The Bayesian approach to poverty measurement

Lecture 3: Modelling the income distribution using mixtures

Michel Lubrano

November 2022

Contents

1 Introduction

For many purposes, we need to be able to model the income distribution. Just because, as we shall see later, many poverty measurements are just transformations of the parameters of the income distribution. In this chapter, we shall develop a specific model which is the mixture of parametric distributions. A mixture is a very flexible model, much more than the usual parametric forms. The traditional functional forms that are adopted for modelling the income distribution are the lognormal for low and median incomes and the Pareto distribution for high incomes. These are two parameter distributions. More recent distribution include the Singh-Maddala distribution with three parameters and the Generalised Beta II distribution with four parameters. However, all those distributions have a single mode. With more parameters, a mixture of two lognormal densities, for instance, manages to fit bimodal income distributions.

2 Two useful parametric densities

Two densities are particularly useful for modelling the income distribution: the *lognormal* distribution and the *Pareto* distribution. Bayesian inference in these two processes is rather simple, while being slightly more demanding for the Pareto case. Both densities have two parameters.

2.1 The lognormal distribution

The log-normal density was introduced in the economic literature for modelling small to medium range incomes. For instance it is widely used for modelling wages (with the Mincer equation), except for top wages, where a Pareto density has to be used. A random variable X has a log normal distribution if its logarithm $\ln X$ has a normal distribution. If Y is a random variable with a normal distribution, then $X = \exp(Y)$ has a log-normal distribution; likewise, if X is log-normally distributed, then $Y = \ln X$ is normally distributed.

Let us suppose that $y \sim N(\mu, \sigma^2)$ and let us consider the change of variable $x = \exp y$. The Jacobian of the transformation from y to x is given by:

$$J(y \to x) = \frac{\partial y}{\partial x} = \frac{\partial \ln x}{\partial x} = \frac{1}{x}.$$

So, the probability density function of a log-normal distribution is:

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp -\frac{(\ln x-\mu)^2}{2\sigma^2}, \quad x>0.$$

The cumulative distribution function has no analytical form and requires an integral evaluation, but it is directly related to the Gaussian CDF with:

$$F_X(x;\mu,\sigma) = \Phi\left(\frac{\ln x-\mu}{\sigma}\right),$$

Φ being the standard normal cdf. This integral is easy to evaluate on a computer and built-in functions are standard.

The moment are easily obtained as functions of μ and σ. If X is a log-normally distributed variable, its expected value, variance, and standard deviation are

$$\begin{aligned} \mathrm{E}[X] &= e^{\mu+\frac{1}{2}\sigma^2}, && (1)\\ \mathrm{Var}[X] &= (e^{\sigma^2}-1)e^{2\mu+\sigma^2}, && (2)\\ \mathrm{s.d}[X] &= \sqrt{\mathrm{Var}[X]} = e^{\mu+\frac{1}{2}\sigma^2}\sqrt{e^{\sigma^2}-1}. && (3) \end{aligned}$$

Note the expectation depends on both parameters μ and σ^2 when for the Normal density the mean is simply μ. We have also analytical expressions for the Gini coefficient and the Lorenz curve with:

$$G = 2\Phi(\sigma/\sqrt{2})-1, \qquad L(p) = \Phi(\Phi^{-1}(p)-\sigma).$$

2.2 Maximum likelihood for lognormal samples

The likelihood function is rather simple to write once we note that this pdf is just the normal pdf times the Jacobian of the transformation which is $1/x$. We have

$$f_L(x;\mu,\sigma) = \prod_{i=1}^{n}\left(\frac{1}{x\,i}\right) f_N(\ln x_i;\mu,\sigma),$$

where by f_L we denote the probability density function of the log-normal distribution and by f_N that of the normal distribution. Therefore, using the same subscripts, we can write the log-likelihood function in the following way:

$$\begin{aligned} \ell_L(\mu,\sigma|x_1,x_2,\ldots,x_n) &= -\textstyle\sum_i \ln x_i + \ell_N(\mu,\sigma|\ln x_1,\ln x_2,\ldots,\ln x_n)\\ &= \text{constant} + \ell_N(\mu,\sigma|\ln x_1,\ln x_2,\ldots,\ln x_n). \end{aligned}$$

Since the first term is constant with regard to μ and σ, both logarithmic likelihood functions, ℓ_L and ℓ_N, reach their maximum with the same μ and σ. Hence, using the formulas for the normal distribution maximum likelihood parameter estimators and the equality above, we deduce that for the lognormal distribution it holds that

$$\widehat{\mu} = \frac{1}{n}\sum_i \ln x_i, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_i \left(\ln x_i - \widehat{\mu}\right)^2 .$$

This means that in a lognormal sample, the two parameters can be estimated by the sample mean of the logs and the variance of the logs.

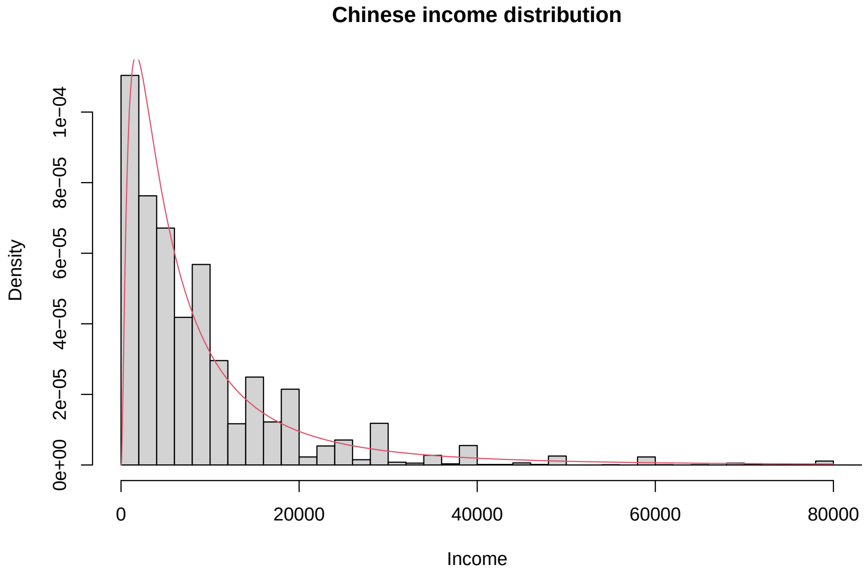

Figure 1: China income distribution with a lognormal adjustment

We have in Figure 1 an histogram of the Chinese income distribution in 2006 and the adjusted lognormal distribution. The estimated parameters are $\hat{\mu} = 8.65$ and $\hat{\sigma} = 1.104$. The corresponding estimated Gini coefficient is 0.565. A direct estimation of the Gini without a parametric assumption would have given 0.527. If we do the same operation on the wages, we get $\hat{\mu} = 6.727$ and $\hat{\sigma} = 0.811$. The implied Gini from the lognormal is 0.434 while the sample Gini is 0.459.

2.3 Bayesian inference for the Lognormal

The likelihood function is the same as in the classical case, but some rewriting is convenient for combining it with the prior:

$$\begin{aligned} L(\mu,\sigma^2|x) &= \left(\prod_{i=1}^{n}(x_i)^{-1}\right)(2\pi)^{-n/2}\sigma^{-n}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log x_i-\mu)^2 \\ &\propto \sigma^{-n}\exp -\frac{1}{2\sigma^2}\sum_{i}(\log x_i-\mu)^2 \\ &\propto \sigma^{-n}\exp -\frac{1}{2\sigma^2}\left(s^2+n(\mu-\tilde{x})^2\right), \qquad (4) \end{aligned}$$

where:

$$\tilde{x}=\frac{1}{n}\sum_{i}\log x_i \qquad s^2=\sum_{i}(\log x_i-\tilde{x})^2.$$

As we can neglect the Jacobian ($\prod_{i=1}^{n}(x_i)^{-1}$), Bayesian inference in the log normal process proceed in the same way as for the usual normal process. In particular, we have natural conjugate prior densities for μ and σ^2. We select a conditional normal prior on $\mu|\sigma^2$ and an inverted gamma2 prior on σ^2:

$$\begin{aligned} \varphi(\mu|\sigma^2) &= f_N(\mu|\mu_0,\sigma^2/n_0)\propto\sigma^{-1}\exp -\frac{n_0}{2\sigma^2}(\mu-\mu_0)^2, \qquad (5) \\ \varphi(\sigma^2) &= f_{i\gamma}(\sigma^2|\nu_0,s_0)\propto\sigma^{-(\nu_0+2)}\exp -\frac{s_0}{2\sigma^2}. \qquad (6) \end{aligned}$$

The prior moments are easily derived as:

$$\begin{aligned} \mathrm{E}(\mu|\sigma^2) &= \mu_0, \qquad \mathrm{Var}(\mu|\sigma^2)=\frac{\sigma^2}{n_0} \qquad (7) \\ \mathrm{E}(\sigma^2) &= \frac{s_0}{\nu_0-2}, \qquad \mathrm{Var}(\sigma^2)=\frac{s_0^2}{(\nu_0-2)^2(\nu_0-4)} \qquad (8) \end{aligned}$$

Let us now combine the prior with the likelihood function to obtain the joint posterior probability density function of (μ, σ^2) in such a way that isolates the conditional posterior densities of each parameter:

$$\varphi(\mu,\sigma^2|x)\propto\sigma^{-(n+\nu_0+3)}\exp -\frac{1}{2\sigma^2}\left(s_0+s^2+n\left(\mu-\tilde{x}\right)^2+n_0(\mu-\mu_0)^2\right).$$

As we are in the natural conjugate framework, we must identify the parameters of the product of an inverted gamma2 in σ^2 by a conditional normal density in $\mu|\sigma^2$. After some algebraic manipulations, the conditional normal posterior is:

$$\begin{aligned} \varphi(\mu|\sigma^2,x) &\propto \sigma^{-1}\exp -\frac{1}{2\sigma^2}\left((n_0\mu_0+n\tilde{x})/n_*\right), \\ &\propto f_N(\mu|\mu_*,\sigma^2/n_*), \end{aligned}$$

with:

$$n_* = n_0 + n, \qquad \mu_* = (n_0\mu_0 + n\tilde{x})/n_*.$$

Then the marginal posterior density of μ is Student with:

$$\begin{aligned} \varphi(\mu|x) &= f_t(\mu|\mu_*, s_*, n_*, \nu_*), \\ &\propto [s_* + n_*(\mu - \mu_*)^2]^{-(\nu_*+1)/2}, \end{aligned} \tag{9}$$

where:

$$\nu_* = \nu_0 + n, \qquad s_* = s_0 + s^2 + \frac{n_0 n}{n_0 + n}(\mu_0 - \tilde{x})^2.$$

The posterior density of σ^2 is given by:

$$\begin{aligned} \varphi(\sigma^2|x) &\propto \sigma^{-(n+\nu_0+2)} \exp -\frac{1}{2\sigma^2}\left(s_0 + s^2 + \frac{n_0 n}{n_0+n}(\mu_0 - \tilde{x})^2\right), \\ &\propto f_{i\gamma}(\sigma^2|\nu_*, s_*). \end{aligned} \tag{10}$$

The posterior densities of μ and σ^2 belong to well-known families. Their moments are obtained analytically and no numerical integration is necessary. We recover the classical results under a non-informative prior.

