The econometrics of inequality and poverty
Chapter 5: Modelling the income distribution

Michel Lubrano

October 2017

Contents

1 Introduction

In this chapter, we enter into deep statistical questions concerning the types of samples we are confronted to (surveys) and the statistical analysis which are involved. Those methods can be quite simple when they rely on order statistics. However, samples are designed in a complex way and inference has to take into account weights to compute means, standard deviation and any other indices. When we want to make inference on densities, we confronted a simple choice: with minimum of prior information on the shape of the density, we have access to non-parametric statistics and smoothing. If we are ready to impose more information, we have to select a parametric form and make inference on the parameters. With a parametric approach, we have a better precision, but we can miss some details of the income distribution. A compromise between efficiency and flexibility is to use mixture of distributions. In this case, a Bayesian approach can be valuable. A complementary reading to this chapter can be found in first chapter of Deaton (1997) which contains a lot of valuable material.

2 Types of survey samples

The data we are interested in are survey data concerning households. Many types of information can be asked to household such as unemployment, wages, education, health status. Here we are mainly concerned with income and sometime consumption. We have a finite population of size N, like the French, the UK or the Chinese population. We want to draw a sample of a smaller size n from that population. How can we proceed? The design of a survey has to follow precise rules. We want to get information on a population and it is too costly to ask the entire population every year (especially in China!). A census occurs at most every five years and gives information on the whole population. The coverage of the population is usually not complete: homeless people, armed forces,...

2.1 Random samples

A survey has to be framed, which means that we have to know the size and composition of the true population. A census is useful to frame a survey, other administrative data can be used too. The census for instance provide a list of households to sample. Or social security numbers.

Then we have to decide about the size n of the survey. The sample survey is then drawn at random. The sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i,$$

is a good estimator for the population mean. As we can obtain different samples for the same population, this estimator has a variance estimated by:

$$\text{Var}(\bar{x}) = \frac{1}{n(n-1)}\sum_{i}^{n}(x_i - \bar{x})^2.$$

Remember the classical result about the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$.

2.2 Using weights

Let us now suppose that we want to get more information on a particular group. That group will be more sampled that the other groups on purpose. It will be over represented: for instance to study the economic impact of AIDS, it is useful to sample in regions where AIDS is more present. If we compute the mean using the simple above formula, the mean will be biased. In this case the sample has to be reweighted to make it representative of the population.

Suppose that we have a population of N households and a sample of n observations. Each household has a probability π_i of being drawn in a sampling scheme with replacement (simplification assumption). For each household, we define a weight:

$$w_i = \frac{1}{n\,\pi_i}.$$

In the usual random case, $\pi_i = 1/N$, so that all the weights are the same and equal to N/n and the sum of the weights is equal to N. We can now compute the weighted mean:

$$\hat{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i}.$$

This is an unbiased estimator of the population mean. The variance of this estimator is

$$\text{Var}(\hat{x}_w) = \frac{n}{n-1}(\sum_{i=1}^{n} \nu_i^2 (x_i - \bar{x}_w)^2)$$

where $\nu_i = w_i / \sum w_i$ are the normalized weights. This variance is minimum when the sampling probabilities are chosen proportional to x_i.

Taking into account weights or not can make a difference. Let us consider again the CGSS. This is a weighted sample with the variable *weight*. Let us consider the income variable and its summary statistics. Table 1 show that there can be large differences for mean and quantiles.

Table 1: Weighted and un-weighted summary statistics

	Min	$Q_{0.25}$	$Q_{0.50}$	Mean	$Q_{0.75}$	Max	Gini
Un-weighted	20	3 000	6 000	9 972	12 000	250 000	0.527
Weighted	20	2 000	5 000	8 186	10 000	250 000	0.538

Of course, minimum and maximum ae unaffected. The involved packages in `R` are `weights` for sufficient statistics and `reldist` for the gini with weights. Weights are sometimes directly available as in the `density` command.

2.3 Stratified samples*

The effect of stratification is to break up a single survey into multiple independent surveys. This is interesting to do when sub-populations vary considerably. Members of the population are grouped into relatively homogeneous subgroups before sampling. The strata should be mutually

exclusive: every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive: no population element can be excluded. Then random sampling is applied within each stratum.

Suppose that we have S strata, that the population size is N while the population in each strata is N_s. The mean of the population is now estimated by

$$\bar{x} = \sum_{s=1}^{S} \frac{N_s}{N} \bar{x}_s,$$

where $\bar{x}_s$ is the estimated mean for each strata. In each strata, we can of course have a particular weighting scheme which is superimposed to the stratification. Stratification often improves the representativeness of the sample by reducing sampling error. It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population. In fact

$$\text{Var}(\bar{x}) = \sum_{s=1}^{S} (\frac{N_s}{N})^2 \text{Var}(\bar{x}_s),$$

because the strata are independent. It can be shown that this variance is lower than the variance of

$$\bar{x}_{srs} = \sum_{s=1}^{S} \frac{n_s}{n} \bar{x}_s,$$

where the weights are formed not using the population size, but the sample size and is finally just the sample mean of the unstratified sample.

2.4 Grouped data

Survey data report private information on households. These data are politically sensitive depending on their content. For instance, there are in France questionings about the use of racial information to study discrimination. In Belgium, it is forbidden to ask question on the language used at home (French or Flemish). So for a long time, these data were simply not available. Researcher had access to data that were so aggregated, that they were presented in groups. The treatment of these grouped data needed special tools and estimation techniques. For instance, Singh and Maddala or McDonald use grouped data for the US income. The remaining columns represent the class frequency. We reproduce here these data in Table 2 as given in McDonald (1984). We have percentages summing 100% in all the columns with dates. The first column represent the end of class for each group. It is presumably in thousands dollars per year per household. This lead to an histogram that has to be drawn by hand.

There is a case when data are given in the form of classes. It is when those data concern very small geographical areas. Giving the exact income would make it too easy to find back the concerned person. We have in mind income data given at the school district level in the US which were used for instance in Benzidia et al. (2017)

Table 2: US Data on income

Endpoints	1970	1975	1980
2.5	6.6	3.5	2.1
5.0	12.5	8.5	4.1
7.5	15.2	10.6	6.2
10.0	16.6	10.6	6.5
12.5	15.8	11.4	7.3
15.0	11.0	10.9	6.9
20.0	13.1	18.8	14.0
25.0	4.6	11.6	13.7
35.0	3.0	9.5	19.8
50.0	1.1	3.2	12.8
∞	0.5	1.4	6.7

Source: McDonald (1984).

3 Natural estimators and resampling methods

In this section, we give indications on how to estimate usual quantities such as cumulative distributions, Lorenz curves, Gini indices using order statistics. The method can be extended so as to consider FGT poverty indices, poverty deficit curves and dominance curves. Most of the time, standard errors or small sample distributions are difficult to obtain so that resampling techniques such as the bootstrap are very useful.

3.1 The use of order statistics

The first estimation techniques that we shall present now are relatively simple. They use order statistics which come from the ordering of the observations. Suppose that the observations from X are ordered by increasing value and let us note this ordering as

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

$x_{(1)}$ represents the smallest observation and $x_{(n)}$ the largest one. In this case, it becomes easy to estimate a cumulative distribution and its quantiles. As a matter of fact, a distribution is defined as $F(x) = \text{Prob}(X < x)$. It can be approximated by

$$\text{Prob}(X \leq x_{(i)}) \simeq i/n$$

when we have enough observations.

The first **decile** of this distribution corresponds to the value $x_{0.10}$ such that $\text{Prob}(X \leq x_{0.10}) = 0.10$. It will be enough to find the observation which rank i corresponds roughly to $i/n = 0.10$ in the ordered sequence X. In the general case, let us note $Q(p)$ the quantile of order p; it can be estimated as

$$Q(p) = x_{(s)} \qquad s - 1 \leq np \leq s.$$

This means that the quantile of order p is the observation having rank s^{th} so that the above inequality is verified. This solution is valid in large samples. In smaller samples, an interpolation can be needed.

The estimated quantiles can lead to the computation of the dispersion measure known as the interquartile range $(x_{0.75} - x_{0.25})/x_{0.50}$.

Using the same order statistics, we can define an estimator for the **generalized Lorenz curve**. The generalized Lorenz curve is defined by the partial sum of the ordered quantiles. Thus

$$Lc(p = i/n) = \frac{1}{n} \sum_{j=1}^{i} x_{(j)}.$$

We have used here partial sums of order statistics. The usual Lorenz curve obtains by normalizing this curve by the sample mean.

Finally, the **Gini coefficient** can be estimated as seen in the previous chapter using a simple weighted sum of order statistics. Which is simpler than just evaluation the double sum of the original definition based on the mean of the absolute difference between each possible pair of observations:

$$\hat{I}_G = \frac{2}{n(n-1)\hat{\mu}} \sum_i i\, x_{(i)} - \frac{n+1}{n-1}.$$

This type of computation can also be used to for Sen-Schorrocks-Thon poverty index:

$$\hat{I}_{SST} = \frac{1}{n^2} \sum_{i=1}^{q} (2n - 2i + 1) \frac{z - x_{(i)}}{z}.$$

where q corresponds to the rank of the poverty line z in the distribution of X.

3.2 Bootstraping

Thus we have simple estimators, but we do not know all the time how to compute standard deviations. For instance it was rather easy to compute the variance of the mean. But the variance of the mode is much more difficult to establish, especially when the sampling design is more complex. The bootstrap is a method for assessing sampling variability of an estimator.

There are two sources of randomness:

1. We have samples from a finite population. We must know the sample design, which can be quite complicated in order to appreciate the source of randomness. Not always easy. For instance N might not be known precisely.

2. There are errors of observations, or simply the nature of the variable which is observed is random as it results from decision making under uncertainty.

The bootstrap is resampling technique designed to simulate the small sample distribution of a given statistics. The bootstrap resamples with replacement n data from the original sample. For each bootstrap sample, the statistics is computed, so that with m replications of it, a mean and a

variance can be evaluated. The resampling technique can be quite complicated, because it has to mimic the data generating process.

