IRISS at CEPS/INSTEAD

An Integrated Research Infrastructure in the Socio-Economic Sciences

STABILITY OF HOUSEHOLD INCOME IN EUROPEAN COUNTRIES IN THE 1990'S

by

Nicholas T. Longford & Maria Grazia Pittau

IRISS WORKING PAPER SERIES

No. 2003-08

Please pay a visit to http://www.ceps.lu/iriss

IRISS-C/I

An Integrated Research Infrastructure in the Socio-Economic Sciences at CEPS/Instead, Luxembourg

Project supported by the European Commission and the Ministry of Culture, Higher Education and Research (Luxembourg)

RESEARCH GRANTS

for individual or collaborative research projects
(grants awarded for periods of 2-12 weeks)

What is IRISS-C/I?

IRISS-C/I is a project funded by the European Commission in its 'Access to Major Research Infrastructures' programme. IRISS-C/I funds short visits at CEPS/Instead for researchers willing to undertake collaborative and/or internationally comparative research in economics and other social sciences.

Who may apply?

We encourage applications from all interested individuals (doing non-proprietary research in a European institution) who want to carry out their research in the fields of expertise of CEPS/Instead.

What is offered by IRISS-C/I?

Free access to the IRISS-C/I research infrastructure (office, computer, library…); access to the CEPS/Instead archive of micro-data (including e.g. the ECHP); technical and scientific assistance; free accommodation and a contribution towards travel and subsistence costs.

Research areas

Survey and panel data methodology; income and poverty dynamics; gender, ethnic and social inequality; unemployment; segmentation of labour markets; education and training; social protection and redistributive policies; impact of ageing populations; intergenerational relations; regional development and structural change.

Additional information and application form

IRISS-C/I
B.P. 48 L-4501 Differdange (Luxembourg)
Email: iriss@ceps.lu
Homepage: http://www.ceps.lu/iriss

Stability of household income in European countries in the 1990's

N. T. Longford and M. G. Pittau

De Montfort University, Leicester, UK, and
University of Rome "La Sapienza", Rome, Italy

Summary

This paper explores the patterns of change in the annual household income in the countries of the European Community during the years 1994–1999. The income is modelled by mixtures of multivariate log-normal distributions, and the mixture components are interpreted as representing one subpopulation with steady increments and others with various levels of volatility. The method is extended to models for a combination of log-normal and categorical variables. An index of income stability is defined for the countries. Throughout, we emphasize graphical summaries of the results.

Keywords: *EM algorithm, Household income; Kernel estimation; Log-normal distribution; Mixture models.*

Address for correspondence: N. T. Longford, James Went Building 2–8, De Montfort University, The Gateway, Leicester LE1 9BH, England. Email: ntl@dmu.ac.uk
Tel.: +44 (116) 250 6120, FAX: +44 (116) 250 6114

1 Introduction

Income of individuals and households is an extensively studied subject in economics. After appropriate adjustments, income is variously interpreted as a measure of the productivity of labour, inflation and welfare of the society. Principal sources of information about income are surveys, although administrative registers, such as records of tax returns, have a largely untapped potential that could be realised if confidentiality issues were resolved. Income received at a time point is expended on goods, services, investments, repayment of debts, and the like, not necessarily immediately. The expenditure is affected not only by the quantity, but also the regularity and certainty of future income.

In most countries, income varies substantially, not only across individuals and households, but also over time, even after adjusting for inflation. Individuals have spells of unemployment and additional income, change jobs, and some are only occasional members of the labour force. The labour-force related behaviour of one member is influenced by the behaviour and employment status of the other members of the household. Income, or its distribution, at any time point provides only a limited insight into the economic affairs of an individual or a household. History of income and potential for future are important factors. This suggests that income can be studied more comprehensively in longitudinal surveys, in which it is recorded over an extended period of time. Then not only the levels but also the changes over time can be analysed.

The European Community Household Panel (ECHP) is a collection of surveys of European countries, with similar designs and, given the linguistic and cultural differences, as identical questionnaire instruments as is feasible. ECHP was started in 1994, and at the time of the analysis presented in this paper, data were available from six years, 1994–1999, referred to as years or *waves* 1–6 (European Commisssion, 2002a). Access to the database can be obtained by application to the website

`http://www.ceps.lu/iriss`

The elementary sampling units in the ECHP surveys are households, but information is collected from adult individuals. In longitudinal surveys of households, the common difficulties of the temporary nature of the observational units are encountered: some households are formed, others are dissolved, and numerous households are altered by one member leaving or joining. Further, nonresponse, combining attrition, unavailability and migration, contributes to the non-rectangularity of the data and raises profound issues about the nature of the incompleteness of the database and its impact on the inferences

made. In the ECHP database, we can distinguish between nonresponse of an existing household and the absence of a record for an individual or a household that was present in an earlier wave. In the case of such absence, it cannot be established whether the household has been dissolved (and no longer belongs to the target population) or failed to respond or was not contacted (is associated with missing data). The analyses conducted in Section 3 are based on complete records, as their target is the population of intact households. We note that intact households are also subject to nonresponse, so the issue of missing data is not resolved. The analysis presented here can be regarded as the complete-data analysis that could be incorporated in a multiple-imputation procedure.

We focus on the progression of household income over the period covered by ECHP and the description of typical patterns of income. Our analysis relies on multivariate log-normal mixture models which assume that the studied population of households comprises a small number of groups, each with relatively homogeneous units. For the six-variate data on income, the homogeneity has to be interpreted as similarity of the levels and patterns of change of income. Our analysis concludes that the majority of households in each country had income with small annual increases — these correspond to a distribution with relatively small variances and high between-year correlations. The remainder, modelled as one or two groups, have substantially larger variances and smaller correlations. The differences in the average incomes for the mixture components are secondary to the substantial within-component variation.

The next section gives further details of ECHP, and it is followed by a section reviewing the literature on methods for analysis of income data. Section 2 describes multivariate mixture models and the EM algorithm for fitting them. The analysis of the ECHP data is presented in Section 3, with details for an arbitrarily selected country, Spain, and summaries for all the surveyed countries. Section 4 evaluates the results and discusses several peripheral issues.

1.1 The ECHP database

ECHP has been designed and is conducted by the Statistical Office of the European Community (Eurostat). The survey collects information about demographics, labour-force behaviour, income, health, housing, education and training, and other issues. It responds to the increasing demand for socio-economic information comparable across the countries of the European Community and over time.

In 1999, fifteen countries participated in ECHP; most of them joined the Panel at its inception. Austria joined in 1995, Finland in 1996 and Sweden in 1997. In Germany, Luxembourg and the UK, the first three years of the panel ran parallel with existing national household surveys. The ECHP database for them contains the amalgam of the

Table 1: The sample-size information about the countries participating in ECHP.

Country	Acronym	Records	Households	Complete households	Years
Austria	AU	15 587	3919	2244	2–6
Belgium	BE	18 690	4005	2159	1–6
Denmark	DK	17 290	4111	1705	1–6
Finland	FI	15 985	4949	3002	3–6
France	FR	38 291	8335	4423	1–6
Germany	DE	39 845	12 054	4933	2–6†
Greece	GR	28 450	6151	3220	1–6
Ireland	IR	18 857	4519	2097	1–6
Italy	IT	41 028	8500	4669	1–6
Luxembourg	LU	15 073	4697	1879	2–6†
the Netherlands	NL	30 511	6881	3407	1–6
Portugal	PT	28 823	5789	3613	1–6
Spain	ES	36 488	8228	4021	1–6
the UK	UK	33 241	10 456	4178	2–6†

Note: † The data for the first year (1994) are excluded because they are not part of the panel.

data for the national and the ECHP-designed surveys. The data for 1994 are not part of the panel for either of these countries. In Sweden, independent random samples were drawn in each year 1997–1999. As they do not form a panel, we do not consider them here.

The target population of ECHP are the households within each country. The surveyed households are selected by a probability sampling design specific to each country; see European Commission (1996a). Most common is the two-stage sampling design, with geographical areas as primary sampling units. The national sample sizes, both planned and achieved, vary a great deal. For example, the household-level database for Italy comprises 41 028 records, with 8500 unique households and 4669 households with entries for income for each year 1994–1999 (years 1–6). Table 1 gives details of the sample sizes. We are concerned only with the variable containing the household income and with the employment status of its adult members. From this status we define a dichotomous variable that indicates whether the household had experienced a spell of unemployment of one of its members, or the whole household had at some point no income from employment.

The key outcome variable, total net household annual income, is defined as the total of the net annual incomes of the adult members of the household (aged 16 or over). The annual income of an individual comprises five main components: employment; self-employment; income from property; net interest and dividends; and pensions and other benefits. The questionnaire and related details are given in European Commission (1996b).

Ambiguities may arise in households whose composition has changed during the reference year. In the survey, the income is added up for the current members of the household for the previous calendar year. Most national surveys took place early in the year when some subjects submit tax returns, and most can easily recall their annual income. See Peracchi (2002) for further information about ECHP.

In Section 3.1, we use a dichotomous variable that indicates whether the household has experienced a spell of unemployment (of one of its members) or had in any wave no income from employment. The ILO definition of employment status is used. Each adult subject is classified as *employed* (at work or with a job, but temporarily not at work), *unemployed* (seeking and/or available for work) and *economically inactive* (not seeking and not available for work).

For each household, we define the following summary of the employment status of its members:

1. all adults employed;
2. some adults employed and some inactive, but nobody unemployed;
3. at least one adult each employed and unemployed;
4. no employed adults.

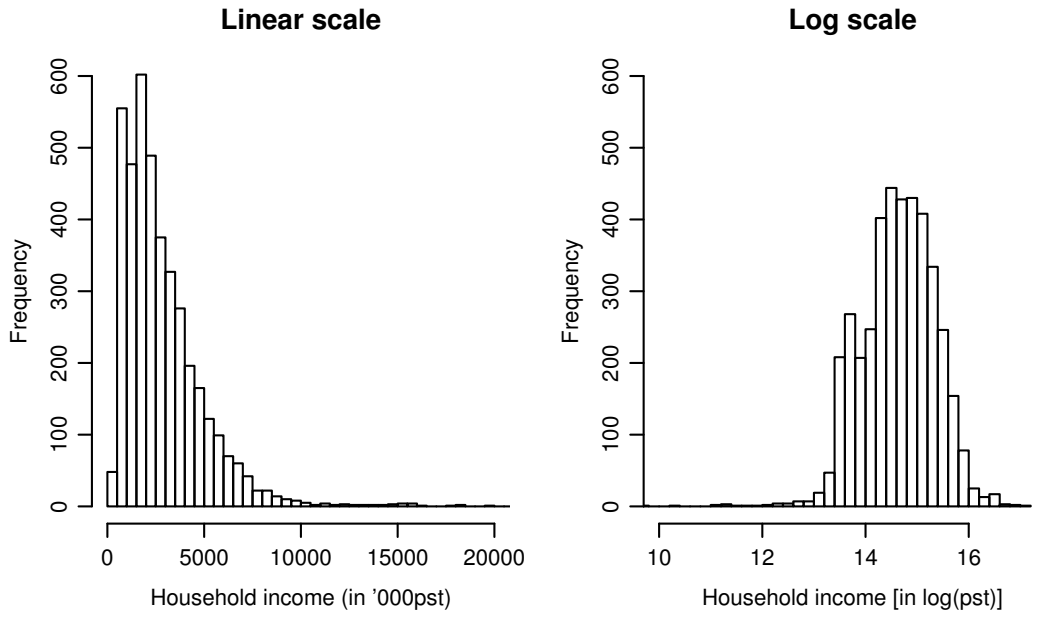
These definitions refer to the week preceding the interview. Finally, we define a dichotomous variable z which indicates whether the household has in at least one year been classified in category 3 or 4. We admit that this definition is problematic because it does not cover the entire period of the survey (4–6 years), but we want to illustrate with it a fuller potential of multivariate mixture models.