In fact, all these results should be familiar to you. This inference process is exactly the same as the inference process in the linear regression model. We have simply to consider the log of the observations as the endogenous variable and reduce the explanatory variables to a constant term.

We can illustrate the procedure, using the Chinese income data again. We have simulated 10 000 draws from the posterior density of the parameters of the lognormal model. We found:

$$\mathrm{E}(\mu|y) = 8.654, \quad \mathrm{SD}(\mu|y) = 0.0102, \qquad \mathrm{E}(\sigma^2|y) = 1.218, \mathrm{SD}(\sigma^2|y) = 0.0198.$$

We can then transform the posterior draws of σ^2 into draws of the Gini, using:

$$G^{(j)} = 2\Phi(\sigma^{(j)}/\sqrt{2}) - 1.$$

The posterior expectation of the Gini is 0.565 and a 90% confidence interval is $[0.559, 0.571]$. We can plot the posterior density of the Gini and visualize the 90% HPD region in Figure 2.

The same exercise done with wages gives:

$$\mathrm{E}(\mu|y) = 6.727, \quad \mathrm{SD}(\mu|y) = 0.0120, \qquad \mathrm{E}(\sigma^2|y) = 0.658, \mathrm{SD}(\sigma^2|y) = 0.0137.$$

The posterior expectation of the Gini is 0.434 and a 90% confidence interval is $[0.427, 0.441]$. These values are very near from those found for MLE, except that now we have posterior confidence intervals.

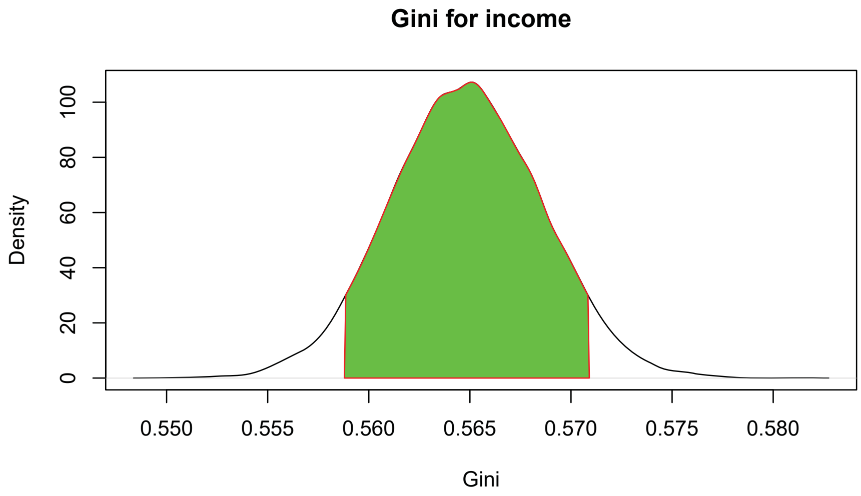

Figure 2: Posterior density of the Gini coefficient for Chinese income in 2006

2.4 Pareto distribution

Pareto (1897) observed that in many populations the income distribution was one in which the number of individuals whose income exceeded a given level x could be approximated by Cx^{α} for some choice of C and α. More specifically, he observed that such an approximation seemed to be appropriate for large incomes, i.e. for x above a certain threshold. If one, for various values of x, plots the logarithm of the income level against the number of individuals whose income exceeds that level, Pareto's intuition suggests that an approximately linear plot will be encountered. In formal terms, a random variable X follows a Pareto distribution if its survival function is:

$$\bar{F}(x) = \Pr(X > x) = \left(\frac{x}{x_m}\right)^{-\alpha}, \quad x > x_m,$$

corresponding to the intuitive characterization of the Pareto. The cumulative function is simply $1 - \bar{F}$ which implies:

$$F(x) = \Pr(X < x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}.$$

We shall verify Pareto's intuition using income and wages for China in 2006 in a subsection. The density is obtained by differentiation of the CDF:

$$f(x) = \alpha x_m^{\alpha} x^{-\alpha-1}, \quad x > x_m.$$

Table 1: Moments of the Pareto distribution

parameters	value	domain
scale	x_m	$x_m > 0$
shape	α	$\alpha > 0$
support	$x \in [x_m; +\infty)$	
median	$x_m \sqrt[\alpha]{2}$	
mode	x_m	
mean	$x_m \frac{\alpha}{\alpha - 1}$	$\alpha > 1$
variance	$x_m^2 \frac{\alpha}{(\alpha-1)^2(\alpha-2)}$	$\alpha > 2$
Gini	$[2\alpha - 1]^{-1}$	$\alpha > 0.5$
Lorenz	$L(p) = 1 - (1-p)^{(\alpha-1)/\alpha}$	$\alpha > 1$

Moments are regrouped in Table 1. They exist only for certain values of α. This is the price to pay for its long tails. This density has a special shape. It is always decreasing, its mode coincide with its origin x_m. So it is valuable only to model high or medium incomes. Note that in Figure 1 we have given the histogram of the Chinese income distribution in 2006 coming from the data collected in the Chinese social survey. Paradoxically, it has the shape of a Pareto density, but the Pareto would certainly provide a poor fit for the whole sample as we shall see below that the Pareto density fits the Chinese income distribution above 13 000 yuans, which corresponds to the top 22% of the distribution.

2.5 Maximum likelihood for Pareto samples

Classical inference is quite easy for the usual Pareto I model. There exists three other variants of the Pareto density that we shall not detail here, see for instance in Arnold (2008), on top of the Generalised Pareto distribution.

Let us suppose that we have an IID sample of X which is drawn from a Pareto I model. The likelihood function is:

$$L(x; x_m, \alpha) = \alpha^n x_m^{n\alpha} \left(\prod x_i\right)^{-(\alpha+1)} \mathbb{1}(x_i \geq x_m).$$

It is easy to see that we have two sufficient statistics which gives immediately the MLE as:

$$\begin{aligned} \hat{x}_m &= x_{[1]} \\ \hat{\alpha} &= \left[\tfrac{1}{n} \sum \log(x_i/x_{[1]})\right]^{-1}. \end{aligned}$$

As underlined by Arnold (2008), these estimators are positively biased in a small sample as we have:

$$\begin{aligned}
\mathrm{E}(\hat{x}_m) &= x_m(1-1/(n\alpha))^{-1} \\
\mathrm{Var}(\hat{x}_m) &= x_m^2 n\alpha(n\alpha-1)^{-2}(n\alpha-2)^{-1} \\
\\
\mathrm{E}(\hat{\alpha}) &= \alpha n/(n-2) \\
\mathrm{Var}(\hat{\alpha}) &= \alpha^2(n-2)^{-2}(n-3)^{-1}.
\end{aligned}$$

Knowing the bias, it is easy to propose unbiased estimators by simply correcting the initial maximum likelihood estimators.

2.6 Bayesian inference for the Pareto process

Bayesian inference in the Pareto process is quite simple if x_m is known. In the case where x_m is also an unknown parameter, inference becomes more delicate and a Gibbs sampler is needed, as will be detailed later on. We treat here only the case where x_m is known.

In the natural conjugate framework, the prior $\varphi(\theta)$ is chosen in such a way that it combines easily with the likelihood function $l(y;\theta)$. The natural conjugate framework relies on the exponential family where sufficient statistics of two samples combine easily. We have to show how the Pareto distribution is related to the exponential family. Suppose X is Pareto-distributed with known minimum x_m and unknown parameter α. Let us consider the following transformation:

$$Y = \log\left(\frac{X}{x_\mathrm{m}}\right).$$

Then Y is exponentially distributed with intensity parameter α, or equivalently with expected value $1/\alpha$:

$$\Pr(Y > y) = e^{-\alpha y}.$$

The cumulative density function is thus $1 - e^{-\alpha y}$ and the pdf:

$$f(y;\alpha) = \begin{cases} \alpha e^{-\alpha y}, & y \geq 0, \\ 0, & y < 0. \end{cases}$$

The likelihood function for α, given an independent and identically distributed sample $y = (y_1, ..., y_n)$ drawn from that variable, is

$$L(\alpha; y) = \prod_{i=1}^{n} \alpha \; \exp(-\alpha y_i) = \alpha^n \; \exp\left(-\alpha \sum_{i=1}^{n} y_i\right) = \alpha^n \exp\left(-\alpha n \overline{y}\right),$$

where

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i/x_m),$$

is the sample mean of y. The conjugate prior for the exponential distribution is the gamma distribution (of which the exponential distribution is a special case). The following parametrisation of the gamma pdf is useful in this case:

$$\varphi(\alpha) = \frac{s_0^{\nu_0}}{\Gamma(\nu_0)}\,\alpha^{\nu_0-1}\,\exp(-\alpha\, s_0),$$

with moments given by:

$$\mathrm{E}(\alpha) = \nu_0/s_0 \qquad \mathrm{Var}(\alpha) = \nu_0/s_0^2.$$

The posterior distribution of α is proportional to the product of the likelihood function defined above and of a gamma prior:

$$\begin{aligned} \varphi(\alpha|y) &\propto L(\alpha; y) \times \varphi(\alpha) \\ &= \alpha^n \exp(-\alpha\, n\overline{y}) \times \frac{s_0^{\nu_0}}{\Gamma(\nu_0)}\,\alpha^{\nu_0-1}\,\exp(-\alpha\, s_0) \\ &\propto \alpha^{(\nu_0+n)-1}\,\exp(-\alpha\,(s_0 + n\overline{y})). \end{aligned}$$

The posterior density has been specified up to a missing normalising constant. Since it has the form of a gamma pdf, this can easily be filled in, and one obtains:

$$\varphi(\alpha|y) = \mathrm{Gamma}(\alpha\,;\, \nu_0 + n, s_0 + n\overline{y}).$$

Here the parameter ν_0 can be interpreted as the number of prior observations, and s_0 as the sum of the prior observations. A non-informative prior corresponds to $\nu_0 = s_0 = 0$.