The bootstrap is available in R with the package `boot`. We must first call the library `boot`. Then define a function with two arguments: the fist argument represents the original data, the second argument indicates the weights of the bootstrapping generated by the package. Here we have given an example with the Gini coefficient, asking for 1000 replications.

```
library(boot,Gini)
r = boot(y79, function(d,i){a=Gini(d[i])}, R=1000)
 hist(r$t, probability=T, col='light blue',
             main="Distribution of the Gini")
 lines(density(r$t),col="red")
 print(r)
 boot.ci(r, type = "norm")
```

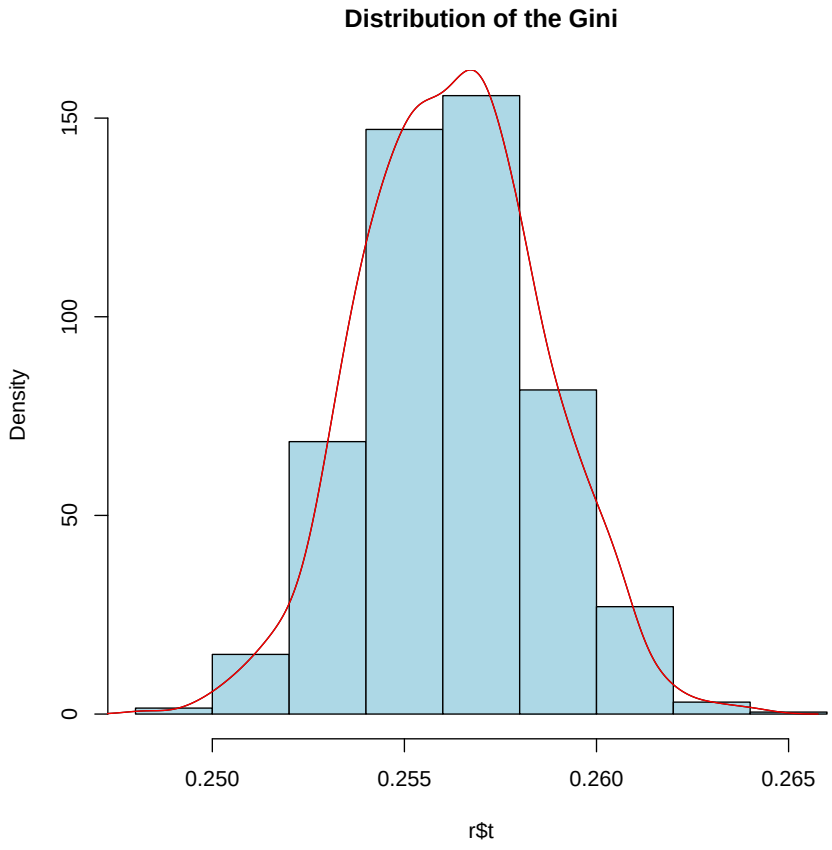

Figure 1: Bootstrapping the Gini

The `boot.ci` function generates 5 different types of equi-tailed two-sided nonparametric confidence intervals. These are the first order normal approximation, the basic bootstrap interval, the studentized bootstrap interval, the bootstrap percentile interval, and the adjusted bootstrap percentile interval. The type of interval is selected in the calling list. In the example, type = "norm" is selected.

The bootstrap gives us a standard deviation and a 95% confidence interval in Table 3. In

Table 3: Bootstrap results for the Gini coefficient using the 1979 FES and the CGSS

	Gini	Bias	std. error	95%
UK	0.256	-7.55e-05	0.00233	[0.252, 0.261]
China	0.500	-0.000124	0.00328	[0.494, 0.507]

Figure 1, we give a graphical representation of the small sample dispersion of the Gini coefficient for the UK. We do not claim that this is the right way to bootstrap the Gini coefficient. This is just an illustration.

4 Non parametric estimation of densities

Densities are much complex to estimate than distributions, just because the above natural estimate of a distribution is not differentiable. Some smoothing has to be used, so this section is devoted to nonparametric estimation using kernels. Most of the material presented in this section and the next ones comes from the book by Pagan and Ullah (1999) which is a valuable reference.

4.1 Histograms

If X is a continuous random variable, we define a neighbourhood of x by $x \pm h/2$ and we count the number of observations x_i that belong to this neighbourhood. Let us define the transformation $\psi_i = (x - x_i)/h$, then

$$f_1(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \mathbb{1}(-1/2 \leq \psi_i \leq 1/2).$$

We notice that x is the centre of the class and that h implicitly defines the number of classes. The indicator function integrates up to 1 as well as $f_1(x)$. Intuitively, we understand that the number of classes can grow with the number of observations, so that $h \to 0$ when $n \to \infty$.

This is a rather crude way of estimating a density. But this is the only way when using group data as the one given above for the US income. In `R`, this can be programmed directly using the function `hist`. In Figure 2, we have used data coming from the Family Expenditure Survey for 1979. The code is:

```
hist(y79,breaks=50)
```

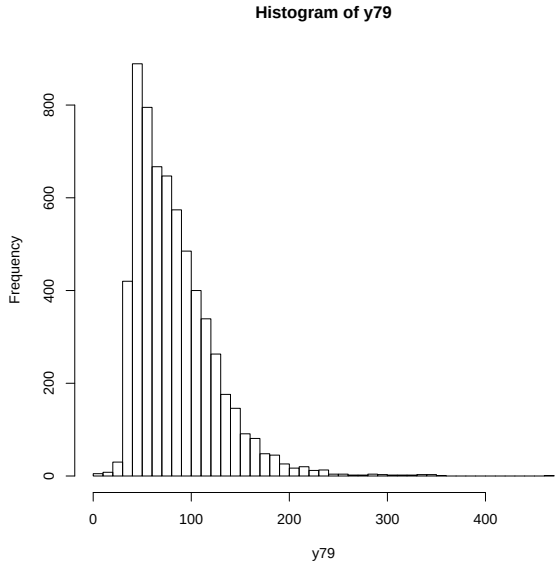

Figure 2: Histogram with 50 cell of FES 1979

where y79 is the FES data for 1979. This graph is relatively regular and gives a good idea of the UK income distribution in 1979. Let us now the same approach, using this time the CGSS income data for 2006. The shape of the Chinese income distribution is quite different. We did

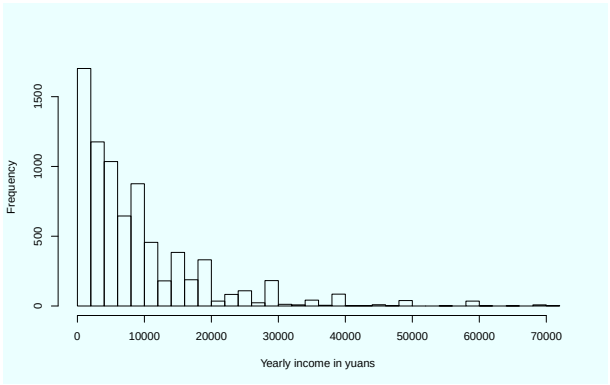

Figure 3: Histogram with 50 cell of Chinese Yearly Income

not use weights. We truncated the data, discarding incomes greater than 80 000 yuans. It i much more like a Pareto distribution, when the UK distribution had the shape of a lognormal.

4.2 Kernel estimation

The histogram has the bad property of being a step function: it is discontinuous and not differentiable. We would like to get a smooth representation, and we feel that this is possible when we have a full sample and not grouped data. Rosenblatt (1956) had the idea of replacing the indicator function by a kernel K which integrates to one like the indicator function. We thus have the new estimator:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\psi).$$

We can deduce some of the properties of a Kernel estimator from those of the indicator function associated with the histogram.

- $\int K(\psi)\, d\psi = 1$,
- $h \to 0$ when $n \to \infty$,
- $K(\pm\infty) = 0$,
- A common choice for K is the standardized normal density. Then $K(|\psi| \geq 3) \simeq 0$.
- The value chosen for h is capital for defining the neighbourhood $|x - x_i|/h \leq 3$.

It is very important to understand the role played by h in determining the shape of the obtained density. We have simulated 500 observations drawn from a mixture of normals N(μ_i,1) with $\mu_1 = 1$, $\mu_2 = 5$ and $p = 0.75$.

$$f(x) = 0.75f(x|1,1) + 0.25f(x|5,1).$$

We then have estimated the density of these random draws using the kernel approach and three values form the window size h. We give the resulting graphs in Figure 11. For the while, we accept the fact that the optimal value of h is given by

$$h = c\hat{\sigma} \times n^{-1/5}.$$

We have selected three values for c in the following graphs. The bimodal nature of the density is well captured in the central graph; it disappears in the first graph where we have over-smoothing while sampling errors are well visible in the last graph where we have under-smoothing.

4.3 Density estimation with weighted samples

When there are weights w_i, we must first impose that the weights sum to unity. The usual formula is simply modified into

