We also provide the average population over the period. For the rest of the analysis, it is better to exclude Bulgaria, Romania, Croatia and Malta which are starting later than 2005. Other countries includes countries out of the EU, Switzerland, Iceland, Norway, Serbia. If we are interesting in economic convergence inside the EU, it is better to exclude those countries. So, we are going to consider two groups: EU15, the original 15 members (including the UK) and NMS, the new-comers, essentially the former eastern countries, so excluding Cyprus. In total, we have 23 countries for 14 years. For further analysis, we can distinguish five groups of countries among those 23 countries with Kranzinger (2020), grouping European countries according to their welfare state regimes. We have:

1. Scandinavian countries: Denmark (DK), Finland (FI), Sweden (SE),
2. Anglo-saxon: Ireland (IE), the UK,
3. Northern Europe: Austria (AT), Belgium (BE), Germany (DE), France (FR), Luxembourg (LU), Netherlands (NL),
4. Southern Europe: Greece (EL), Spain (ES), Italy (IT), Portugal (PT),
5. Baltic countries: Estonia (EE), Lithuania (LT), Latvia (LV),
6. Eastern Europe: Czech Rep (CZ), Hungary (HU), Poland (PL), Slovenia (SI), Slovakia (SK)

This classification will help us to present our empirical results.

Table 8: Old EU-15. Founding members plus joiners

Country	Full name	Min	Max	Source	Population
DK	Denmark	2004	2019	Register	5.8
FI	Finland	2004	2019	Register	5.5
SE	Sweden	2004	2019	Register	10.4
IE	Ireland	2004	2019	Register	5.0
UK	United Kingdom	2005	2018	Survey	67.0
AT	Austria	2004	2019	Mixed	8.9
BE	Belgium	2004	2019	Mixed	11.6
DE	Germany	2005	2019	Survey	83.2
FR	France	2004	2019	Mixed	67.4
LU	Luxembourg	2004	2019	Mixed	0.6
NL	Netherlands	2005	2019	Register	17.5
EL	Greece	2004	2019	Survey	10.7
ES	Spain	2004	2019	Mixed	47.4
IT	Italy	2004	2019	Mixed	59.3
PT	Portugal	2004	2019	Survey	10.3

The nature of the source of income comes from *The use of registers in the context of EU–SILC: Challenges and opportunities* by Jantti et al. (2013).

Table 9: New member states

Country	Full name	Min	Max	Source	Joining	Population
EE	Estonia	2004	2019	Survey	2004	1.3
LT	Lithuania	2005	2019	Mixed	2004	2.8
LV	Latvia	2005	2019	Mixed	2004	1.9
CZ	Czech Rep	2005	2019	Survey	2004	10.7
HU	Hungary	2005	2019	Survey	2004	9.7
PL	Poland	2005	2019	Survey	2004	37.8
SI	Slovenia	2005	2019	Register	2004	2.1
SK	Slovakia	2005	2019	Survey	2004	5.5

Source from *The use of registers in the context of EU–SILC: Challenges and opportunities* by Jantti et al. (2013).

A.2 The income variables

For each country and year, three files are available in the EU-SILC: P-file for individuals, H-file for households and D-file for identifiers. The variable HX090 corresponds to household income, normalised by the OECD equiv-

alence scale. This scale assigns value 1 to the first adult, 0.5 to any other person aged 14 or older, and 0.3 to each child younger than 14. HY020 is also total (disposable) household income, but without normalisation and HY010 corresponds to total household gross income. HX050 is the OECD-modified scale equivalised household size. HX090 is the variable that is mostly used in empirical studies with $HX090 = HY020/HX050$. It corresponds to the notion of HGDI (household gross disposable income) of national accounts. The series are made comparable for all EU countries. They are adjusted for cost-of living differences both across countries and over time. We have excluded negative and missing values.