2.7 A graphical device to determine x_m

Because we have considered x_m as fixed, it becomes crucial to have a rule of thumb to determine a plausible value for it. The survival function of the Pareto is a power function with:

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

So its logarithm is a linear function, suggesting a linear regression:

$$\log(1 - \hat{F}(x)) = -\alpha\log(x) - \alpha x_m + \epsilon.$$

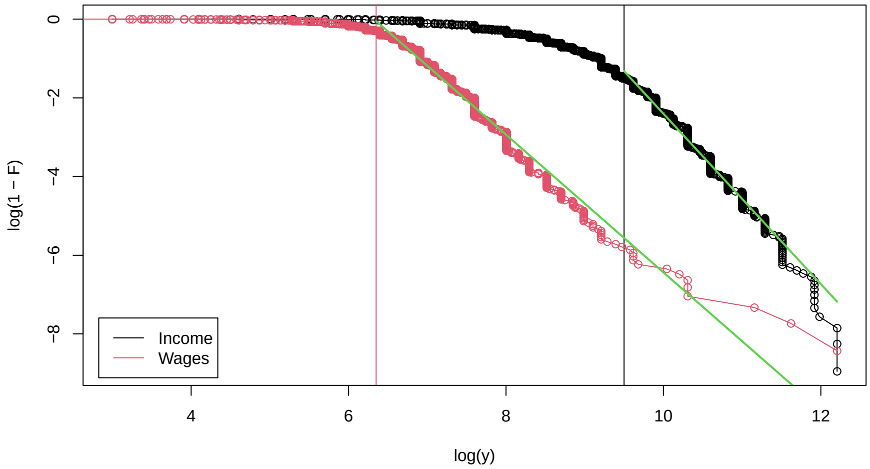

Figure 3: Pareto plot for Chinese income and wages in 2006

This is not a valid procedure to estimate α, but a convenient way of checking if X follows a Pareto process, and starting at which value. The part of the sample corresponding to a Pareto process will correspond to a straight line on the plot of $\log(x)$ against $\log(1-\hat{F}(x))$. We can try this device on two Chinese series: an income series and a wage series where all zeros were excluded. These are variables `qc34a` (wages) and `qd35a` (income) in the Chinese Social Survey of 2006. Figure 3 shows that the Pareto assumption can be assumed for high wages over 700 yuans (top 58% of wages) and for high incomes 13 000 yuans (top 23% of incomes). This graphical device is useful for determining x_m, but it does not provide a feasible way for making inference on α.

From these graphs, we can choose $x_m = 14000$ for incomes and $x_m = 1000$ for wages. Let us $m = 10000$ draws form the posterior density of α given x_m, which is a gamma density. The posterior moments of α and Gini are for top incomes:

$$\mathrm{E}(\alpha|y, x_m) = 1.911, (0.0468), \qquad \mathrm{E}(G|y, x_m) = 0.355, (0.0118).$$

and for top wages

$$\mathrm{E}(\alpha|y, x_m) = 1.640, (0.0416), \qquad \mathrm{E}(G|y, x_m) = 0.439, (0.0160).$$

The Gini is quite different among top incomes and top wages.

3 Mixture of distributions

3.1 Definition

A finite mixture of distributions is simply a weighted sum of densities. We give a simple example involving two lognormal distributions:

$$f(x) = \eta f_\Lambda(x|\mu_1, \sigma_1^2) + (1 - \eta) f_\Lambda(x|\mu_2, \sigma_2^2).$$

In this example, we have twice two parameters of the lognormal, which makes four parameters and the mixing parameter η, which makes a total of five parameters. Mixtures are used when the observed sample comes from several different sub-samples, depicting such heterogeneity. If we know who is who, i.e. which part of the sample has been generated by which member of the mixture, then the inference problem is very simple as we shall see for the World Income Distribution. If we do not have this knowledge, more sophisticated methods have to be used, for instance the Gibbs sampler.

3.2 Properties

Mixture have very nice properties due to linearity. The cumulative distribution is just the weighted sum of the cumulative of each member, so:

$$F(x) = \eta F_\Lambda(x|\mu_1, \sigma_1^2) + (1 - \eta) F_\Lambda(x|\mu_2, \sigma_2^2).$$

The uncentered moments are just the weighted sum of the uncentered moments of each member. This is wrong of course for the variance.

3.3 Mixtures where the weights are known

The World Income Distribution (WID) is modelled using mixtures. The case is rather simple because we know who is who. More precisely, we have a collection of samples and each sample belong to a specific country. The second thing we know are the weights. The usual practice is to compute the total population and then to attribute to each country a weight which is its population divided by the total population of the world or of the reference group if we restrict for instance our attention to the European Union. The inference problem is then limited to a series of separated inference problems, one per country for which we have to estimate a distribution. The most simple case is to assume a lognormal distribution for each country, but this a rather restrictive assumption. More complex distributions have been used in the literature with Gamma densities, or generalised beta II densities.

3.4 Mixtures with unknown weights

This case is much more complex. To relate it to the previous case of the WID, suppose that we have a collection of data samples, but we do not know to which country they belong. We know that our observed sample corresponds to a collection of countries, but everything is mixed up. We can then suppose that a mixture is needed to represent this collection of sample. But the number of members for the mixture is not necessarily equal to the number of countries. Either because two countries look the same, or because inside a single country, the population has very heterogeneous incomes. In this case, the mixture is just a convenient way to represent heterogeneity and it is difficult to identify a member to a specific country.

4 A simple estimate of the world income distribution

In the previous chapter, we have left unfinished the question of the IPL. We have determined a posterior distribution for the IPL and the probability of a country to belong to the small reference group of poor countries. But we have not determined the final number of poor in the world as well as its geography. For this, we have estimate a world income distribution (WID). This is a research question in itself. We shall review here a very simple method based on the assumption that in each country, the income distribution is a lognormal. The WID is then obtained as the weighted sum of these distribution, the weights being given by the population shares.

4.1 A parametric income distribution for the developing world

Atkinson and Bourguignon (2001) were the first to propose to implement this simple idea. In each country, the income distribution can be represented by a lognormal density $f_\Lambda(y|\mu,\sigma^2)$. We know that the mean of a lognormal distribution is equal to $\exp(\mu+\sigma^2/2)$. Atkinson and Bourguignon (2001) propose to calibrate the mean using the PPP figure for daily consumption per capita, while pegging σ^2 at different prior values. This is a very crude method, due to the lack of availability of adequate data.

Holzmann et al. (2007) have proposed to calibrate σ^2 using data on the Gini coefficient, as the formula for the Gini in the lognormal density is $2\Phi(\sigma/\sqrt{2})-1$, which depends only on σ. Data for the Gini coefficients are available from the World Bank, presumably using different sources, such as survey data. In both papers, the world income distribution is obtained by aggregating national income distributions using population shares.

In Xun and Lubrano (2018), we wanted to use more information, noting that the Gini coefficient may not be sufficient to provide a precise indication on the shape of the left tail of the distribution if uncertainty concerning the value of the Gini coefficient is high. For instance, the Gini for consumption and the Gini for income might not be the same. The World Bank provides extra information that can be used to model the left tail of the income distribution, in the form of headcount poverty rates for two values of the poverty line, namely \$1.25 and \$2.00, using 2005 PPP. The theoretical headcount poverty rate corresponds to $F_\Lambda^{-1}(1.25|\mu,\sigma)$ for a \$1.25 poverty line for instance, where $F_\Lambda^{-1}()$ represents the quantile function of the lognormal distribution. We have collected these supplementary data in order to construct a loss function for each country:

$$\begin{aligned} Loss &= (pv_{1.25} - F_\Lambda^{-1}(1.25,\mu,\sigma))^2 + (pv_{2.00} - F_\Lambda^{-1}(2.00,\mu,\sigma))^2 \\ &\quad + (C - \exp(\mu + \sigma^2/2))^2 + (Gi - 2 * \Phi(\sigma/\sqrt{2}) + 1)^2. \end{aligned}$$

Here $pv_{1.25}$ is the empirical headcount for \$1.25 and $pv_{2.00}$ the corresponding value for \$2.00. C is the empirical daily mean consumption per capita and Gi is the empirical Gini coefficient for one country.[1] This is a method of moments. We propose to minimise this loss function for each country separately.

We managed to minimise our loss function for each of the 74 countries of our sample with no significant outlier. Using this method, we aim to obtain an income distribution for each country that is consistent with both the macro data of mean consumption per capita and with some microeconomic measures of dispersion, in particular for the left tail of the distribution. Then we aggregate these national adjusted distributions using population pop_i as a weight, imposing that the weights sum to one to get the world distribution of income (WDI) $f_W(y)$:

$$f_W(y) = \sum_{i=1}^{74} \eta_i f_\Lambda(y_i|\mu_i,\sigma_i^2), \qquad \eta_i = pop_i / \sum pop_i.$$

Figure 4 shows the graph of this estimated mixture of 74 lognormals, together with two poverty lines, the old \$1.00 a day and our revised proposal of \$1.48 (without weighting). China and India represent 53% of the population of

[1] An update using the 2011 PPP would mean of course considering a totaly new data set, with updated Gini coefficients and new poverty headcounts for \$1.90 and \$3.10 which are the two values of the poverty now documented in Povcal.

