UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 03-02
Statistics and Econometrics Series 01
January 2003

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

ESTIMATION OF INCOME DISTRIBUTION AND DETECTION OF SUBPOPULATIONS: AN EXPLANATORY MODEL

Emmanuel Flachaire and Olivier G. Núñez*

Abstract

Inequality and polarization analyses are complementary but conceptually different. They are usually implemented independently in practice, with different a priori assumptions and different tools. In this paper, we develop a unique method to study simultaneously these different and complementary concerns. Based on mixture models, the method we develop includes at the same time : an estimation of income distribution with no a priori assumptions - a decomposition in several homogeneous subpopulations - an explanatory model to study the structure of the income distribution.

Keywords: inequality, polarization, income distribution, mixture models.

*Flachaire, Eurequa, University Paris I Panthéon-Sorbonne, France; Núñez, Dept. of Statistics and Econometrics, Universidad Carlos III de Madrid, C/Madrid, 126, 28903 Getafe (Madrid). Spain. E-mail: nunez@est-econ.uc3m.es. Tel: 34 916249826.
Financial support from Spanish DGES, grant BEC2002-03720, is acknowledged.

1 Introduction

In this paper, we develop an estimation method of income distribution, with no *a priori* assumptions, which leads us to detect the number and the constitution of subpopulations and, at the same time, includes an explanatory model with a set of explanatory factors that could explain differences between the distinct subpopulations.

In income distribution analysis, inequality and polarization are two conceptually different approaches used as complements in empirical studies.

Inequality analysis can be used for different purposes, see Cowell (2000) and Maasoumi (1997) for relevant surveys. For instance, the shape of the income distribution can be of primary interest. In such cases, parametric or non-parametric methods are used to estimate income density functions. On the one hand, parametric estimation imposes *a priori* strong assumptions such as unimodality. On the other hand, non-parametric estimation is less restrictive but is only *descriptive* and no inference can be implemented. To go further, ranking inequality levels between several countries or periods in time can be investigated. In such cases, inequality measures or Lorenz curves are computed for each income distribution estimated from different samples, then they are compared. Distribution-free inference for inequality measures and Lorenz dominance is now well known, see Beach and Davidson (1983), Davidson and Duclos (1997, 2000). Finally, the structure of inequality can be of primary interest and is often analysed by decomposing a population. In such cases, the class of additively decomposable inequality measures is widely used in practice, see Shorrocks (1980): subpopulations are defined by a set of individual characteristics chosen *a priori*, such as age, ethnicity, sex (e.g. Cowell and Jenkins 1995). Studying between-group inequality leads one to measure the impact of these characteristics on inequality.

Polarization analysis is conceptually different. For instance, if we are interested in the question of the "disappearing middle class" (e.g. Kuttner 1983, Thurow 1984), no inequality measure is appropriate, see Levy and Murnane (1992). In fact, this question is fundamentally different from the notion of inequality, and takes place in polarization theory developed by Esteban and Ray (1994) and Wolfson (1994). A society is said to be polarized if its population is divided into different groups or clusters. Studying the formation of groups and the gap between richest and poorest is closely related to understanding tension and social conflict. In empirical studies, the number and the location of groups has to be fixed *a priori*, then polarization measures and curves can be calculated. The main difference between inequality and polarization measures is that most of inequality measures respect the Pigou-Dalton condition of transfers, equivalent to the Lorenz curve criterion, which is the most basic axiom of inequality analysis. This condition says that any transfer of income from an individual to a richer one must increase inequality. However, this criterion is inconsistent with the concept of polarization. As argued by Wolfson (1994), it reopens questions about the Pigou-Dalton condition as the axiomatic foundation of inequality measures. In that sense, based on a questionnaire experiment, Amiel and Cowell (1999, 2001) noted that a majority of persons interviewed reject this condition as a part of their representation of inequality.

All these different concerns of inequality and polarization analyses are usually implemented independently in practice, with different *a priori* assumptions and different tools. In this paper, based on mixture models, we develop a unique method to study simultaneously these different and complementary concerns. From a theoretical point of view, mixture models reunify standard estimation methods, from parametric to non-parametric. From a practical point of view, these models benefit from the advantage of parametric estimation as the interpretation of the parameters, and the advantage of non-parametric estimation, namely that any distribution can be estimated without restrictive hypothesis. But the success of mixture models is largely due to the decomposition into different components, that can be easily interpreted as the inter-group and intra-group variability.

Our starting point is as follows: when we look at a population which is not fairly homogeneous, the observed population can be viewed as a mixture of several fairly homogeneous subpopulations. From empirical studies on income distribution analysis (Aitchison and Brown 1957, Weiss 1972), we know that the Lognormal distribution fits homogeneous subpopulations well. From theory on mixture models we know that, under regularity conditions, any probability density can be consistently estimated by a mixture of normal densities (see Ghosal and van der Vaart 2001 for a recent result about rates of convergence). Note that non-parametric Gaussian kernel density estimation is a particular case of Gaussian mixture estimation. From the relationship between the Normal and Lognormal distributions, it follows that any probability density with a positive support (as for instance income distribution) can be consistently estimated by a mixture of Lognormal densities. Then, with a finite mixture of Lognormal distributions, we expect to estimate closely the true income distribution as the number of observations tends to infinity. In this paper, we consider estimation of income distribution by mixtures of Lognormal distributions. In addition, we supplement this estimation method by an explanatory model for income distribution based on personal characteristics.

Finally, with personal income data and a set of personal characteristics, our explanatory mixture estimation gives us at the same time

1. an estimation of income distribution with *no a priori* assumptions.
2. a decomposition into several distinct homogeneous subpopulations.
3. an explanatory model to study the structure of income distribution.

In section 2, we compare mixture estimation of income distribution to standard parametric and non-parametric estimation. In section 3, we develop our explanatory mixture estimation. Then, in section 4, we apply our method to study inequality and polarization changes in Great Britain in the 1980s and 1990s.

2 Estimation of income distribution

In general, income distribution is estimated with parametric or non-parametric estimation. In this section, we consider a mixture of Lognormal distributions to estimate income distribution. Let us assume that for a homogeneous subpopulation (a proportion p_k of the population) the logarithmic-transformation of the income is distributed as a Normal distribution with mean μ_k and standard deviation σ_k. Then, the density function of the income distribution in the whole population is a finite mixture of Lognormal densities, defined as,

$$f(y) = \sum_{k=1}^{K} p_k \, \Lambda \left(y \, ; \mu_k, \sigma_k \right) \tag{1}$$

For a fixed number of components K, we can estimate $f(y)$ by maximum likelihood, see Titterington, Makov, and Smith (1985) and Lindsay (1995). We estimate the number of components K as the K which minimises a criterion, such as the BIC (Schwarz, 1978) or the AIC (Akaike, 1973). Typically, K is substantially less than the sample size. This method has been recently developed in the statistical literature, see Titterington, Makov, and Smith (1985) and Lindsay (1995). Details of estimation can be viewed as a simple case of the method developed in the next section. In the following, we study the goodness of fit of mixture estimation compared to parametric and non-parametric estimation.

2.1 Mixture vs parametric estimation

Let us assume that it is possible to fit an income distribution with a particular density function that we can write as $f(y; \theta)$, where θ is a k-vector of unknown parameters. It is of particular interest to find such a functional form to fit income distributions. Indeed, it is very easy to compare two distributions, as for example the income distribution of a population at two different periods in time, because we can explain the whole change by noting the change in the k-vector of parameters θ. A great many parametric functional forms have been employed in social science. One of them has been of particular interest when studying income distribution: the Lognormal distribution. Cowell (1977) gives many reasons to justify the use of this distribution. Firstly, this distribution has a simple relationship to the Normal distribution and many convenient properties, such as easy interpretation of parameters, non-intersecting Lorenz curves and invariance under log-linear transformations. Secondly, the Lognormal distribution fits income distribution well for a fairly homogeneous population. For instance, Aitchison and Brown (1957) and Weiss (1972) show that the Lognormal fits particularly well data sets of different fairly homogeneous sectors of the labour market. In many empirical analyses, observed income distributions have an upper tail which is not well approximated by a Lognormal. From these results, more flexible functional families have been developed with more parameters. McDonald (1984) describes the link between different parametric functions and show that the Singh-Maddala distribution fit data well in many cases.

