DIW Berlin

German Institute
for Economic Research

Discussion Papers

325

Conchita D'Ambrosio,
Pietro Muliere and
Piercesare Secchi

Income Thresholds and Income Classes

Berlin, February 2003

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

DIW Berlin

German Institute
for Economic Research

Königin-Luise-Str. 5
14195 Berlin,
Germany

Phone +49-30-897 89-0
Fax +49-30-897 89-200

www.diw.de

ISSN 1619-4535

Income Thresholds and Income Classes*

Conchita D'Ambrosio†, Pietro Muliere‡and Piercesare Secchi§

November 2002

Abstract

This paper proposes a method for detecting income classes based on the change-point problem. There is an increasing demand for such a method in the literature. Computation of polarization indices requires a pre-grouping of the incomes. Similarly, indices of social exclusion and sometimes indices of income inequality require detection of thresholds. The estimation procedure is implemented using a bootstrap technique. Finally, an application of the method to EU member states and to the United States is also considered.

Keywords: income distribution, change-point, thresholds.
JEL-codes: D31

*We would like to thank David S. Johnson and Stephan Klasen for suggestions and comments, Carsten Kuchler, Ingo Sieber, and Roberta Tordi for technical assistance, and the participants at the 27^{th} General Conference of the IARIW, the 4^{th} Annual Meeting of LivinTaX, the Brown Bag seminars at DIW Berlin, and at the International Workshop organized for Università Bocconi Centennial.

†Istituto di Economia Politica, Università Bocconi, Italy and DIW Berlin, Germany. E-mail: conchita.dambrosio@uni-bocconi.it

‡Istituto di Metodi Quantitativi, Università Bocconi, Milano, Italy. E-mail: pietro.muliere@uni-bocconi.it. Muliere acknowledges financial support from CNR (Progetto Coordinato - Agenzia 2000: il Ruolo Economico dell'Informazione).

§Dipartimento di Matematica, Politecnico di Milano, Milano, Italy. E-mail: secchi@mate.polimi.it

1 Introduction

In the recent income distribution literature there is an increasing demand for a method capable of detecting groups that constitute the underlying distribution. The indices of polarization (Esteban and Ray, 1994), for example, require a pre-grouping of the incomes in order to be computed. In the same way, the detection of a threshold is at the basis of the indices of social exclusion (Tsakloglou and Papadopoulos, 2001), and of the inequality measures (Castagnoli and Muliere, 1990, Mosler and Muliere, 1996 and 1998) which are consistent with weaker versions of the Pigou-Dalton principle of transfers that restrict the class of admissable transfers and relate those to threshold incomes which separate classes of richer from poorer people.

The present paper builds directly on the latter. We propose a method for determining endogenously the threshold incomes as those where the population under analysis comes to be distributed in a different way. An example might help to clarify our basic idea. Let us imagine that the underlying mechanism that generates incomes (and their distribution) all of a sudden is subject to an abrupt change. For simplicity, let us assume that each income is generated according to a Pareto distribution but, whenever there is an abrupt change in the generating mechanism, the parameters of the distribution vary in such a way that different classes of incomes come to be generated by different Pareto distributions. We are interested in estimating the income thresholds separating these classes, each composed of incomes generated by the same Pareto. The number of income thresholds will indicate the degree of heterogeneity in the total income distribution. The lower thresholds could be viewed as poverty lines in that they divide the poor from the better off. The poverty line, and each income thershold more in general, will be hence determined endogenously based on the *change-point problem*.

The estimation procedure is implemented using a bootstrap technique. The empirical analysis is carried on EU member states and on the United States. The datasets used are respectively the European Community Household Panel and the Current Population Survey. For the EU member states we apply the 1984 European Council Decision, hence the unit of analysis is the individual and the definition of income is equivalent income obtained by applying the modified OECD equivalence scale. For the US, on the contrary, we follow the official procedure for measuring poverty of the Census Bureau, hence we use family income and we allow the thresholds to vary depending on family size. Secondly, we compare the estimated poverty line with the one adopted at European Union level as a working definition of the 1984 Council Decision, namely 60 percent of the median of the distribution, and for the US we compare our results with the official poverty line. Results show that there is enough heterogeneity in the data to estimate income classes for all the analyzed countries. The dimension of the various classes, and its changes over time depend dramatically on the underlying distributional

assumption made in order to implement the method.

The rest of the paper is organized as follows: Sections 2 and 3 contain respectively the method proposed to estimate the threshold incomes, and its application to the EU member states and to the US. In Section 3, we suggest the derivation of an endogenous equivalence scale for the US. Conclusions are drawn in Section 4.

2 The proposed method

Let $X = (X_1, X_2, ..., X_n)$ be a sequence of random variables (henceforth r.v.'s) indexed by time and let $x = (x_1, x_2, ..., x_n)$ be a realization of X. Suppose that, at unknown time instant τ, the underlying mechanism that regulates the distribution of the variables of the sequence X is subject to an abrupt change. We are interested in the detection of τ; this is known, in the statistical literature, as the *change-point problem*. Generally defined, change-point methods deal with sets of sequentially ordered observations (as in time) and undertake to determine whether the fundamental mechanism generating the observations has changed along the sequence. Often observations are assumed to be independent and with the same parametric distribution, the change involving one or more parameters defining it. There is a wide range of applications of these problems. The traditional example is that of quality control, in which a sequence of measurements from a production process is analyzed for a change in, say, the thickness of a manufactured part. The time τ is said *change-point* (henceforth CP) of the sequence if the random variables $X_{\tau+1}, X_{\tau+2}, ..., X_n$ are distributed somehow differently from $X_1, ..., X_\tau$.