1.2 Modelling income

The distribution of income in most populations is highly skewed, with a long right-hand-side tail and high density at the lower percentiles. The logarithm is the natural transformation for such data. Comparisons of income are more practical on the multiplicative scale, such as by annual changes expressed in percentages. The distribution of log-transformed income is usually much closer to normality, and to symmetry in particular, but deviations from normality can be perceptible in large-scale data. See Figure 1 for the histograms of the data from Spain in 1999.

A single normal distribution may be inadequate for describing income. Mixtures of normal distributions are a much more general class. Any absolutely continuous distribution can be approximated by a finite mixture of normals with arbitrary precision (Marron and Wand, 1992). We are interested in mixtures of a few components, three or four, which

Figure 1: Histograms of the annual household incomes in Spain in 1999, on the linear and logarithm scales.



yield a substantial improvement over the fit by a single normal distribution, and yet they can be easily interpreted as associated with groups of relatively homogeneous households.

In the recent literature, kernel density estimation has been extensively applied in modelling income (e. g., Bianchi, 1997, and Burkhauser *et al.*, 1999), and the main goal has been the testing of certain hypotheses related to economic theories. Kernel density estimation can model the data in lesser of finer detail, depending on the extent of smoothing applied. In that respect, mixture models appear to be much less flexible, because the model choices (the number of components) are discrete. However, kernel density estimation is practicable only for uni- or bivariate data, although some tri-variate features can be inferred by modelling conditional distributions (Quah, 1996). The constraint on the number of dimensions is undesirable when studying the income over several time points. Mixture models can be fitted to multivariate data without any profound difficulties, beyond those encountered with univariate data. Also, they are better suited for inferences about distinct subpopulations. For applications of mixture models to univariate income data, see Paap and van Dijk (1998), Flachaire and Nuñez (2002), and Pittau (2003).

The likelihood ratio (LR) is commonly used for setting the number of components of the mixture. The criterion based on it identifies few components for small data sets and usually many for large ones because it is related to hypothesis testing and search for evidence supporting more complex models. We regard the LR test as not relevant in our modelling of gross features of the income distribution, and base our inferences on models

with fewer components than are suggested by LR. In particular, we are not interested in small groups of households with unusual (patterns of) income. A more appropriate criterion for our purposes is to stop increasing the number of mixture components when fewer than a given percentage, such as 5%, of the households are estimated to belong to one component. In principle, mixtures with more components may allocate subjects to components more evenly, but that represents a very esoteric situation.

The membership of each group is not identified with certainty. In our approach based on the EM algorithm (Dempster, Laird, and Rubin, 1977; McLachlan and Krishnan, 1997; and McLachlan and Peel, 2000), we do not assign a household to a component, but estimate the conditional multinomial distribution of belonging to the components, given the model parameters. See Richardson and Green (1997) for a Bayesian approach to model fitting and selection, based on Markov chain Monte Carlo simulations, and Meng and van Dyk (1998) for extensions of the EM algorithm and ways of accelerating its convergence.

The multinomial probabilities, specific to each unit, are evaluated at the parameter estimates. In mixtures of univariate distributions, these probabilities are a function of the location of each household's values. As the component distributions overlap, there is considerable uncertainty about the households' components, unless the distributions are well separated. For mixtures of multivariate distributions, the uncertainty can be much smaller, even when the univariate marginal distributions are not well separated. This is the case in our analyses, and it is due to the presence of very distinct patterns of income over the six years.

In the analyses for most countries, we identify a component characterised by regular small annual increases of income — the corresponding normal distribution of log-income has a variance matrix with relatively small variances and very high correlations. In most countries, this component accounts for more than 50% of the households. The variance matrices for the other components have greater variances and smaller covariances. Their expectations are smaller than for the dominant component, although the difference of the means is a feature secondary to the substantial variation of households' income from one year to the next.

2 Multivariate mixtures

Mixture models have been identified by Dempster *et al.* (1977), as an application of the EM algorithm, with the assignment of the component for each observation regarded as the missing information. Let $f_k(x)$ and p_k be the densities and marginal probabilities associated with component $k = 1, \ldots, K$. The conditional probability that household

$i = 1, \ldots, n$ belongs to component k is

$$r_{ik} = \quad \mathrm{P}\{C(i) = k \mid \mathbf{f}(x_i); \mathbf{p}\} = \frac{p_k f_k(x_i)}{\sum_{h=1}^{K} p_h f_h(x_i)}, \tag{1}$$

where $\mathbf{f} = (f_1, f_2, \ldots, f_K)$, $\mathbf{p} = (p_1, p_2, \ldots, p_K)$, and $C(i)$ indicates the component to which household i belongs. The naive estimate of r_{ik} is denoted by $\hat{r}_{ik}$.

The M-step comprises estimation of the densities f_k and the (marginal) probabilities p_k. When f_k are univariate normal densities, their estimation entails computing sample means and variances. As the complete-data analysis requires the indicator values $C(i)$, the E-step evaluates their conditional expectations, that is, the probabilities $\hat{r}_{ik}$ given by (1), evaluated at the current solution $(\hat{\mathbf{f}}, \hat{\mathbf{p}})$. In the M-step, the densities f_k are estimated by the weighted version of the complete-data procedure, with weights equal to the probabilities $\hat{r}_{ik}$ obtained in the preceding E-step. The marginal probabilities p_k are estimated as the means of the probabilities $\hat{r}_{ik}$. The pairs of E and M steps are applied iteratively. Convergence problems are encountered with the EM algorithm when the fraction of missing information is substantial, or the log-likelihood has ridges and similar features. The number of iterations increases with the complexity of the model (number of components). In our analyses, we have had no problems with convergence or multiple extremes. In a few instances, more than 100 iterations were required, but the E- and M-steps are simple.

Mixtures of multivariate distributions are conceptually no more complex than univariate mixtures. Equation (1) remains valid, although the densities f_k now have multidimensional arguments. Mixtures are well defined for distributions comprising a mix of continuous and categorical variables. In Section 3.1, we analyse data about income and employment, $(\mathbf{x}, z)$, where z summarises the employment status of the adult members of the household over the past six years. The joint distribution of $\mathbf{x}$ and z can be decomposed in two ways:

$$f(\mathbf{x}, z) = f(\mathbf{x} \mid z) f(z) \tag{2}$$

$$f(\mathbf{x}, z) = f(z \mid \mathbf{x}) f(\mathbf{x}), \tag{3}$$

where f is the generic notation for a density of an absolutely continuous or discrete distribution. In (2), $f(\mathbf{x} \mid z)$ are the (multi-normal) distributions of income within each category of employment status. In (3), $f(z \mid \mathbf{x})$ are the (multinomial) probabilities of the employment states given income. This can be modelled by logistic regression.

3 Analysis

We describe in detail the analysis for Spain because it represents a typical case. In Section 3.2, we discuss all the other countries, focusing on their deviations from the stereotype.

The fit to the two-component mixture model is given by the estimates of the means and variances listed in Table 2, and the correlation matrices

$$\frac{1}{1000}\begin{pmatrix} 1000 & 794 & 761 & 698 & 699 & 666 \\ 794 & 1000 & 822 & 748 & 755 & 729 \\ 761 & 822 & 1000 & 810 & 774 & 768 \\ 698 & 748 & 810 & 1000 & 801 & 747 \\ 699 & 755 & 774 & 801 & 1000 & 784 \\ 666 & 729 & 768 & 747 & 784 & 1000 \end{pmatrix}$$

and

$$\frac{1}{1000}\begin{pmatrix} 1000 & 182 & 106 & 71 & 81 & 81 \\ 182 & 1000 & 218 & 143 & 147 & 139 \\ 106 & 218 & 1000 & 259 & 176 & 183 \\ 71 & 143 & 259 & 1000 & 309 & 182 \\ 81 & 147 & 176 & 309 & 1000 & 289 \\ 81 & 139 & 183 & 182 & 289 & 1000 \end{pmatrix}$$

for the respective components 1 and 2. The component 1 accounts for $\hat{p} = 91.8\%$ of the households. The high correlations imply that the annual incomes are strongly associated. The means show steady annual increases, by between 0.02 and 0.06 on the log scale, which convert to increases between 2% and 6%. However, the variation between the households is much greater; the component accounts for households with both high and low incomes. For illustration, the standard deviation of $\sqrt{0.49} = 0.70$ corresponds to the 5th and 95th percentiles about a quarter and quadruple of the mean income.

In contrast, the mean incomes of the second component, accounting for 8.2% of households, display an irregular pattern with substantial increases and reductions from one year to the next. The fitted variances are much greater than for the first component, and all the correlations are much smaller. For example, the estimated variance of the difference between the incomes in year 1 and 2 is 2.12. Hence, a $\exp(\sqrt{2.12}) = 4.3$-fold increase or reduction of annual income of a household in this component is not unusual. This may appear incredible, but numerous such households can be identified in the data. The main difference between the two component distributions is not in their means, but in the variance structure — the pattern of annual changes.

As the fitted distributions differ so substantially, many households are identified with their component (pattern) with a high degree of certainty. This is easiest of all to summarise by a tabulation of the estimated conditional probabilities $\hat{r}_{i1} = 1 - \hat{r}_{i2}$. These values are greater than 0.9 for 3583 (89.1%) and smaller than 0.1 for 218 households (5.4%). For the remainder, 5.5% of households, the mixture component is uncertain, as the conditional probabilities are in the range 0.1–0.9.

Figure 2 depicts the pattern of household income for random samples drawn from households that are very likely to belong to component 1 or 2 ($\hat{r}_{i1} > 0.9$ or $\hat{r}_{i2} > 0.9$,

Table 2: The two-component mixture model fit to the annual household incomes in Spain, 1994–1999.

	Year						
Component	1	2	3	4	5	6	
	Means						
1	14.43	14.49	14.53	14.55	14.60	14.65	
2	13.69	13.97	13.92	13.82	14.06	14.29	
	Variances						**Correlations**
1	0.48	0.43	0.43	0.51	0.50	0.49	0.67 — 0.82
2	2.00	1.46	1.46	2.29	1.75	1.38	0.07 — 0.31

respectively), those that cannot be assigned ($\hat{r}_{ik}$ in the range 0.4–0.6), and for a random sub-sample from the entire data set. Each sample is of size 40. The vertical axes are on the log scale. The thick lines indicate the (arithmetic) means of the incomes for the entire sample (solid line) and the relevant component (dashes). The latter is calculated as the weighted mean of incomes, with the weights equal to the estimated probabilities $\hat{r}_{ik}$. The two lines are indistinguishable for component 1. The diagram confirms that the mixture components are easy to interpret, even though the model fit can be improved substantially, as we discuss below. Note that the households that are difficult to assign to a component appear to have some feature of each component: segments of nearly constant annual incomes, with one or two aberrations.

The fit of the three-component mixture model is summarised in Table 3. The estimated marginal probabilities are $(0.84, 0.14, 0.02)$. The dominant component 1 is very similar to its counterpart in the two-component solution. It represents households with nearly constant annual (percentage) increments. The second component in the three-component solution is similar to the second component in the two-component solution, although its estimated variances are somewhat smaller. The third component comprises households with extremely varying and very weakly correlated annual incomes. Note that the moments of the third component are estimated with very low precision, as they are based on an effective sample size of only $4021 \times 0.019 \doteq 76$, and the observations involved are widely dispersed.

It is instructive to explore how the households' components, and the associated uncertainty, are related for the two- and three-component solutions. For each solution, we define the pattern of the probabilities by rounding downward, to one decimal place, the estimated probabilities $\hat{\mathbf{r}}_i = (r_{i1}, r_{i2})$ or (r_{i1}, r_{i2}, r_{i3}). Thus, the pattern for the two-component solution comprises two digits, such as 90 for a household very likely to belong to component 1

Figure 2: The typical patterns of household income in Spain, 1994–1999, with reference to the solution given in Table 2. The overall mean is drawn by a thick solid line, and the means for the relevant component by thick dashes. The patterns of probabilities are indicated in the subtitles.