$$f(x) = \frac{1}{n\,h}\sum w_i K\left(\frac{x - x_i}{h}\right)$$

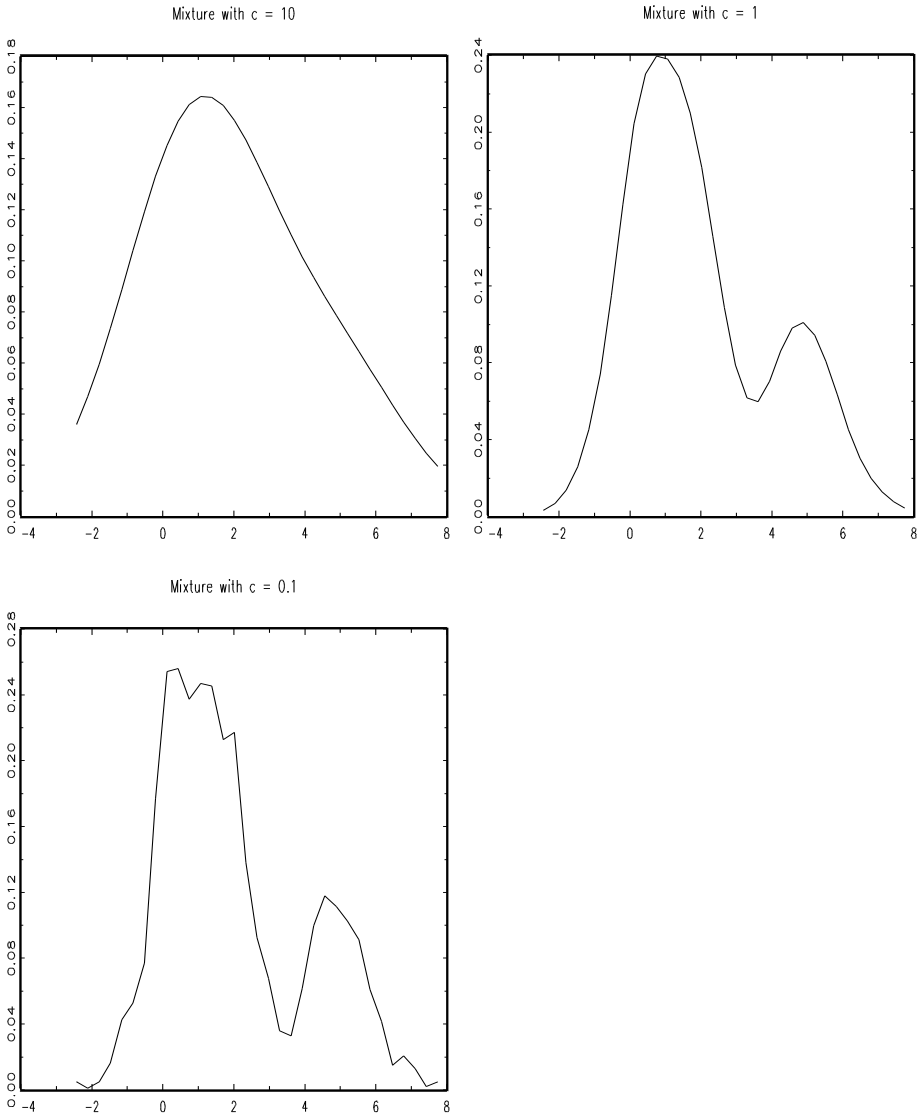

Figure 4: Over-smoothing and under-smoothing in density estimation

5 Sampling properties of kernel estimates

We have investigated many factors that influenced the final aspect of a non-parametric density estimate. The two basic ingredients are the choice of the kernel and the choice of the smoothing window size. How could we measure exactly their influence on the precision of the final result? The basic question is to find a way to measure the distance between the true density and the estimated density. A natural distance measure between an estimator and a true value is the Mean Squared Error:

$$\text{MSE}_x(\hat{\theta}) = \text{E}[\hat{\theta} - \theta]^2$$

that can be easily decomposed into:

$$\text{MSE}_x(\hat{\theta}) = \text{Biais}[\hat{\theta}]^2 + \text{Var}[\hat{\theta}].$$

But this indicator concern a point estimator and not a complete density. We are thus looking for a global measure valid for the whole range of x. We are thus going to integrate over x in order to get the MISE, or Mean Integrated Squared Error:

$$\text{MISE}_x(\hat{f}) = \text{E} \int [\hat{f}(x) - f(x)]^2 dx.$$

This corresponds to a notion of risk. If we want to minimize the loss, we simply have to consider the ISE (Integrated Squared Error):

$$\text{ISE}_x(\hat{f}) = \int [\hat{f}(x) - f(x)]^2 dx.$$

The MISE is the most commonly used indicator, but it might be difficult to compute. So that most of the time we rely on approximations that are found by noting that the MISE can be decomposed into:

$$\text{MISE}_x(\hat{f}) = \int [\text{E}(\hat{f}(x)) - f(x)]^2 dx + \int \text{Var}[\hat{f}(x)]\, dx.$$

It is then sufficient to find approximations for the bias and the variance and report those values in this expression.

5.1 Assumptions and notations

We already made some assumptions concerning the Kernel and the window size. We recall them and introduce some useful notations:

- $\int K(t)\, dt = 1$
- $\int K^2(t)\, dt = c_K < \infty$
- $\int tK(t)\, dt = 0$
- $\int t^2 K(t)\, dt = \mu_2$

The quantity μ_2 is going to play an important role in the sequel. Finally, concerning the window size, we have the following assumptions:

- $h \to 0$ when $n \to \infty$
- $n\,h \to \infty$ when $n \to \infty$

The window size has to go to zero as the sample size grows, but at a speed which is not too high.

5.2 Bias and variance of a kernel estimate

The bias and the variance of an estimator can be computed as expectations with respect to the true and unknown distribution $f(.)$. Let us start from the usual kernel density estimator

$$\mathrm{E}(\hat{f}(x)) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)\, dy$$

in order to compute the bias. For the variance we have:

$$\begin{aligned} n\mathrm{Var}\hat{f}(x) &= \int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)\, dy \\ &\quad - \left\{\int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)\, dy\right\}^2. \end{aligned}$$

5.3 Approximating the bias and the variance

The exact formulae that we have just given includes integrals that cannot readily be evaluated and thus are of a direct practical interest. We have to find approximations, using a first order Taylor expansion, reduced to the first order.

Let us first propose the change of variable $y = x - ht$ with Jacobian h. With this change of variable, the bias becomes:

$$\text{biais} = \int K(t)[f(x-ht) - f(x)]dt.$$

Let us develop $f(x-ht)$ around $h = 0$:

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f"(x) + \ldots.$$

Using the fact that a kernel is of zero expectation and of variance μ^2,

$$\text{biais} \simeq \frac{1}{2}h^2f"(x)\mu_2 + \ldots.$$

Similar computations for the variance show that

$$\mathrm{Var}(\hat{f}(x)) \simeq \frac{1}{nh} f(x)\, c_K,$$

supposing that n is big and h small. The approximation for the MISE is thus:

$$\text{AMISE} \simeq \frac{1}{4}h^4\mu_2^2 \int f"(x)^2 dx + \frac{1}{nh}.c_K$$

The bias depends on the window size and not on the sample size. On the contrary, the variance is a function of the sample size. Moreover, we can minimize the bias by decreasing the window size h, but at the same time we increase the variance. Choosing a value for h implies a trade-off between systematic error and random errors, between bias and variance. If we want to minimize the MISE (or the AMISE here), we see that the first term is of the same order as h^4, when the second term is of the same order as $1/(nh)$. Bias and variance are of the same order for

$$h \propto n^{-1/5}.$$

This rate of convergence for the window size is quite general for the whole non-parametric inference.

5.4 What are the ideal kernel and window size?

We are going to differentiate the approximate MISE with respect to h in order to find the ideal h by setting this expression to zero. We have:

$$\begin{aligned} h_{opt} &= \mu_2^{-2/5} c_K^{1/5} \left\{ \int f"(x)^2 dx \right\}^{-1/5} n^{-1/5} \\ &= \left[\frac{c_K}{n\,\mu_2^2 \int f"(x)^2 dx} \right]^{1/5} \end{aligned}$$

The ideal window size is a function of quite different things:

- It tends to zero at a very low speed
- It depends on the fluctuations of f. If f fluctuates a lot beaucoup, a small h will be needed. Some methods will determine h with respect to a known density like the Normal (Silverman's rule of thumb).
- Finally, h depends on the kernel. The latter can always be normalized so that $\mu_2 = 1$. So that the kernel takes part to the final result only with $c_K = \int K^2(t)\,dt$. Silverman's rule will again take advantage of this result.

Let us plug the optimal h into the expression of the MISE. We get:

$$\text{MISE} \simeq \frac{5}{4}\mu_2^{2/5} c_K^{4/5} \left\{ \int f"(x)^2 dx \right\}^{1/5} n^{-4/5}$$

The ideal kernel is the one that minimizes the MISE for a given$f$. In order to find it, we have to minimize c_K under the provision that this kernel is a density, that is to say integrates to one and is

normalized, which means that $\mu_2 = 1$. One can show that this ideal kernel is the Epanechnikov kernel that has a very simple expression:

$$K(t) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - t^2/5) & \text{if } |t| \leq \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

We can compare the efficiency of the other kernels with respect to the Epanechnikov kernel by defining the ratio:

$$Ef = \frac{\sqrt{\int t^2 K_e(t)\, dt} \int K_e(t)^2 dt}{\sqrt{\int t^2 K(t)\, dt} \int K(t)^2 dt}.$$

And using the properties of the Epanechnikov kernel, this ratio is simplified into:

$$Ef = \frac{2/(5\sqrt{5})}{\sqrt{\int t^2 K(t)\, dt} \int K(t)^2 dt}.$$

Let us now compute the efficiency of the usual kernels. The most inefficient kernel is the rectangular kernel which leads to the histogram. With this kernel, we have an efficiency which is very near from one. It is thus not very useful to spend much time finding an efficient kernel. To justify the search for an efficient kernel, we have to take into account other criteria than efficiency. For instance, the Epanechnikov kernel is not differentiable at an order greater than one, when the biweight kernel is differentiable at the order two and when the Gaussian kernel is infinitely differentiable. Some kernels have a finite support, while others have an infinite support. This makes a difference in term of numerical efficiency. With the Gaussian, a lot of time can be spend computing very small weights.

Table 4: Efficiency loss in density estimation

Kernel	$K(t)$	efficiency
Epanechnikof	$\frac{3}{4\sqrt{5}}(1 - t^2/5)$	1
Biweight	$\frac{15}{16}(1 - t^2)^2$	0.99
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2$	0.95
Rectangular	$\frac{1}{2}$ pour $\lvert t \rvert < 1$	0.93

6 Choosing the window size

The choice of the window size is crucial for the final aspect of the graph of the density. This choice can be driven by the final aim of the study. If we want to present the empirical content of a data set, a subjective choice is convenient. If we want to derive statistical conclusions, some under-smoothing could be necessary, as the reader is able to smooth visually when he

cannot rebuild details that would have been smoothed out by using a too large h. When many results have to be presented, an automatic method can be useful. If we want to compare results, a standardised method will be preferable. We must note that automatic methods cannot be qualified of being objective as they al rely on particular assumptions.

6.1 Subjective choices

We consider several graphs of the density, each one corresponding to a given choice for the window size. We chose the window size which produces the more aesthetics graph. Just have a look at previous Figures where under or over smoothing are easily detected.