The easiest formulation of a CP problem is the following. Consider two statistical models M_0 and M_1: according to M_0 the random variables of the sequence X are independent with the same distribution $F(\cdot|\delta_0)$, whereas M_1 says that the same variables are independent and identically distributed with distribution $F(\cdot|\delta_1)$. Distributions $F(\cdot|\delta_0)$ and $F(\cdot|\delta_1)$ are different and their expression might depend on the values of the parameters δ_0 and δ_1 respectively. Finally let $\tau \in \{0, 1, ..., n\}$ and assume that, for all $x_1, ..., x_n \in \mathbf{R}$,

$$P\left(X_1 \leq x_1, ..., X_n \leq x_n | \tau, \delta_0, \delta_1\right) =$$

$$= \begin{cases} \prod_{i=1}^{n} F\left(x_i|\delta_0\right) & \text{for } \tau = n; \\ \prod_{i=1}^{\tau} F\left(x_i|\delta_0\right) \prod_{i=\tau+1}^{n} F\left(x_i|\delta_1\right) & \text{for } \tau = 1, ..., n-1; \\ \prod_{i=1}^{n} F\left(x_i|\delta_1\right) & \text{for } \tau = 0. \end{cases} \qquad (1)$$

This means that the r.v.'s of the sequence X are independent; they are identically distributed according to $F\left(\cdot|\delta_0\right)$ up to time τ, while they are identically distributed according to $F\left(\cdot|\delta_1\right)$ from time $\tau + 1$ on; τ is the

change-point. In other words, we assume the statistical model M_0 for the first τ random variables $(X_1, X_2, ..., X_\tau)$ of X, and the model M_1 for the remaining $(n-\tau)$ random variables $(X_{\tau+1}, X_{\tau+2}, ..., X_n)$. When the distribution functions $F(\cdot|\delta_0)$ and $F(\cdot|\delta_1)$ have densities $f(\cdot|\delta_0)$ and $f(\cdot|\delta_1)$ respectively, the joint density of X computed in $x_1, ..., x_n \in \mathbf{R}$ becomes:

$$p(x_1, ..., x_n|\tau, \delta_0, \delta_1) = \prod_{i=1}^{\tau} f(x_i|\delta_0) \prod_{i=\tau+1}^{n} f(x_i|\delta_1), \tag{2}$$

where the products $\prod_1^0 \cdot$ and $\prod_{n+1}^n \cdot$ are conventionally set equal to 1.

In most applications τ, and possibly the parameters δ_0 and δ_1, are unknown and must be estimated based on a realization $x = (x_1, ..., x_n)$ of X. For this purpose, it is common to look for the maximum likelihood estimates, $\widehat{\tau}(x), \widehat{\delta_0}(x)$ and $\widehat{\delta_1}(x)$, i.e. for those values of τ, δ_0 and δ_1, functions of x, that maximize the log-likelihood function:

$$l(\tau, \delta_0, \delta_1|x_1, ..., x_n) = \log p(x_1, ..., x_n|\tau, \delta_0, \delta_1). \tag{3}$$

It could be generally quite difficult to maximize (3) and it could be even more cumbersome to determine the distribution of the statistic $\widehat{\tau}$ (Hinkley, 1970). The literature on CP problems is enormous. For a survey, see Shaban (1980), Krishnaiah and Miao (1988), Muliere and Scarsini (1993).

In this paper, the CP problem is applied to the income distribution. The sequentially ordered observations are the ordered incomes of the population under analysis, and the CP is the income level (threshold) where the distribution changes. The underlying idea is that individuals that belong to different income classes come from different populations, i.e. that incomes of the poor are generated by one mechanism, and incomes of the rich by another one. There are various underlying factors of the economy that could give rise to these differences: the minimum wage, the minimum income guarantee, different characteristics of the individuals in terms of educational level attained, social background, occupation, sector of employment, different characteristics of the households in terms of number of children, number of employed members, sex of the head of the household. These very important issues will be the focus of further research. For the moment, we deal with the detection of CP's.

Let then $X = (X_1, ..., X_n)$ be the (unordered) sequence of the n random incomes of a population, and $x = (x_1, ..., x_n)$ be a realization of X. Indicate with θ a threshold income and assume that incomes $X_i \leq \theta$ are generated by the model M_0, while incomes $X_i > \theta$ are generated by the model M_1. According to M_0 the incomes X_i are i.i.d. with absolutely continuous distribution with density $f(\cdot|\delta_0)$, whereas according to M_1 the X_i's are i.i.d. with absolutely continuous distribution with density $f(\cdot|\delta_1)$; for any choice of δ_0 and δ_1, we assume the existence of real numbers $0 \leq a < b < c \leq \infty$ such

that the support of $f(\cdot|\delta_0)$ is contained in $[a,b]$ while the support of $f(\cdot|\delta_1)$ is contained in (b,c). The threshold θ is generally unknown and must be estimated based on x. For this purpose, indicate with X_τ the largest income, among the incomes $X_1,...,X_n$, that is smaller than or equal to θ; then τ is a change point for the sequence of ordered incomes $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Based on a realization x of X, we estimate τ with its maximum likelihood estimate $\widehat{\tau}(x)$, and θ with the $\widehat{\tau}(x)$-th smallest income among the incomes in x. In order to obtain $\widehat{\tau}(x)$, we need to write the log-likelihood of τ, δ_0, δ_1 relative to the sequence of ordered incomes. Indeed, letting $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ be the order statistic of X, according to (2), its joint density is:

$$p'\left(x_{(1)},...,x_{(n)}|\tau,\delta_0,\delta_1\right) = \tau! \prod_{i=1}^{\tau} f\left(x_{(i)}|\delta_0\right) (n-\tau)! \prod_{i=\tau+1}^{n} f\left(x_{(i)}|\delta_1\right), \quad (4)$$

for all $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)} \in \mathbf{R}$. Correspondingly, the log-likelihood of τ, δ_0 and δ_1 relative to a realization $x = (x_1,...,x_n)$ of X becomes:

$$l(\tau,\delta_0,\delta_1|x_1,...,x_n) = \log p'\left(x_{(1)},x_{(n)},...,x_{(n)}|\tau,\delta_0,\delta_1\right), \quad (5)$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ is the ordered sequence of the incomes in x.