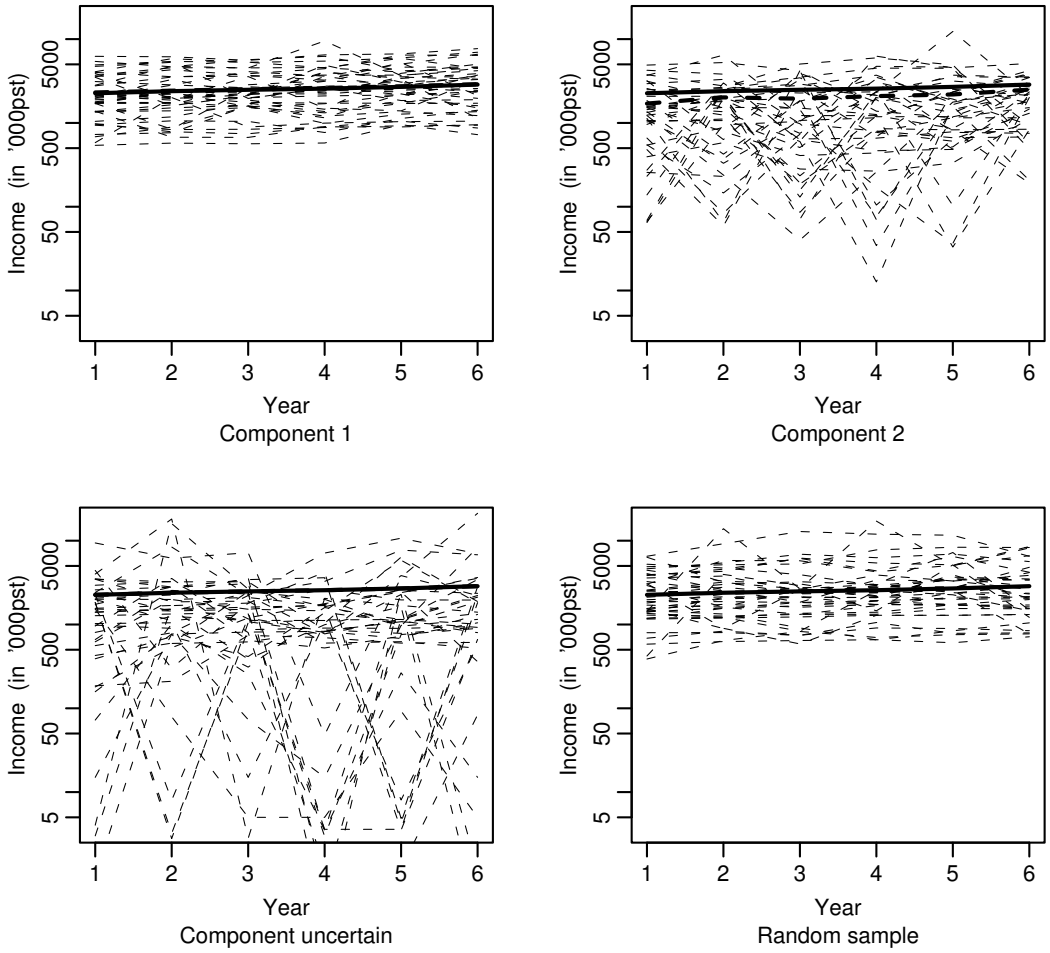


Table 3: The three-component mixture model fit for annual household incomes in Spain, 1994–1999.

Component	Year						Correlations
	1	2	3	4	5	6	
Means							
1	14.45	14.51	14.55	14.57	14.62	14.67	
2	13.61	13.16	13.97	12.67	13.02	14.00	
3	14.02	14.29	14.15	14.25	14.36	14.45	
Variances							
1	0.45	0.41	0.41	0.45	0.44	0.47	0.72 — 0.89
2	1.30	0.75	1.04	1.06	0.86	0.90	0.14 — 0.51
3	2.62	2.73	1.16	5.17	4.69	2.15	–0.38 — 0.25

(the probability of 1.0 is 'rounded' downward to 0.9, so that we have only 10 groups). The patterns for the three-component solution comprise three digits. In principle, almost 200 distinct patterns are possible, but only a fraction of them are likely to occur more than a few times.

Table 4 cross-tabulates the patterns for the two- and three-component solutions. For conciseness, patterns that occur fewer than ten times are excluded. The table shows a remarkable consistency in the allocations of households to components. Households almost certain to belong to the dominant group in the three-component solution (pattern 900) are almost certain to belong to the dominant group also in the two-component solution (pattern 90). The second group in the three-component solution 'recruits' most of the households that are not assigned to the dominant group in the two-component solution with near-certainty (see the row for pattern 090). The third group of the three-component solution recruits a small fraction of households from the minority group of the two-component solution, and a small fraction of households that could not be assigned to a group in the two-component solution. Most transitions between the two- and three-component patterns do not take place. Note that for the vast majority of households there are only two likely components; the only exception with an appreciable number of households is pattern 333. In general, more complex models entail more uncertainty about the allocation of households to components. But the near-certainty patterns 900, 090 and 009 in the three-component solution still account for 88.3% of the households.

The estimated probabilities of belonging to the components can be graphically summarised by a composition (triangular) plot (Aitchison, 1986), as shown in Figure 3. In the plot, each household is represented by a point. Vertices C1, C2 and C3 correspond to certainty that the household belongs to the respective component 1, 2 and 3. Proximity

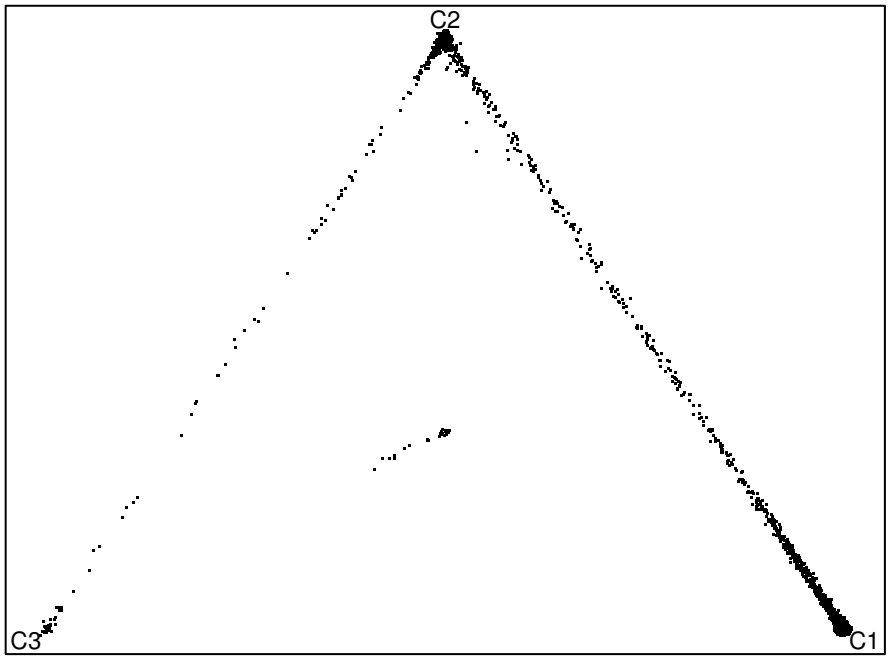
Table 4: Patterns of the estimated probabilities $\hat{\mathbf{r}}_i$ for the two- and three-component solutions for Spain. The labels for the near-certainty patterns are printed in italics. Patterns with fewer than 10 households in the sample are omitted from the table.

3-comp.	Pattern for the 2-component solution										
pattern	*90*	81	72	63	54	45	36	27	18	*09*	**Total**
900	3166	0	0	0	0	0	0	0	0	0	3166
810	137	0	0	0	0	0	0	0	0	0	137
720	60	0	0	0	0	0	0	0	0	0	60
630	28	0	0	0	0	0	0	0	0	0	28
540	38	0	0	0	0	0	0	0	0	0	38
450	26	0	0	0	0	0	0	0	0	0	26
360	28	0	0	0	0	0	0	0	0	0	28
333	0	0	0	0	0	20	0	0	0	0	20
270	30	0	0	0	0	0	0	0	0	0	30
180	34	3	0	0	1	0	0	0	0	0	38
090	30	39	18	30	16	12	16	16	26	141	344
072	0	0	0	0	0	0	0	0	0	14	14
009	0	0	0	0	0	14	1	0	0	27	42
Total	3583	43	19	30	17	51	18	16	26	218	**4021**

to the vertices reflects the probability $\hat{r}_{ik}$ that the household belongs to the corresponding component. A small amount of random noise (jigging, Gelman *et al.*, 1995) is added to the points, so that multiplicities become transparent. The graph confirms that most households can be allocated to a component with a high degree of certainty. Moreover, most points are on or near the two sides of the triangle, C1–C2 and C2–C3. For these households, either component 1 or component 3 can be ruled out. The side C1–C3 has no points in its proximity, apart from the points at or close to the vertices. This implies an ordering of the components, 1–2–3, according to their variances (and correlations).

The third component is strongly supported by the LR criterion; the value of the LR statistic is 2790. This suggests that we should continue with the modelling exercise and look for solutions with four or more components. This we regard as not constructive, as more complex solutions are likely to identify small groups of households with strange progressions of income that are not relevant for the gross description of income. The summary of the model fit in Table 3 is useful despite the fact that the solution can be 'improved', as judged by the traditional LR criterion. Figure 4 displays the income progression for random samples of 40 households that are very likely to belong to each of the mixture components, and those that are unlikely to belong to one of the components. In the bottom panel, all the households with pattern 333 are displayed, as there are only

Figure 3: The composition plot of the estimated probabilities of belonging to the mixture components; three-component solution for Spain.



20 of them. The panels for the components 1–3 show that typical households have very different patterns of annual income. Component 3 appears to contain many households that had a substantially reduced income for one year. The means for the sample and the components are drawn by thick lines. The main features in the panels are the distinct patterns of variation; the differences in the means are unimportant in comparison.

Fitting the mixture models required 30 and 121 iterations for the respective two- and three-component solutions. The criterion for convergence was that the square root of the average change in the parameter estimates and the deviance (-2 log-likelihood) be smaller than 10^{-5}.

3.1 Employment status and income

In this section, we consider the household-level summary of unemployment and economic inactivity, z, defined in Section 1.1, as another outcome variable, and fit mixture models to the seven-variate data. For the M-step of the EM algorithm, we have the two options given by (2) and (3); they correspond to vectors of mean log-income for each category of z,

Figure 4: The typical patterns of household income in Spain, 1994–1999, for the solution given in Table 3. The overall mean is drawn by a thick solid line, and the mean for the relevant component by thick dashes.