6.2 Reference to a known distribution

We have seen that the optimal h was given by:

$$h_{opt} = \mu_2^{-2/5} c_K^{1/5} \left\{ \int f"(x)^2 dx \right\}^{-1/5} n^{-1/5} \tag{1}$$

Some of the elements of this expression are known as n and $K(.)$. But f is of course unknown, as we want to estimate it. We have to compute $\int f"(x)^2 dx$. If we suppose that the true distribution f is Normal of zero mean and of variance σ^2, then

$$\int f"_{N(0,\sigma^2)}(x)^2 dx = \sigma^{-5} \frac{0.375}{\sqrt{\pi}} \simeq 0.212 \sigma^{-5}$$

Let us now choose a normal, we can verify that $\mu_2 = 1$ and $c_K = 0.5/\sqrt{\pi}$. Gathering all these small bits, we have an expression for the optimal h:

$$h \simeq 1.06\, \sigma\, n^{-1/5}.$$

The only remaining question is to find a consistent estimate for the variance of the sample to get an estimate for the optimal h. This is the rule of Silverman which is the most popular way of finding easily a window size.

This procedure is very efficient as soon as we are not far from the Normal case, but lacks efficiency when we are far from it. In particular, if the true distribution f is a mixture, the rule of Silverman will tend to over smooth the density as soon as the modes of the mixture get apart. Different articles have also shown that we have over smoothing when f is asymmetric, but no over smoothing in the case of kurtosis. In particular if f is Student, the rule of Silverman is rather efficient.

6.3 Estimating the curvature

In (1), we have an expression for the optimal window size. It depends on several quantities which are function, of the sample, of the Kernel and of the true density. It is possible to find direct values

or estimates for those quantities, except of course for those which depend on the true density. The rule of Silverman assumes that the true density is a Normal, so it is easy to compute a direct value for $\int f"(x)^2 dx$ which measure the average curvature of the true density. The idea of Sheather and Jones (1991) was to propose a non-parametric estimator for this quantity. The procedure gives in general quite good results.

6.4 Least squares cross validation*

Instead of considering a pseudo likelihood function as a criterion to optimize, we shall consider this time the Integrated Squared Error:

$$ISE(h) = \int (\hat{f}(x,h) - f(x))^2 dx.$$

Let us develop the square. This resulting expression can be simplified as one of its terms does not depend on h:

$$ISE(h) \propto \int \hat{f}(x,h)^2 \, dx - 2 \int \hat{f}(x,h) \, f(x) \, dx$$

We have to find the value of h that minimizes as estimation of the $ISE(h)$. Here again, the cross-validation method is the right solution for evaluating this criterion. We have

$$\hat{f}_{-i}(x,h) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - x_j}{h}\right)$$

The notation $-i$ means that we drop observation i for evaluating $f(x_i)$. We can now notice that $\int \hat{f}(x,h) \, f(x) \, dx$ is the expectation of $\hat{f}(x,h)$. An unbiased estimator of this expectation is given by the empirical mean of $\hat{f}_{-i}(x,h)$, or in other terms

$$\mathrm{E}(\hat{f}(x,h)) \simeq \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x_i,h).$$

We have now to compute the first element of the ISE by means of

$$\int \hat{f}^2 dx = \frac{1}{n^2 h^2} \sum_i \sum_j \int_x K\left(\frac{x_i - x}{h}\right) K\left(\frac{x_j - x}{h}\right) dx,$$

with a solution given by

$$\int \hat{f}^2 dx = \frac{1}{n^2 h^2} \sum_i \sum_j \tilde{K}\left(\frac{x_i - x_j}{h}\right)$$

$\tilde{K} = K \circ K$. If the kernel (0,1), then $\tilde{K} = N(0,2)$.

The method is rather intensive in term of computer time. For every value of h, we have to evaluate ISE(h) which contains a double sum. Moreover, the function can have several local minima. Pagan and Ullah mention the "binning" technique which is used for instance in the software `Xplore` for reducing computer time.

6.5 Using R

The standard `stats` package includes a routine for estimating densities. The density object is created by simply calling `density(x)` where x represents the data set, assuming that the data are presented in a column. By default a Gaussian kernel is used and the classical rule of Silverman for the bandwidth. Of course many options are possible which can be found on the help. We present these options in Table 5. To obtain a graph, it suffices to use the routine plot

Table 5: R options for density estimation

Bandwidth	Kernel	Weight
bw = nrd0(x)	kernel = "gaussian"	weights = rep(1/nx, nx)
bw=bw.ucv(x)	kernel = "epanechnikov"	
bw=bw.SJ(x)	kernel = "triangular"	

together with the output object of density. For instance `plot(density(x))`. If we want to change the default method for determining the bandwidth, using for instance the cross validation method, we can use

```
plot(density(y79,bw=bw.ucv(y79)))
```

We are not obliged to use the same sample for estimating the density and for computing the bandwidth. In particular, we can use a sub-sample for computing the bandwidth. We can draw a sub-sample at random for instance.

In the column Bandwidth of Table 5, $bw = nrd0(x)$ is a slight modification of the rule of Silverman as it uses an improved estimator of the sample variance. $bw = bw.ucv(x)$ was already explained as being the unbiased cross validation. $bw = bw.SJ(x)$ is the implementation of the Sheather and Jones (1991) plug-in rule. It estimates non-parametrically the integral of the squared second order derivative of the true density. This method is very popular, as it is a robust plug-in rule which in general gives better results than the simple Silverman rule. But is requires the fourth order derivative of the Kernel. So it cannot be used with the Epanechnikov kernel. But it is safe with a Gaussian kernel.

7 General estimation methods for parametric models

Anon-parametric approach is nce for getting an idea about the general shape on an income density. However, the methods requires a lot of observations because the rate of convergence is only of $n^{-1/5}$ instead of the usual rate of $n^{-1/2}$. Moreover, the method is rather imprecise in the tail of the distribution where there are by definition fewer observations. So, if we are sure that the true distribution is uni-modal, the temptation is great to adjust a parametric density. The question becomes how to estimate its parameters. There are several principles which can be applied, depending on the available data and on the complexity of the parametric density.

7.1 Adjusting a parametric density with grouped data

Grouped data used to be very common because they solve the question of anonymity when individual data are involved. Considering grouped data can also be a way to solve difficult estimation problems. For instance, it is quite impossible to use the maximum likelihood principle to make inference with the Generalized Gamma density due to its awkward parameterizations (see Johnson et al. 1995).

When data are grouped into clusters, inference is based on the comparison of two quantities:

- $p_i(\theta)$ is the theoretical probability to belong to cluster i^{th} among the g possible clusters of the population:
$$p_i(\theta) = \int_{I_i} f(x;\theta)dx.$$
This probability is given by integrating the density to be estimated over the range of cluster i. The cluster corresponds to the interval $[x_{i-1}, x_i]$, the integral is computed over this range.

- n_i/n are the observed frequencies, they are given by the data. For instance, the cluster frequencies in an histogram. n is the total sample size, while n_i is the number of observations in cluster i.

McDonald and Ranson (1979) give different ways two confront these two quantities.

In a likelihood framework, we have to represent the multinomial process generating the histogram. The likelihood function is thus:

$$L(\theta) = n! \prod_{i=1}^{g} \frac{p_i(\theta)^{n_i}}{n_i!}.$$

They call this approach a scoring method because we have to compute the first derivative of the likelihood function in order to find its maximum.

The Pearson minimum chi-squared estimator minimizes a chi-squared distance between the theoretical probability and its empirical counterpart

$$n \sum_{i=1}^{g} \frac{\left(n_i/n - p_i(\theta)\right)^2}{p_i(\theta)}.$$

This quantity is distributed as a χ^2 with $g-k-1$ degrees of freedom which give a direct way for testing the adequation between the data and the model. This a goodness-of-fit test. This method of estimation is asymptotically equivalent to the maximum likelihood.

The least squares estimator minimizes a simpler distance between theoretical and empirical probabilities with

$$\sum_{i=1}^{g} \left(\frac{n_i}{n} - p_i(\theta)\right)^2$$

This last method gives often different results than the previous ones and is not recommended. The Pearson method corresponds to a weighted least-squares.

On US grouped data for 1970, 1972, 1974, 1975, McDonald and Ranson (1979) found that in general the Singh-Maddala distribution gave the better fit, much better than the logNormal. Scoring and Pearson methods gave very similar results either for the parameters or the implied Gini coefficient. Least squares gave sometimes rather different results.

7.2 A regression based on the empirical distribution

When the data are not grouped, it is possible to use other methods to fit a density. The method we examine here is used for instance in Singh and Maddala (1976). It is still based on the comparison between a statistics and its theoretical counterpart. But here, Singh and Maddala (1976) take advantage of the fact that the distribution has an analytical form. They confront it to the natural nonparametric estimator of the distribution. For the SM distribution, we have

$$F(x) = 1 - \frac{1}{(1 + a_1 x^{a_2})^{a_3}}.$$

The estimation procedure consists in minimizing the least squares distance between $F(x, a)$ and $\hat{F}(x)$ computed either for each sample value or for a given grid. Only $\hat{F}$ has to make use of the whole sample. The minimization problem is:

$$\hat{a} = \text{argmin} \sum [\log(1 - \hat{F}) + a_3 \log(1 + a_1 x_i^{a_2})]^2.$$

This is a nonlinear regression problem which has to be solved by numerical optimization in a quite simple way.

We can make two comments concerning this method:

- it uses a least squares distance and not a χ^2 distance. We can have a first source of errors by not using weighted least squares as underlined in the previous subsection.
- We have a problem at the right infinite boundary as we cannot compute $\log(1-F)$ because $F(x_{max}) = 1$. This problem does not exist when probabilities are confronted to their empirical counterparts.

The same regression method can be used for making inference on the Pareto parameter because we have then a linear regression. For the Pareto density, this was in fact the original method. We have

$$(1 - F(x_i)) = (x_i/x_m)^{-\alpha}.$$

Taking the logs each side and using a natural estimate for F leads to the regression

$$\log(1 - \hat{F}(x_i)) = cste - \alpha \log(x_i) + \epsilon_i.$$

If we do not get a straight line when plotting the two logs, it is a test that the sample does not come from a Pareto distribution. We can also estimate α in a similar way using the empirical

Lorenz curve. These estimators are consistent.

Finally, let us consider the Weibull case. The cumulative distribution is

$$F(x) = 1 - \exp(-(kx)^{\alpha}).$$

Taking log twice and paying attention to the signs, we have the following regression

$$\log(-\log(1 - \hat{F}(x_i)) = \alpha \log k + \alpha \log x_i + \epsilon_i.$$

This regression is similar to that obtained for the Pareto case, except that we have to take twice the logs for the left hand side. A graphical device is also a good test for the adequacy of the Weibull model to the data.