Equivalently, we may focus on the proportion q of incomes that are generated by the model M_0. For a sequence of incomes $X = (X_1,...,X_n)$ of size n, the change point corresponding to a given $q \in [0,1]$ is then $\tau = [qn]$, with $[qn]$ indicating the largest integer smaller than qn. The log-likelihood of q, δ_0 and δ_1 corresponding to a realization x of X is:

$$\begin{aligned} &l'(q,\delta_0,\delta_1|x_1,...,x_n) = \\ &\quad = \quad l([qn],\delta_0,\delta_1|x_1,...,x_n) = \log p'\left(x_{(1)},x_{(n)},...,x_{(n)}|[qn],\delta_0,\delta_1\right), \end{aligned} \quad (6)$$

where the density p' and the log-likelihood l are defined in (4) and (5) respectively. In order to find an estimate of θ based on x we then proceed as follows: by maximizing (6) we find the maximum likelihood estimates $\widehat{q}(x), \widehat{\delta_0}(x)$ and $\widehat{\delta_1}(x)$. Then $\widehat{\tau}(x) = [n\widehat{q}(x)]$ and an estimate of the threshold θ is $\widehat{\theta}(x) = x_{(\widehat{\tau}(x))}$, the $\widehat{\tau}(x)$-th income among the ordered incomes.

The method is implemented using a bootstrap procedure that has the advantage of providing estimates for the distribution of the statistic $\widehat{q}$.[1] The procedure runs as follows:

(i) Generate a random sample of dimension N from the empirical distribution of the observed incomes $x = (x_1,...,x_n)$ and sort the sample in a nondecreasing order.

[1] The bootstrap procedure will hence allow to have all the information needed, for example, to compute the probability that the estimator $\widehat{q}$ falls in a given interval, or is smaller than a given value.

(ii) Obtain an estimate of q, δ_0 and δ_1 by maximizing the log-likelihood (6) corresponding to the sample obtained in (i). The maximization is performed stochastically along the following steps:

(ii.a) A starting value is randomly selected, and the function (6) is maximized with the use of the Powell's algorithm;

(ii.b) Step (ii.a) is replicated K times;

(ii.c) Of the K vectors of estimates for q, δ_0 and δ_1 obtained at step (ii.b) only that corresponding to the largest value for (6) is retained; let $\widehat{q}_1$ be the value for the estimate of q appearing in this vector.

(iii) Repeat steps (i) and (ii) for S times, thereby generating the estimates $\widehat{q}_1, \widehat{q}_2, ..., \widehat{q}_S$.

(iv) The empirical distribution function $\widehat{Q}$ of $(\widehat{q}_1, \widehat{q}_2, ..., \widehat{q}_S)$ is considered to be an estimate of the distribution Q of the statistic $\widehat{q}$.

(v) Finally q is estimated with the median $\text{Me}(\widehat{Q})$ of $\widehat{Q}$. This value estimates the proportion of incomes $x_1, ..., x_n$ generated by M_0.

(vi) The threshold θ is estimated with the $[n\text{Me}(\widehat{Q})]$-th income among the ordered observed incomes $x = (x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)})$.

Note that, given a sequence $x = (x_1, ..., x_n)$ of observed incomes, the procedures requires the specification of the parameters N, K, S.[2]

The analysis of income distribution carried in this paper implement two versions of the method described above. The method is, though, very general and can be implemented with any distribution. The results that we provide below should be taken as a practical example.

2.1 Example 1: the Pareto-Pareto case

In our first example, we assume that $f(\cdot|\delta_0)$ is the density of a Pareto distribution with parameters (α_0, λ_0) truncated at $\alpha_1 > \alpha_0$, while $f(\cdot|\delta_1)$ is the density of a Pareto distribution with parameters (α_1, λ_1). Hence the

[2]Once the first income threshold is estimated, the second is obtained by considering only incomes above the threshold and applying to them the same algorithm. And so forth for all the thresholds.

log-likelihood (6) becomes:

$$
\begin{aligned}
& l'(q,(\alpha_0,\lambda_0),(\alpha_1,\lambda_1)|x_1,...,x_n) = \\
& = \log\left[\tau![1-(\frac{\alpha_0}{\alpha_1})^{\lambda_0}]^{-\tau}[\prod_{i=1}^{\tau}\frac{\lambda_0}{\alpha_0}(\frac{\alpha_0}{x_{(i)}})^{\lambda_0+1}I[\alpha_0<x_{(i)}<\alpha_1]] \cdot\right. \\
& \qquad \left.\cdot(n-\tau)![\prod_{i=\tau+1}^{n}\frac{\lambda_1}{\alpha_1}(\frac{\alpha_1}{x_{(i)}})^{\lambda_1+1}I[x_{(i)}>\alpha_1]]\right].
\end{aligned}
$$

In actual fact, we obtain the estimates $\hat{q}(x), \hat{\lambda}_0(x)$ and $\hat{\lambda}_1(x)$ by maximizing, through the procedure described in steps (i)-(vi), the log-likelihood:[3]

$$
l'\left(q,\left(x_{(1)},\lambda_0\right),\left(x_{([qn])},\lambda_1\right)|x_1,...,x_n\right). \tag{7}
$$

In order to evaluate the performance of the bootstrap procedure described above, we run for 100 times the following experiment. We simulated a population of $n = 8000$ incomes; $q \cdot 100$ per cent of the incomes were generated from a Pareto P_0 with parameters (α_0, λ_0), and the remaining $(1-q)\cdot 100$ per cent from a Pareto P_1 with parameters (α_1, λ_1). The parameters q, α_0, λ_0, λ_1 where independently generated. In particular, q was generated from a Uniform distribution over $[0, 0.15]$, both λ_0 and λ_1 from a Uniform distribution over $[1, 10]$, α_0 from a Uniform distribution over $[1000, 5000]$. The parameter α_1 was set equal to the maximum income among the $q \cdot 8000$ incomes generated by P_0. For each of the 100 replicates of the experiment, we applied the bootstrap procedure (i)-(vi) to the log likelihood (7) (run with N=4000, K=50 and S=100) and we estimated q as the median and as the mean of the empirical distribution $\widehat{Q}$ computed at step (iv). Hence, we compared the 100 estimates of q thus obtained with the corresponding true values: this is shown in Figures 1 and 2 where we plot the true value of q against its estimate obtained as the median of $\widehat{Q}$ and against the estimate obtained as the mean of $\widehat{Q}$ respectively. Qualitatively, the median performs better and this justifies the choice made at step (v) of the procedure.