Most references focus on the household equivalised **disposable** income variable (`HX090`) readily available in EU-SILC data (e.g., Van Kerm 2007, Hlasny and Verme 2018, Brandolini and Rosolia 2021). This variable is defined as:

$$\texttt{HX090} = \frac{\texttt{HY020}}{\texttt{HX050}} \tag{32}$$

where the OECD-modified scale equivalised household size (`HX050`) is defined as[6]

$$\texttt{HX050} = 1 + 0.5 \times (\#\text{ household members aged 14 and over} - 1) \tag{33}$$

$$+ 0.3 \times (\#\text{ household members aged 13 or less}) \tag{34}$$

and the total disposable household income (`HY020`) is defined as

$$\texttt{HY020} = \underbrace{\texttt{HY010}}_{\text{Total household gross income}} \tag{35}$$

$$- \underbrace{\texttt{HY120G}}_{\text{Regular taxes on wealth}} \tag{36}$$

$$- \underbrace{\texttt{HY130G}}_{\text{Regular inter-household cash transfers paid}} \tag{37}$$

$$- \underbrace{\texttt{HY140G}}_{\text{Regular taxes on income and social ins. contributions}} \tag{38}$$

An equal-split (i.e., per capita) household income variable can be computed as:

$$\texttt{HX090}_{es} = \frac{\texttt{HY020}}{\texttt{HX040}} \tag{39}$$

The composition of total household gross incomes (`HY010`) considered is as follows:

[6]This scale gives a weight of 1 to the household head, of .5 to each and every other adult household member and of .3 to each and every child in the household.

$$\texttt{HY010} = \underbrace{\texttt{HY040G}}_{\text{Income from rental of a property or land}} \tag{40}$$

$$+ \underbrace{\texttt{HY050G}}_{\text{Family/children related allowances}} \tag{41}$$

$$+ \underbrace{\texttt{HY060G}}_{\text{Social exclusion not elsewhere classified}} \tag{42}$$

$$+ \underbrace{\texttt{HY070G}}_{\text{Housing allowances}} \tag{43}$$

$$+ \underbrace{\texttt{HY080G}}_{\text{Regular inter-household cash transfers received}} \tag{44}$$

$$+ \underbrace{\texttt{HY090G}}_{\text{Interests, dividends, profit from capital investments in unincorporated business}} \tag{45}$$

$$+ \underbrace{\texttt{HY110G}}_{\text{Income received by people aged under 16}} \tag{46}$$

$$+ \sum_{i \text{ in household}} \underbrace{\texttt{PY010G}_i}_{\text{Gross employee cash or near cash income}} \tag{47}$$

$$+ \underbrace{\texttt{PY021G}_i}_{\text{Company car}} \tag{48}$$

$$+ \underbrace{\texttt{PY050G}_i}_{\text{Gross cash benefits or losses from self-employment (including royalties)}} \tag{49}$$

$$+ \underbrace{\texttt{PY080G}_i}_{\text{Pensions received from individual private plans (other than those covered under ESSPROS)}} \tag{50}$$

$$+ \underbrace{\texttt{PY090G}_i}_{\text{Unemployment benefits}} \tag{51}$$

$$+ \underbrace{\texttt{PY100G}_i}_{\text{Old-age benefits}} \tag{52}$$

$$+ \underbrace{\texttt{PY110G}_i}_{\text{Survivors' benefits}} \tag{53}$$

$$+ \underbrace{\texttt{PY120G}_i}_{\text{Sickness benefits}} \tag{54}$$

$$+ \underbrace{\texttt{PY130G}_i}_{\text{Disability benefits}} \tag{55}$$

$$+ \underbrace{\texttt{PY140G}_i}_{\text{Education-related allowances}} \tag{56}$$