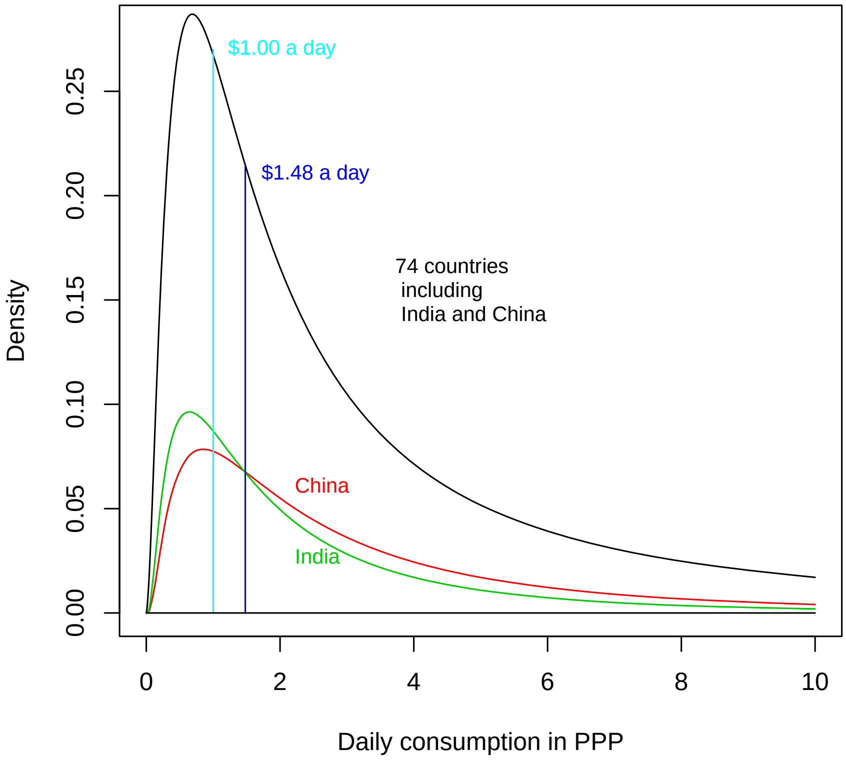

Figure 4: Income distribution for 74 developing countries around 2001

our sample. They have different income distributions: China is richer, but with more inequality. The overall distribution is fairly smooth, probably because we have only 74 countries representing the developing world. Very rich countries or regions like the US and Europe are excluded, so no income polarisation at world level can be detected.

Collecting the results of the previous chapter, we are now equipped with all we need to obtain a posterior distribution of the number of poor people. We have a probabilistic way to determine the composition of the reference group, the distribution of the IPL and a representation (albeit imperfect) of the income distribution of each of the countries. First, we characterise the number of poor people inside the reference group and then we generalise these computations to the whole group of 74 countries in our data base.

4.2 Modelling the poverty count in the reference group

We have provided a proper way of defining the group of countries for which a single poverty line can be used, the IPL for which we have a posterior density. What is the number of poor people in that group? We shall work conditionally on our estimated WDI. First, for each draw of the parameters of our two regime model, we get a sample separation depending on the value of $\theta^{(j)}$. The countries in the reference group are those for which $C_i < \theta^{(j)}$. We then deduce a draw for the poverty line $z^{(j)}$. For each country of the reference group, we compute a poverty rate by inverting its lognormal income distribution:

$$h_i^{(j)} = F^{-1}(z^{(j)}|\hat{\mu}_i, \hat{\sigma}_i^2).$$

We then multiply this rate by the national population N_i to get the corresponding number of poor people in that country. By aggregation over the countries of the reference group, we get a draw for the posterior density of the number of poor people in the reference group:

$$np^{(j)} = \sum_{i \in [C_i < \theta^{(j)}]} h_i^{(j)} N_i.$$

For M draws, we have an estimation of the posterior density of the number of poor people in the reference group which takes into account the stochastic composition of that group. In the left panel of Figure 5, we provide a graph of our posterior density for three different models of the IPL. With the unweighted model, we get a number of poor people averaging to 1 448 million (and a standard deviation of 105). If we weight by population, the average number of poor people rises slightly to 1 505 (standard deviation of 89). Weighting the regression by the official number of poor people increases the mean slightly to 1 584 (78). These figures, presented in Table 2, differ greatly according to the method of weighting. This is mainly because we only focus on the reference group, whose composition changes with the method of weighting. There is a strong discontinuity effect. The rightmost curve corresponds to the highest number of poor people derived from a mean poverty line of \$1.63, while the curve in the middle is derived from a slightly higher poverty line (\$1.65) but indicates a slightly lower number of poor people. This is simply because the type of weighting chosen affects wether or not China and Indonesia are included in the reference group. We are thus looking for a poverty line definition and an evaluation of the number of poor people that would be less sensitive to discontinuity.

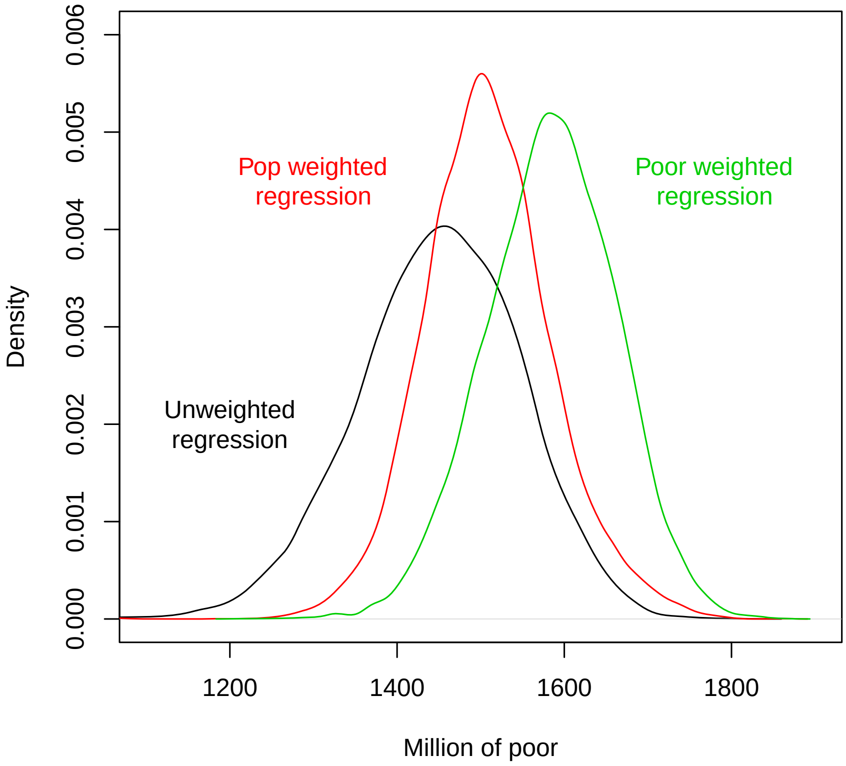

Figure 5: Posterior density of the number of poor people in the reference group around 2001

4.3 Modelling the poverty count in the developing world (74 countries)

The objective of Deaton (2010) was to find a mechanism to determine an international poverty line which did not include any discontinuity. That meant not having a reference group. However, the argument led by Atkinson and Bourguignon (2001) supports having at least two different poverty lines, depending on the income level of the different countries: "*...to provide a framework which unifies the measurement of poverty in developing and developed countries*". In this subsection, we try to simultaneously analyse the determination of a poverty line and the determination of the number of poor people for our 74 countries, representing most of the developing countries and a sample of moderately developed countries.

Our two regime model assumes two types of poverty lines, an IPL common to all the countries in the reference group and a collection of relative poverty lines, each specific to a country outside this group. For this second poverty line, we can take the national poverty line, provided it is greater than the random draw of the IPL. So for each draw of the parameters of our two regime model, we have:

$$\tilde{z}_i^{(j)} = \mathbb{1}(C_i \leq \theta^{(j)})\, IPL^{(j)} + \mathbb{1}(C_i > \theta^{(j)}) \max(z_i, IPL^{(j)}). \tag{11}$$

From this draw $\tilde{z}_i^{(j)}$, we determine 74 poverty rates:

$$h_i^{(j)} = F^{-1}(\tilde{z}_i^{(j)} | \hat{\mu}_i, \hat{\sigma}_i^2),$$

which are aggregated into

$$np^{(j)} = \sum_{i=1}^{74} h_i^{(j)} N_i,$$

in order to get a draw from the posterior density of the number of poor people in the world. This procedure involves no specific discontinuity, but rather a comprehensive definition of the number of poor people, mixing both absolute poverty and inclusion. Figure 6 contains three graphs of this posterior density, depending on the method of weighting. There is still a difference between weighting or not weighting, but the method of weighting is now less of a factor. Moreover, the ordering of the posterior densities of the number of poor people is now consistent with the ordering of the level of the mean poverty lines.

Table 2: Poverty count in the developing world around 2001 (millions)

Group	Reference	World	China	India
Poverty line	IPL	$\max(IPL, z_i)$	IPL	IPL
Unweighted	1 448 (105)	1 698 (95)	409 (32)	498 (32)
Pop weighted	1 505 (89)	1 846 (72)	459 (25)	547 (24)
Poor weighted	1 584 (78)	1 833 (75)	455 (26)	543 (25)
Official figures	Reference	World	China	India
	1 195	1 599	360	416

Official figures were computed using the official poverty rate at the national poverty line. No figures exist for 43 countries in the World Bank data set. So we determined which of the normalised poverty lines of the World Bank ($1.25, $2.00, $2.50, $4.00 and $5.00) was closest to the national poverty line and took the corresponding poverty rates.