To compare mixture vs. parametric estimation, we make a simple simulation experiment. Let us consider a sample y of $N = 2{,}000$ individual incomes drawn independently from a Singh-Maddala density function,

$$f(y) = \frac{abc\, y^{b-1}}{(1 + ay^b)^{c+1}} \tag{2}$$

We use the parameters values $a = 100$, $b = 2.8$, $c = 1.7$, which closely mirrors the net income distribution of German households, apart from a scale factor, see Brachman, Stich, and Trede (1996). We obtain, from a mixture estimation based on y,

$$\hat{f}(y) = 0.3197\, \Lambda(-2.0992, 0.7604) + 0.6803\, \Lambda(-1.8420, 0.4317) \tag{3}$$

Furthermore, an estimation based on a single parametric Lognormal distribution gives

$$\tilde{f}(y) = \Lambda(-1.9242, 0.5710) \tag{4}$$

Figure 1 show the Singh-Maddala distribution (2), the income distribution estimated by mixture (3) and by Lognormal distribution (4). It is not surprising to see that a single Lognormal distribution does not estimate a Singh-Maddala distribution well. However, we can see that the Singh-Maddala distribution is very well fitted by a mixture of two Lognormal distributions, especially the upper tail. This result suggests that we can closely estimate the Singh-Maddala distribution with a mixture of several Lognormal distributions. In our case, we can regard our population following a Singh-Maddala distribution as an aggregation of two distinct fairly homogeneous subpopulations. Mixture distribution estimation allows us to estimate the number of groups, K, the proportion of persons by groups, p_k, and the parameters of Lognormal distributions, μ_k and σ_k. This result can be easily interpreted in practice. For example, if we wish compare income distributions of one population at two different periods in time with parametric estimation and we show that Lognormal fits data well at period one, and that Singh-Maddala distribution fits data well at period two. With mixture estimation, we could analyse this evolution as the formation of several distinct groups in time.

2.2 Mixture vs non-parametric estimation

A basic hypothesis made by the use of a parametric function is that income distribution belongs to the parametric family considered. When we study income distribution with standard parametric family functions, an underlying hypothesis is that the distribution is unimodal. This hypothesis has been shown to be very restrictive in some recent empirical studies. Marron and Schmitz (1992) study the distributions of net income in Great Britain in the 1970s. They show that modelling income distributions by a parametric Singh-Maddala family leads to misleading conclusions. They use non-parametric methods and show with kernel density estimation that the densities of all years have a bimodal structure. Their results make it clear that non-parametric estimation is very useful to describe a density function. However, with these methods which are more

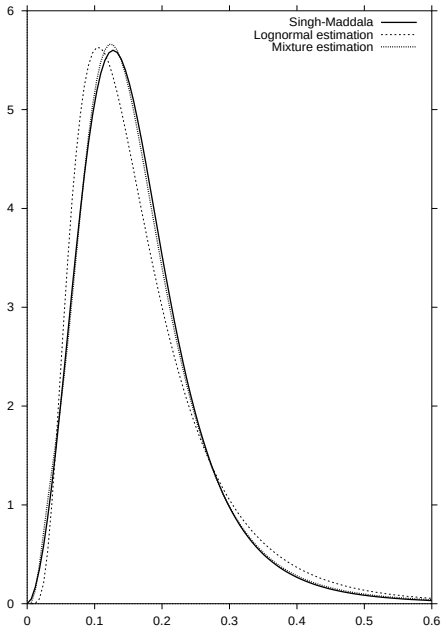

Figure 1: Mixture estimation of S-M

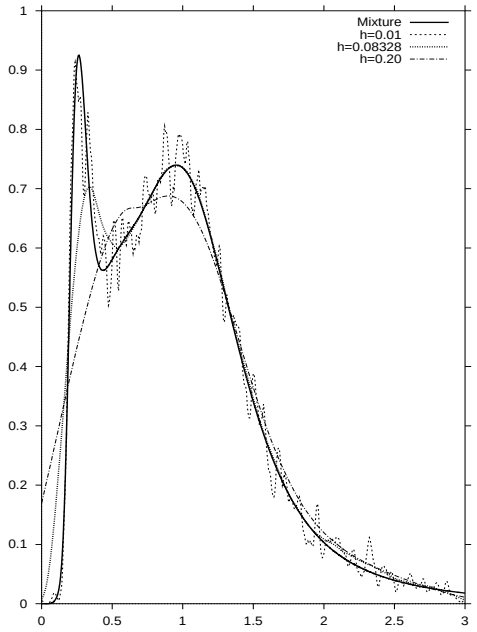

Figure 2: Kernel vs. Mixture estimation

robust, the lack of parametrisation can complicate the interpretation of the shape of the distribution: estimation function is purely descriptive. Moreover, estimation is crucially dependent on the choice of the smoothing parameter. Note that a Gaussian kernel estimator is nothing but a mixture of $K = n$ (n being the sample size) normal components with the same variance h^2 (h being the bandwidth).

Let us take the same data as Marron and Schmitz (1992) available in the ESCR Data Archive at the University of Essex: Family expenditure Survey of the United Kingdom, or FES. They use household incomes, with no use of equivalence scales, which are normalized by the arithmetic mean of the year. Based on these data, figure 2 shows the distribution of net income in Great Britain in 1973 estimated by Epanechnikov kernel density with bandwidth $h = 0.01, 0.2$ and with an *optimal* bandwith $h = 0.08328$ that would minimize the mean integrated square error if the data were Gaussian and a Gaussian kernel were used. We find similar results to Marron and Schmitz (1992, figure 2). We observe a bimodal density function, and a crucial dependence of the estimator in the amount of smoothing: the first mode is higher than the second with $h = 0.01$ and smaller than the second with h *optimal* and $h = 0.2$. Our estimation by mixture of Lognormal distribution gives the following results,

$$\hat{f}(y) = 0.106\,\Lambda(-1.321, 0.232) + 0.652\,\Lambda(-0.141, 0.607) + 0.241\,\Lambda(0.145, 0.259) \qquad (5)$$

and is plotted in figure 2 too; the observed population is a mixture of three fairly homogeneous groups. The mixture curve shows that the first mode is higher than the second, this curve is close to a smooth version of the kernel density with $h = 0.01$.

Actually, it is well known that the *optimal* bandwith used ($h = 0.08328$) is usually too wide and oversmooths the density for multimodal and highly skewed densities (Silverman 1986). Income distribution is usually highly skewed and in our case, the data generate multimodality. Then, we can suspect that the curve of kernel density with optimal bandwidth oversmooths the income distribution and reduces the first mode too much. As we have seen from figure 2, mixture estimation is similar to a smooth version of a kernel density estimation with a smaller bandwith $h = 0.01$ which does not reduce the first mode so much. Then, our results make clear that the mixture distribution estimate income distribution well, without any problem with the choice of the smoothing parameter and with a parametric form easy to interpret.

3 Explanatory mixture model

We have seen in the last section that we can closely estimate income distributions with a mixture of Lognormal densities. However, this estimation technique is unidimensional and has no explanatory power. In this section, we extend this estimation method in order to explain the structure of income distribution, based on individual characterictics.

Mixture estimation decomposes income distribution into several distinct Lognormal distributions. We make the hypothesis, justified by the previous empirical studies of Aitchison and Brown (1957) and Weiss (1972), that each Lognormal component defines a homogeneous subpopulation. Note that, as with the number of modes used to detect heterogeneity, the number of components in the mixture is invariant under a continuous and monotonic transformation of income Y. It follows that, if Y is a mixture of K Lognormal densities, then $\log(Y)$ is a mixture of K Normal densities. Then, our hypothesis is equivalent to supposing that a homogeneous subpopulation is defined by a Normal density in the distribution of the logarithmic transformation of income $\log(Y)$.

We now explain the differences between these distinct homogeneous subpopulations. We suppose that an individual's belonging to a specific subpopulation is not purely random, and can be explained by some individual characteristics. For instance, households with no adult working are expected to be more represented in the bottom of the income distribution, compared to households with all adults working. In other words, it means that individuals do not necessarily have the same probability to belong to each subpopulation and these differences can be explained by individual characteristics. In our model, it follows that, conditionally on a vector of individual characteristics X_i, or explanatory variables, the income of the i^{th} individual is distributed as the mixture

$$f\left(y_i|X_i\right) = \sum_{k=1}^{K} p_{ik}\ \Lambda\left(y_i; \mu_k, \sigma_k\right) \tag{6}$$

where p_{ik} is the probabilitity of individual i to belong to the homogeneous subpopulation k. We define p_{ik} as the probability of a random variable to belong to an interval defined with the characteristics X_i of this individual. Details of the model and its estimation

are given in the following subsections. A simple case is previously developed to illustrate and justify this approach.