$$\tag{57}$$

Ederer et al. (2020) gives a framework allowing to understand the links between these EU-SILC income aggregates and National Accounts aggregates under the 2010 European System of Accounts (ESA2010). For most EU-SILC countries, they find that the Adjusted Disposable Income (ADINC) aggregate from the OECD-Eurostat Expert Group on Disparities in National Accounts (EG-DNA) methodology very closely resembles the Post-tax National Income (POTNI) aggregate from the WID.world DINA methodology, which are the key aggregates for studying the National Accounts-consistent post-tax distribution of incomes. In their methodology for EU-SILC data, out of 22 ESA2010 items in total which define the ADINC, 10 items must be imputed to the micro-data based only on National Accounts, and 4 are imputed through simulations from EUROMOD (employer social contributions, employee social contributions, taxes on employment income, taxes on simulated property income, & taxes on non-simulated property income) (Ederer et al. 2020, Table 1). Additionally, they don't take taxes from EU-SILC but rather from EUROMOD simulations on the micro-data and they consider some income items which are not included in `HY010`: Gross non-cash employee income (`PY020G`$_i$. Includes within it the imputed value for company car `PY021G`$_i$), gross imputed rents (`HY030G`), and gross value of goods produced for own-consumption (`HY170G`). Mixed income and capital income are particularly not well covered in EU-SILC data in general while they contribute significantly to National Account aggregates and income inequality.

B Generalised Pareto versus Pareto II

The Generalised Pareto distribution (GPD) introduced by Pickands (1975) is slightly more general than the Pareto II as the latter corresponds to a parametric restriction ($\xi > 0$) and to a re-parameterisation. The original formulation of the GPD was:

$$F(x|\xi,\varsigma) = 1 - (1 + \xi x/\varsigma)^{-1/\xi}. \tag{58}$$

It reduces to the exponential distribution for $\xi = 0$. The Pareto II model corresponds to positive values of ξ which induce a heavy tail behaviour. For negatives values of ξ, the random variable is bounded above with $0 \leq x \leq -\varsigma/\xi$. The uniform distribution on $[0, \varsigma]$ is obtained for $\xi = -1$. Recovering the original Pareto II implies considering the following re-parameterisation:

$$\alpha = 1/\xi, \qquad \beta = \varsigma/\xi.$$

The `R` packages `SpatialExtremes` and `eva` provide an implementation of an extension of this distribution, with three parameters (location μ, scale ς and

shape ξ:

$$F(x) = 1 - \left[1 + \xi \frac{x-\mu}{\varsigma}\right]^{-1/\xi}, \qquad f(x|\xi,\varsigma) = \frac{1}{\varsigma}\left(1 + \xi \frac{x-\mu}{\varsigma}\right)^{-1-1/\xi}.$$

Arnold (2008) underlines that the statistical literature on the GPD has become quite extensive and thus provides a useful source of information for the Pareto II. We can nevertheless notice a certain confusion of appellation. Mokrani et al. (2016) use the notation of the Pareto II for their Generalised Pareto. Hu and Gui (2018) use a notation similar to the Pareto II, but with $\lambda = 1/\sigma$ which gives $F(x) = 1-(1+\lambda x)^{-\alpha}$. Jenkins (2017) called the Pareto II the following survival function:

$$S(x) = \left[1 + \xi \frac{x-\mu}{\sigma}\right]^{-1/\xi}, \xi > 0,$$

which is nothing but the usual Generalised Pareto, restricted to positive values of ξ. Finally, it is worth noting that the Singh-Maddala (or Burr XII) distribution could be classified also among distributions having a Pareto tail as its CDF corresponds to:

$$F(x) = 1 - [1 + (x/\sigma)^b]^{-\alpha},$$

which is the parameterisation of the Pareto IV with $\mu = 0$. For $b \leq 1$, we have a zero mode distribution.