8 Using the likelihood function for making inference

When individual data are available, it is possible to write the likelihood function of the model and use it for making inference. In this section, we shall apply this principle of inference for two standard processes the Pareto density and the lognormal density.

8.1 Maximum likelihood for Pareto samples

Inference is quite easy for the usual Pareto I model. It is detailed for instance in Arnold (2008). Let us suppose that we have an IID sample of X which is drawn from a Pareto I model. The likelihood function is:

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

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

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

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

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

Knowing the bias, it is easy to propose unbiased estimators by simply correcting the initial maximum likelihood estimators. Once we know the estimates of x_m and of α, it is easy to produce an estimate for the needed transformations of these parameters such as for instance the Gini coefficient and to find their standard deviation using the delta method (which is not very precise, however).

8.2 Bayesian inference for the Pareto*

Instead of using the frequentist estimation approaches discussed above, we may consider a Bayesian formulation of the problem. See for instance the summary available in Arnold (2008). If x_m is known, the problem is quite simple. In the case where x_m is also an unknown parameter, inference becomes more delicate and a Gibbs sampler is needed. We treat here only the case where x_m is known.

Let us recall that in a classical framework, the sample space is probabilized and that one looks for the value of the parameter θ that gives the maximum probability to get the observed sample. In a Bayesian framework, the parameter space is also probabilized. It is endowed with a prior $p(\theta)$ possibly non-informative and the product of inference is a posterior density obtained by applying Bayes' theorem:

$$p(\theta|y) = \frac{l(y;\theta)p(\theta)}{\int l(y;\theta)p(\theta)d\theta},$$

where the denominator is the integrating constant of the posterior density. It is usually the case to work up to a constant of proportionality as the denominator does not depend on the parameters (they are integrated out). So that the posterior is defined as:

$$p(\theta|y) \propto l(y;\theta)p(\theta).$$

In the natural conjugate framework, the prior $p(\theta)$ is chosen in such a way that it combines easily with the likelihood function $l(y;\theta)$. The natural framework relies on the exponential family where sufficient statistics of two samples combine easily.

The Pareto distribution is related to the exponential distribution as follows. Suppose X is Pareto-distributed with minimum x_m and index 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

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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 parametrization of the gamma pdf is useful:

$$\mathrm{Gamma}(\alpha\,;\,\nu, s) = \frac{s^\nu}{\Gamma(\nu)}\,\alpha^{\nu-1}\,\exp(-\alpha\, s),$$

with moments given by

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

The posterior distribution p can then be expressed in terms of the likelihood function defined above and a gamma prior:

$$\begin{aligned} p(\alpha|y) &\propto L(\alpha; y) \times \mathrm{Gamma}(\alpha\,;\,\nu, s) \\ &= \alpha^n\,\exp(-\alpha\, n\overline{y}) \times \tfrac{s^\nu}{\Gamma(\nu)}\,\alpha^{\nu-1}\,\exp(-\alpha\, s) \\ &\propto \alpha^{(\nu+n)-1}\,\exp(-\alpha\,(s + n\overline{y})). \end{aligned}$$

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

$$p(\alpha|y) = \mathrm{Gamma}(\alpha\,;\,\nu + n, s + n\overline{y}).$$

Here the parameter ν can be interpreted as the number of prior observations, and s as the sum of the prior observations.

Knowing the posterior parameters, we can compute easily the posterior moments by applying simply the above analytical formulae. We can draw the graph of the posterior density of α. More interestingly, we can generate random numbers from the posterior density in order to find the distribution of any inequality index such as the Gini coefficient or the Atkinson index or of any of the other transformation of α. We have in this way np draws of transformations of α for which we can compute a mean, a standard deviation and estimate a density using a nonparametric kernel estimate.

8.3 Maximum likelihood for Lognormal samples

The probability density function of a log-normal distribution is:

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

where μ and σ are the mean and standard deviation of the variable's natural logarithm. This means for instance that $\mu = \mathrm{E}(\log(x))$. 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 indices to denote distributions, 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 log-normal distribution it holds that

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

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.

8.4 Bayesian inference for the Lognormal*

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

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

where:

$$\bar{x} = \frac{1}{n}\sum_i \log x_i \qquad s^2 = \frac{1}{n}\sum_i (\log x_i - \bar{x})^2.$$

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:

$$\pi(\mu|\sigma^2) = f_N(\mu|\mu_0,\sigma^2/n_0) \propto \sigma^{-1}\exp -\frac{n_0}{2\sigma^2}(\mu-\mu_0)^2, \tag{3}$$

$$\pi(\sigma^2) = f_{i\gamma}(\sigma^2|\nu_0,s_0) \propto \sigma^{-(\nu_0+2)}\exp -\frac{s_0}{2\sigma^2}. \tag{4}$$

The prior moments are easily derived as:

$$\mathrm{E}(\mu|\sigma^2) = \mathrm{E}(\mu) = \mu_0, \qquad \mathrm{Var}(\mu|\sigma^2) = \frac{1}{n_0}\sigma^2 \qquad \mathrm{Var}(\mu) = \frac{1}{n_0}\frac{s_0}{\nu_0 - 2} \tag{5}$$

$$\mathrm{E}(\sigma^2) = \frac{s_0}{\nu_0 - 2}, \qquad \mathrm{Var}(\sigma^2) = \frac{s_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)} \tag{6}$$

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.

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

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

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

with

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

Then the marginal posterior density of μ is Student with

$$\begin{aligned} \pi(\mu|x) &= f_t(\mu|\mu_*, s_*, n_*, \nu_*), \\ &\propto \left[s_* + n_*(\mu - \mu_*)^2\right]^{-(\nu_*+1)/2} \end{aligned} \tag{7}$$

where

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

The posterior density of σ^2 is given by

$$\begin{aligned} \pi(\sigma^2|x) &\propto \sigma^{-(n+\nu_0+2)} \exp -\frac{1}{2\sigma^2}\left(s_0 + s^2 + \frac{n_0 n}{n_0 + n}\left(\mu_0 - \bar{x}\right)^2\right), \\ &\propto f_{i\gamma}(\sigma^2|\nu_*, s_*). \end{aligned} \tag{8}$$

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

8.5 Estimating the income distribution of California using grouped data

In the data base which is provided by the American Community Survey,[1] information of the household income distribution is provided at the level of each of 724 school districts of California in the form of grouped data represented by ten unequal classes, with top coding for the largest. The lowest class represents the number of households with an income plus benefits below \$10 000 per year, while the largest class corresponds to the number of households with a year income and benefit greater than \$200 000. It is supposed that this distribution concerns households with two kids, so a family of four persons. By aggregating all the 724 school districts of California, we get the income distribution represented in Figure 5. We note two things.

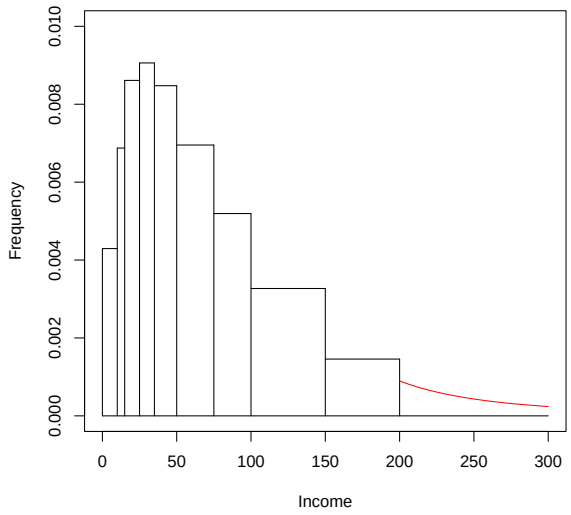

Figure 5: Household income distribution for California

First, the ten classes are unequal. Second, as the last class is open, it is represented as a Pareto distribution, drawn in red in this graph. The Pareto parameter was estimated to have a value of $\hat{\alpha} = 2.28$). This value was found using the method described in Quandt (1966).

Let us call x_i the lower bound of each of the 10 classes. We then define nc_i as the number of households in each of the ten income classes while nc_{10} represents the number of household in the last open income class with an income greater than $x_{10} = \$200\ 000$. The formula given in Quandt (1966) and applied here gives:

$$\hat{\alpha} = \frac{\log(nc_9 + nc_{10}) - \log(nc_{10})}{\log(x_{10}) - \log(x_9)}. \tag{9}$$

The rationale for this formula is quite simple to find. From the open class, we have first that

$$1 - F(x_{10}) = (x_{10}/x_m)^{-\alpha}.$$

[1] Available on: http://nces.ed.gov/programs/edge/demographicACS.aspx"

This expression shows that we have to chose $x_m = x_9$. Equating this expression with the empirical frequency, taking the logs and doing the same with the previous class gives the result. The complete proof is in Quandt (1966).

8.6 Using R for Pareto and lognormal fit

Using the same data set as before (UK family expenditure survey in real terms), we shall here compare the fit obtained by using a Pareto density and a lognormal density.

We first try to fit a Pareto density. There is a simple way to test the Pareto assumption. We just have to plot the graph of $\log(y)$ against $\log(1 - F)$. For this the following R routine is convenient. It assumes that the observations are ordered. The boundary problem is solved by dropping the last observation :

```
pareto = function(y){
  n = length(y)
  F = (1:n)/n
  F = F[1:n-1]
  y = y[1:n-1]
  plot(log(y),log(1-F))
  lines(log(y),log(1-F))
}
```

Figure 6 shows that the Pareto assumption might be valid only above a certain income level. The black line represents 1979, while the red line corresponds to 1988, blue to 1992 and green to 1996.

The Pareto model does not fit correctly the complete sample. Using the 1979 FES data, the MLE for α in the Pareto process is 1.974 when the complete sample is used. If we now turn to the lognormal process, the MLE estimate for σ is 0.459, also for the complete sample. We can now plug these two values into the expression of the Lorenz curves for the two models and compare the result to the natural estimate of the Lorenz curve. This is done in Figure 7 using the following R code

```
plot(Lc(y79))
p = seq(0,1,0.05)
lines(p,Lc.pareto(p, parameter=2),col="red")
  text(0.9,0.6,"Pareto 2.0")
lines(p,Lc.lognorm(p, parameter=0.45),col="blue")
 text(0.45,0.4,"Lognormal 0.45")
```

The lognormal seems to fit the data quite well when of course the Pareto is not able to produce a good account of the whole sample. So, we could perform the same exercise as we did for the Gini coefficient with the Pareto process. The posterior density of σ is an inverted gamma2 with hyperparameters ν_* and s_* based on sample mean and variance of the log variable under a non-informative prior. We could then simulate σ^2 and compute the Gini as $2\Phi(\sigma/\sqrt{2}) - 1$ for each draw. This is done in a next chapter.