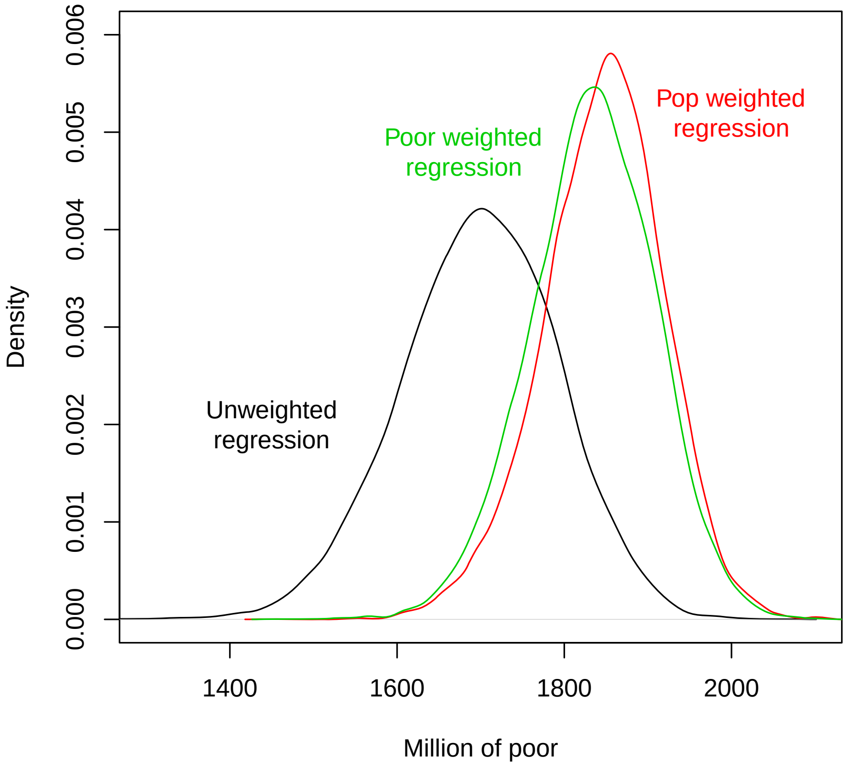

Figure 6: Posterior density of the number of poor people in the world around 2001

Using the comprehensive definition (11) of a poverty line, we reach an estimation of the number of poor people in the developing world, reported in Table 2, corresponding to a period around 2001. Table 3 details how these individuals are divided among the six traditional regions of the World Bank. Most of the poor are located in East Asia (China) and South Asia (India). Weighting does not have much of an influence on the ranking of poverty counts. Using our IPL defined in (11) (which is \$1.48 in the unweighted case for the reference group), we find 1 698 million. We collected the official poverty rates on the web site of the World Bank. When we multiply the official poverty rates by the population and sum up the countries, we find a total of 1 599 million poor people. Our unweighted evaluation of 1 698 million appears to be consistent with the information contained in our data base. We provided posterior densities and posterior confidence intervals.

Table 3: The location of poor people in the developing world around 2001

Region	Unweighted	Pop weighted	Poor weighted
Africa	245	263	262
East Asia	576	639	634
East. Europe	36	36	36
Latin America	177	177	177
MENA	26	29	28
South Asia	639	702	697

Figures are in millions. It was not possible to obtain feasible standard deviations because the poverty line is fixed outside the reference group.

These are of course conditional on the available data and we did not include in our estimation the possibility of measurement error. However beyond the question of measurement error, different types of data lead to different evaluations of consumption and inequality. More precisely, there can be huge differences between consumption evaluated with national account data and consumption evaluated with survey data (see e.g. Deaton, 2005).

5 Inference for the mixture model

The case of the WID that we presented above is a very simplified case for two reasons. First, we know the sample separation which are the countries. Second inside each country, the income distribution is modelled using the simple lognormal assumption. And its parameters were more calibrated than estimated because we had very little information. We are going here to present a more general case, using individual data.

5.1 Why the lognormal is not such a good idea

We use the UK Family Expenditure Survey of 1979. Which kind of density can we fit to the income data reported in this survey? Does the lognormal provides a good fit as would for instance suggest a comparison between the parameter free Lorenz curve and the Lorenz curve implied by a lognormal model? In fact, Figure 7 displays the limitations of the lognormal model. It was obtained with the following `R` code:

```
plot(density(y79))
lines(dlnorm(seq(0,350,1), meanlog=mean(ly79),
      sdlog=sd(ly79)),col="red")
```

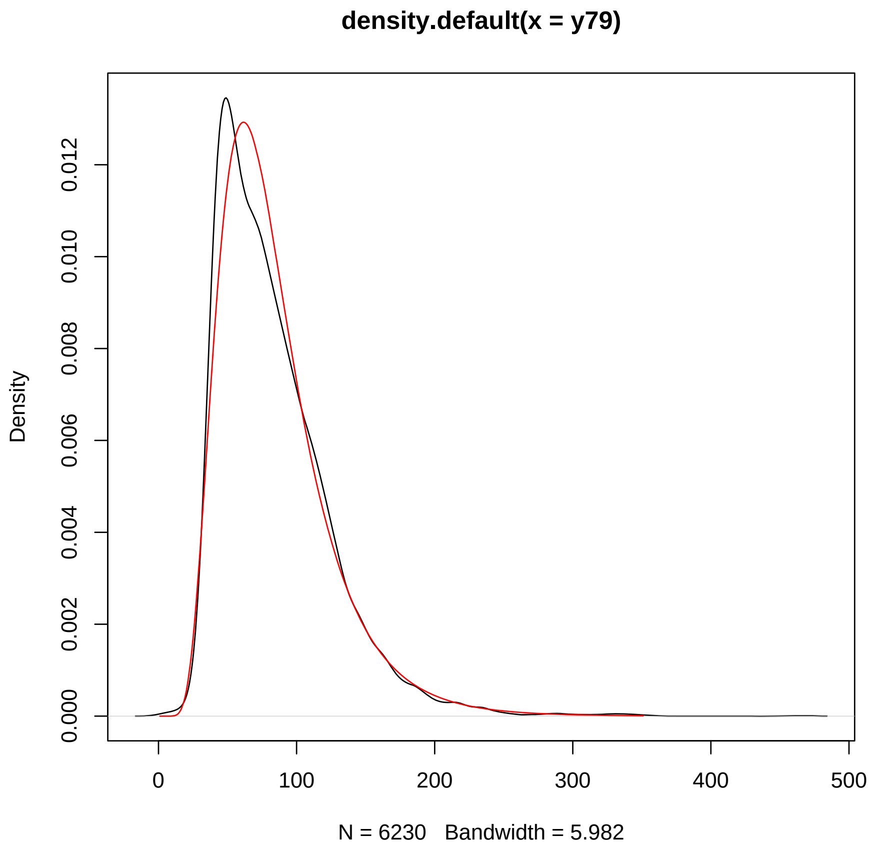

Figure 7: Non parametric estimate of the density for FES79 compared to a lognormal fit

We see clearly that if the overall fit of the lognormal could pass for being nice, the two modes are of course smoothed into something with is even not in between, while the right tail seems to be fitted quite well. So the lognormal model is not adequate to describe completely the sample.

5.2 Estimation procedure

Let us consider a mixture of two normal distributions, $f_N(x|\mu_i, \sigma_i^2)$. It is convenient to introduce a new random variable called Z that will be associated to each observation x_i and that will say if x_i belongs to the first component of the mixture $z_i = 1$ or to the second component of the mixture $z_i = 2$. Suppose that we know the n values of z. We can compute easily the following conditional sufficient statistics:

$$\begin{array}{ll}
n_1(z) = \sum \mathbb{1}(z_i = 1) & n_2(z) = \sum \mathbb{1}(z_i = 2) \\
\bar{x}_1(z) = \frac{1}{n_1} \sum x_i \times \mathbb{1}(z_i = 1) & \bar{x}_2(z) = \frac{1}{n_2} \sum x_i \times \mathbb{1}(z_i = 2) \\
\bar{s}_1(z) = \frac{1}{n_1} \sum (x_i - \bar{x}_1(z))^2 \times \mathbb{1}(z_i = 1) & \frac{1}{n_2} \bar{s}_2(z) = \sum (x_i - \bar{x}_2(z))^2 \times \mathbb{1}(z_i = 2)
\end{array}$$

These statistics give direct estimates for the parameters of the two members. Of course we do not know the z_i, but we can compute the following probabilities for each observation:

$$\Pr(z_i = 1|x, \bar{\theta}) = \frac{\hat{\eta} \times f_N(x_i|\bar{\theta}_1)}{\hat{\eta} \times f_N(x_i|\bar{\theta}_1) + (1 - \hat{\eta}) \times f_N(x_i|\bar{\theta}_2)}$$

provided we have evaluated η as $\hat{\eta} = n_1(z)/n$ and $\bar{\theta}$ using the conditional sufficient statistics. We have then two solutions for allocating the observations between the two regimes:

- We allocate observation i to the first member if $\Pr(z_i = 1|x, \bar{\theta}) > 0.5$.
- We randomly allocate observation i to one regime according to a binomial experience with probability $\Pr(z_i = 1|x, \bar{\theta})$.

Once we have chosen between the two possibilities, we iterate the process. A deterministic allocation corresponds to the EM algorithm of Dempster et al. (1977) while a random allocation corresponds to an algorithm which is not far from a Bayesian Gibbs sampler.

5.3 Difficulties of estimation

As we have already said, estimating a mixture of densities is not a simple task. In the above writing of the data density, all the parameters are free to move in their domain. The likelihood function

$$L(x; \theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} \eta_j \times f(x|\mu_j, \sigma_j^2)$$

goes to infinity if one of the σ_j goes to zero which happens if there are less observations in one cluster than there are parameters to estimate. So only a local maximum can be found.