2.2 Example 2: the LogNormal-Pareto case

The second example assumes that $f(\cdot|\delta_0)$ is the density of a LogNormal distribution with parameters (μ_0, σ_0) and truncated at $\alpha_1 > 0$, while $f(\cdot|\delta_1)$ is the density of a Pareto distribution with parameters (α_1, λ_1). Hence the

[3] That is we set $x_{(1)}$ and $x_{([qn])}$ as estimates of α_0 and $\alpha_1$respectively.

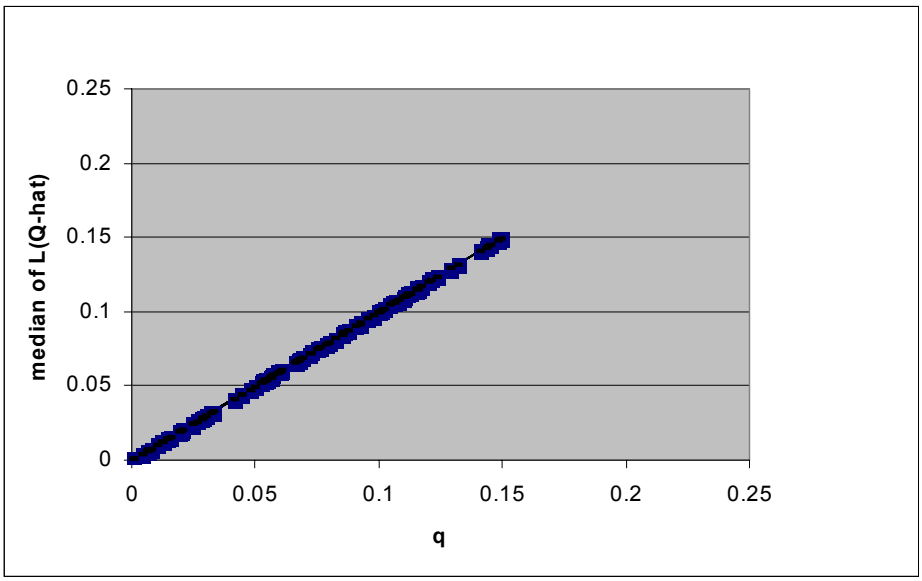

Figure 1: Results of the Pareto-Pareto simulation. Plot of "true" q against estimated q as the median of the empirical distribution, $\widehat{Q}$.

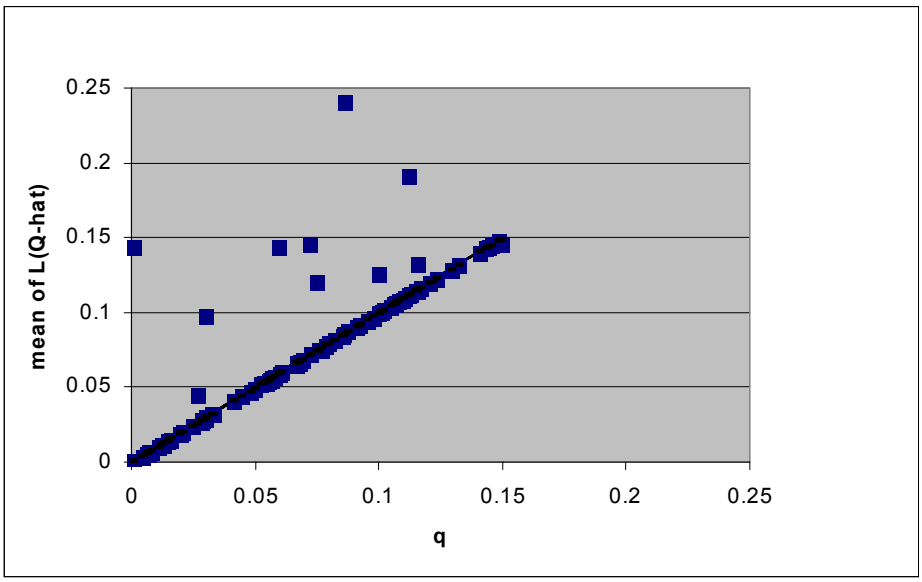

Figure 2: Results of the Pareto-Pareto simulation. Plot of "true" q against estimated q as the mean of the empirical distribution, $\widehat{Q}$.

log-likelihood (6) becomes:

$$l'(q,(\mu_0,\sigma_0),(\alpha_1,\lambda_1)|x_1,...,x_n) =$$
$$= \log\left[\tau!\prod_{i=1}^{\tau}\frac{1}{x_{(i)}\Phi(\sigma_0^{-1}(\log\alpha_1-\mu_0))\sqrt{2\pi\sigma_0^2}}\exp[-\frac{1}{2\sigma_0^2}(\log x_{(i)}-\mu_0)^2]\cdot\right.$$
$$\left.\cdot(n-\tau)![\prod_{i=\tau+1}^{n}\frac{\lambda_1}{\alpha_1}(\frac{\alpha_1}{x_{(i)}})^{\lambda_1+1}I[x_{(i)}>\alpha_1]]\right],$$

where Φ represents the cumulative distribution function of the standard Normal. In actual fact, we obtain the estimates $\hat{q}(x), \hat{\mu}_0(x), \hat{\sigma}_0(x)$ and $\hat{\lambda}_1(x)$ by maximizing, through the procedure described in steps (i)-(vi), the log-likelihood:

$$l'(q,(\mu_0,\sigma_0),(x_{([qn])},\lambda_1)|x_1,...,x_n). \tag{8}$$

In order to evaluate the performance of the bootstrap procedure described above, we run for 100 times the following experiment. We simulated a population of $n = 8000$ incomes; $q\cdot 100$ per cent of the incomes were generated from a LogNormal $LogN_0$ with parameters (μ_0, σ_0), and the remaining $(1-q)\cdot 100$ per cent from a Pareto P_1 with parameters (α_1, λ_1). The parameters q, μ_0, σ_0, λ_1 where independently randomly generated. In particular, q was generated from a Uniform distribution over $[0, 0.20]$, both σ_0 and λ_1 from a Uniform distribution over $[1, 10]$, μ_0 from a Uniform distribution over $[10, 100]$. The parameter α_1 was set equal to the maximum income among the $q \cdot 8000$ incomes generated by $LogN_0$. For each of the 100 replicates of the experiment, we applied the bootstrap procedure (i)-(vi) to the log likelihood (8) (run with N=4000, K=50 and S=100) and we estimated q as the median and as the mean of the empirical distribution $\widehat{Q}$ computed at step (iv).[4] Hence, we compared the 100 estimates of q thus obtained with the corresponding true values: this is shown in Figures 4 and 5 where we plot the true value of q against its estimate obtained as the median of $\widehat{Q}$ and against the estimate obtained as the mean of $\widehat{Q}$ respectively. As for the previous model, the median performs qualitatively better and this confirms the choice made at step (v) of the procedure.

3 The application to the EU member states and the USA

In the application that we are here proposing, we extracted a sample of approximately half of the income observations in each EU member state (N=$\frac{1}{2}$ number of observations) while the dimension of the US sample, more

[4] The numerical implementation of the second model was slower and more complex.

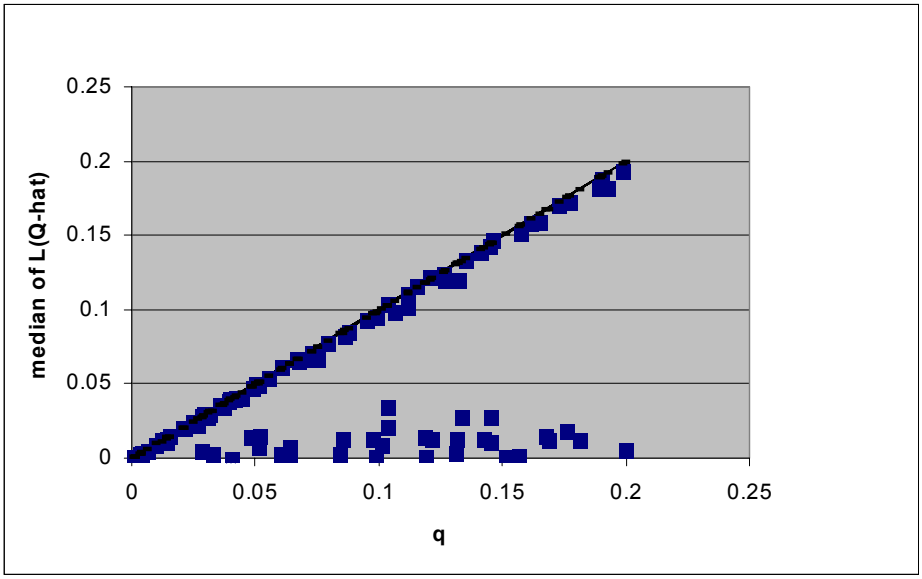

Figure 3: Results of the LogNormal-Pareto simulation. Plot of "true" q against estimated q as the median of the empirical distribution, $\widehat{Q}$.

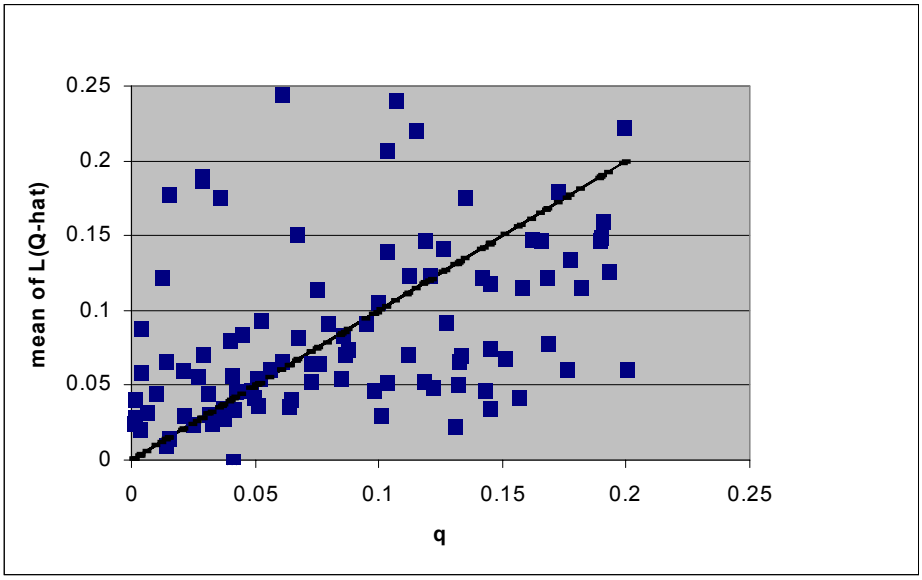

Figure 4: Results of the LogNormal-Pareto simulation. Plot of "true" q against estimated q as the mean of the empirical distribution, $\widehat{Q}$.

than 121,000 observations, combined with the power of the computer we had available, forced us to extract a much lower proportion. The value of K has been set to 50, while S to 100. Results are provided for the models illustrated in the two examples of the previous section. According to the first, we assume that each income group is distributed according to a Pareto, owing to its traditional application in representing the income distributions; in the second, instead, the first income group is distributed according to a LogNormal while the others according to various Pareto distributions.

The paper uses data from the European Community Household Panel (henceforth ECHP) which is the only dataset that provides comparable data on EU member states in the early 1990s. The Panel was conducted at the European national level under the supervision of Eurostat. The information was collected by means of questionnaires. We restrict the analysis to the first four waves of ECHP, which cover the period 1994-1997. For the United States, we use the Current Population Survey of the Census Bureau of the same years. The concept of income that we use is "net income from all sources during the previous year" adjusted with the modified OECD equivalence scale for the EU member states. For the EU member states we apply the 1984 European Council Decision, hence the unit of analysis is the individual and the definition of income is equivalent income obtained by applying the modified OECD equivalence scale. For the US, on the other hand, we follow the official procedure of measuring poverty of the Census Bureau, hence we use family income and we allow the thresholds to vary depending on family size. Only positive incomes have been used in the analysis.