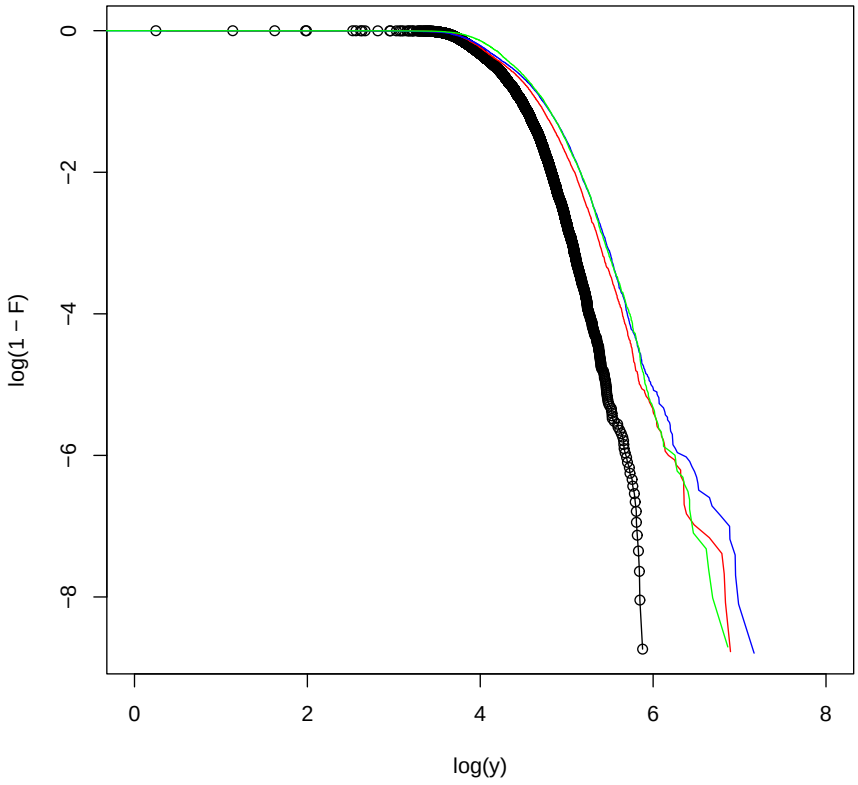

Figure 6: Pareto tail for the income distribution

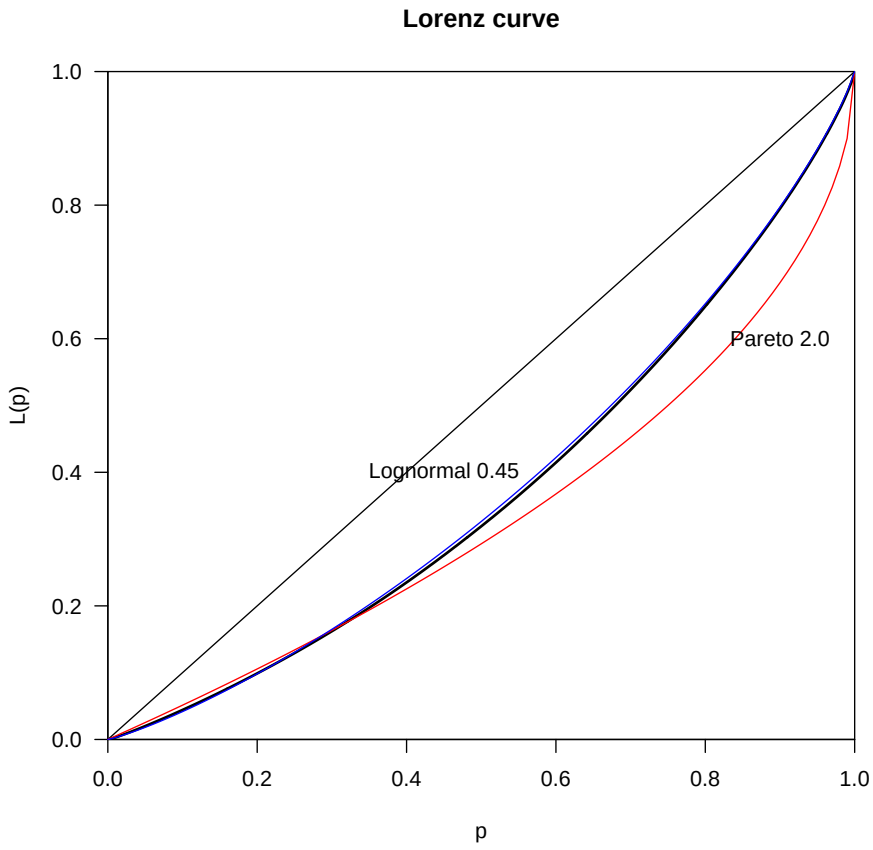

Figure 7: Lorenz for Pareto and Lognormal

Let us now turn to the Chine data of the CGSS. We know that we have definitively to use weights to treat those income data. So the previous `pareto` function has to be changed into:

```
pareto = function(y,w){
  n = length(y)
  F = (1:n)/n
  F = F[1:n-1]
  yw = y*w
  ys = sort(yw)
  ys = ys[1:n-1]
  plot(log(ys),log(1-F))
  lines(log(ys),log(1-F))
}
```

The data are first weighted and then ordered. The histogram we had made of these data led us to think that the income distribution would be represented by a Pareto. Figure 8 shows that this is not the case. Here again, we have a Pareto tail, but only after a certain level. In fact, when we try

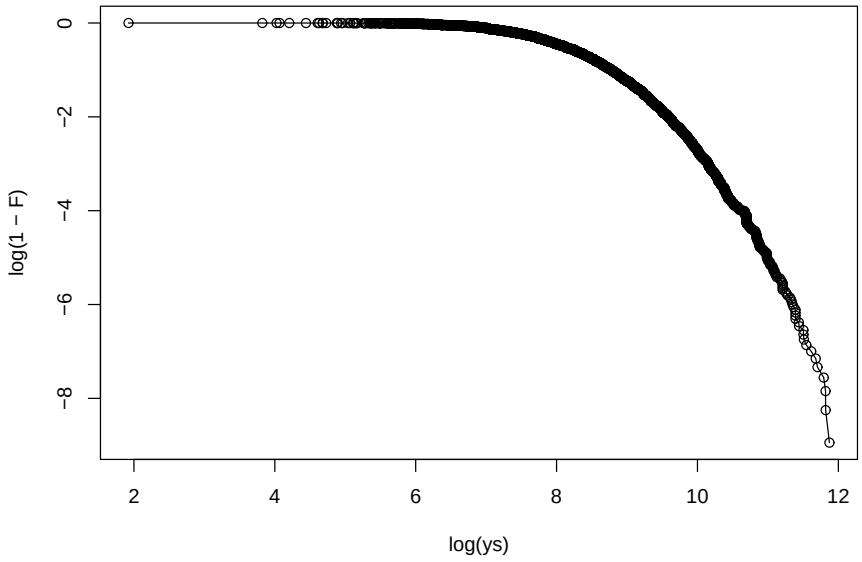

Figure 8: Pareto lines for the CGSS

to estimate the Pareto coefficient using a linear regression of $\log(1-F)$ over $\log(y * weights)$ (the mean of the weights has to be equal to 1), we find a coefficient equal to 0.807 (0.0047) which is much to low to be able to draw a Lorenz curve, because the latter exist only for $\alpha > 1$.

8.7 Using R for Bayesian inference on the Gini*

A Bayesian inference for the α parameter of the Pareto is easy to program, provided we take into account the way the Gamma distribution is parameterized in R. The shape parameter corresponds to the sample size and the scale parameter corresponds to $1/(n\bar{x})$ when using a non informative prior. This can be implemented in the following routine which include the computation of the Gini index together with its small sample properties.

```
Bayes = function(x,np){
# Bayesian inference for alpha when xm is known.
# Simulation of the Gini
 yb = sum(log(x/min(x)))
 n = length(x)
 alpha = rgamma(np,scale = 1/yb,shape = n)
 a = alpha[alpha>0.6]
 g = 1/(2*a-1)
 cat("Gini = ",mean(g)," S.D. =", sd(g),"\n")
 plot(density(g))
```

}

The answer given by the Bayesian inference using a Pareto model depends heavily on the truncation point. We have chosen 120, which leaves only 971 observations out of 6230 for 1979. But do not forget that Pareto is for high incomes. Bayesian inference produced an α with

Table 6: Alternative inferences for the Gini index

Method	Mean	Standard deviation
Bayesian with Pareto	0.129	0.00486
Bootstrap parameter free	0.118	0.00426

posterior mean of 4.365 and a standard deviation of 0.143. Fitting a Pareto model leads to a Gini coefficient which is slightly greater than that obtained when computing it directly using the sole sub-sample.

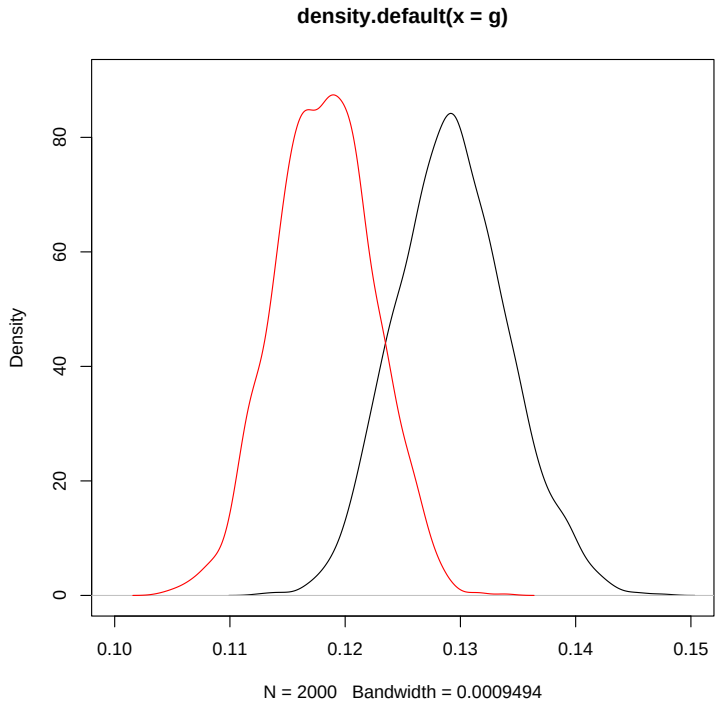

Figure 9: Comparing Bayes and bootstrap estimates for the Gini

The bootstrap produces a density which is slightly more concentrated that its Bayesian counterpart as shown in Figure 9 where the Bayesian estimate is in black while the bootstrap is in red.