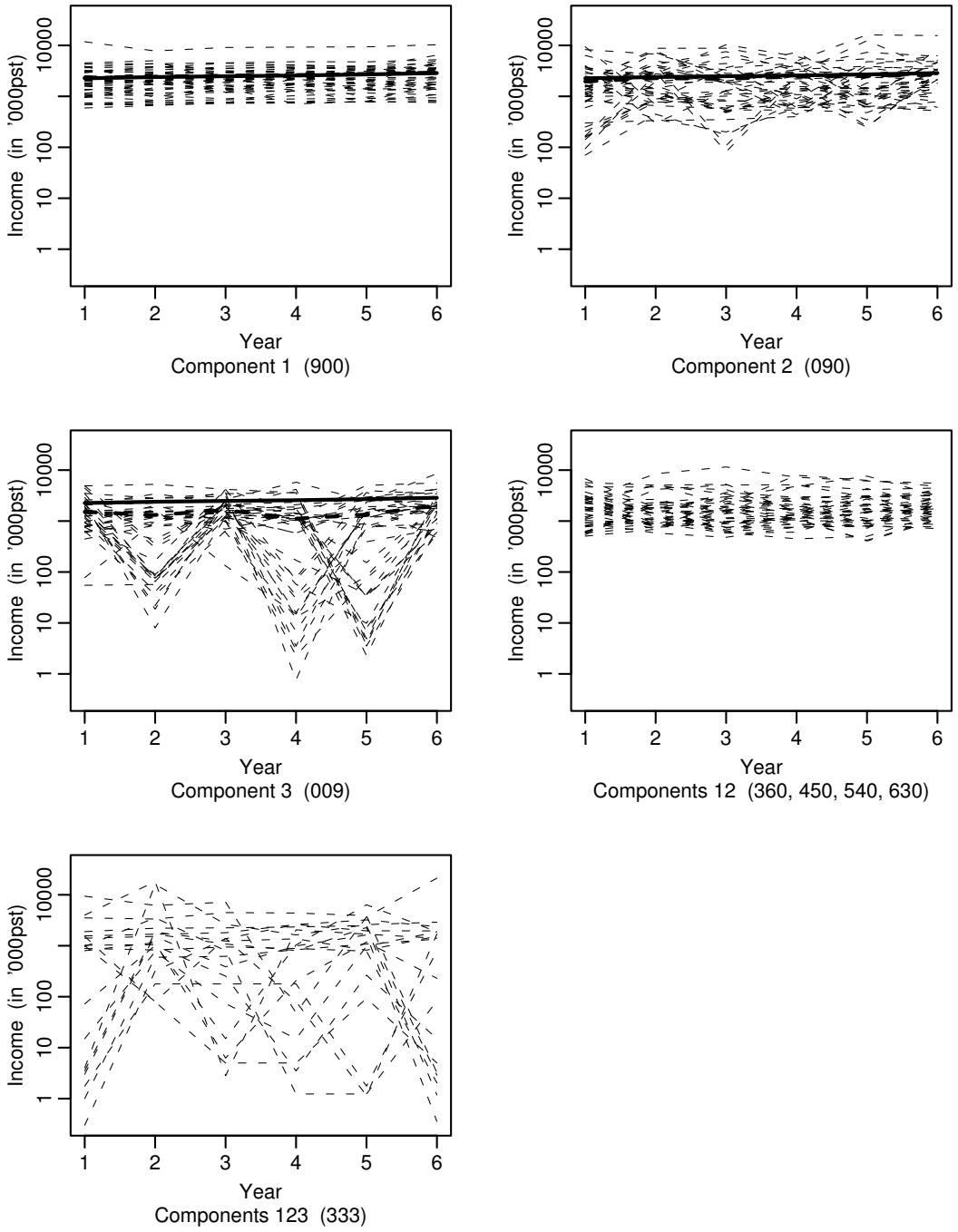


Table 5: The three-component mixture model fit to the annual household income and employment data; Spain, 1994–1999.

	Year						
Component	1	2	3	4	5	6	
	Logistic regression						**Intercept**
1	−0.33	−0.03	−0.73	−0.03	−0.04	−0.66	27.94
2	−0.17	0.06	0.08	−0.27	−0.42	−0.10	13.92
3	−0.08	−0.49	0.10	−0.17	−0.13	−0.35	16.53
	Means						
1	14.46	14.52	14.57	14.58	14.63	14.69	
2	14.38	14.47	14.46	14.47	14.52	14.57	
3	13.56	13.83	13.82	13.68	13.96	14.26	
	Variances						**Correlations**
1	0.44	0.42	0.41	0.47	0.48	0.49	0.81 — 0.91
2	0.54	0.47	0.51	0.62	0.61	0.57	0.28 — 0.63
3	2.06	1.45	1.24	2.28	1.65	1.20	0.01 — 0.30

and a regression of z on the log-income, respectively. The two options are not equivalent; the former is more general, unless we allow for a very flexible regression in the latter. Since the covariates in the regression are highly correlated, we use only linear regression, with no interactions. Of the choices for the link function, we apply logit; the other links are unlikely to yield different results, as most fitted probabilities are distant from zero and unity.

We prefer the approach based on the (logistic) regression because its results are easier to summarise, by a fit of the logistic regression and the marginal moments of the six-variate normal distribution. The model fit with three mixture components is given in Table 5. The top part of the table gives the three logistic regression fits $\hat{f}_k(z \mid \mathbf{x})$. Owing to multicolinearity, the parameter estimates have no straightforward interpretation. However, the predominance of negative estimates indicates that higher income is associated with lower probability of an experience of no income from employment ($z = 1$). The purpose of using as regressors the log-incomes from all the earlier years is mainly for uniformity. The nominal standard errors of the regression parameter estimators, derived from the last M-step, are not an appropriate reflection of the sampling variation because they are *conditional* on the allocation of households to components. This issue is addressed in Section 3.1.2.

The components have estimated marginal probabilities 0.67, 0.26 and 0.07; they differ from their counterparts in the analysis without the employment variable. Therefore,

different subpopulations are identified by the two models. Nevertheless, the estimated means, variances and correlations have similar features in the two model fits. The third components in both solutions correspond to households with very high variation and low correlations across the years. In Table 5, the second component differs from the first mainly in the correlations; the means and variances are very similar. This highlights the importance of jointly modelling all the variables of interest.

For brevity, we refer to the two models, with and without the employment variable, by their respective generic densities $f(\mathbf{x})$ and $f(\mathbf{x}, z)$. The allocation to the components by the two models is best summarised by plots of the corresponding estimated probabilities. These are given in Figure 5, with the dichotomous variable indicated by the plotting symbol. To avoid excessive clutter, only a 25% random sample of points is plotted; the plotted points are 'jigged' to resolve multiplicities. Although in general there is an ambiguity about the order of the components, in the two model fits compared there is a natural ordering according to the variances and covariances, or the estimated marginal probabilities $\hat{p}_k$.

The plot for component 1 can be interpreted as follows: households that do not belong to component 1 for $f(\mathbf{x})$ do not belong to component 1 for $f(\mathbf{x}, z)$ either, and households that belong to component 1 for $f(\mathbf{x}, z)$ also belong to component 1 for $f(\mathbf{x})$. In other words, by moving from $f(\mathbf{x})$ to $f(\mathbf{x}, z)$, the first component loses some households or, more precisely, the probabilities of belonging to component 1 are reduced, with only a few exceptions. The change in the probabilities for component 2 between models $f(\mathbf{x})$ and $f(\mathbf{x}, z)$ is more difficult to interpret, but the low density of points in the middle of the plot indicates that the probabilities for a household are distant from zero and unity for at least one model. The plot for component 3 shows that its version for $f(\mathbf{x}, z)$ 'inherits' all the households from component 3 for $f(\mathbf{x})$, and acquires most of the households for whom component 3 for $f(\mathbf{x}, z)$ cannot be ruled out.

Figure 6 displays the plots of typical income progressions for the components and for households with uncertain assignment to the components. The layout of the diagram is the same as for Figure 4, except for the additional panel (components 2 and 3) and for the symbols distinguishing between households with $z = 0$ and $z = 1$. In the sample, 76% of the households have $z = 1$; the respective estimated percentages within the components 1–3 are 73, 86 and 69. For the corresponding 'typical' patterns 900, 090 and 009, these percentages are 68, 96 and 66.

The component-specific logistic regressions yield different probabilities of experiencing no income from employment ($z = 1$), especially for households for which this probability is much smaller than 1.0. Figure 7 shows this by plotting the component-specific fits of the logistic regressions. In contrast, the distributions of the probabilities within the typical

Figure 5: The probabilities of belonging to the mixture components in the models $f(\mathbf{x}, z)$ and $f(\mathbf{x})$.

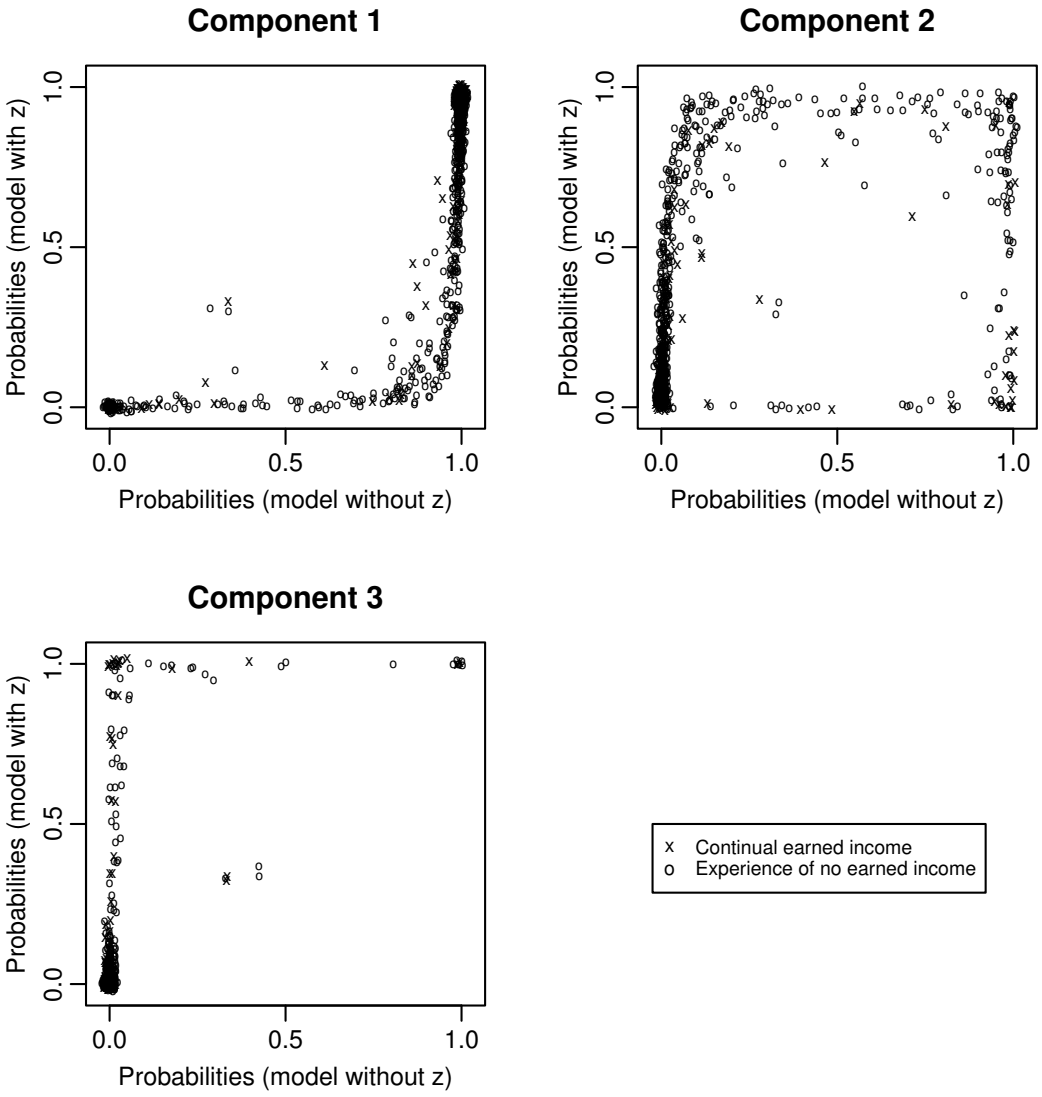


Figure 6: Typical patterns of income progression for the mixture components in the model $f(\mathbf{x}, z)$.