C Bernstein polynomials for truncated distributions

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 center 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 $x \in [0,1]$:

$$B_k(x,j) = C_k^j x^j (1-x)^{k-j}, \tag{59}$$

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:

$$\hat{f}_{n,k}(x) = (k+1)\sum_{j=0}^{k} \frac{n_j}{n} B_k(x,j). \tag{60}$$

We propose to approximate the coefficients in (60) 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 midsts 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:

$$\hat{f}_{n,k}(x) = \exp(\sum_{j=0}^{k} B_k(x,j)\,\hat{\delta}_j). \tag{61}$$

It has to be normalised 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. We compare in Figure 15 three different estimators for the truncated French income distribution in 2018. For $k = 12$, the Bernstein estimator (61) and the kernel estimator provide very similar results for the bulk of the distribution. But the kernel estimate shows its deficiencies in the right tail, which is of particular concern in our context. The fit of the lognormal is adequate for the right tail, but not for the rest of the distribution, which has motivated Cabras and Castellanos (2011) to propose a semi-parametric estimator for the lower part of the distribution.

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 $Fn = (1,\cdots,n)/(n+1)$.[7] Babu et al. (2002) propose the following estimator:

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

We have the same problem as before for the dimension of the polynomial that we shall solve by introducing a new regression. Because an estimated

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

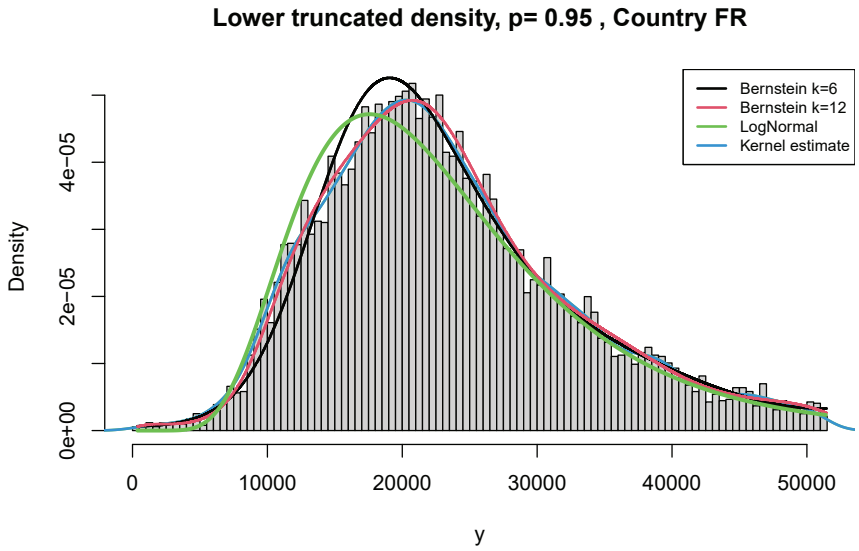

Figure 15: Comparing various density estimators

cumulative is not only a positive, but also an increasing function of x, we have to impose this supplementary restriction. 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:

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

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

$$\hat{F}_{n,k}(x) = \frac{1}{1 + \exp(Z_k(x)\hat{\delta})}. \tag{62}$$

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.

Let us now compare the performance of these estimators in Table 10. We have computed the RMSE between an estimate and the natural estimator alternatively for the whole sample and for the top 10% of French and UK SILC income data for 2018. The quality of the adjustment increases with k as expected. Figure 16 shows that even with a very low value for k, the estimator manages to reproduce very well the shape of the CDF.

Table 10: Comparing CDF estimators

Method	France full	France top 10%	UK full	UK top 10%
Bernstein $k = 3$	0.547	0.097	0.779	0.176
Bernstein $k = 6$	0.470	0.040	0.264	0.070
Bernstein $k = 12$	0.162	0.036	0.253	0.049