3.1 Simple case

Let us taken a simple case: a population is a mixture of two homogeneous subpopulations, in the sense that the distribution of this population is a mixture of two Normal distributions with different parameters. We can express this model with a binary variable Z_i, equal to 0 if individual i belongs to the first group and equal to 1 if individual i belongs to the second group, $i = 1 \dots n$. Conditionally on $Z_i \in \{0, 1\}$, the logarithmic transformation of income Y_i of individual i follows the distribution

$$F\left(y_i|Z_i\right) = Z_i N_1\left(y_i\right) + \left(1 - Z_i\right) N_2\left(y_i\right), \tag{7}$$

where N_1 et N_2 are respectively two Normal distribution of the first and of the second groups. Then, we have two cases.

The simplest case is if we can observe Z for each individual $i, i = 1 \dots n$. In such cases, we can create two distinct samples for each group and estimate the two Normal distributions from them independently. Then we can study variability in each group independently with explanatory factors, without any bias from heterogeneity in the whole population.

However, in general, we don't know which group each individual belongs to: we observe only the result of the mixture of several groups. If we don't know anything about Z, we can express it as a random variable following a Binomial distribution with parameter p, this parameter is a probability which can be viewed as the proportion of individuals belonging to one of the two groups. Then, conditional model (7) cannot be observed, only the following marginal model is observed

$$F\left(y_i\right) = pN_1\left(y_i\right) + \left(1 - p\right) N_2\left(y_i\right), \tag{8}$$

The main point we are interested in is to explain the distribution of individuals accross groups by means of explanatory factors, or individual characteristics, as in regression analysis. Let us denote by X_i a $1 \times l$ vector of explanatory factors, β a l-vector of unknown parameters and $X_i\beta$ a linear combination of these factors. With Z_i a continuous variable, we could have used a simple linear regression $Z_i = X_i\beta + \varepsilon_i$ where the error term ε_i is white noise. However, Z_i is binary, and so we have

$$Z_i = \begin{cases} 1 \text{ if } X_i\beta + \varepsilon_i \geq \gamma \\ 0 \text{ if } X_i\beta + \varepsilon_i < \gamma \end{cases},$$

where γ is an unknown bound to be estimated. Without loss of generality, the distribution of the error term ε_i is with expectation zero and variance equal to one. If X_i contains a constant term, it is impossible to identify the constant along with γ. A solution to this problem of identification is to replace X_i by the vector of the centered

explanatory variables $X_i^c = X_i - \bar{X}$, with $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. We adopt this solution. As in standard regression, estimation and inference on parameters β leads us to select explanatory factors which significantly explain variability between groups. Each individual i, with $i = 1 \ldots n$, has now his/her own probability to belong to the first group

$$\begin{aligned} P(Z_i = 1) &= P(\varepsilon_i \geq \gamma - X_i^c \beta) \\ &= 1 - \mathrm{G}(\gamma - X_i^c \beta), \end{aligned}$$

where $\mathrm{G}(.)$ is a continuous cumulative distribution function (cdf) of a probability distribution, with expectation zero and variance unity. Consequently, for each individual, the probability to belong to the first group is equivalent to the probability that a random variable belongs to an interval with bounds which depend on the values of individual explanatory factors.

In the following, we choose G(.) as the cumulative standard normal distribution function $\Phi(.)$, as used in ordered probit models, and we extend this model to the general case of K groups.

3.2 Model

Let $U_i = X_i^c \beta + \varepsilon_i$, $(i = 1, 2, \ldots, n)$, where X_i^c is a centered vector of explanatory factors for the i^{th} individual, β a l-vector of parameters and ε_i are i.i.d. random variables, with the common distribution $N(0, 1)$. Now, for $k = 1, 2, , \ldots, K$, let

$$Z_{ik} = \begin{cases} 1 & \text{if} \quad U_i \in \left[\gamma_{k-1},\ \gamma_k\right[\\ 0 & \text{if} \quad U_i \notin \left[\gamma_{k-1},\ \gamma_k\right[\end{cases},$$

where $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_{K-1} < \gamma_K = +\infty$.
It is assumed that, given the vectors $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iK})$, the observed logarithmic transformations of income Y_i are independent and distributed according to the density

$$f(y_i | Z_i) = \sum_{k=1}^{K} Z_{ik}\ \varphi(y_i; \mu_k, \sigma_k), \tag{9}$$

where $\varphi(.; \mu, \sigma)$ is the density function of the Normal distribution with mean μ and standard deviation σ. To avoid problems of non identifiability, we assume that $\mu_1 < \mu_2 < \ldots < \mu_K$.
Note that the components of the vector Z_i are independent and distributed according to the multinomial distributions $M(1; p_{i1}, p_{i2}, \ldots, p_{iK})$, where

$$p_{ik} \equiv \mathrm{E}(Z_{ik}) = \Phi(\gamma_k - X_i^c \beta) - \Phi\left(\gamma_{k-1} - X_i^c \beta\right), \tag{10}$$

From the previous model, it follows that marginally, the Y_i are independent and distributed according to the mixture densities

$$f(y_i | X_i) = \sum_{k=1}^{K} p_{ik}\ \varphi(y_i; \mu_k, \sigma_k). \tag{11}$$

Moreover, given y_i, it can be shown that the Z_i are independent and Multinomial $M\left(1; p'_{i1}, p'_{i2}, \ldots, p'_{iK}\right)$, where

$$p'_{ik} \equiv \mathrm{E}\left(Z_{ik}|y_i\right) = \frac{p_{ik}\ \varphi\left(y_i; \mu_k, \sigma_k\right)}{\sum_{j=1}^{K} p_{ij}\ \varphi\left(y_i; \mu_j, \sigma_j\right)}. \tag{12}$$

Let $\mu = (\mu_k)_k$, $\sigma = (\sigma_k)_k$, and $\gamma = (\gamma_k)_k$. The log-likelihood function of the parameters $\theta = (\mu, \sigma, \gamma, \beta)$ is equal to

$$\ell_n(\theta, y) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} p_{ik}\ \varphi\left(y_i; \mu_k, \sigma_k\right) \right] \tag{13}$$

The maximum likelihood estimator (MLE) can be found by equating to zero the first derivatives of $\ell_n(\theta, y)$ with respect to the different parameters. There is no explicit solution to this system of equations and an iterative algorithm may be used.

3.3 Estimation

The log-likelihood function (13) is not necessarily globally concave with respect to the unknown parameters θ, and so Newton's methods can diverge. Another approach is often used to estimate mixture models: for a fixed K, an easy scheme for estimating θ is the EM algorithm (Dempster *et al.*, 1977), the "missing data" being $Z = (Z_{ik})_{i,k}$. However, one key feature of the EM algorithm is that it commonly displays a very slow linear rate of convergence. We choose to use the EM algorithm initially to take advantage of its good global convergence properties and to then exploit the rapid local convergence of Newton-type methods by switching to a direct Maximum Likelihood (ML) estimation method, see for instance Redner and Walker (1984) and McLachlan and Peel (2000).

Let us define the EM algorithm: assume for a moment, that Z is observed, then the full log-likelihood of the observations is

$$\ell_n\left(\theta, Z, y\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \left(\log \varphi\left(y_i; \mu_k, \sigma_k\right) + \log p_{ik} \right)$$

The first derivatives of this log-likelihood with respect to θ are

$$\frac{\partial \ell_n\left(\theta, Z, y\right)}{\partial \mu_k} = \sum_{i=1}^{n} Z_{ik} \frac{(y_i - \mu_k)}{\sigma_k^2}, \quad k = 1, 2, \ldots, K, \tag{14}$$

$$\frac{\partial \ell_n\left(\theta, Z, y\right)}{\partial \sigma_k} = \sum_{i=1}^{n} Z_{ik} \left[\frac{(y_i - \mu_k)^2}{\sigma_k^3} - \frac{1}{\sigma_k} \right], \quad k = 1, 2, \ldots, K, \tag{15}$$

Then, for $j = 1 \ldots l$,

$$\frac{\partial \ell_n\left(\theta, Z, y\right)}{\partial \beta_j} = -\sum_{i=1}^{n} X_{ij}^c \sum_{k=1}^{K} \frac{Z_{ik}}{p_{ik}} \left[\varphi\left(\gamma_k; X_i^c \beta, 1\right) - \varphi\left(\gamma_{k-1}; X_i^c \beta, 1\right) \right] \tag{16}$$

and since γ_0 and γ_K are fixed, for $k = 1, 2, \ldots, K-1$,

$$\frac{\partial \ell_n(\theta, Z, y)}{\partial \gamma_k} = \sum_{i=1}^{n} \varphi\left(\gamma_k; X_i^c \beta, 1\right) \left[\frac{Z_{ik}}{p_{ik}} - \frac{Z_{i(k+1)}}{p_{i(k+1)}}\right] \tag{17}$$

But Z is unobserved. In the iterative EM procedure, the conditional expectation of the full likelihood given the observations y is first evaluated (E step), then this "predicted" log-likelihood is maximised with respect to θ (M step). Applying this procedure to our case, we obtain the following double step.