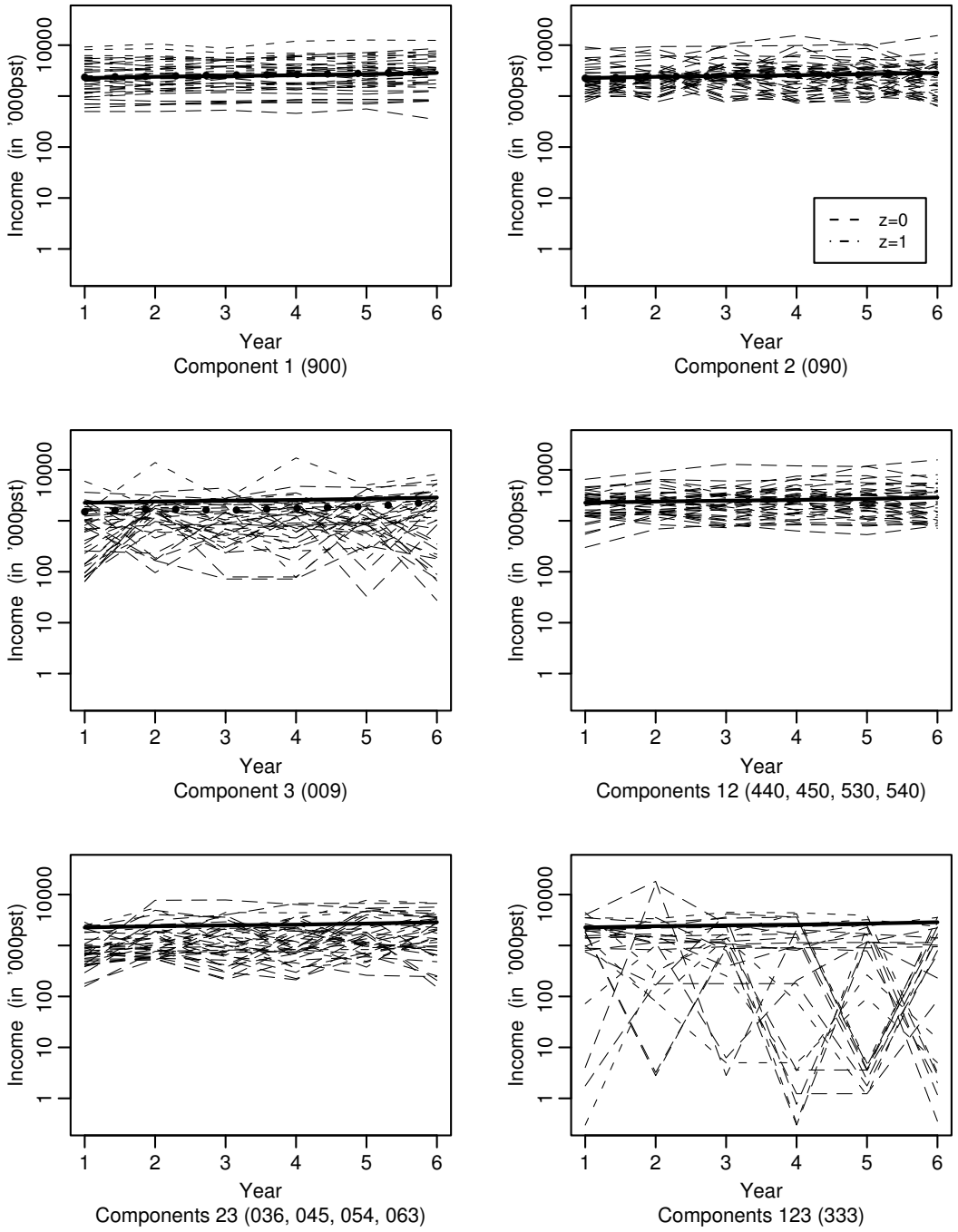
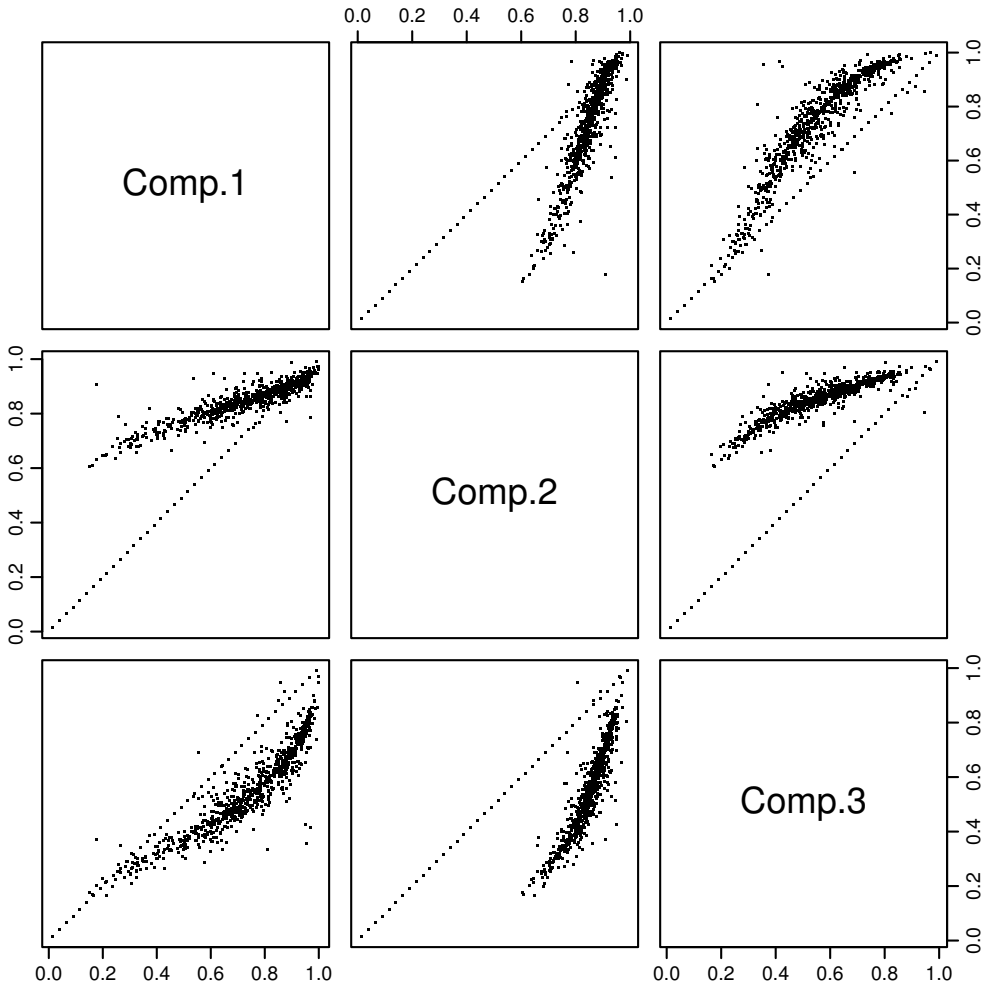


Figure 7: Probabilities of experiencing no income from employment, based on the three-component mixture model fit.



patterns are very similar to the distribution without conditioning on the component. In general, the probability of experiencing no income from employment is negatively associated with income. This association is very strong for component 1, less so for component 2, and is weakest for component 3. Figure 8 illustrates this by plotting the fitted probabilities against the mean annual log-income. No causal association is implied by these regressions. Even when one member's loss of employment brings about a reduction of the household's income, other members of the household may adopt labour-force behaviour that aims to compensate for this. Also, through one of its members, the household may qualify for other income, such as social security benefits and a lower level of taxation.

Figure 8: The fitted probability of experiencing no income from employment ($z = 1$) and the average annual log-income for typical households in the three mixture components.

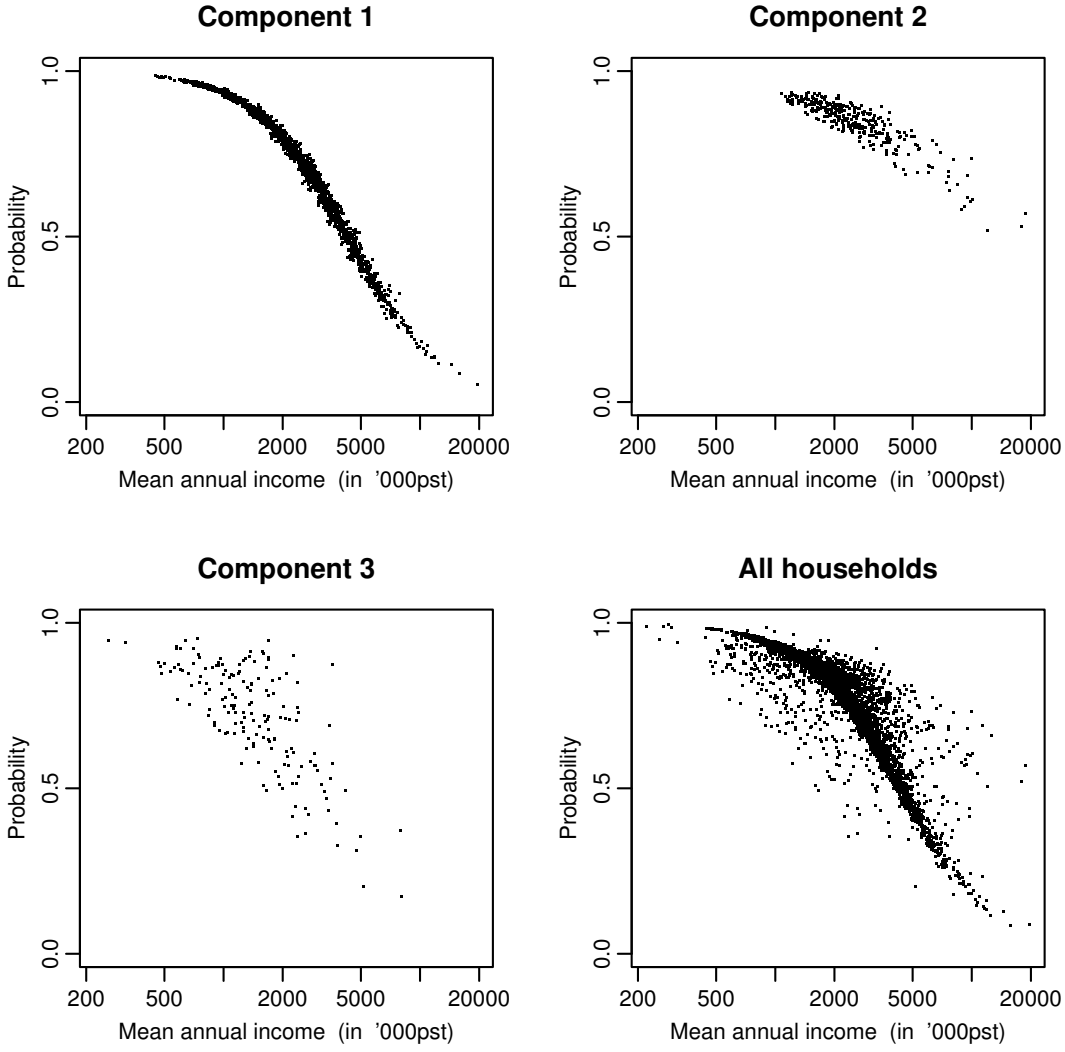


Table 6: The three-component mixture model fit with conditioning on the dichotomous variable z.