The EM algorithm or the Gibbs sampler have global convergence properties. The EM algorithm converges to the maximum likelihood estimator. But both algorithms are sensitive to starting values.

There is a fundamental identification problem which is called a labelling problem. The likelihood function does not change is we change the order of the parameters. So, a usual way of identifying the parameters consists in imposing an ordering, either on the means or the variances. But this ordering should not go against the sample properties. So some checks have to be done.

5.4 Estimating mixtures in R

The complexity of the estimation procedures is reflected in the packages proposed in R. One of the many different available packages is `mixdist`. We shall now detail its use. In order to simplify the problem, the program start by considering an histogram, which means grouped data. So we have first to select the number of cells in the histogram. Then we have to give starting values for the parameters, and first of all the number of components. It it is quite safe to start by estimating a two component mixture. Mixture of a higher order are difficult to manipulate and many references in the empirical literature indicate that they are rarely successful. Usually an equal weight is given as a starting value for the η_i. A visual inspection of the histogram gives clues about plausible values for the mean. The prior variance is small when the prior mean correspond to a sharp part of the histogram and much larger for the prior mean corresponding to the tail.

```
library(mixdist)
  FES.mix = function(y){
  chist = hist(y,breaks=100)
  y.gd = mixgroup(y,breaks=chist$breaks)
  y.par = mixparam(mu = c(50,80), sigma = c(10,50))
  y.res = mix(y.gd,y.par,"lnorm")
  print(y.res)
  plot(y.res)
}
FES.mix(y79)
```

In this code, we first determine break points with the instruction `hist`. Then, `mixgroup` is used for grouping the observations using the previously computed break points. `mixgroup` creates a data frame containing grouped data, a data frame being a special type of object in *R*. `mixparam` creates a data frame containing starting values for the mean and the standard deviation. If no other argument is given, it is assumed that the starting p are all equal while summing to one. `mix` is the proper function for estimation. It has at least three arguments: two data frames for the observatons and the parameters. The third arguments give the density which is used. The choices for continuous densities are "*norm*", "*lnorm*", "*gamma*" and "*weibull*". Note that the last case `weibull` needs special type of entry for its parameters. The function `weibullpar` takes as an entry the prior mean and the prior standard deviation and creates a data frame containing the shape, scale and location parameters of the Weibull.

For FES 1979, we could not estimate a mixture of more than two components. We fitted two lognormals. The estimated parameters are given in Table 4. We must note that the estimation gives values for the mean and the standard deviation of the sample, and not for the parameters of the lognormal. This is the same for the starting values. For recovering the parameters of the two underlying lognormal distributions, we have

Table 4: Output for a twin mixture

member	η	μ	σ
1	0.1369	45.42	6.764
2	0.8631	89.14	40.811

$$\mu = \ln(\mathrm{E}(X)) - \frac{1}{2}\ln\left(1+\frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2}\right), \tag{12}$$

$$\sigma^2 = \ln\left(1+\frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2}\right), \tag{13}$$

from which we compute the parameters of the underlying two lognormal distributions in Table 5.

Table 5: Parameter estimates for a twin mixture

member	η	μ	σ^2
1	0.1369	3.805	0.0225
2	0.8631	4.398	0.1878

The graph show that the fit is rather good. It is rather difficult to identify a particular to group to each of these members. The second group seems to correspond to the large segment of the population as $\eta_2 = 0.85$ and the corresponding mean is not too large with $\mu_2 = 90$. The first group correspond to poorer people. A poverty line of half the mean is equal to 41.54.

Figure 9 compares three estimates for the 1979 UK income distribution. If we take the NP estimate as the truth, we see the large bias provided by the simple lognormal and how this bias is strongly reduced by simply considering a mixture of two lognormal distributions.

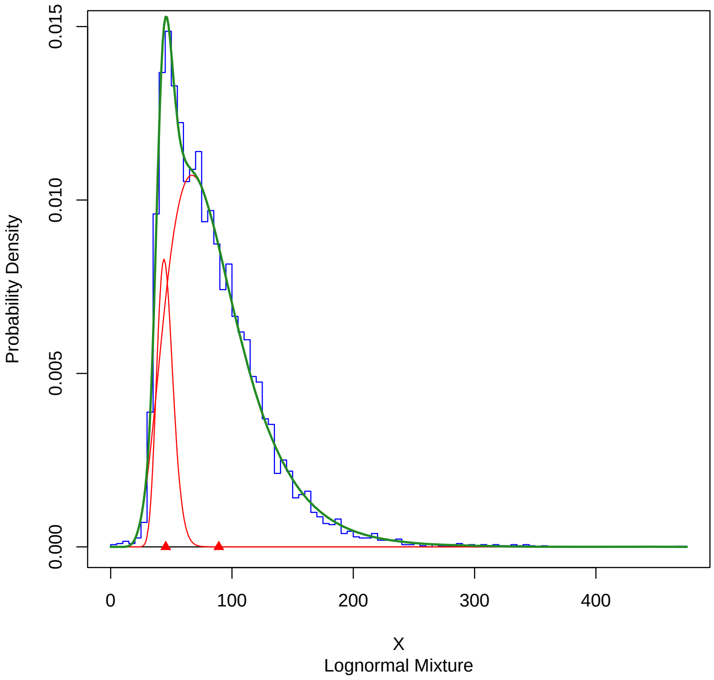

Figure 8: Mixture of two lognormal densities

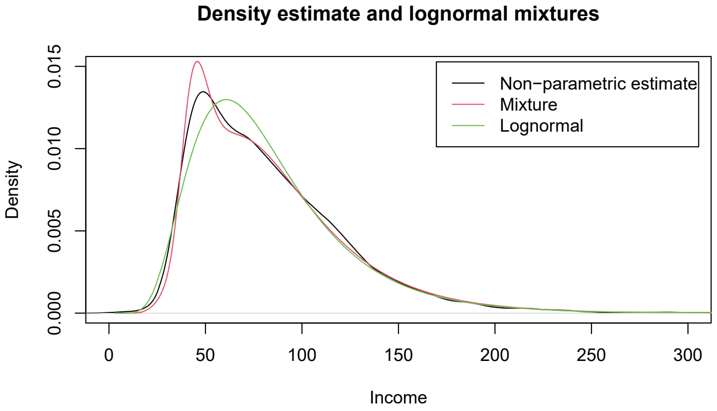

Figure 9: Mixture versus non-parametric density estimate

6 Bayesian inference for mixtures of log-normals using survey data

Fourrier-Nicolaï and Lubrano (2020) made use of mixture of distributions in order to investigate child poverty in Germany. The mixture was estimated using survey data and weights. It builds also on earlier work with Lubrano and Ndoye (2016) who introduced the use of a mixture of lognormal densities to make inference on an income distribution in a Bayesian framework. We can recall that mixtures of gamma densities were also considered in Chotikapanich and Griffiths (2008) for modelling the income distribution. Fourrier-Nicolaï and Lubrano (2020) introduced specifically sampling weights and zero income observations.

6.1 Finite mixture of log-normals

A finite mixture $f(y|\vartheta)$ of lognormal densities is a linear combination of k parametric densities $f_\Lambda(y|\theta_j)$ such that:

$$f(y|\vartheta) = \sum_{j=1}^{k} \eta_j f_\Lambda(y|\theta_j), \qquad 0 \leq p_j < 1, \quad \sum_{j=1}^{k} \eta_j = 1, \tag{14}$$

where $\vartheta = (\eta, \theta)$ and the parameter vectors are $\theta = (\theta_1, ..., \theta_j)$ and $\eta = (\eta_1, ..., \eta_k)$ with η_j and θ_j being, respectively, the weight and the parameters of the j-th component. We assume that all components arise from the univariate log-normal distribution $f_\Lambda(y; \mu_j, \sigma_j)$. The log-normal has two parameters, and its pdf is given by:

$$f_\Lambda(y; \mu, \sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp \frac{-(\ln y - \mu)^2}{2\sigma^2},$$

with $\sigma \in [0; +\infty[$ being the shape parameter and $\mu \in]-\infty; +\infty[$ the location parameter.

6.2 Mixtures as an incomplete data problem

Bayesian inference in this mixture model looks very similar to the previous classical procedure explained for a mixture of two normal densities. We have to deal with two issues. First, the classification of observations into the k different components with probability η_j. Second, the estimation of the parameters for every component density. The problem would simplify greatly if the classification of the observations were known. This led Diebolt and Robert

(1994) to consider a mixture problem as an *incomplete data problem.* Each observation y_i has to be completed by an unobserved variable z_i taking a value in $\{1, ..., k\}$, indicating from which member of the mixture each y_i comes. The model has to explain the couple (y_i, z_i). The EM algorithm in a classical framework and the Gibbs sampler in a Bayesian framework start from an initial hypothetical sample separation $[z_i]$ and conditionally on $[z_i]$ make inference on the parameters ϑ. Once the sample allocation is known, we can treat each component separately meaning that μ_j, σ_j are estimated for all $j = 1, ..., k$ from the observations in group j only, whereas estimation of p is based on the number $n_1(z), ..., n_k(z)$ of observations allocated to each group. This means that with this approach we have simplified the global problem of inference into k separate inference problems, that are simple to treat because they are identical to what was treated above. Once we have these first results, we can determine a new sample separation $[z_i]$, given the previous values found for μ_j, σ_j and η_j. This approach is particularly well suited in a Bayesian framework because given $[z_i]$ we can manage to find conjugate prior for each sub-model $f_\Lambda(y|\mu_j, \sigma_j k)$ and for η_j.

As explained for instance in Lubrano and Ndoye (2016), the natural conjugate priors for each member of a mixture of log-normals are a conditional normal prior on $\mu_j|\sigma_j^2 \sim f_N(\mu_j|\mu_0, \sigma_k^2/n_0)$, an inverted gamma prior on $\sigma_j^2 \sim f_{i\gamma}(\sigma_j^2|v_0, s_0)$. A Dirichlet prior is used for $\eta \sim f_D(\gamma_1^0, ..., \gamma_k^0)$. The hyper-parameters of these priors are $v_0, s_0, \mu_0, n_0, \gamma_k^0$.