- *E-step:* Given θ, the missing data Z_{ik} are replaced by their conditional expectation

$$p'_{ik} \equiv \mathrm{E}\left(Z_{ik} | \theta, y_i\right) = \frac{p_{ik}\ \varphi\left(y_i; \mu_k, \sigma_k\right)}{\sum_{j=1}^{K} p_{ij}\ \varphi\left(y_i; \mu_j, \sigma_j\right)},$$

- *M-step:* Given the previous predictions of the missing data, the estimates of θ are obtained by maximising the expression $\ell_n(\theta, p', y)$: the equations $\partial \ell_n(\theta, p', y)/\partial \mu = 0$ and $\partial \ell_n(\theta, p', y)/\partial \sigma = 0$ give the explicit estimates

$$\widehat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} p'_{ik} y_i, \qquad \text{and} \qquad \widehat{\sigma}_k = \sqrt{\frac{1}{N_k} \sum_{i=1}^{n} p'_{ik} \left(y_i - \widehat{\mu}_k\right)^2},$$

  where $N_k = \sum_{i=1}^{n} p'_{ik}$, is the current estimate of the number of observations in the k^{th} cluster, $k = 1, 2, \ldots, K$.
  Current estimates of β and γ are computed via an iteration of a Newton algorithm based on the first derivatives

$$\frac{\partial \ell_n(\theta, p', y)}{\partial \beta_j} = -\sum_{i=1}^{n} X_{ij}^c \sum_{k=1}^{K} \frac{p'_{ik}}{p_{ik}} \left[\varphi\left(\gamma_k; X_i^c \beta, 1\right) - \varphi\left(\gamma_{k-1}; X_i^c \beta, 1\right)\right]$$

  for $j = 1 \ldots l$, and

$$\frac{\partial \ell_n(\theta, p', y)}{\partial \gamma_k} = \sum_{i=1}^{n} \varphi\left(\gamma_k; X_i^c \beta, 1\right) \left[\frac{p'_{ik}}{p_{ik}} - \frac{p'_{i(k+1)}}{p_{i(k+1)}}\right]$$

  for $k = 1, 2, \ldots, K-1$.

These two steps are iterated until some convergence criterion is met.

The next step is the use of the Maximum Likelihood estimation, based on Newton's methods. These methods are well known and largely used in practice to maximise multidimensional functions, see Press et al. (1986) for algorithmic details. Note that after each iteration, it is necessary to sort parameter estimates $(\widehat{\mu}_k, \widehat{\sigma}_k, \widehat{\gamma}_k)$ by increasing $\widehat{\mu}_k$; standard errors of the parameter estimates are given by the square root of the diagonal components of the inverse of the information matrix.

3.4 Simulations

In mixture models, and consequently in our "explanatory" mixture model, the presence of significant multimodality in finite sample has a number of important consequences (Lindsay 1995).

The first implication is that the solution of the algorithm employed can greatly depend on the initial values chosen. Starting values can be chosen in different ways, for instance Finch, Mendell, and Thode (1989) suggest the use of multiple random starts, Furman and Lindsay (1994) investigate the use of moment estimators. However, there is no best solution. In our experiments, we estimate initial values of the mean μ and of the standard deviation σ with robust statistics: from a sorted subsample, we compute the median and the interquartile range in K subgroups with the same number of observations. This choice works fine in many simulation experiments.

The second implication is that a simulation study can be highly dependent on the stopping rules and search strategies employed. Then, it can be difficult to compare simulation studies. In mixture models, a problem of convergence can be encountered when the proportion of observations in a subgroup is too small: it can come from initial values too far from the true values of parameters, or when K, the number of components chosen, is too large. We decide to reduce the number of components when the current estimation of the number of observations in a subpopulation is equal to zero ($N_k = 0$).

In our simulations, we consider the explanatory mixture model defined in (6) and (10) with the following values,

$$\mu_k = 2\,k \qquad \sigma_k = 0.5 + (k/100)(-1)^k \qquad \gamma_k = -3 + 6\,k/K \quad \text{and} \quad \beta_j = (-1)^j \quad (18)$$

for $j = 1, \ldots, l$. These values are chosen to have distinct Lognormal distributions with quite similar, but different, variances and proportions of individuals in each distribution. We define the $n \times l$ matrix of regressors X by drawing observations from the Normal distribution $N(0, 1)$. In our experiments, the number of observations ($n = 500$) and the number of regressors ($l = 5$) are fixed, the number of component is respectively equal to $K = 2, 4, 6, 8$. For each value of K, we draw 5,000 samples and we estimate μ_k, σ_k, γ_k and β_j with a mixture model with K components[1]. Then, we compute the mean and the standard deviation of the 5,000 realisations obtained for each parameter.

Results are given in table 1, with true values given in the second column, note that the true values of γ_k are not given because they are not the same for different values of K. From this table, we can see that the unknown parameters are very well estimated with the explanatory mixture model: means are very close to the true values and standard deviations are small. Additional experiments could be done. However, our goal is not to address a complete simulation study, because of the preceding reasons and because there are many experiments in the unidimensional case already done, as for instance Finch, Mendell, and Thode (1989), Furman and Lindsay (1994). From our experiments, the

[1]We fix the number of components K in the mixture estimation equal to the number of components in the data simulation process. We do not address the issue of the choice of K in these simulations.

	true	$K=2$		$K=4$		$K=6$		$K=8$	
$\hat{\mu}_1$	2	2.000	(0.033)	2.001	(0.045)	2.000	(0.053)	1.998	(0.056)
$\hat{\mu}_2$	4	4.000	(0.035)	3.999	(0.062)	3.998	(0.089)	3.996	(0.125)
$\hat{\mu}_3$	6			6.002	(0.054)	6.000	(0.074)	6.000	(0.101)
$\hat{\mu}_4$	8			7.999	(0.051)	8.001	(0.085)	7.997	(0.110)
$\hat{\mu}_5$	10					10.001	(0.072)	10.000	(0.087)
$\hat{\mu}_6$	12					12.000	(0.061)	11.995	(0.117)
$\hat{\mu}_7$	14							14.000	(0.090)
$\hat{\mu}_8$	16							16.000	(0.069)
$\hat{\sigma}_1$	0.49	0.489	(0.024)	0.487	(0.034)	0.486	(0.039)	0.485	(0.043)
$\hat{\sigma}_2$	0.52	0.519	(0.025)	0.519	(0.056)	0.520	(0.090)	0.525	(0.132)
$\hat{\sigma}_3$	0.47			0.468	(0.048)	0.468	(0.073)	0.469	(0.109)
$\hat{\sigma}_4$	0.54			0.537	(0.038)	0.538	(0.091)	0.543	(0.137)
$\hat{\sigma}_5$	0.45					0.450	(0.069)	0.446	(0.093)
$\hat{\sigma}_6$	0.56					0.555	(0.046)	0.561	(0.135)
$\hat{\sigma}_7$	0.43							0.431	(0.106)
$\hat{\sigma}_8$	0.58							0.575	(0.054)
$\hat{\gamma}_1$		-0.019	(0.098)	-1.548	(0.126)	-2.058	(0.137)	-2.319	(0.144)
$\hat{\gamma}_2$				-0.019	(0.097)	-1.034	(0.108)	-1.541	(0.122)
$\hat{\gamma}_3$				1.508	(0.121)	-0.020	(0.099)	-0.784	(0.113)
$\hat{\gamma}_4$						0.997	(0.106)	-0.017	(0.100)
$\hat{\gamma}_5$						2.014	(0.129)	0.739	(0.106)
$\hat{\gamma}_6$								1.509	(0.121)
$\hat{\gamma}_7$								2.279	(0.139)
$\hat{\beta}_1$	-1	-1.037	(0.137)	-1.019	(0.082)	-1.017	(0.072)	-1.020	(0.067)
$\hat{\beta}_2$	1	1.035	(0.128)	1.019	(0.080)	1.016	(0.071)	1.017	(0.068)
$\hat{\beta}_3$	-1	-1.038	(0.137)	-1.020	(0.083)	-1.017	(0.072)	-1.019	(0.069)
$\hat{\beta}_4$	1	1.036	(0.125)	1.020	(0.077)	1.018	(0.067)	1.020	(0.065)
$\hat{\beta}_5$	-1	-1.034	(0.136)	-1.019	(0.083)	-1.017	(0.072)	-1.019	(0.071)

Table 1: Simulation results: mean and standard deviation of 5,000 realisations

main result is that explanatory mixture model estimation works fine when the observed population is defined as a mixture of sufficiently distinct subpopulations.