	Year						
Component	1	2	3	4	5	6	
	Means		$z = 0$				$\hat{p}_k(z)$
1	14.86	14.91	14.97	14.99	15.04	15.10	0.23
2	14.12	14.46	14.37	14.46	14.55	14.71	0.30
3	13.83	13.93	13.11	13.32	14.03	14.36	0.35
			$z = 1$				
1	14.33	14.38	14.42	14.43	14.47	14.52	0.77
2	13.88	14.08	14.14	14.12	14.28	14.36	0.70
3	13.45	13.31	13.60	12.52	12.75	13.95	0.65
	Variances		$z = 0$				**Correlations**
1	0.32	0.28	0.29	0.35	0.33	0.33	0.69 — 0.85
2	1.58	0.63	0.75	1.32	0.99	0.70	0.21 — 0.50
3	2.50	1.95	2.77	4.40	2.29	2.33	−0.13 — 0.36
			$z = 1$				
1	0.43	0.39	0.38	0.43	0.41	0.44	0.67 — 0.84
2	1.32	1.04	1.00	1.19	1.01	1.05	0.12 — 0.45
3	2.67	2.79	1.69	5.07	5.34	2.40	−0.38 — 0.10

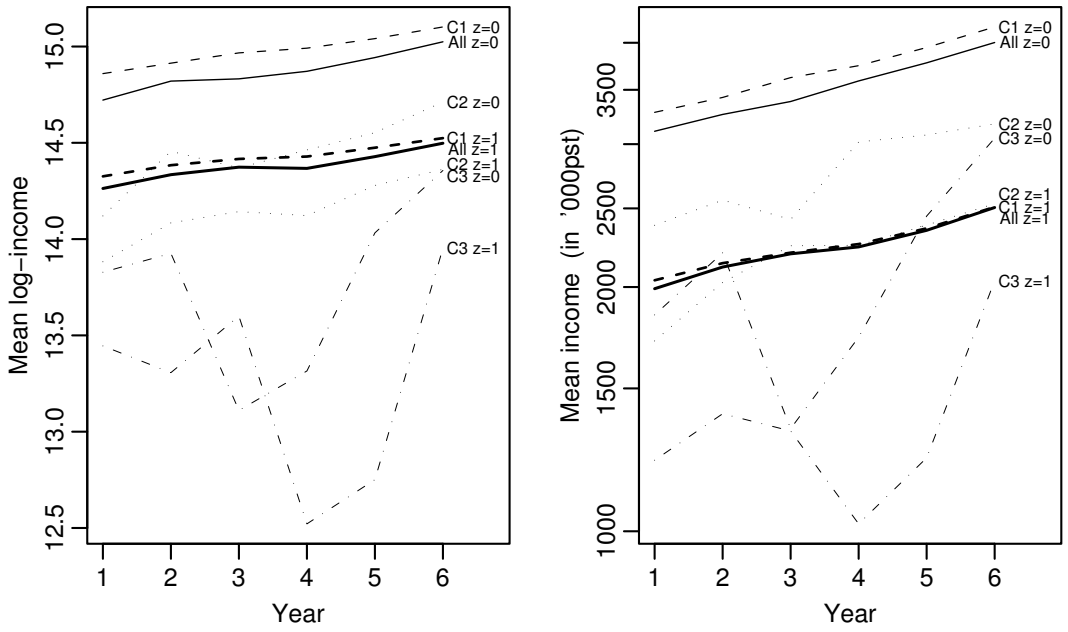
In summary, the categorical variable has an influence secondary, but not negligible, on the fit of the mixture model, and on the composition of the components in particular.

3.1.1 Conditioning on z

In this section, we briefly summarize the mixture-model fit with conditioning on the categorical variable, based on (3). The fit is described by three objects, each comprising mean vectors and variance matrices, one for each category of z. The model fit for Spain, with three components, is displayed in Table 6. The right-hand-side column, labelled $\hat{p}_k(z)$, gives the estimated probabilities of the categories $z = 0$ and $z = 1$, given the mixture components 1–3. To save space, the correlation matrices are summarised by the ranges of their off-diagonal values.

The solution in Table 6 has similar features to those in Table 5; the first component comprises households with steadily increasing incomes (small variances and large correlations), and the other two components comprise households with large variances and small between-year correlations. However, most correlations in the two variance matrices for the first component are smaller than their counterparts in Table 5. The component-specific probabilities of experiencing no income from employment ($z = 1$), 0.77, 0.70 and 0.65,

Figure 9: The estimated mean household incomes in the three-component mixture model with conditioning on the dichotomous variable z. The thickness of each line is proportional to the percentage of the households belonging to the combination of the component and category of z.



for the respective components 1, 2 and 3, are also substantially different from the earlier solution.

The differences between the conditional means of $\mathbf{x}$ given $z = 0$ and $z = 1$ tend to vary little within the components; the income of those who have at some point experienced no income from employment is lower. The six vectors of estimated means are plotted in panel A of Figure 9, together with the sample means for the two categories of z. To illustrate the impact of the log-scale, panel B plots the corresponding means on the linear scale. The means on the linear scale are relatively higher when associated with greater variances.

3.1.2 Standard errors

The sampling variation of any parameter comprises the variation estimated by the M-step and the variation associated with the missing information (assignment to the components, $C(i)$). As the latter is difficult to establish, the bootstrap provides a more practical solution. Although it requires extensive computing, the programming effort is insubstantial. In our implementation, the data-based estimates are the starting solution for each replication. This reduces the number of iterations. We do not require the precise standard errors for the many estimated parameters, merely an indication of the uncertainty entailed. For

this, 200 replications suffice, and there is no need to apply them for every country, or indeed for each analysis.

It is essential to establish that the ordering of the components in the replications is unambiguous. In the replications, the components are ordered according to the average of their six variances. The estimated standard errors of the marginal probabilities are 0.023, 0.016 and 0.009, for the respective estimates 0.67, 0.26 and 0.07, so no ambiguity arises.

The standard errors of the logistic regression parameters are in the range 0.11–0.35, except for the intercepts. Thus, most of the years do not contribute to the three logistic regressions significantly. The standard errors for the means are in the ranges 0.012–0.014, 0.032–0.039 and 0.087–0.133 for the respective components 1–3, so the inference of steady increases is well supported, although the mean percentage increase is poorly estimated for each pair of years.

The standard errors of the variances in the log-normal distributions are strongly affected not only by the effective sample sizes but also by the actual sizes of the variances. The standard errors for the respective components 1–3 are in the ranges 0.012–0.019, 0.043–0.080 and 0.140–0.401, so the comparisons of the variances across the components are not challenged, although the sizes of the variances in the third component are subject to substantial uncertainty.

Typical standard errors for the correlations for the three components are 0.025, 0.050 and 0.070. Thus, the comparisons of the correlations across the components are well supported, although the occasional occurence of a negative estimated correlation in the third component is more likely due to chance than a non-positive underlying correlation.

Although there are indications of moderate bias in estimating some of the parameters, the contribution of the bias to the mean squared error is unimportant.

3.2 Results for the other countries

In this section, we describe the main features in which the results for the other countries in ECHP differ from those for Spain.

In the model fit for $f(\mathbf{x})$, the dominant component for Austria accounts for only 67% of the sample and the second component is much more substantial (29%) than for Spain. The between-year correlations for the first component, 0.86–0.96, are much higher than for Spain, as are the correlations in the second component (0.46–0.73). The mixture model points to a subpopulation with greater stability of income over the years, but this may be a consequence of applying a stricter 'criterion' for allocating households to the first component. In the model with the dichotomous variable z, the component with stable income accounts for only 52%. The second component has slightly higher means and slightly lower variances and differs substantially from the first component only by having

lower correlations (0.47–0.77). The lower inflation in Austria is reflected in more moderate mean increases in the incomes, but it has no impact of the variation.

For Belgium, the second components in the models considered have somewhat higher correlations and they account for higher percentages of households. The means for the second component for $f(\mathbf{x})$ are only slightly smaller than the means for the first component, and for $f(\mathbf{x}, z)$ the means for the second component are greater for the first four years. The variances in the third components are less extreme, between 1.2 and 2.6, and the correlations (0.26–0.55) are higher than for Spain.

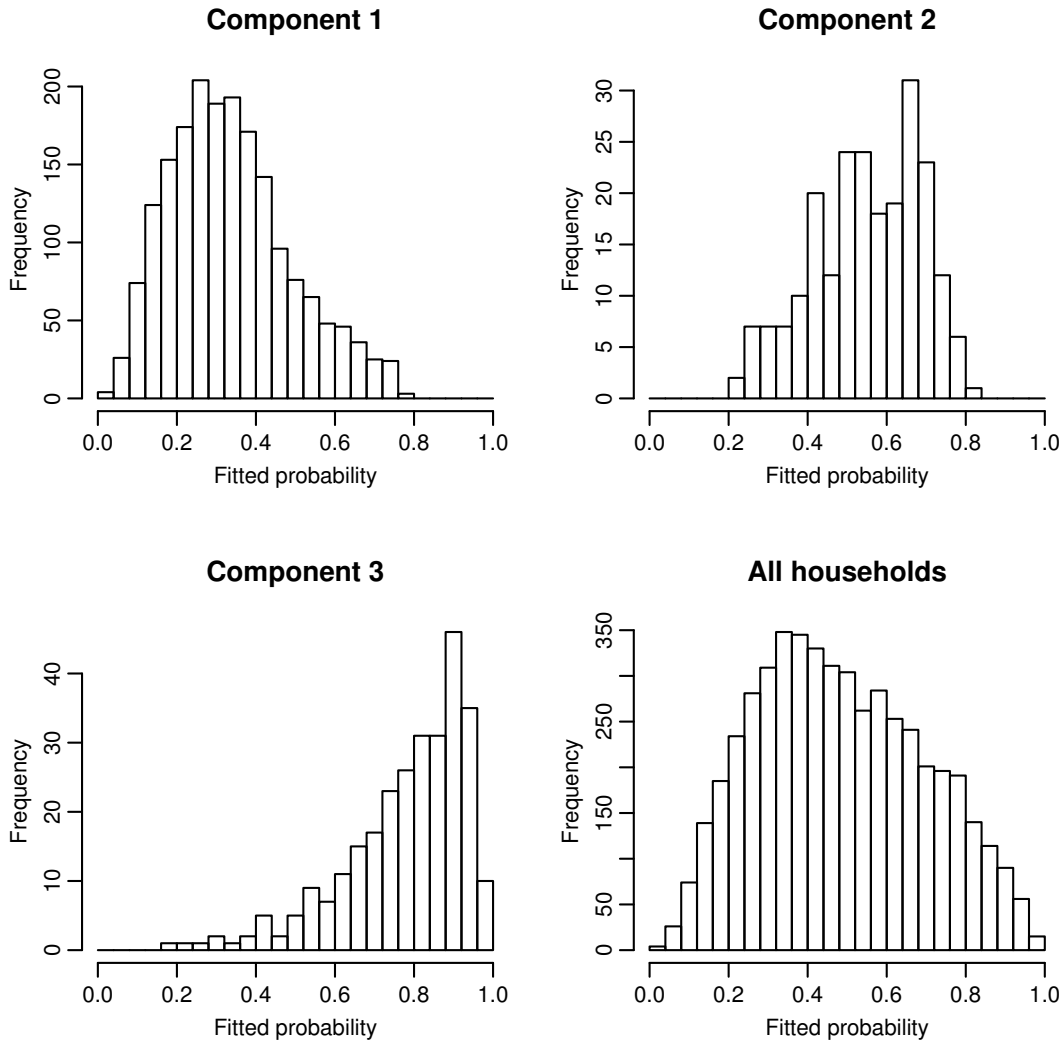
The results for Germany closely resemble those for Spain, both in the fitted distributions of the components and the marginal probabilities. A distinct departure from the general pattern is in the distribution of the dichotomous variable z and of the probabilities fitted by the mixture components. The latter are summarised in Figure 10 by the histograms of the fitted values for the households typical for each component, supplemented by the histogram of the values for all households, without conditioning on the component. Although component 2 is quite sizeable (26%), its conditional probability exceeds 0.9 for only 65 households (1.3%). To avoid small counts, we have included in the histogram for component 2 all the 223 households with $\hat{r}_{i2} > 0.85$. Unlike for the other countries, the fitted probabilities of $z = 1$ tend to be higher for the second and third components. Overall, only 48% of the households have an experience of no income from employment.

Denmark can be characterised as having very low dispersion of the log-incomes for each mixture component. This reflects the small (unconditional) variation of the household incomes. But the three-component solutions have the same ordering as for the other countries, with increasing variances and decreasing correlations from the dominant to the minority component. Unlike for the other countries, the mean annual increases are small and positive not only for the first but also for the second component.

The analysis for Finland is based on only four years. The first two components in both three-component model fits have very similar means and variances. The difference is only in the correlations, and even there it is not extreme. The fitted correlations for component 1 are in the range 0.94–0.98, and for component 2 0.65–0.83, much higher than for most other countries. The mean vector for the third component is uniformly lower than the mean vector for the other two components. The means decrease over time; for the model $f(\mathbf{x})$, they drop by 0.60 over four years, representing an average 1.83-fold reduction for the component. This component comprises over 7% of the households.