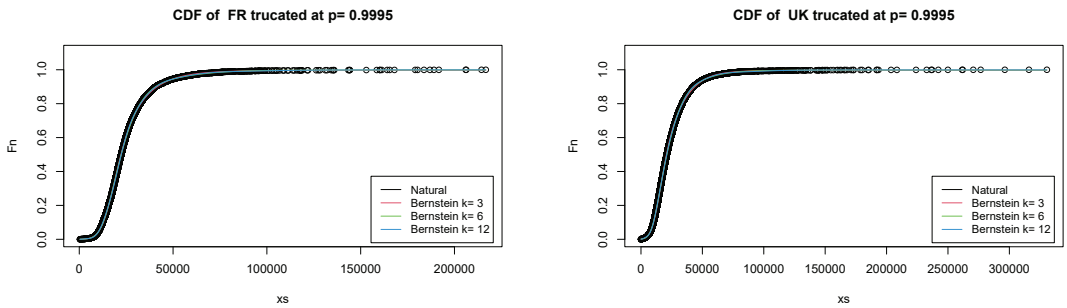

Figure 16: Two estimators for the CDF

D The need of a Pareto II

Is it necessary to adopt a Pareto II tail, instead of a simple Pareto I which is easier to estimate. The answer is given by looking at the posterior density of the difference $\beta - h$. The Pareto I corresponds to the restriction $\beta = h$. So if 0.00 belong to a 90% HPD confidence interval of the posterior density of $\beta - h$, the Pareto I restriction becomes valid. A second information can be drawn from the posterior densities of $\beta - h$. From Figure 2. Depending on the sign of $\beta - h$, the Gini coefficient associated to the Pareto II will be lower ($\beta - h < 0$) or greater ($\beta - h > 0$) than the Pareto I Gini. This means that imposing a Pareto I tail can over-estimate or under-estimate inequality.

Table 11 indicates that a Pareto II is needed in all EUR-15 countries, except for Ireland in 2008, Luxembourg, and Greece in 2008. In all other cases, a Pareto I would induce a positive bias, except for Portugal where the bias would be negative. Note the particular case of Luxembourg where the Pareto II is not needed according to this test but for which the estimated h corresponds to a quite low value of p. Such a low value would not have been possible with a Pareto I model.

For the New Member States, Table 12 indicates that a Pareto I could be valid for more cases (Lituania 2008 and 2018, Czech Republic 2018, Poland 2008). The Pareto II tail induces a greater measure of inequality for Estonia,

Table 11: Posterior distribution of $\beta - h$: EUR-15

	2008		2018	
	$0 \in 90\% HPD$	$\Pr(\beta > h)$	$0 \in 90\% HPD$	$\Pr(\beta > h)$
DK	No	0.00	No	0.00
FI	No	0.00	No	0.00
SE	No	0.00	No	0.00
IE	Yes	0.92	No	0.00
UK	No	0.00	No	0.00
AT	No	0.00	No	0.00
BE	No	0.00	No	0.00
DE	No	0.00	No	0.00
FR	No	0.00	No	0.00
LU	Yes	0.82	Yes	0.90
NL	No	0.00	No	0.00
EL	Yes	0.45	No	0.00
ES	No	0.00	No	0.00
IT	No	0.00	No	0.00
PT	No	1.00	No	1.00

YES means that a 90% HPD of the posterior density of the difference $\beta - h$ contains the value 0.00. In this case, a Pareto I model is enough for modelling the right tail. If the posterior probability that $\beta > h$ is greater than 0.50, the Pareto II model provides a higher measure of inequality than the Pareto I.

Table 12: Posterior distribution of $\beta - h$: NMS

	2008		2018	
	$0 \in 95\% HPD$	$\Pr(\beta > h)$	$0 \in 95\% HPD$	$\Pr(\beta > h)$
EE	No	1.00	No	1.00
LT	Yes	0.62	Yes	0.94
LV	No	1.00	No	1.00
CZ	No	0.02	Yes	0.15
HU	No	0.00	No	1.00
PL	Yes	0.91	No	1.00
SI	No	0.00	No	0.00
SK	No	0.00	No	1.00

YES means that a 90% HPD of the posterior density of the difference $\beta - h$ contains the value 0.00. In this case, a Pareto I model is enough for modelling the right tail. If the posterior probability that $\beta > h$ is greater than 0.50, the Pareto II model provides a higher measure of inequality than the Pareto I.

Latvia and Hungary 2018, Slovakia 2018.