For a given sample separation, we get the following sufficient statistics:

$$\begin{aligned} n_j &= \sum_{i=1}^{n} \mathbb{1}(z_i = j), \\ \bar{y}_j &= \frac{1}{n_j} \sum_{i=1}^{n} \log(y_i) \mathbb{1}(z_i = j), \\ s_j^2 &= \frac{1}{n_j} \sum_{i=1}^{n} (\log(y_i) - \bar{y}_j)^2 \mathbb{1}(z_i = j). \end{aligned}$$

Let us combining these sufficient statistics with the prior hyperparameters, we get :

$$\begin{aligned} n_{*j} &= n_0 + n_j, \\ \mu_{*j} &= (n_0\mu_0 + n_j\bar{y}_j)/n_{*j}, \\ v_{*j} &= v_0 + n_j, \\ s_{*j} &= s_0 + n_j s_j^2 + \frac{n_0 n_j}{n_0 + n_j}(\mu_0 - \bar{y}_j)^2, \end{aligned}$$

which are used to index the conditional posterior densities of first σ_j^2 which is still an inverted gamma:

$$\varphi(\sigma_j^2|y,z) = f_{i\gamma}(\sigma_j^2|v_{*j}, s_{*j}), \tag{15}$$

and second of $\mu_j|\sigma_j^2$, which is a conditional normal:

$$\varphi(\mu_j|\sigma_j^2, y, z) = f_N(\mu_j|\mu_{*j}, \sigma_j^2/n_{*j}). \tag{16}$$

The conditional posterior distribution of η_j is a Dirichlet with:

$$\varphi(\eta|y,z) = f_D(\gamma_1^0 + n_1, ..., \gamma_k^0 + n_k) \propto \prod_{j=1}^{k} \eta_j^{\gamma_j^0 + n_j - 1}. \tag{17}$$

We can then determine the posterior probability that the i-th observation comes from the j-th component $z_i = j$ conditionally on the value of the parameters. It is given by:

$$Pr(z_i = j|y, \theta) = \frac{\eta_j f_\Lambda(y_i|\mu_j, \sigma_j^2)}{\sum_j \eta_j f_\Lambda(y_i|\mu_j, \sigma_j^2)}. \tag{18}$$

6.3 Label switching and prior information

A recurrent problem when estimating mixture models is due to label switching. Label switching comes from the fact that the likelihood function does not change if the labels of the parameters of two members of the mixtures are switched. The likelihood function has $k!$ equivalent modes due to label switching. This is not a problem for maximum likelihood estimation as only one maximum is selected among $k!$. But it becomes a problem for Bayesian inference, particularly when estimating posterior marginal densities because we do not know the exact behaviour of the Gibbs sampler which can explore alternatively several regions of the likelihood function, corresponding to several maxima. An extensive discussion of this question is provided in (Fruhwirth-Schnatter, 2006, p. 78). There are common rules to reduce this problem and ensure identification of the mixture model. We can impose the ordering of one of the component parameters, for instance we can impose for each MCMC draw that the μ_j or the σ_j must be ordered. These solutions are not equivalent and the limitations of these practices are discussed in Fruhwirth-Schnatter (2001).

How to build a sample based prior information? We have understood that it is difficult to make inference for the parameters of a mixture without prior information. The usual practice is to provide the same prior information for each member of the mixture in the form of a normal-inverted gamma2 prior centered on the sample mean and the sample variance. However, Lubrano and Ndoye (2016) note that this is not the best way to solve the label switching question. Lubrano and Ndoye (2016) prefer to introduce a separate prior for each member which centered on different parts of the sample, with for instance increasing prior means or increasing prior variances. This is the counterpart of the starting values that have to be provided in the classical approach. With this type of prior, it is less necessary to order the MCMC draws as traditionally suggested for avoiding label switching.

6.4 A Gibbs sampler algorithm

Let us propose the following Gibbs sampler algorithm:

Algorithm 1 Gibbs sampler for mixtures

```
Set $k$ the number of components, $m$ the number of draws, $m_0$ the number of warming draws and initial values of the parameters $\vartheta^{(0)} = (\mu^{(0)}, \sigma^{(0)}, \eta^{(0)})$ for $l = 0$.
for $l = 1, ..., m_0, ..., m + m_0$ do
    Generate a classification $z_i^{(l)}$ independently for each observation $y_i$ according to a multinomial process with probabilities given by equation (18), using the value of $\vartheta^{(l-1)}$.
    Compute the sufficient statistics $n_j, \bar{y}_j, s_j^2$.
    Generate the parameters $\sigma^{(l)}, \mu^{(l)}, \eta^{(l)}$ from the posterior distributions given in equations (15), (16) and (17) respectively, conditionally on the classification $z^{(l)}$.
    Order $\sigma^{(l)}$ such that $\sigma_1^{(l)} < ... < \sigma_k^{(l)}$ and sort $\mu^{(l)}$, $\eta^{(l)}$ and $z^{(l)}$ accordingly.
end for
Discard the first $m_0$ stored draws to compute posterior moments and marginals.
```

There are packages in `R` where this is programmed. `BayesMix` is an example, well suited to be used with the book Fruhwirth-Schnatter (2006). But it is restricted to Gaussian mixtures.

6.5 Introducing survey weights

In population studies, it is common to sample individuals through complex sampling designs in which the population is not adequately represented in the sample: some individuals or groups can be over or under-represented. Analysing data from such designs is tricky, since the collected sample is not representative of the overall population. To correct for discrepancies between sample and population, survey weights are constructed. However, literature on the estimation of mixtures most of the time ignores this issue, or is concerned with specific cases as Kunihama et al. (2016) and their quoted references for stratification. We shall propose a simple method, easy to implement within a Gibbs sampler, to introduce sampling weights.

Consider that n individuals are sampled from the whole population with survey weights $w_i = c/\pi_i$ with c being a positive constant and π_i the inclusion probability that individual i belongs to the survey. A mixture estimate of the income distribution representative of the genuine population can be obtained by using the weighted sufficient statistics in step 2.(b) of the Gibbs sampler such that:

$$\begin{aligned} n_j &= \sum_{i=1}^{n} w_i \mathbb{1}(z_i = j), \\ \bar{y}_j &= \frac{1}{n_j} \sum_{i=1}^{n} w_i \log(y_i) \mathbb{1}(z_i = j), \\ s_j^2 &= \frac{n_j}{n_j^2 - \sum_{i=1}^{n} w_i^2 \mathbb{1}(z_i = j)} \sum_{i=1}^{n} w_i (\log(y_i) - \bar{y}_j)^2 \mathbb{1}(z_i = j). \end{aligned}$$

The other steps of the Gibbs sampler are left unchanged. Re-weighting the conditional sufficient statistics is enough to modify the sample allocation performed in step 2.(a). The method in fact simply consists in introducing an unbiased weighted estimator for the j-th component sample mean $\bar{y}_j$ and the sample variance s_j^2.

In Figure 10, we compare two non-parametric estimator of a density, one without using sample weights, the second using sample weights. The difference is striking.

6.6 Modelling zero-inflated income data

In household survey data we observe an excess number of zeros (greater than expected under the distributional assumptions). Particularly in income studies, zero incomes are numerous when measured before taxes and redistribution. Actually, a large part of the population has no market income: elderly persons, unemployed workers, children, ... This is a problem when estimating the income distribution in both a parametric approach and a non-parametric approach using smoothing techniques. As the log-normal is defined on the

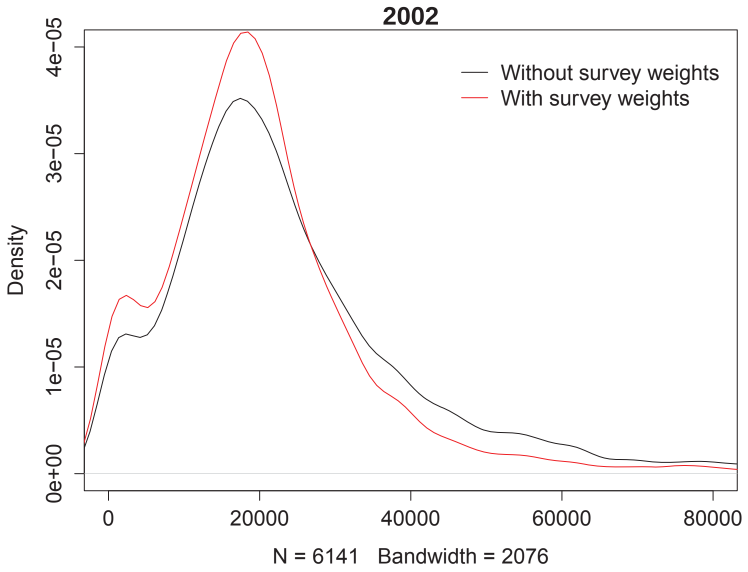

Figure 10: The influence of weight for density estimation

strict positive support, we have to add an extra-component for modelling the zero incomes:

$$f(y|\vartheta) = \mathbb{1}(y=0)\omega + \mathbb{1}(y>0)(1-\omega)\sum_{j=1}^{k}\eta_j f(y|\theta_j), \qquad (19)$$

where $\omega = \Pr(y=0) \simeq (\sum_i \mathbb{1}(y_i = 0)w_i)/\sum w_i$. This is a zero-inflated mixture model. ω is estimated as the (weighted) proportion of zeros in the sample, while inference on the other parameters is made on the sample excluding the zeros. Hence zeros are not a problem for inference. But we have to take them into account when modelling the income distribution.

Figure 11 is particularly interesting. It present the income distribution in Germany. Inference is made using the German Socio Economic Panel (GSOEP). It concerns gross income, before redistribution. So there are households with a zero income which causes difficulties on the left part of the graph. The non-parametric estimate is not at ease with this feature as shown with the black line. However, this estimator is using sampling weights. The blue line is the Bayesian estimator for a mixture of three lognormal densities, taking into account the zero incomes.