3.1 The EU member states

The results for the EU member states are contained in Tables 1 and 2 for the years 1994 and 1997 respectively. In the columns of the tables are indicated the proportion of the population (number above), and the corresponding income level (number below), that belong to each income class. The first three columns show the results of the model proposed in the first example, the Pareto-Pareto case, where we assumed that all the groups are distributed according to various Pareto. Column four and five, on the other hand, contain the results of the second example, The LogNormal-Pareto case, that is based on the assumption that the first income class is distributed according to a LogNormal while all the other classes continue to be distributed according to a Pareto. The minimum and the maximum income level are reported in column six and seven respectively. The last column of Tables 1 and 2 reports the poverty rate (number above) and the poverty line (number below), where the latter is the one adopted at European Union level as a working definition of the 1984 Council Decision, namely 60 percent of the median of the distribution.

According to the Pareto-Pareto case, in both years analyzed, Ireland

and Greece lie at the opposite side of the range of the estimated values of the poverty line, the latter being for Ireland 6.5% and 4.3% in 1994 and 1997 respectively, while the corresponding values for Greece are 17.4% in both years. In addition, two digits values are found in Portugal, UK (BHPS), Germany, Italy, and Greece in 1994, while in 1997 the proportion of the population that belong to the first income class increased since two digits values are registered in all the countries but Ireland, France, Germany (SOEP) and Portugal. The values of the first income threshold that we estimate are always lower than the EU poverty line, but in Denmark, Austria and Finland in 1997. Furthermore, the rank of the countries according to the poverty rates differs from the one obtained by looking at the proportion of the population below the first threshold.

The results of the second example are dramatically different. The first threshold is always above the EU poverty line. A clear grouping emerges between the countries in 1994 while the values are more spread in 1997. In the first wave, Luxembourg is the unique country where less that 30% of the population belongs to the first income class; the next group is composed of Denmark, the Netherlands, Belgium and France with values close to 23%; Germany, Greece and Spain are in the third group showing approximately 29% of the population below the threshold; to the last group belong the UK, Italy and Ireland where the first 50% or above of the whole population is distributed according to a LogNormal. In 1997, the percentage of the population belonging to the first income class shrinks in all the EU member states, but in France and Greece. Denmark and the UK are the countries showing the extreme values, 12% and 47% respectively.

We then focused on the incomes above the first threshold and asked the same question: is it possible to find an income threshold such that the observations below and the observations above are distributed differently? Results for the second threshold are contained in column two,[5] for the first model and in column five for the second model, while values estimated for the third threshold are provided only for the first model[6] and are in column three. In almost all the EU member states the majority of the population belongs to the second income class, as expected. The latter can be hence interpreted as the middle class in the income distribution. When we add to the population below the first threshold the estimates for those below the second we reach an average for the member states of 69% in 1994 and of 67% in 1997 for the first model,[7] while the values for the second model are

[5] The values of the proportion of the population for the second/third threshold include all the population below the threshold. In other words, the reported value for the second/third threshold indicates all the income levels in the population under analysis that are below that value and not the incomes beteween two thresholds.

[6] For the moment, we did not estimate the third thershold for the second model since, according to it, it is often the case that above 77% of the population belong to the first two income classes leaving very few observations to proceed further.

[7] Denmark, Belgium and Portugal behave differently in 1994 since less than 50% of the

respectively 79% in 1994 and 72% in 1997. If we compare these values with the average of the population that belongs to the first income class according to the first model, 11% in both years, we could not confirm for the EU as a whole the well known phenomenon of the "shrinkage of the middle class" that has characterized the last decade.[8] In addition, France, Ireland, and the Southern European countries (Greece, Portugal, and Spain) experience an increase of the middle class.

According to the second example, in the majority of the member states there is a shift of density from the first two income classes upwards, but in Greece where there is an increase in both; the Netherlands, Belgium, and Italy where there is a decrease in the first class and an increase in the middle class; and France that experience the phenomenon opposite to the latter.

3.2 The US and the endogenous equivalence scale

The results of the method described above for the United States are contained in Table 3 and 4. For the US we further compare the equivalence scale that we obtain (columns three for the first model, and six for the second model) with the one implicit in the official poverty line (the last column). The US official poverty line was developed in the Social Security Administration in 1963/64 by Mollie Orshanky and since then it has been updated only for inflation. The line differs by the number of family members.

According to the Pareto-Pareto case, the income level of the first threshold (column one) that we estimated with the CP method for families with two members is surprisingly similar, in both years, to the official poverty line. On the contrary in both years, the estimated value is lower for one, six, and nine plus members' families, and higher for three, and four members' families. For the other types of families the results depend on the year of analysis. The equivalence scale that we obtain is, as expected from the previous results, very different form the one implicit in the official poverty line. The former increases more sharply up to families with four (in 1994)/five (in 1997) members, the value being almost twice the official, and it drops afterwards with a strong increase in 1997 for families with seven and eight members. It is worthwhile noticing the differences between the distributions of income of families with 7 or more members both in the medians (column seven) and in the maximums (column nine). The CP's are estimated on the basis of the income data while the official poverty line is not, hence these sharp differences in the distributions have a great influence on the former and none on the latter. For the same reason, the EU poverty line is not reasonably applicable to the US, if the distinction between families has

population is below the second threshold, while in 1997, slightely more than 50% is below the second threshold in the last two countries.

[8] And an increase of the "poor" as well in all the member states but Germany, France, Ireland, and Portugal.

to be maintained. It is indeed the case that 60% of the median decreases sharply from families with 4 members to more numerous families, more than the presence of economies of size justifies, and it is lower than the official poverty line for families with 7 or more members.

The results of the application of the LogNormal-Pareto case are presented in columns four to six of Table 3 and 4. In 1994 more than 50% of the population is distributed according to a LogNormal distribution and the estimated poverty line is generally larger (up to two and three times larger) than the corresponding official poverty line, but in the case of families with 9 or more members in 1997. Focusing exclusively on the incomes above the first threshold, we were able to estimate the second threshold. Results show that more than 87% of the population belongs to the first two income classes in 1994, while in 1997 there is a wider range of values, from 60% in the case of families with 8 members to 99% in that of families with 7 members.