For France, the first component has steadily increasing means, while the third component has substantial increases in the first two years and changes within the range ± 0.02 in the following years. Several estimated correlations are negative for the third component, but their estimation is based on a very small effective sample. The three components fitted

Figure 10: Fitted probabilities $\hat{r}_{ki} = \mathrm{P}(z = 1 \mid k)$ for households typical for each component, and for all the households; Germany, 1995–1999.



are very well separated; 91% of the sample is assigned to a component with probability exceeding 0.90. The corresponding figures for most other countries are 75% (Spain 88%), although less separation might be expected for countries with data for only five or fewer years.

Greece and Portugal have the highest sampling variances of log-incomes for each year. In Greece, the three-component mixture model fit for the incomes, model $f(\mathbf{x})$ has the usual pattern, but the variances in the first component are higher than for the other countries. With the dichotomous variable z, the variances of the first component are much smaller, but the component accounts for only 47% of the households. The model fit departs from the usual pattern: the second and third components share the remainder of the households much more evenly (31% and 22%), and the ranges of their fitted variances overlap (0.74–1.19 *vs.* 0.72–0.91). The correlations of the incomes in the first component are relatively low (0.76–0.89) and in the other two components relatively high (0.62–0.88 and 0.43–0.54). The mixture components are rather poorly separated; only 63% of the households are assigned to a component with a probability exceeding 0.90.

The results for Portugal are very similar, although less extreme in the features discussed for Greece. The variances in the first components of both model fits (with and without z) are very high and in the other two components very low in comparison to the other countries. In fact, the solution for the model $f(\mathbf{x}, z)$ has slightly higher variances for component 1 than for component 2 for each year (0.62–0.65 *vs.* 0.52–0.61). However, the correlations have the usual pattern.

For Ireland, the second component accounts for somewhat greater percentage of households for both three-component mixture solutions. The estimated means and variance matrices for $f(\mathbf{x})$ have sizes and patterns similar to their counterparts for Spain, but for $f(\mathbf{x}, z)$, the second component has higher means and lower variances than the first component. The correlations have the usual pattern.

The results for Italy also conform to the stereotype. The vector of estimated means for the first component is greater by around 0.25 than the vector for the second comoponent. The estimated mean log-income in the third component is smaller by 1.40 than for the first component in year 1, but increases at a much higher rate, so that in year 6 it is lower by only 0.45. The associated variances also drop quite dramatically, from 2.32 to 0.56. The estimated mean of third component in the fit to $f(\mathbf{x}, z)$ lags behind the first two components somewhat less in year 1 (lower by 0.92) and grows over the years only slightly faster than component 1 (difference 0.49 in year 6).

For Luxembourg, data are available only for years 2–6 and the sample size, 1879, is smaller than for all other countries, except for Denmark. However, the EM algorithm does not require an excessive number of iterations, suggesting that the mixture components

are well identified. This is confirmed by exploring the estimated probabilities $\hat{r}_{ik}$; the model fits for $f(\mathbf{x})$ and $f(\mathbf{x}, z)$ assign 71% and 45% of the households into a component with probabilities greater than 0.9. In both analyses, the first two components comprise proportions of households (53% and 40% for $f(\mathbf{x})$ and 42% and 34% for $f(\mathbf{x}, z)$) that are smaller than the stereotype for the first component and greater for the second component. The second components have means and variances very similar to the respective first components for both models; the components differ substantially only in their correlations. All the correlations in both solutions exceed 0.95 for the first component, and are in the range 0.56–0.94 for the second component.

The Netherlands stands out as having exceptionally small variances in the first two components for both three-component solutions. They are in the range 0.21–0.22 in the first components in both solutions, and 0.25–0.38 in the second components. The marginal probabilities of the first two components are not exceptional (24% and 30% of households in the second components), and so the small variances are not an artefact of the allocation of households to components. In Section 3.4, we identify the Netherlands as the country with the highest stability of household income based on a criterion designed without a reference to any models. The mixture components are particularly well separated for the Netherlands; the allocation of the households to components is with probabilities exceeding 0.9 for 84% and 73% for the respective three-component mixture models for $f(\mathbf{x})$ and $f(\mathbf{x}, z)$.

The three components in the solution for the UK are well separated for $f(\mathbf{x})$, and conform to the stereotype. However, in the solution for $f(\mathbf{x}, z)$, the first two components are separated much less well. Their fitted means are very similar, and the variances for the second solution are only slightly higher than for the first (0.37–0.39 *vs.* 0.36–0.48). Only 37% of the households are assigned to a component with probabilities greater than 0.90. The lack of separation can be attributed to the second component in particular. Although the second component is much more numerous than the third (30% *vs.* 18%), only 36 households (0.9%) have pattern 090, compared to 326 households (7.8%) who have pattern 009. Of course, the patterns that indicate that the second component is likely are quite frequent, especially the patterns in which the first two components are vying for the household and the third component can be ruled out.

3.3 Solutions with more components

Greater numbers of components yield much better fits to the data for all the countries, but they are more difficult to interpret and have fewer features in common. As the number of estimated parameters grows with the number of components, their tabular display is not very useful for recognising patterns or comparing countries. Figure 11 gives a graphical

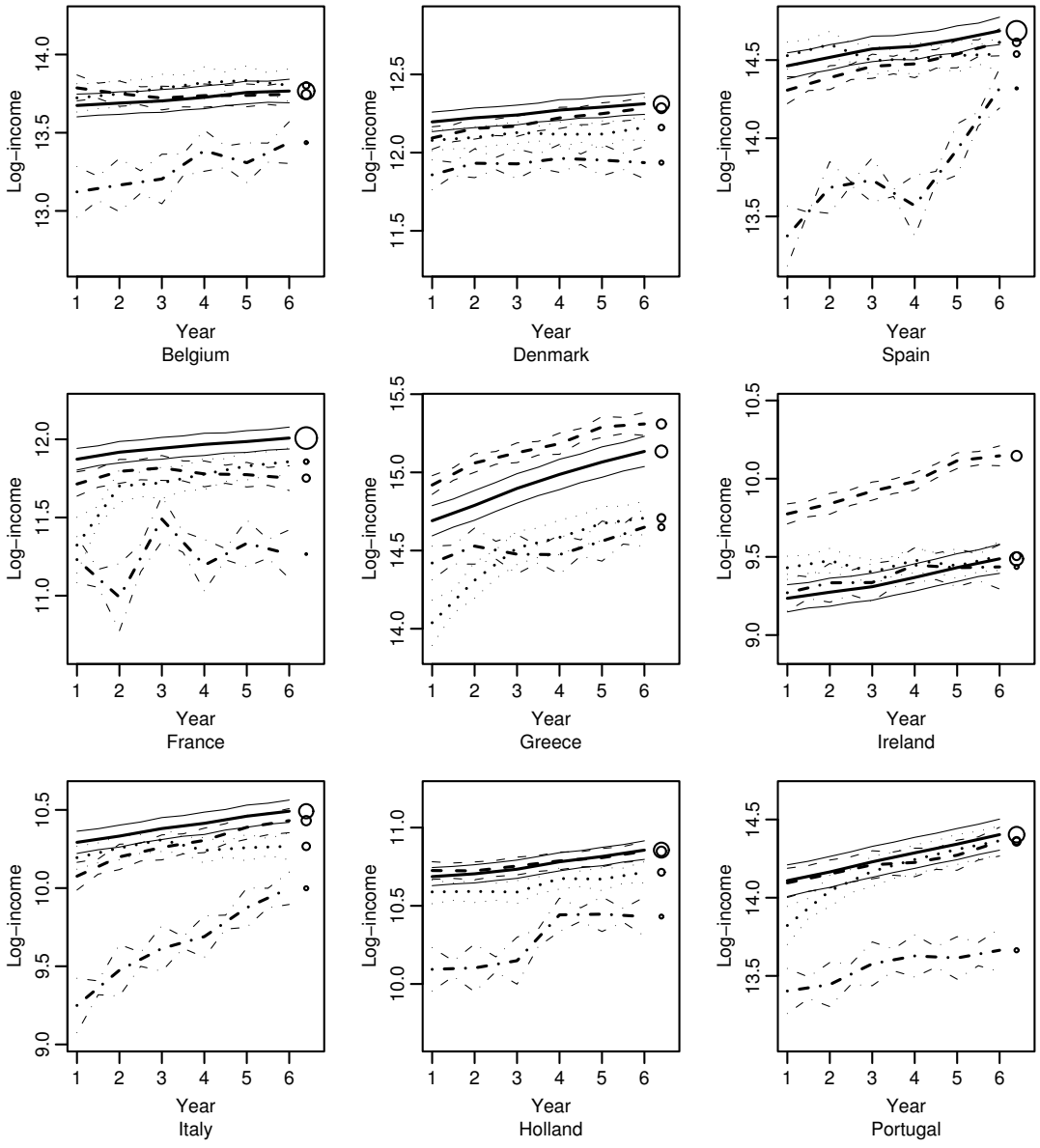
summary of the four-component model fits for $f(\mathbf{x}, z)$ for the nine countries that have data from each year of ECHP. In each panel, the estimated means are drawn by thick lines, each component with a different line type. The thin lines of the same type are drawn one-eighth of the estimated standard deviation above and below the mean at points (years) 1, 2, ..., 6, but are 'cut-in' at 1.5, 2.5, ..., 5.5, to indicate the covariance of the log-incomes for the consecutive years by the distance between either thin line and the mean. For example, if the estimated correlation between years 1 and 2 is close to unity, the lines are nearly straight between the years. If the estimated correlation is zero, the thin lines approach the mean linearly interpolated at 1.5. Finally, the circles on the extreme right of each plot have diameters proportional to the estimated marginal probabilities. The panels have the same scale on the vertical axis, equal to 1.67; the top represents 5.3 times higher income than the bottom of the axis. Relative homogeneity of the incomes can be inferred from the panels, although the size of the components has to be taken into account in any comparisons. Thus, households in Spain, Portugal, Belgium, and the Netherlands have relatively homogeneous incomes, especially when the small fourth components in the former two countries are discarded. Households in Greece and Ireland have the most heterogeneous incomes. The correlations among the non-consecutive years are the only information about the model fit that is not represented in the graphs.

The diagram shows that the fourth component has the smallest means for most countries (Ireland and Luxembourg are the only exceptions) and the smallest estimated correlations. For five and more components, these diagrams become rather cluttered when presented in the same format as in Figure 11, but are more presentable when each panel (country) is on a separate page. In some multi-component solutions, the group with very weakly correlated incomes is immediately recognised. For most countries, it has the smallest marginal probability, the smallest means and the largest variances for each year. Problems with model fitting arise for six and more components. They can be traced to singularities in the logistic regression for a component, for example, for Luxembourg in the six-component solution.

3.4 From models back to the data

An important finding of our analysis is that each country contains a large proportion of households whose incomes differ little from the pattern of small annual increases, akin to adjustment for (wage) inflation. The size of this subpopulation may be of interest as an indicator of the income stability of the households. Note that it is affected not only by the changes in income but also by the rates and extent of changes in the composition of households. Characterising this subpopulation by the first component of the fitted mixture models is problematic because its composition (and probabilities of belonging to it) are

Figure 11: The four-component mixture model fit for $f(\mathbf{x}, z)$ for countries with six years of data. In each panel, the distinct line types (solid, dashes, dots, and dots-and-dashes) are used for the components in the descending order of the marginal probabilities.



affected by the other two components. If the two 'minor' components are very distant from the dominant component the dominant component attracts more households into its domain. Also, the three-component solution is not 'valid', in the sense that further components would substantially improve the model fit for every country. A definition of income stability that is independent of models might therefore be preferred.