9 Using mixtures for IID samples

We are presenting in this section an intermediate approach between a fully parametric model for the income distribution and a fully nonparametric density estimation. It is a semiparametric approach as it is based on the combination of parametric densities where the number of needed densities has to be determined by the sample.

9.1 Informal introduction

Let us go back to the FES data sets. Which kind of density can we fit to these data? We have illustrated several stylized facts

- The Pareto does not fit the data as shown by the Lorenz curve
- The lognormal seems to fit the data better as shown again by the Lorenz curve
- The high incomes, greater than 120, seem to behave like a Pareto

Does the lognormal fit really well the data as the Lorenz curve would suggest? In Figure 10, we compare the adjusted parametric lognormal density with a non-parametric estimate of the density using the following R code:

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

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.

9.2 Mixture of distributions

When a single density is not enough to represent correctly the distribution of a sample, a simple explanation is that the observed sample is heterogenous and this result from the mixing of different populations, each being represented by a particular density indexed by a given parameter. The trouble is that we do not know first how many different sub-populations there are and second what is their proportion. This lack of knowledge makes the problem difficult. For a simplification, let us suppose that we have only two sub-populations, each one being described by a density indexed by θ_i and in unknown proportion p. The density of one observation is

$$f(x|\theta) = p \times f_N(x|\mu_1, \sigma_1^2) + (1-p) \times f_N(x|\mu_2, \sigma_2^2)$$

if we suppose as a simplification that the two members of the mixture are normal densities. If we knew the sample separation, i.e. which observation belongs to group 1 or 2, the inference problem would be very simple. But of course, the allocation of the observations is unknown.

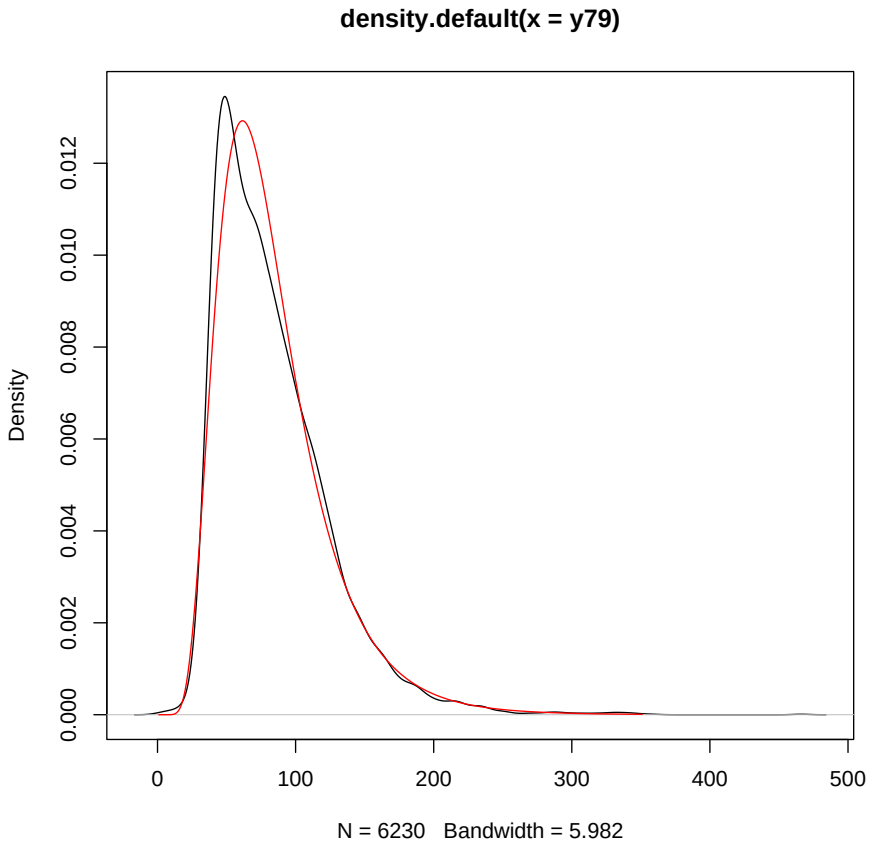

Figure 10: Non parametric estimate of the density for FES79 compared to a lognormal fit

9.3 Estimation procedures

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 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 that we shall call θ_1 and θ_2. Of course we do not know the z_i, but we can compute the following probabilities for each observation:

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

provided we have estimated p as $\hat{p} = n_1/n$. 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.

9.4 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

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

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.

9.5 Estimating mixture 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 p_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 were We must note that the estimation gives values for

Table 7: Parameter estimates for a two members mixture of lognormals

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

the mean and the variance of the sample, and not for the parameters of the lognormal. This is the same for the starting values.

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 $p_2 = 0.85$ 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.

We can try to do the same exercise for the Chinese income data. One simple way of dealing with weights is to multiply each observation by its weight, provided the mean of the weights is one. We have then to cut the observations above 50 000 yuan, because otherwise the right tail is too long for a nice display. In order to find starting value for the mean and the standard deviations of the observations, we have to compute these values for this truncated weighted sample. We find 6986 and 7527. This justifies the following code:

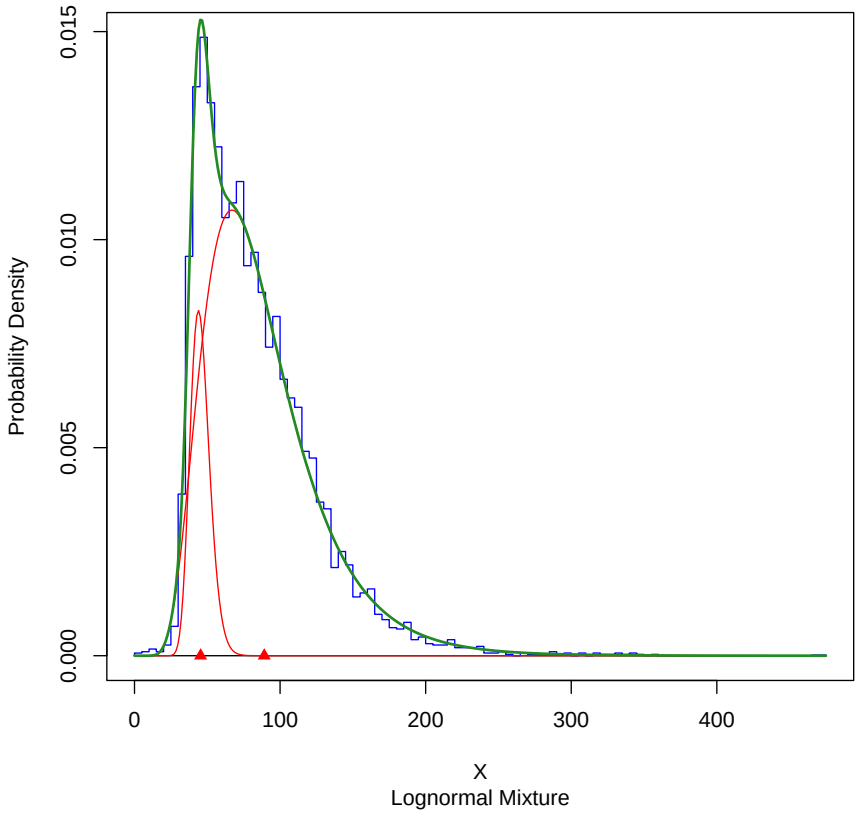

Figure 11: Mixture of two lognormal densities

```
library(mixdist)
  ywc = yw[yw<50000]
  mean(ywc)
  sd(ywc)
  y.gd = mixgroup(ywc,breaks=100)
  y.par = mixparam(mu = c(5000,10000), sigma = c(4000,6000),pi=c(0.6,0.4)
  y.res = mix(y.gd,y.par,dist="lnorm")
  print(y.res)
  plot(y.res)
  lines(density(ywc))
```

This produces estimates reported in Table 8. We have added on Figure 12 the nonparametric density estimate in black. We see that there are differences, compared to the histogram. Because of smoothing, there are negative values for income. The green line of the two member mixture reproduces quite well the shape of the histogram. It is interesting to compare the two mixtures, the one estimated for the UK in 1979 and the one estimated for China in 2006. For the UK,