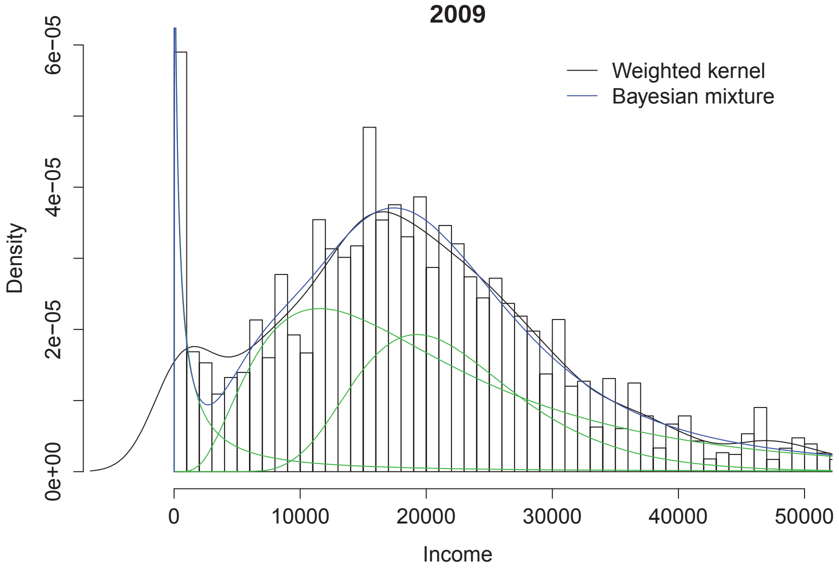

Figure 11: Income distribution before redistribution in Germany in 2009 using the GSOEP

7 Return to the Pareto process

We have now a new tool which is the Gibbs sampler. We are going to use it to make inference for the two parameters of the Pareto I process. Bayesian inference for the Pareto process has been treated in a number of papers as detailed in Arnold (2008). The usual practice is to use a Gibbs sampler after a re-parametrisation in $\tau = 1/x_m$ where τ is called a precision parameter. This parametrisation is convenient for Bayesian inference because both a Pareto prior on τ and a gamma prior on α are natural conjugates priors. The conditional posterior of τ is itself a Pareto density, while the conditional posterior of α is a gamma density. However, a Pareto prior on τ is difficult to interpret while x_m has a natural sample interpretation. We found that it is possible to keep the usual parametrisation of the Pareto process if we choose the prior on x_m as a power function density.

7.1 Power functions

A random variable X is said to have a power function distribution if its probability density function is defined as:

$$p(x) = \alpha x_m^{-\alpha} x^{\alpha-1} \mathbb{1}(x < x_m), \qquad \alpha > 0, \quad x_m > 0.$$

It is an increasing function of x for $\alpha > 0$ and is defined over $[0, x_m]$. Its moments are:

$$\mathrm{E}(x) = \frac{\alpha}{\alpha+1} x_m \qquad \mathrm{Var}(x) = \frac{\alpha}{(\alpha+1)^2(\alpha+2)} x_m^2.$$

They always exist, contrary to the Pareto process. The cumulative distribution function is:

$$F(x) = x_m^{-\alpha} x^{\alpha} \mathbb{1}(x < x_m).$$

Two sufficient statistics are provided by $\mathrm{Max}(x)$ and $\sum \log(x_i/x_m)$.[2] If x has a power function distribution in $(\alpha,\ x_m)$, then $y = 1/x$ is distributed according to a Pareto(α, y_m) where $y_m = 1/x_m$. We have chosen to present separately this distribution even if it corresponds to a simple transformation of the Pareto I because we shall use its properties and moments to elicit a prior information.

7.2 Likelihood function and prior densities

Let us consider a series of n observed incomes $(y_1, \dots, y_n)$ coming from a Pareto distribution. The associated likelihood function is:

$$L(y; \alpha, y_m) = \alpha^n y_m^{\alpha n} \prod y_i^{-(\alpha+1)} \mathbb{1}(y_{(1)} > y_m),$$

where $y_{(1)}$ is the first order statistics, i.e. the minimum of the sample. The two parameters are y_m the location parameter and α the shape parameter. It is convenient to rewrite this likelihood function as:

$$L(y; \alpha, y_m) = \alpha^n \exp\left\{-(\alpha+1) \sum \log(y_i) + \alpha\, n \log(y_m)\right\} \mathbb{1}(y_{(1)} > y_m).$$

It is clear that the Pareto distribution does not belong to the exponential family when its two parameters are unknown, just because the support depends on one of the parameters, namely y_m. However, from this writing, we can find that $y_{(1)}$ and $\sum \log(y_i)$ are two sufficient statistics. In fact, conditionally on y_m, the Pareto does belong to the exponential family.

We propose an independent prior $p(\alpha, y_m) = p(\alpha)p(y_m)$, which simplifies greatly the discussion. When y_m is known, $\log(y/y_m)$ is distributed according to an exponential distribution. In this case, the natural conjugate prior for α is the Gamma density with ν_0 degrees of freedom and as scale parameter α_0:

$$p(\alpha|\nu_0, \alpha_0) \propto \alpha^{\nu_0 - 1} \exp(-\alpha\alpha_0), \quad \mathrm{E}(\alpha) = \nu_0/\alpha_0, \mathrm{Var}(\alpha) = \nu_0/\alpha_0^2.$$

[2] It is simple to draw random numbers using the inverse transform method with $x = x_m\, u^{1/\alpha}$ and $u \sim U(0,1)$. For a Pareto process, we have $y = y_m\, u^{-1/\alpha}$.

A non-informative prior corresponds to letting the prior parameters go to the limit of their domain of definition with $\alpha_0 = 0$ and $\nu_0 = 0$:

$$p(\alpha) \propto 1/\alpha.$$

When α is known, it is also possible to find a convenient conjugate prior for y_m. The conjugate prior is a Power function distribution with shape parameter α_0 and scale parameter y_{m0}:

$$p(y_m|\gamma_0, y_{m0}) = \gamma_0 \, y_{m0}^{-\gamma_0} y_m^{\gamma_0 - 1} \mathbb{1}(y_m < y_{m0}).$$

A non-informative prior is obtained for $\gamma_0 = 0$ and letting y_{m0} go to infinity:

$$p(y_m) \propto 1/y_m.$$

7.3 Conditional posteriors

Conducting Bayesian inference jointly on the two parameters is quite cumbersome. It is possible to derive the two marginal distribution, but they do not belong a class of known densities. So it is better to try to implement a Gibbs sampler. For that, it is enough to derive the conditional posterior densities of the two parameters, just remembering that the marginal posterior of y_m is ill behaved when the prior density of this parameter is non-informative. So for making inference on (y_m, α), we have to be informative at least on y_m.

The conditional posterior of α given y_m is

$$p(\alpha|y_m, y) \propto \alpha^{n+\nu_0-1} \exp -\alpha(\sum \log(y_i) + \alpha_0 - n \log(y_m)).$$

This is a Gamma density $\mathrm{G}(\alpha_*, \nu_*)$ where:

$$\nu_* = \nu_0 + n \qquad \alpha_* = \alpha_0 + \sum \log(y_i/y_m).$$

The conditional posterior of y_m given α is obtained by neglecting all the elements which are independent of y_m in the product of the likelihood function times the prior:

$$p(y_m|y, \alpha) \propto y_m^{\alpha n + \gamma_0 - 1} \mathbb{1}(y_m < y_i) \mathbb{1}(y_m < y_{m0}).$$

We identify a Power function density $\mathrm{PF}(\gamma_*, y_{m*})$ with parameters:

$$\gamma_* = \gamma_0 + n\alpha \qquad y_{m*} = \mathrm{Max}(\mathrm{Min}(y_i), y_{m0}).$$

We note that the support of the conditional posterior density y_{m*} depends on the minimum value of the sample and on the value of y_{m0}. Collecting these results, inference on α and y_m is conducted using a Gibbs sampler. If y_m were given, inference would rely only on the Gamma posterior density $p(\alpha|y_m, y)$.

8 Conclusion

Modelling the income distribution is essential if we want to analyse various curves that we are going to detail in the next lectures. In particular TIP curves, growth incidence curves and dominance curves. For finding the posterior density of these curves, we have simply to transform posterior draws from the posterior density of the parameters of the income distribution. It is then essential to have a precise modelling of the income distribution and mixture of distributions become a very precious tool.

References

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.

Atkinson, A. B. and Bourguignon, F. (2001). Poverty and inclusion from a world perspective. In Stiglitz, J. E. and Muet, P.-A., editors, *Governance, Equity and Global Markets.* Oxford University Press.

Chotikapanich, D. and Griffiths, W. (2008). Estimating income distributions using a mixture of gamma densities. In *Modeling Income Distributions and Lorenz Curves*, volume 5 of *Economic Studies in Equality, Social Exclusion and Well-Being*, pages 285–302. Springer, New York.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics*, 87:1–19.

Deaton, A. (2010). Inequality, and the measurement of world poverty. *American Economic Review*, 100:3–34.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375.

Fourrier-Nicolaï, E. and Lubrano, M. (2020). Bayesian inference for TIP curves: an application to child poverty in Germany. *The Journal of Economic Inequality*, 18:91–111.

Fruhwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer-Verlag New York.

Holzmann, H., Vollmer, S., and Weisbrod, J. (2007). Perspectives on the world income distribution - beyond twin peaks towards welfare conclusions. Technical Report 158, Ibero America Institute for Econ. Research (IAI).

Kunihama, T., Herring, A. H., Halpern, C. T., and Dunson, D. B. (2016). Nonparametric Bayes modelling with sample survey weights. *Statistics & Probability Letters*, 113:41–48.

Lubrano, M. and Ndoye, A. A. J. (2016). Income inequality decomposition using a finite mixture of log-normal distributions: A Bayesian approach. *Computational Statistics and Data Analysis*, 100:830 – 846.

Xun, Z. and Lubrano, M. (2018). A Bayesian measure of poverty in the developing world. *Review of Income and Wealth*, 64(3):649–678.