3.5 Interpretation

From our explanatory mixture model, we can make few remarks about its use in practice.

• Let us consider model (6), with individual probabilities p_{ik} defined in (10). Under the null hypothesis $H_0 : \beta_j = 0$, the individual characteristic X_{ij} is not significant in p_{ik}. A t-test can be easily computed: we divide the parameter estimate by its standard deviation, as is done in standard linear regression. If we reject the null hypothesis

$\beta_j = 0$, it means that individual probabilities are not the same and therefore, that the characteristic X_{ij} is statistically significant to explain "inter-subpopulation" variability.

• A nice feature of model (6) is that we estimate conditional distributions for each individual. In our model, individual characteristics appear only in the probability to belong to a subpopulation. Then, it is clear that if we reject the null hypothesis $\beta_j = 0$, we can conclude that individual characteristic X_{ij} is statistically significant in the conditional distribution $f(y_i|X_i)$. Non-parametric methods are used in the literature to have a plot of a conditional distribution, based on bivariate distributions. For instance, Pudney (1993) study the relationship between age and income distributions. He uses plot and contour plot of the conditional distribution to have an idea of this relationship and he computes inequality measures based on conditional distributions. These non-parametric methods are often restricted to the use of two dimensions, that is to say, income with one additional characteristic. Note that similar studies could be done, based on conditional distributions estimated by mixture (6), with more than two dimensions.

• An interesting interpretation of parameter β_j, $j = 1, \ldots, l$, is to explain individual position in the income distribution, based on individual characteristics X_{ij},

If $\widehat{\beta}_j > 0$ (respectively $\widehat{\beta}_j < 0$), then the individual position moves in the direction of the upper part of the income distribution (respectively to the bottom) when X_{ij} increases.

To describe this result formally, we define the individual "position" in the income distribution as $P_i = \sum_{k=1}^{K} \widehat{p}_{ik}\, \widehat{\mu}_k$, recall that $\widehat{\mu}_k$ are sorted in increasing order. Then, the partial derivative of P_i with respect to X_{ij} measure the influence on P_i of a change in the value of X_{ij},

$$\begin{aligned}
\frac{\partial P_i}{\partial X_{ij}} &= -\widehat{\beta}_j \left[\sum_{i=1}^{K} \left(\varphi\left(\widehat{\gamma}_k; X_i\widehat{\beta}, 1 \right) - \varphi\left(\widehat{\gamma}_{k-1}; X_i\widehat{\beta}, 1 \right) \right) \widehat{\mu}_k \right] \\
&= \widehat{\beta}_j \left[\sum_{i=1}^{K-1} \varphi\left(\widehat{\gamma}_k; X_i\widehat{\beta}, 1 \right) \left(\widehat{\mu}_{k+1} - \widehat{\mu}_k \right) \right]
\end{aligned}$$

The right term, in brackets, is always positive. Then, we can see that, if β_j is positive, P_i increases if X_{ij} increases. In addition, we can see that the first term $\widehat{\beta}_j$ does not depend on the component k, and the last term, in brackets, is specific to the component k. Then, we can view $\widehat{\beta}_j$ as the overall influence of the characteristic j on the position of the individual i in the income distribution. A large negative value, relatively to its standard deviation, shows an income position of individuals with characteristic j clearly in the bottom of the income distribution. A large positive value shows an income position clearly in the top of the income distribution.

• To have a plot of the whole income distribution, we can use an estimate of the

marginal distribution,

$$\hat{f}(y) = \sum_{k=1}^{K} \bar{p}_k \, \Lambda\left(y \,; \widehat{\mu}_k, \widehat{\sigma}_k\right) \quad \text{with} \quad \bar{p}_k = \frac{1}{n} \sum_{i=1}^{n} \widehat{p}_{ik} \tag{19}$$

where $\bar{p}_k$ is the average proportion of individuals in subpopulation k, calculated as the mean of the estimated individual probabilities to belong to this subpopulation.

4 Application

We analyse the position of particular households in the income distribution and relative income changes between 1979 and 1996. The data used have been derived from the Family Expenditure Survey (FES), which is a continuous survey of samples of the UK population living in households. Data were made available by the ESCR Data archive at the University of Essex: Department of Employment, Statistics Division. We take disposable household income (i.e., post-tax and transfer income) before housing costs. To compare household income between households with different sizes, we divide household income by an adult-equivalence scale defined by McClements. Furthermore, we exclude the self-employed from the data, as recommended by the methodological review produced by the Department of Social Security (1996): some evidence suggests that the survey questions prior to 1996/7 lead us to an under estimation of self-employed household income that can distort the income distribution for the whole population. To restrict the study to relative effects, the data for each year have been normalized by the arithmetic mean of the year. In addition, the data give us the composition of households: for each person of a household we know its sex, age, labour force status (employee, selfemployed, unemployed, inactive, student). For a detailed description of data, known as HBAI-like data, and McClements equivalent scale, see the annual report produced by the Department of Social Security (1998).

Based on these data, Jenkins (2000) and the annual report produced by the Department of Social Security (1998) show that having increased during the 1980s, inequality appears to have fallen slightly during the 1990s. Table 2 shows Theil, Mean Logarithmic

	Theil		MLD		Gini	
1979	0.1066	(0.0023)	0.1056	(0.0020)	0.2563	(0.0023)
1988	0.1619	(0.0053)	0.1542	(0.0036)	0.3074	(0.0034)
1992	0.1794	(0.0065)	0.1743	(0.0046)	0.3214	(0.0037)
1996	0.1507	(0.0046)	0.1457	(0.0036)	0.2976	(0.0033)

Table 2: Inequality measures over years

Deviation and Gini indexes, with their standard deviations in parentheses, for the years 1979, 1988, 1992 and 1996. All these inequality measures increase considerably from 1979 to 1988 and decrease from 1992 to 1996.

In this section, we analyse this evolution of inequality over the years with the use of the method we proposed in the preceding section, a mixture estimation with explanatory variables. We define an adult as a person aged 19 or over, or a 16 to 18 year old not student, otherwise it is a child. Then, we consider the following characteristics:

X_{i1} - *Pensioner* : the head of the family is a person of state pension age or above (65 for men, 60 for women).

X_{i2} - *Lone parent family* : a single non-pensioner adult with children.

X_{i3} - *All-working* : non-pensioner household with all adults working.

X_{i4} - *Non-working* : non-pensioner household with all adults not working.

X_{i5} - *Number of children*

Note that X_{i1}, X_{i3} and X_{i4} are mutually exclusive variables (a pensioner household cannot be a non-working or all-working household), not X_{i2} and X_{i5} (a lone parent family is a non-working or all-working household too). We use the explanatory mixture estimation with the dummy variables X_{i1}, X_{i2}, X_{i3}, X_{i4} and X_{i5} as a set of explanatory factors. Our estimation by a mixture of Lognormal distributions with explanatory variables leads to the following results,

$$\hat{f}(y|X_i) = \sum_{k=1}^{K} \widehat{p}_{ik}\, \Lambda\left(\widehat{\mu}_k, \widehat{\sigma}_k\right) \tag{20}$$

Numerical results are given in appendix and in table 3 for the years 1979, 1988, 1992 and 1996. From these results, we begin by studying changes in the shape of the income distribution. Then, we study changes in the structure of the income distribution through parameter estimates of the explanatory variables X_{i1}, X_{i2}, X_{i3}, X_{i4} and X_{i5} as defined above.