4 Conclusions

The change-point problem, borrowed from the statistical literature, is here applied to the income distribution. We have shown that it provides a powerful method to estimate income thresholds and income classes endogenously based on the degree of heterogeneity of the population under analysis. The underlying assumption is that the distribution of income of each group differs. The proposed method is implemented using a bootstrap procedure that has the advantage of providing estimates for the distribution of the estimators of the threshold. We evaluated the performance of the bootstrap procedure by simulations, and the results of the latter justify the choice we made in estimating the thresholds. A practical application of the method to the EU member states and the US is contained in the two examples presented. Results show that there is enough heterogeneity in the data to estimate income classes for all the analyzed countries. The dimension of the various classes, and its changes over time depend dramatically on the underlying distributional assumption made in order to implement the method.

References

[1] Castagnoli, E. and P. Muliere (1990): "A Note on Inequality Maesures and the Pigou-Dalton Principle of Transferse," in C. Dagum and M. Zenga, eds., *Income and Wealth Distribution, Inequality and Poverty,* 171-182, Springer Verlag, Berlin.

[2] Esteban, J.M., and D. Ray (1994): "On the Measurement of Polarization," *Econometrica*, 62, 4, 819-851.

[3] Hinkley, D.V. (1970): "Inference about the Change Point in a Sequence of Random Variables," *Journal of Economic Theory*, 2, 244-263.

[4] Krishaiah, P. R. and B.Q. Miao (1988): "Review about estimation of change points," in P.R. Krishnaiah and C.P. Rao (eds) *Handbook of Statistics,* 7, 375-402, Elsevier, New York.

[5] Mosler, K. and P. Muliere (1996): "Inequality Indices and the Star-shaped Principle of Transfers," *Statistical Papers*, 37, 343-364.

[6] Mosler, K. and P. Muliere (1998): "Welfare Means and Equalizing Transfers," *Metron*, LVI n.3-4, 1-52.

[7] Muliere, P. and M. Scarsini (1993): "Some Aspects of Change-Point Problems," in R.E. Barlow, C.A. Clarotti and F. Spizzichino, eds., *Reliability and Decision Making,* 273-285, Chapman and Hall, London.

[8] Shaban, S.A. (1980): "Change Point Problem and a Two Phase Regression: an Annotated Bibliography," *International Statistical Review*, 48, 83-93.

[9] Tsakloglou P. and F. Papadopoulos (2001) "Identifying population groups at high risk of social exclusion", IZA Discussion Paper No 392 (forthcoming in R. Muffels, P. Tsakloglou and D. Mayes (eds), *Social exclusion in European welfare states*, Edward Elgar, Cheltenham).

[10] Zacks, S. (1983): "Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation," in M.H. Rivzi et al. (eds), *Recent Advances in Statistics*, 245-269, Academic Press, New York.

Tables

	PaPa (third, second) and PaPa (first)			PaPa (second) and LnPa (first)				
WAVE 1 ECHP	**First Threshold**	**Second Threshold**	**Third Threshold**	**First Threshold**	**Second Threshold**	**min**	**Max**	**Eurostat 60% median**
Germany	13.80 14760.00	62.29 30952.38	89.69 49725.33	28.48 20174.50	79.71 39739.50	46.11	592656.10	16.31 15867.75
Germany SOEP	14.50 15459.00	74.19 38118.33	79.79 41511.60	30.13 21423.55	75.05 38640.37	81.50	331324.00	16.76 16464.33
Denmark	6.75 65800.00	42.25 113148.60	85.14 171900.00	22.72 90131.07	56.44 127338.00	504.00	1518000.00	10.42 72514.17
Netherlands	9.52 14376.00	81.81 38271.33	88.55 43942.67	23.13 17956.19	80.67 37596.67	128.67	320000.00	10.13 14692.73
Belgium	7.52 246000.00	48.43 525748.60	96.25 1319291.00	23.79 367060.00	60.79 600000.00	664.00	18700000.00	16.48 320282.81
Luxembourg	8.10 372000.00	79.55 1174000.00	na	18.96 460258.50	85.66 1349333.00	2343.33	8513574.00	15.63 432000.00
France	8.23 37276.00	57.04 85777.78	93.96 176883.30	22.62 53073.21	77.01 113450.60	24.67	2382957.00	16.30 47041.43
UK	8.54 2903.33	86.17 13585.00	97.36 23927.14	51.74 7346.50	88.05 14200.00	7.00	383150.00	22.36 4278.86
UK BHPS	12.75 3669.00	77.66 11620.00	94.82 18408.00	50.24 7792.00	84.37 13273.33	50.00	75402.00	21.12 4654.00
Ireland	6.46 2757.50	76.05 8853.50	96.18 15797.60	56.57 6365.48	94.12 14196.19	28.67	448938.00	18.18 3384.00
Italy	13.95 6996.00	76.76 21113.33	82.04 22900.00	49.01 13734.50	75.69 20741.11	33.33	125485.00	20.59 8388.00
Greece	17.37 666666.70	76.83 2080586.00	83.69 2366667.00	29.91 929354.80	78.45 2139916.00	1666.67	40000000.00	23.03 796800.00
Spain	9.33 416000.00	82.10 1670596.00	91.08 2185976.00	28.60 679143.00	86.44 1858971.00	20.50	13200000.00	20.15 572445.30
Portugal	11.35 314285.70	38.92 669196.00	93.43 2247329.00	35.79 629391.30	78.89 1304640.00	846.50	12300000.00	22.78 469989.83

Table 1: Estimated income thresholds and official EU poverty line for EU member states. For each country, the first number indicates the proportion of the population while the second the corresponding income level, expressed in national currency. 1994.