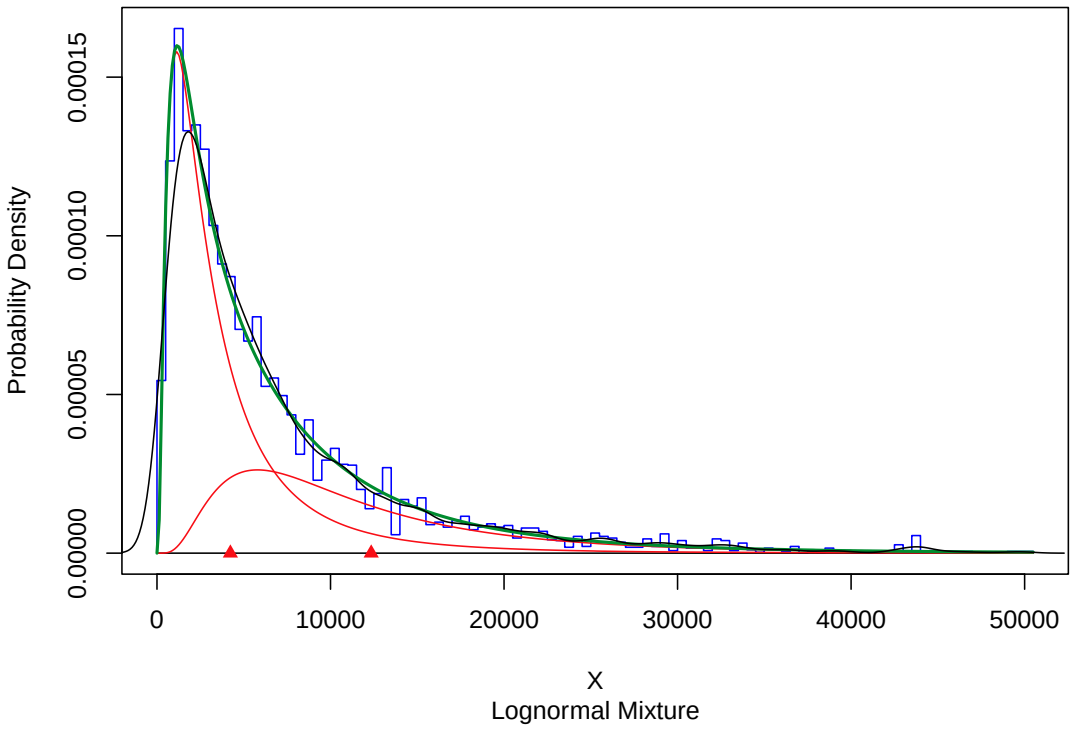

Figure 12: Mixture for Chinese income

Table 8: Parameter estimates for a
two members mixture of lognormals for China

member	p	μ	σ
1	0.652	4 230	5 027
2	0.348	12 345	10 000

the major component is the second one, around medium incomes. The first component is very much concentrated. For China, we have just the reverse configuration. The first component corresponding to lower income is the major one, while the second component has a very long tail and is very asymmetric. Clearly, the two countries have radically different income distributions. Note that for both samples, it was impossible to fit a three component mixture.

10 Bayesian inference for mixtures of log-normals using survey data*

This section comes from a joint paper with Edwin Fourrier. It aims at estimating an income distribution, 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 Duangkamon Chotikapanich (2008) for modelling the income distribution. The joint work with Edwin Fourrier introduces specifically sampling weights and zero income observations.

10.1 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:

$$f(y|\vartheta) = \sum_{j=1}^{k} p_j f_\Lambda(y|\theta_j), \qquad 0 \leq p_j < 1, \quad \sum_{j=1}^{k} p_j = 1, \tag{10}$$

where $\vartheta = (p, \theta)$ and the parameter vectors are $\theta = (\theta_1, ..., \theta_j)$ and $p = (p_1, ..., p_k)$ with p_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.

10.2 A Gibbs sampler algorithm

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 p_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 p_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 p_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 $p \sim f_D(\gamma_1^0, ..., \gamma_k^0)$. The hyperparameters 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:

$$p(\sigma_j^2|y,z) = f_{i\gamma}(\sigma_j^2|v_{*j}, s_{*j}), \tag{11}$$

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

$$p(\mu_j|\sigma_j^2, y, z) = f_N(\mu_j|\mu_{*j}, \sigma_j^2/n_{*j}). \tag{12}$$

The conditional posterior distribution of p_j is a Dirichlet with:

$$p(\eta|y,z) = f_D(\gamma_1^0 + n_1, ..., \gamma_k^0 + n_k) \propto \prod_{j=1}^{k} p_j^{\gamma_j^0 + n_j - 1}. \tag{13}$$

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:

$$Pr(z_i = j|y,\theta) = \frac{\eta_j f_\Lambda(y_i|\mu_j, \sigma_j^2)}{\sum_j p_j f_\Lambda(y_i|\mu_j, \sigma_j^2)}. \tag{14}$$

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).

Let us propose the following Gibbs sampler algorithm:

1. 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 $l = 0$.

2. For $j = 1, ..., m_0, ..., m + m_0$:

   (a) Generate a classification $z_i^{(l)}$ independently for each observation y_i according to a multinomial process with probabilities given by equation (14), using the value of $\vartheta^{(l-1)}$.

   (b) Compute the sufficient statistics $n_j, \bar{y}_j, s_j^2$.

   (c) Generate the parameters $\sigma^{(l)}, \mu^{(l)}, \eta^{(l)}$ from the posterior distributions given in equations (11), (12) and (13) respectively, conditionally on the classification $z^{(l)}$.

(d) Order $\sigma^{(l)}$ such that $\sigma_1^{(l)} < ... < \sigma_k^{(l)}$ and sort $\mu^{(l)}$, $\eta^{(l)}$ and $z^{(l)}$ accordingly.

(e) Increase l by one and return to step (a).

3. Finally 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). It is restricted to Gaussian mixtures.

10.3 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 asKunihama et al. (2014) 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 13, we compare two non-parametric estimator of a density, one without using weight, the second using weights. The difference is striking.

10.4 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

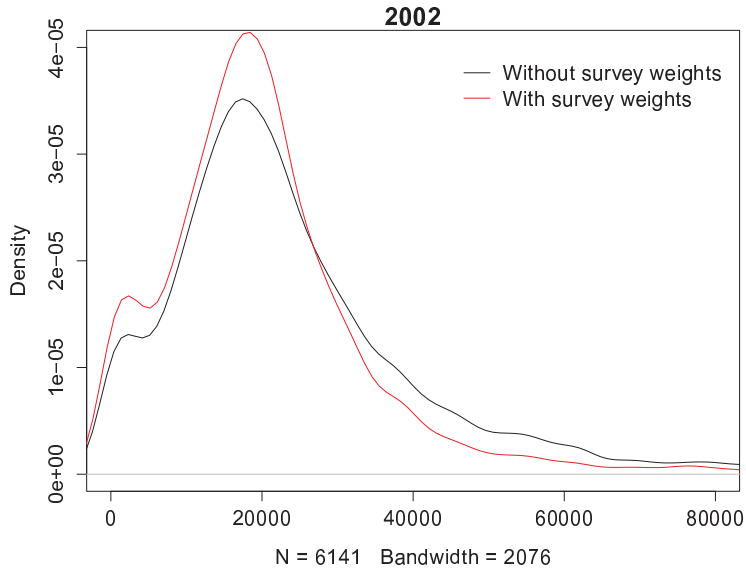

Figure 13: The influence of weight for density estimation

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

$$f(y|\vartheta) = \mathbb{1}(y=0)\omega + \mathbb{1}(y>0)(1-\omega)\sum_{j=1}^{k} p_j f(y|\theta_j), \tag{15}$$

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 14 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 household 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.

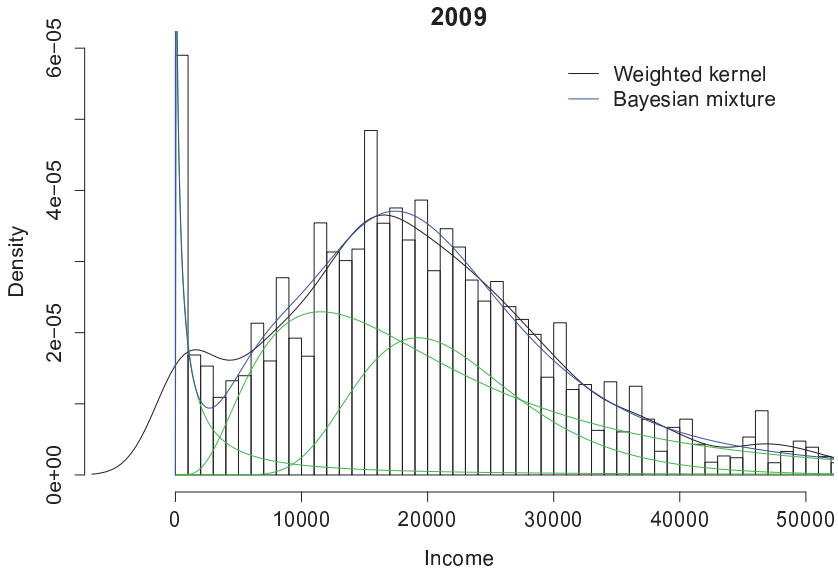

Figure 14: Income distribution before redistribution
Germany in 2009 using the GSOEP

11 Exercises

1. Compare the bootstrap results for the Gini index with the Davidson and the Giles methods which were given in Chapter 4, using the FES data.

2. When estimating an histogram, the number of cells has to be given. Compute the implicit bandwidth which is implied by the number of cells.

3. The Weibull density has an analytical cumulative distribution. Use this property to propose a way to adjust a Weibull density to the grouped data given in Table 2 for the US income distribution. Run the program in R.

4. Propose a regression method for estimating the main parameter of a Pareto distribution using the empirical Lorenz curve.

5. Propose an unbiased estimator for the Pareto I model, starting from the maximum likelihood estimator as given above.

6. Using the FES data set, fit a mixture of normal densities after taking the logs of the observations. Compare your results with the results obtained by considering directly a mixture of two lognormal densities.

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.

Benzidia, M., Melindi Ghidi, P., and Lubrano, M. (2017). Education politics, schooling choice and public school quality: The impact of income polarisation. Technical Report WP 2016 - Nr 42, GREQAM, Marseille.

Deaton, A. (1997). *The Analysis of Household Surveys*. The John Hopkins University Press, Baltimore and London.

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):1–38.

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.

Duangkamon Chotikapanich, W. E. G. (2008). Estimating income distributions using a mixture of gamma densities. In Chotikapanich, D., editor, *Modeling Income Distributions and Lorenz Curves*, pages 285–302. Sprnger.

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.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*, volume 2 of *Wiley series in probability and mathematical statistics: applied probability and statistics*. Wiley, New York,, New-York.

Kunihama, T., Herring, A. H., Halpern, C. T., and Dunson, D. B. (2014). Nonparametric bayes modeling with sample survey weights. *Biometrika*.

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

McDonald, J. (1984). Some generalised functions for the size distribution of income. *Econometrica*, 52(3):647–663.

McDonald, J. B. and Ranson, M. R. (1979). Functional forms, estimation techniques and the distribution of income. *Econometrica*, 47(6):1513–1525.

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge University Press.

Quandt, R. E. (1966). Old and new methods of estimation and the Pareto distribution. *Metrika*, 10(1):55–82.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:642–6.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690.

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