4.1 The Shape of the Income Distribution

Figures 3, 4, 5 and 6 plot the marginal distribution of our estimation by mixture with explanatory variables (`mixture`) and the several Lognormal distributions which constitute the mixture, `pLog`$k = \bar{p}_k\, \Lambda(\widehat{\mu}_k, \widehat{\sigma}_k)$, for $k = 1, \ldots, K$, for the years 1979, 1988, 1992 and 1996, see equation (19) and numerical results in appendix. If we restrict our attention to the global curve, we can see in all figures a multimodal distribution, which is slightly modified over the years. However, from estimation of the income distribution only, no clear conclusion can been drawn to explain inequality evolution. Our method allows us to decompose the income distribution into several distinct Lognormal distributions, that can be associated to several distinct fairly homogeneous subpopulations. Then, we can analyse the relative evolution of these distinct distributions over years.

Before the analyse of the shape of the income distribution, we can make two remarks:

1. Our results show that a mixture of K Lognormal distributions does not necessarily mean that the observed population is composed of K different and fairly homogeneous subpopulations. For instance in 1988, numerical results show that income distribution can be estimated by a mixture of seven Lognormal distributions (see appendix). However, from figure 4 we can clearly see six distinct Lognormal distributions and another one (`pLog7`), close to the x-axis and very difficult to see, which is very flat with a large dispersion ($\widehat{\sigma}_7 = 0.4358$) and a small probability ($\bar{p}_7 = 0.0170$). The role of this *"flat"* Lognormal distribution (`pLog7`) is not to identify another distinct distribution, but to give a better fit of the whole distribution. Then, we can consider two types of Lognormal distribution which constitute a mixture estimation: a first one which is a distinct individual distribution in the mixture ; and a second one which improves the precision of the global estimate. This last type of distribution can be detected by large dispersion and very small probability, compared to the others.

2. We know that the Lognormal distribution fits income distribution well for a fairly homogeneous population, see for instance Aitchison and Brown (1957) and Weiss (1972). Then, we could consider as many fairly homogeneous subpopulations as we can see distinct Lognormal distributions in mixture estimation. For the year 1988 (figure 4), we would consider six fairly homogeneous subpopulations that compose the observed population. Note that if, for instance, we are only concerned by a distinction between "rich" and "poor" in the whole population, figure 4 suggests that we could describe a "poor" subpopulation as a mixture of the first three Lognormal distributions, and a "rich" subpopulation as a mixture of the last three Lognormal distributions. In that way, we define two subpopulations, "rich" and "poor", with no a priori assumption on their respective income distribution. However, we would not suppose these subpopulations to be homogeneous, because their respective income distributions are not estimated by a single Lognormal distribution.

Let us compare income distributions in 1979 and 1988, respectively in figures 3 and 4. Firstly, we detect five distinct homogeneous subpopulations in 1979 and six in 1988: a new small distribution appears in the bottom of the distribution. In addition, we can see that the lowest distributions move to the left ($\widehat{\mu}_3 = 0.6184$ in 1979 and $\widehat{\mu}_4 = 0.5550$ in 1988, see appendix). Secondly, we can see that the upper single Lognormal distribution has significantly increased: more people are in the upper distribution, $\bar{p}_5 = 0.2106$ in 1979 becomes $\bar{p}_6 = 0.3240$ in 1988, which means that the "richest" subpopulation is represented by 21.06% of the whole population in 1979 and by 32.40% in 1988. Finally, we can see two changes in opposite directions: an increasing number of people at the top of the distribution and an increasing gap between upper and lowest distributions. This suggests an increasing number of "rich" people and an increasing gap between the "richest" and the poorest" subpopulations and so, an increasing inequality in the 80s.

Let us compare income distributions in 1988 and 1992, respectively in figures 4 and 5. We detect six homogeneous subpopulations in 1988 and seven in 1992. We can see that

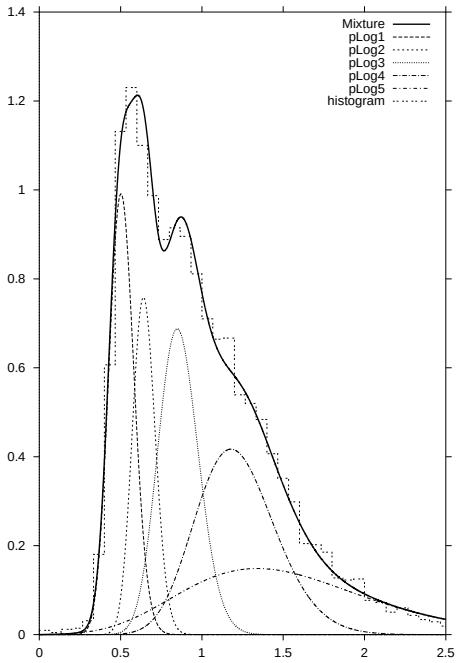

Figure 3: Income distribution in 1979

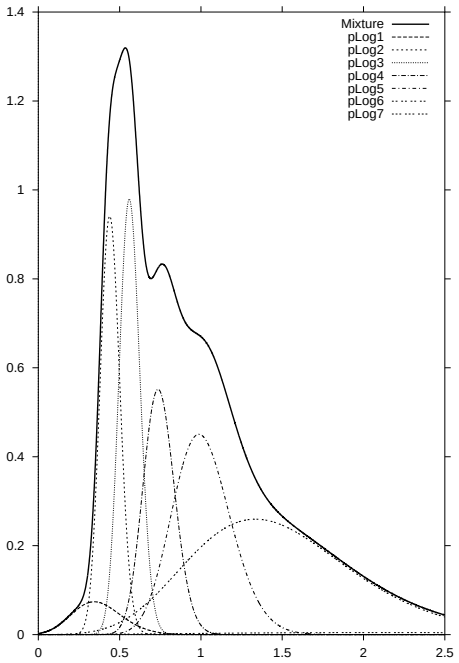

Figure 4: Income distribution in 1988

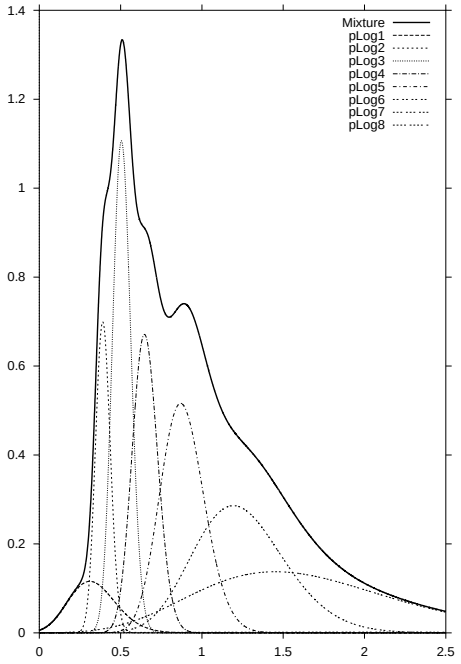

Figure 5: Income distribution in 1992

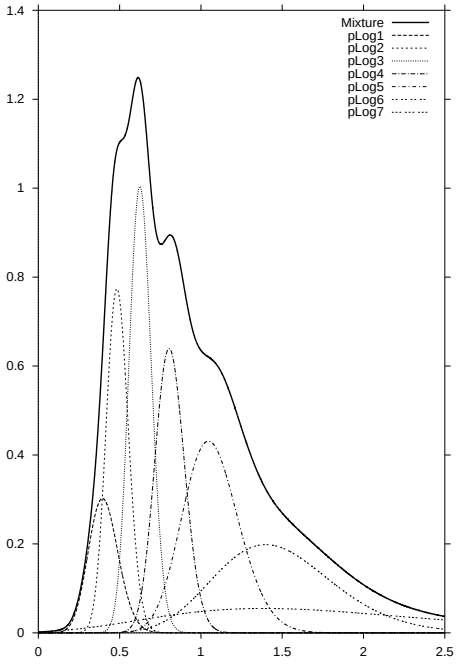

Figure 6: Income distribution in 1996

the lowest distribution has significantly increased ($\bar{p}_1 = 0.0280$ in 1988 and $\bar{p}_1 = 0.0419$ in 1992) and that the upper distribution has significantly decreased ($\bar{p}_6 = 0.3240$ in 1988 and $\bar{p}_7 = 0.2104$ in 1992). It suggests that there is less very "rich" people, but more very "poor" people and so, it can explain an increasing inequality with not so many changes as in the 80s.