	PaPa (third, second) and PaPa (first)			PaPa (second) and LnPa (first)				
WAVE 4 ECHP	**First Threshold**	**Second Threshold**	**Third Threshold**	**First Threshold**	**Second Threshold**	**min**	**Max**	**Eurostat 60% median**
Germany SOEP	7.81 14503.33	43.76 26986.67	66.23 34610.50	28.16 22420.00	68.88 35590.67	200.00	379441.00	14.24 17264.80
Denmark	10.45 82293.34	38.97 121200.00	43.61 122310.00	11.94 84470.59	37.85 119841.10	2000.00	1038879.00	8.00 76755.90
Netherlands	12.01 15309.33	79.44 38504.00	90.84 48444.40	19.74 17928.00	81.73 39991.33	1.00	404968.00	12.99 15787.2
Belgium	13.00 311665.00	51.40 565661.60	53.23 568800.00	22.62 384000.00	62.24 640000.00	564.00	33900000.00	14.92 332217.38
France	6.24 38338.57	78.39 127718.90	85.66 145828.00	30.31 65192.73	79.01 128576.00	74.00	1634588.00	17.41 51098.50
UK BHPS	15.31 4490.33	62.40 10893.33	76.03 13356.00	46.93 8660.00	75.56 13188.50	32.00	187267.30	22.40 5515.20
Ireland	4.28 3202.22	85.01 13217.50	94.96 17740.00	38.82 6052.22	87.10 13688.50	317.00	433160.80	20.01 4354.20
Italy	16.75 8608.70	76.77 23047.92	92.71 32913.04	37.26 12738.10	77.56 15500.00	60.40	142732.80	19.22 9300.00
Greece	17.40 933333.30	81.85 3114867.00	92.04 4166103.00	31.10 1314286.00	87.10 3500000.00	9756.52	31900000.00	22.35 1078571.40
Spain	11.54 494347.80	86.53 2121425.00	95.23 3088000.00	24.51 740000.00	84.70 2073846.00	166.67	12700000.00	19.08 638608.73
Portugal	9.60 403633.30	52.77 996080.00	92.58 2651429.00	33.99 728000.00	53.45 1006808.00	1099.00	12200000.00	23.49 568800.00
Austria	13.86 118613.90	69.26 238554.10	80.26 278294.70	17.61 127539.00	68.66 238554.10	767.00	1776380.00	13.23 116251.33
Finland	10.75 46025.00	72.67 91485.00	91.74 126071.30	16.16 49955.00	72.38 91215.50	614.00	793908.70	8.51 43252.8
Sweden	10.19 70562.50	65.10 140952.4	91.55 203120.00	20.20 122087.00	77.95 161500.00	100.00	1908267.00	11.85 73252.20

Table 2: Estimated income thresholds and official EU poverty line for EU member states. For each country, the first number indicates the proportion of the population while the second the corresponding income level, expressed in national currency. 1997.

	PaPa (second) and PaPa (first)			PaPa (second) and LnPa (first)							
USA CPS 1994	**First Threshold**	**Second Threshold**	**Eq. Scale**	**First Threshold**	**Second Threshold**	**Eq. Scale**	**Median**	**min**	**Max**	**Official Census**	**Eq. Scale (official)**
1 member	8.88 4805	77.37 30000	1.00	62.05 21052	86.80 39000	1.00	16025	1	319400	7547	1.00
2 members	12.22 10005	84.30 64400	2.08	63.10 39160	96.24 108526	1.86	30000	1	599994	9661	1.28
3 members	18.88 15113	89.70 87556	3.15	52.37 39200	94.70 105200	1.86	37500	2	474273	11821	1.57
4 members	18.06 20000	91.38 101800	4.16	50.15 44840	90.28 99620	2.13	44623	12	386774	15141	2.01
5 members	24.38 23460	91.24 97970	4.88	54.19 45925	91.60 99911	2.18	42681	16	417976	17900	2.37
6 members	12.66 13000	91.41 102499	2.71	59.04 47848	89.90 97200	2.27	40000	16	256500	20235	2.68
7 members	15.03 13000	69.89 48565	2.71	57.73 36004	91.71 92840	1.71	31000	600	261753	22923	3.04
8 members	15.50 15004	87.95 90700	3.12	59.05 38665	87.45 88836	1.84	34551	800	242166	25427	3.37
9+ members	9.93 10604	91.84 105632	2.21	57.19 37285	93.01 137000	1.77	29418	600	208311	30300	4.01

Table 3: Estimated income thresholds and official poverty line for the USA. The first number in each column indicates the proportion of the population while the second the corresponding income level. 1994.

	PaPa (second) and PaPa (first)			PaPa (second) and LnPa (first)							
USA CPS 1997	**First Threshold**	**Second Threshold**	**Eq. Scale**	**First Threshold**	**Second Threshold**	**Eq. Scale**	**Median**	**min**	**Max**	**Official Census**	**Eq. Scale (official)**
1 member	7.63 5178	64.87 25000	1.00	63.60 24443	72.68 30000	1.00	18000	1	542869	8183	1.00
2 members	10.00 10020	75.08 58500	1.94	55.76 39031	81.96 68887	1.60	34372	1	723980	10473	1.28
3 members	18.51 17175	65.65 58000	3.32	55.92 48100	93.10 115574	1.97	43234	2	582883	12802	1.56
4 members	16.87 23000	58.08 58600	4.44	41.28 43760	70.32 70908	1.79	51048	1	538351	16400	2.00
5 members	12.89 17856	55.98 53701	3.45	41.18 40983	62.35 59330	1.68	47714	1	678766	19380	2.37
6 members	16.51 18000	74.04 67100	3.48	56.23 46840	84.97 82080	1.92	41172	12	688872	21886	2.67
7 members	41.43 33080	99.06 194220	6.39	47.36 36499	99.07 194220	1.49	38320	3735	321303	24802	3.03
8 members	39.82 36000	58.47 49920	6.95	41.46 39000	60.27 53049	1.60	41573	3440	320982	27593	3.37
9+ members	16.75 13710	70.75 55458	2.65	41.91 28800	94.79 131800	1.18	32251	600	216992	32566	3.98

Table 4: Estimated income thresholds and official poverty line for the USA. The first number in each column indicates the proportion of the population while the second the corresponding income level. 1997.