Let us compare income distributions in 1992 and 1996, respectively in figures 5 and 6. Firstly, the top distribution in 1996 has a large dispersion ($\widehat{\sigma}_7 = 0.3398$) compared to the others, but its probability is not very small ($\bar{p}_7 = 0.1181$): it is not a clear distinct distribution and its role is not clearly to improve the precision of the global estimate only (see remark 1 above). It suggests that 1996 is a year of transition between seven and six homogeneous subpopulations: one subpopulation is in the process of disappearing[2]. Secondly, we can see that the lowest distribution and so, the bottom of the global curve, moves significantly to the right: condition in life of the "poorest" people get better. In addition, from the shape of the global curve we can see a decrease of the gap between the two major modes. All these remarks suggest a decreasing inequality.

Finally, the study of the shape of the income distribution follows increasing inequality in the 80s and slightly decreasing in the 90s, and gives us a better idea of this evolution through the different parts of the distribution.

4.2 The structure of the income distribution

Parameter estimates of explanatory variables X_{i1}, X_{i2}, X_{i3}, X_{i4} and X_{i5}, based on mixture estimation, for years 1979, 1988, 1992 and 1996 are given in table 3, with standard deviations in parenthesis. These results allow us to analyse the position of households in the income distribution.

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$
1979	-1.770 (0.059)	-0.672 (0.106)	0.611 (0.050)	-1.160 (0.086)	-0.439 (0.020)
1988	-1.329 (0.058)	-0.694 (0.106)	0.781 (0.053)	-1.440 (0.068)	-0.352 (0.022)
1992	-1.109 (0.053)	-0.546 (0.083)	0.717 (0.050)	-1.240 (0.060)	-0.345 (0.019)
1996	-0.999 (0.055)	-0.616 (0.078)	0.758 (0.053)	-1.107 (0.062)	-0.384 (0.020)

Table 3: Parameter estimates $\widehat{\beta}_j$ of individual characteristics X_j

In 1979, the largest negative values are successively associated to *pensioners* ($X_{i1} : \widehat{\beta}_1 = -1.770$) and *non-working* ($X_{i4} : \widehat{\beta}_4 = -1.160$), the largest positive value is associated to *all-working* ($X_{i3} : \widehat{\beta}_3 = 0.611$). It means that households with no adult working and pensioners are strongly over-represented in the bottom of the distribution, while households with all adults working are over-represented in the top of the distribution.

If we restrict our attention to the most significant variables, from table 3, major

[2]It is confirmed with additional data for the year 1999: we detect six homogeneous subpopulations

changes over years can be reduced to:

1. An improvement in the income position of *pensioners*: parameter estimates $\widehat{\beta}_1$ decrease over time, from -1.770 in 1979 to -0.999 in 1996.

2. A large increasing gap between the income position of *all-working* and *non-working* households in the 80s and an small decrease in the 90s: $\widehat{\beta}_3 - \widehat{\beta}_4$ is respectively equal to 1.771, 2.221, 1.957, 1.865.

3. The income position of *non-working* households becomes less than that of *pensioners*: respectively -1.160 vs. -1.770 in 1979 and -1.107 vs. -0.999 in 1996.

These results show that in the 80s the polarization between all-working and non-working households increased, then polarization decreased slowly in the 90s. On another side, the position of pensioners increased over years.

We calculate the percentage of the population in the bottom 10% of the income distribution by household type (pensioner, lone parent family and no adult working), and in the top 10% for households with all adult working. Results are given in table 4, with the percentage of the population by household type given in parentheses.

	X_{i1}		X_{i2}		X_{i3}		X_{i4}	
	bottom	(%pop)	bottom	(%pop)	top	(%pop)	bottom	(%pop)
1979	6.5	(29.3)	0.6	(2.8)	7.4	(44.9)	2.3	(5.9)
1988	4.8	(30.7)	1.0	(4.1)	7.2	(40.1)	4.2	(12.2)
1992	3.8	(30.1)	1.3	(5.5)	6.8	(38.0)	4.9	(15.1)
1996	3.6	(29.9)	1.8	(6.6)	7.1	(39.0)	4.9	(15.7)

Table 4: Percentage of households in the bottom/top 10% of the distribution

These results confirm our preceding analysis:

- Representation of *pensioners* (X_{i1}) in the bottom 10% of the income distribution decreases from 6.5% in 1979 to 3.6% in 1996, while its representation in the whole population is still around 30% over the years.

- The percentage of *non-working* households (X_{i4}) in the whole population increases, from 5.9% in 1979 to 15.7% in 1996. Moreover, their representation in the bottom 10% of the distribution largely increases in the 80s and is still stable in the 90s (it is a slight decrease relative to the proportion of this household type, which has increased between 1992 and 1996).

- The bottom 10% of the distribution is represented in the great majority by *pensioners* and *non-working* households: together they represent 8.8% in 1979 and 8.5% in 1996 of the whole population. However, its distribution has been largely modified: *pensioners* are dominant in 1979 (6.5% against 2.3%), but not in 1996 (3.6% against 4.9%).

In addition, we can see that the number of *lone parent families* (X_{i2}) increases over the years and its representation in the bottom 10% of the distribution increases. Finally,

the proportion of *all-working* households decreases over the years but its representation in the top 10% of the distribution is still large (around 7.2%).

From our studies on the shape and on the structure of the income distribution over the years, we can explain increasing inequality in the 1980s by an increasing polarization between working and non-working households: the proportion of non-working households has been multiplied by more than twice (5.9% to 12.2%) and more people moved to the upper part of the distribution. Then, we can explain the slight decrease of inequality in the 1990s by a small decrease of this polarization: the number of people in the upper part of the distribution decreased and the income position of non-working households increased slightly. On the other hand, the income position of pensioners has improved. All these results are supported by previous work in the literature, as for instance Jenkins (2000), or the descriptive statistical studies of the Department of Social Security (1998).

5 Conclusion

In this paper, we have proposed a new method to analyse income distribution, based on mixture models. This method allows us to estimate the density of the income distribution, to detect homogeneous subpopulations and to analyse the position of individuals with specific characteristics. An application to income data in Great Britain in the 1980s and 1990s shows how to analyse the shape and the structure of the income distribution and leads us to study at the same time inequality and polarization changes over years. Our empirical results show that this method can be succesfully used in practice.

Acknowledgment

Financial support from Spanish DGES, grant BEC2002-03720, is acknowledged.

References

Aitchison, J. and J. A. C. Brown (1957). *The Lognormal Distribution*. Cambridge University Press, London.

Amiel, Y. and F. A. Cowell (1999). *Thinking about Inequality*. Cambridge University Press, Cambridge.

Amiel, Y. and F. A. Cowell (2001). "Attitudes to risk and inequality: a new twist on the priciple transfer". DARP-56 Discussion Paper, STICERD, London School of Economics.

Beach, C. M. and R. Davidson (1983). "Distribution-free statistical inference with Lorenz curves and income shares". *Review of Economic Studies 50*, 723–735.

Brachman, K., A. Stich, and M. Trede (1996). "Evaluating parametric income distribution models". *Allgemeines Statistisches Archiv 80*, 285–298.

Cowell, F. A. and S. P. Jenkins (1995). "How much inequality can we explain? A methodology and an application to the U.S.A.". *Economic Journal 105*.

Cowell, F. A. (1977). *Measuring Inequality*. Philip Allan Publishers Limited, Oxford.

Cowell, F. (2000). "Measurement of inequality". In *Handbook of Income Distribution*, Volume 1, pp. 87–166. A. B. Atkinson and F. Bourguignon (eds), Elsevier Science.

Davidson, R. and J. Y. Duclos (1997). "Statistical inference for the measurement of the incidence of taxes and transfers". *Econometrica 65*, 1453–1465.

Davidson, R. and J. Y. Duclos (2000). "Statistical inference for stochastic dominance and for the measurement of poverty and inequality". *Econometrica 68*, 1435–1464.

Department of Social Security (1996). Households Below Average Income: Methodological Review Report of a Working Group. Corporate Document Services, London.

Department of Social Security (1998). Households Below Average Income 1979-1996/7. Corporate Document Services, London.

Esteban, J. M. and D. Ray (1994). "On the measurement of polarization". *Econometrica 62*(4), 819–851.

Finch, S. J., N. R. Mendell, and H. C. Thode (1989). "Probabilistic measures of adequacy of a numerical search for a global maximum". *Journal of the American Statistical Association 84*, 1020–1023.

Furman, D. and B. G. Lindsay (1994). "Measuring the relative effectiveness of moment estimators as starting values in maximizing mixture likelihoods". *Comput. Statist. Data Anal. 17*, 493–507.

Ghosal, S. and A. W. van der Vaart (2001). "Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities". *Annals of Statistics 29*(5), 1233–1263.

Jenkins, S. P. (2000). "Trends in the UK income distribution". In *The personal Distribution of Income in an International Perspective*. Springer-Verlag, Berlin.

Kuttner, B. (1983). "The declining middle". *Atlantic Monthly 252*, 60–71.

Levy, F. and R. J. Murnane (1992). "U.S. earnings levels and earnings inequality: a review of recent trends and proposed explanations". *Journal of Economic Literature 30*, 1333–81.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and applications*. Regional Conference Series in Probability and Statistics.

Maasoumi, E. (1997). "Empirical analyses of inequality and welfare". In *Handbook of Applied Econometrics : Microeconomics*, pp. 202–245. Pesaran, M. H. and P. Schmidt (eds), Blackwell.

Marron, J. S. and H. P. Schmitz (1992). "Simultaneous density estimation of several income distributions". *Econometric Theory 8*, 476–448.

McDonald, J. B. (1984). "Some generalized functions for the size distribution income". *Econometrica 52*, 647–663.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics Section.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1986). *Numerical Recipes*. Cambridge University Press, Cambridge.

Pudney, S. (1993). "Income and wealth inequality and the life cycle: a non parametric analysis for china". *Journal of Applied Econometrics 8*, 249–276.

Redner, R. and H. F. Walker (1984). "Mixture densities, maximum likelihood and the EM algorithm". *SIAM Rev. 26*, 195–239.

Shorrocks, A. F. (1980). "The class of additively decomposable inequality measures". *Econometrica 48*, 613–625.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Thurow, L. (1984). "The disappearance of the middle class". *New York Times*, 5 february 1984, section 3, p. 2.

Titterington, D. M., U. E. Makov, and A. F. M. Smith (1985). *Statistical Analysis of Finite Mixture Distributions*. J. Wiley, New-York.

Weiss, Y. (1972). "The risk element in occupational and educational choices". *Journal of Political Economy 80*, 1203–1213.

Wolfson, M. (1994). "When inequality diverge". *American Economic Review 84*(2), 353–358.

Appendix

In table 5, we present results of mixture estimation with explanatory variables for the income distribution in 1979, 1988, 1992 and 1996.

We estimate the unknown parameters $\theta = (\mu_k, \sigma_k, \gamma_k, \beta)$: estimates of μ_k, σ_k, γ_k and $\bar{p}_k$ are presented in table 5 and estimates of β are presented in table 3. In our data, some values of income are equal to zero, note that values of income close to zero can become extreme values with the logarithmic transformation and can cause problems to estimate γ_k. To take into account observations equal to zero and to avoid problems of convergence, we can translate data with a fixed parameter $y + \xi$. Then, we use the marginal distribution

$$\hat{f}(y) = \sum_{k=1}^{K} \bar{p}_k \, \Lambda\left(y + \xi \,; \widehat{\mu}_k, \widehat{\sigma}_k\right) \quad \text{with} \quad \bar{p}_k = \frac{1}{n} \sum_{i=1}^{n} \widehat{p}_{ik} \tag{21}$$

where $\widehat{p}_{ik} = \Phi(\widehat{\gamma}_k - X_i^c \widehat{\beta}) - \Phi(\widehat{\gamma}_{k-1} - X_i^c \widehat{\beta})$, and

$$\Lambda\left(y + \xi \,; \widehat{\mu}_k, \widehat{\sigma}_k\right) = \frac{1}{(y+\xi)\sqrt{2\pi}\,\widehat{\sigma}_k} \exp\Big[- \frac{1}{2\widehat{\sigma}_k^2} \big(\log\left(y + \xi\right) - \widehat{\mu}_k \big)^2 \Big], \tag{22}$$

to plot an estimate of the income distribution for the years 1979 (figure 3), 1988 (figure 4), 1992 (figure 5) and 1996 (figure 6). Our numerical results are computed with $\xi = 1$.

	1979	1988	1992	1996
$\hat{\mu}_1$	0.4096 (0.0041)	0.3080 (0.0218)	0.2828 (0.0168)	0.3369 (0.0100)
$\hat{\mu}_2$	0.4967 (0.0065)	0.3657 (0.0056)	0.3304 (0.0086)	0.3962 (0.0098)
$\hat{\mu}_3$	0.6184 (0.0070)	0.4458 (0.0068)	0.4102 (0.0090)	0.4869 (0.0075)
$\hat{\mu}_4$	0.7910 (0.0116)	0.5550 (0.0118)	0.5010 (0.0134)	0.5928 (0.0103)
$\hat{\mu}_5$	0.9053 (0.0129)	0.6949 (0.0132)	0.6307 (0.0129)	0.7228 (0.0156)
$\hat{\mu}_6$	-	0.8918 (0.0127)	0.8014 (0.0182)	0.8973 (0.0255)
$\hat{\mu}_7$	-	1.3216 (0.1167)	0.9550 (0.0208)	0.9846 (0.0253)
$\hat{\mu}_8$	-	-	1.4536 (0.1879)	-
$\hat{\sigma}_1$	0.0507 (0.0024)	0.1117 (0.0107)	0.1094 (0.0076)	0.0649 (0.0061)
$\hat{\sigma}_2$	0.0426 (0.0034)	0.0418 (0.0034)	0.0325 (0.0053)	0.0455 (0.0041)
$\hat{\sigma}_3$	0.0668 (0.0044)	0.0407 (0.0038)	0.0372 (0.0036)	0.0421 (0.0046)
$\hat{\sigma}_4$	0.1109 (0.0069)	0.0552 (0.0064)	0.0473 (0.0050)	0.0501 (0.0063)
$\hat{\sigma}_5$	0.2349 (0.0077)	0.0889 (0.0067)	0.0718 (0.0058)	0.0834 (0.0087)
$\hat{\sigma}_6$	-	0.2086 (0.0075)	0.1258 (0.0104)	0.1491 (0.0206)
$\hat{\sigma}_7$	-	0.4358 (0.0443)	0.2419 (0.0113)	0.3398 (0.0280)
$\hat{\sigma}_8$	-	-	0.6068 (0.0781)	-
$\hat{\gamma}_1$	-1.2964 (0.0831)	-2.6619 (0.1500)	-2.3222 (0.1027)	-1.9912 (0.1821)
$\hat{\gamma}_2$	-0.6855 (0.0573)	-1.3767 (0.1060)	-1.5818 (0.1309)	-1.1308 (0.0971)
$\hat{\gamma}_3$	0.1538 (0.0728)	-0.6687 (0.0640)	-0.8137 (0.0932)	-0.4395 (0.0740)
$\hat{\gamma}_4$	1.1937 (0.1098)	-0.1540 (0.0835)	-0.3227 (0.0752)	0.0629 (0.0751)
$\hat{\gamma}_5$	-	0.6188 (0.0772)	0.2897 (0.0794)	0.7316 (0.1116)
$\hat{\gamma}_6$	-	2.8623 (0.1930)	1.0760 (0.1276)	1.6041 (0.2285)
$\hat{\gamma}_7$	-	-	3.0681 (0.1747)	-
$\bar{p}_1$	0.1893	0.0280	0.0419	0.0687
$\bar{p}_2$	0.1328	0.1421	0.0792	0.1309
$\bar{p}_3$	0.2131	0.1559	0.1554	0.1724
$\bar{p}_4$	0.2543	0.1329	0.1310	0.1450
$\bar{p}_5$	0.2106	0.2002	0.1740	0.1850
$\bar{p}_6$	-	0.3240	0.1995	0.1799
$\bar{p}_7$	-	0.0170	0.2104	0.1181
$\bar{p}_8$	-	-	0.0086	-

Table 5: Estimation by explanatory mixture: numerical results.