We regard an income progression with annual increases equal to the national average as the ideal of stability. For two consecutive years, we define the (multiplicative) deviation from this ideal,

$$d_{i,t} \;=\; \frac{x'_{i,t+1}}{x'_{i,t}}\,,$$

where $x'_{i,t}$ is the income of household i in year t, adjusted so that $x'_{i,t}$ have the same mean for every year t. We characterise the stability of income of a household by the maximum of the absolute log-deviations,

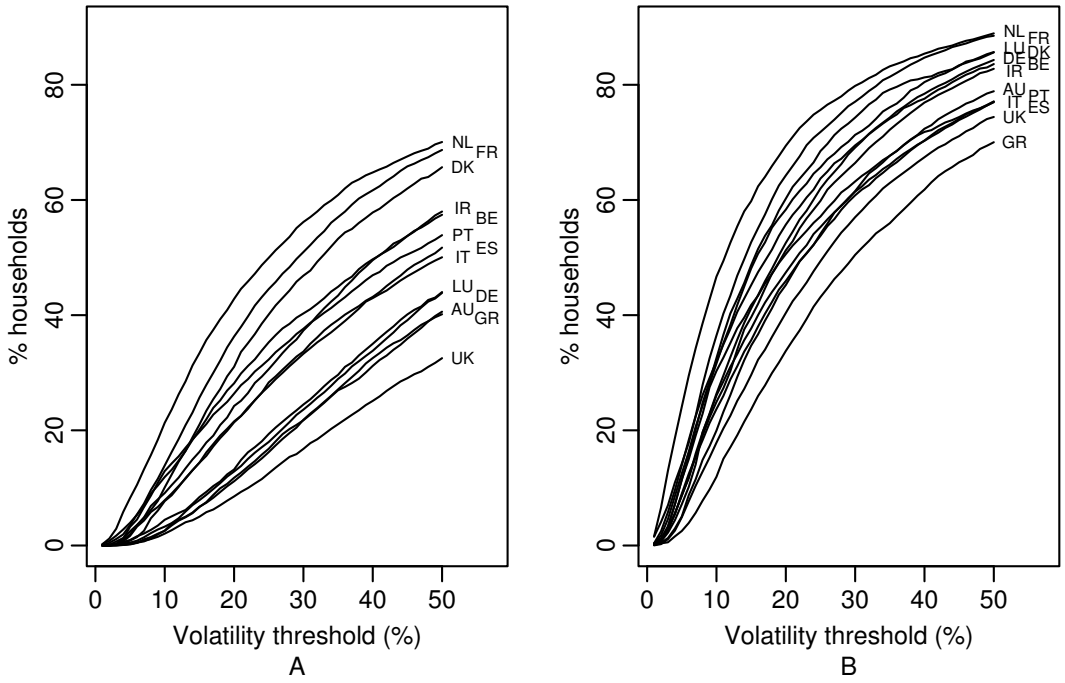
$$D_i \;=\; \max_t \,|\log(d_{i,t})|\;, \tag{4}$$

and define the country's index of stability as the percentage of households whose value of D_i does not exceed a set *volatility threshold*. Instead of setting a single threshold, we consider these percentages for a range of thresholds. The resulting stability functions (curves) for the years 1995–1999 for the 13 countries with data from this period (all except Finland), are plotted in panel A of Figure 12. They show that the Netherlands, France and Denmark had the highest percentages of households with stable income for any reasonable volatility threshold (say, 20–50%) and the UK the lowest. The countries' stability curves do not intersect much, so their ranking is affected only slightly by the choice of the volatility threshold. However, the stability has to be qualified by the number of years.

Similarly defined curves can be plotted for the period 1994–1999 for the nine countries that have panel data also for 1994. The order of the stabilities in that plot is not changed, although the relative differences between the countries are altered somewhat. The curves can be smoothed, although the effect of smoothing would be mainly cosmetic.

Earlier we observed that some households have stable incomes, except for one year. In a less stringent definition of the stability of a household's income, one year with an extreme income is discarded and equation (4) is applied to the remaining values. The resulting stability curves are plotted in panel B of Figure 12. The percentages are much higher than their counterparts in panel A. The curves in panel A fan out; the differences between the extremes grow with the volatility threshold, except for the thresholds near 50%. In contrast, the stability curves in panel B fan out only till the threshold of about 20%, and then their differences tend to decrease. For the volatility thresholds beyond 30%, the curves in panel B are clustered much more closely than in panel A. The order of the

Figure 12: Plot of income stabilities for all years 1995–1999 (panel A) and with one year excluded (panel B).



countries is changed somewhat. At the threshold of 50%, the increases *vis-à-vis* panel A exceed 40% for Germany, Luxembourg and the UK, whereas for Denmark, France and the Netherlands, the countries with the highest levels of stability in panel A, they are below 20%.

A measure of volatility of the household income, complementary to the definition of stability, can be based on the counts of the number of changes in the adjusted annual incomes that are greater than a set threshold.

4 Discussion

Our analysis of household income has emphasised the multidimensional nature of the problem. Exploring the income in one year, or its differences (ratios) in consecutive years, could not point to the presence of a substantial subpopulation with steadily rising household income. Yet, the multivariate analysis is not any more complex than the set of univariate analyses, especially in the frequentist framework when no (multivariate) prior distributions have to be specified for the numerous model parameters.

We have shown that the mixture models identify subpopulations with distinct covariance structures, and differences in the location (means) are secondary. The component with the highest representation in each country corresponds to households with relatively

low variances and very high inter-year correlations. Such inference could not be drawn from (sequences of) lower dimensional analyses.

We have chosen to compose the distribution of income from log-normal densities because of the convenience of the normal distribution — its analytical form, its description by its first two moments, and the full range of correlation structures that it caters for. The log-transformation applied conforms with a strong tradition in analysing income, and the transformed data are much closer to normality. Also, the multiplicative scale is more natural for the various comparisons.

Kernel density estimation is an alternative to mixture modelling, feasible for low-dimensional problems. By assuming a particular dependence structure, the multivariate distribution of the income could be reconstructed from the bivariate kernel density estimates. Kernel estimation in two dimensions is more flexible than mixture modelling with a few components; the bandwidth and the correlation in the kernel lead to estimated densities with a wide range of smoothness. Although there are algorithms for 'optimal' bandwidth, the choice of the bandwidth should be guided by the purpose of the smoothed density. For some purposes, a lot of detail is desirable. In contrast, our focus has been on the gross features of the income distribution, for which a wide bandwidth would be better suited. A distinct disadvantage of the kernel estimation is that the estimated density cannot be related directly to any subpopulations. In mixture models, in which such subpopulations are assumed, with unknown distributions, the assignment of units to subpopulations is subject to uncertainty, described by the conditional probabilities obtained in the concluding E-step.

Mixture model fitting attempts to reconstruct the observed distribution from normal distributions. Their role is similar to kernel smoothing, as discontinuities and sudden changes in the empirical density are smoothed over. In fact, the result of kernel smoothing is also a mixture model fit, with a component corresponding to every fitted point (Scott and Szewczyk, 2001). With distinctly non-normal data, the interpretation of the mixture components is problematic because of the reliance on the normal distribution as the building block for a more complex distribution. After the log-transformation of income, the reconstruction of the empirical distribution is much easier, but uncritical interpretation is still not warranted. Some observed features of the mixture model fit may have an alternative explanation related to the mechanics of model fitting. For example, components accounting for smaller fractions of the sample are likely to have more extreme features, such as unusual patterns of the means, high variances and low correlations.

By interpreting mixture model fits as a form of density smoothing, we set aside the issue of the number of mixture components. The conventional hypothesis testing by LR decisively points to more than three components for all the analysed countries. However,

the model fits with more components only confirm the presence of a majority subpopulation with steadily rising income and separate small subpopulations, some with esoteric patterns of income. These are exceedingly difficult to interpret because, being based on effectively small sample sizes, they are subject to large sampling variation. The standard errors for the parameters in mixture models cannot be estimated straightforwardly. Their completed-data versions, obtained from the concluding M-step, underestimate them. The bootstrap offers a general solution. Difficulties may arise when the mixture components cannot be identified unambiguously in the replications. We have used the mixture model fits informally, relying on the large sample sizes, and designed an index of stability that reflects our findings, but is not related to the models otherwise. The resulting summary in Figure 12 can easily be associated with (pointwise) standard errors derived from the binomial distribution. For example, the estimated percentage of 70% of households for the Netherlands in panel A is associated with the estimated standard error $100\sqrt{0.21/3407} \doteq 0.8\%$.

Outliers could in principle be identified as forming large-variance components with small representation. In our context, outlying vectors of income are assigned to the minority components with high probabilities because they are more typical among vectors with large variances and small correlations. This would not affect the substantial first component. The differences between the two panels in Figure 12 suggests that there are many households each with a single outlying annual income. This feature of the data should be regarded as noteworthy, although such combinations of values may result from an imputation that is subject to substantial uncertainty. In ECHP, imputation is applied for the components of income and the length of the related activity during the year, using a mix of deterministic and stochastic schemes (European Commission, 2002b). In general, deterministic schemes with good prediction (e. g., based on the data from the previous year) cause an over-representation of the stable-income component, and stochastic imputation with large residual variation causes an over-representation of the large-variance small-correlation components.

One or several categorical variables can be modelled simultaneously with a vector of continuous variables by fitting the conditional and marginal distributions as stated in (2) and (3). We defined a dichotomous variable that indicates whether the household had experienced a spell of unemployment of one of its members or nobody in the household derived any income from employment. Our definition is imperfect because it is based on subjects' recall for only one month each year. The variable has a perceptible and fairly consistent impact on the mixture model fit; although the principal features (the ordering of the means, variances and correlations) are largely maintained, the third component for most countries accounts for much greater percentage of households with this variable

than without it. The assignment of the households to components tends to be subject to more uncertainty with the categorical variable. The extension to more than one categorical variable, or a variable with a greater number of categories, is straightforward to implement, especially when no constraints are imposed on the parameters related to different components. However, the orientation among the many parameters becomes more difficult.

Whether a model fit allocates many subjects to a component with a high degree of certainty should not be regarded as a criterion for which variables (and models) to apply. The choice of the variables should be guided solely by the substance and purpose of the study. Different sets of variables define different subpopulations (components), because similarity (within components) has an essentially different meaning.

A problematic aspect of our analysis is the definition of the population of intact households. We define them by the presence in the data for all the relevant years. Thus, a household that failed to respond (or be contacted) in a wave is not regarded as intact. Although this rule resolves some practical difficulties, it lacks validity and gives rise to the problems associated with the analysis of complete cases (Little and Rubin, 1987). Addressing this problem, common to all large-scale surveys, is beyond the scope of this paper, although the method presented can be regarded as the complete-data method in a multiple-imputation approach. A difficulty in any imputation approach is that there is uncertainty even about the existence and intact nature of some of the households with missing information. Although the details of the results may be altered, the principal conclusion is unlikely to change.

Household income is often adjusted for the number of its members. As the adjustment is the same for every wave, the inferences about stability of income of intact households are unlikely to be affected. We have not adjusted the income for inflation or converted the amounts to a single currency. These adjustments are additive on the log-scale, and so they do not affect the variance matrices or the composition of the components. The components' means are adjusted, additively, by a constant vector and the sampling variation of the estimates is unaffected.

5 Conclusion

We have presented a multivariate analysis of data about household income in ECHP using mixture models based on the log-normal distribution. The principal finding is that the majority of households in each country studied have stable incomes, close to the pattern of constant income apart from an adjustment for inflation. This conclusion is based on models much simpler than those arrived at by conventional model selection. We justify this approach by confirming it empirically — identifying many households that have income

with the inferred patterns. More complex models provide a better fit, but they defy a simple description and offer little additional insight. In particular, the sets of parameter estimates (means, variances or correlations) for one component are not uniformly greater or smaller than the corresponding estimates for another component.

The EM algorithm applied to fitting the mixture models does not require an excessive number of iterations because the fraction of missing information (the assignment of households to the mixture components) is relatively small. Informally, this can be assessed by inspecting the fitted probabilities of belonging to the mixture components. The fraction tends to be greater for models with fewer components and for more complex within-component models, such as the combination of the logistic regression and multivariate log-normal marginal distribution.

A mixture model fit cannot be interpreted literally as a combination of subpopulations, unless the number of subpopulations is correctly specified and multivariate log-normality applies to the income in each subpopulation. Although the latter assumption is more palatable than the assumption of normality, the model fit is very sensitive to this assumption, and identifying the number of components with near certainty is a tall order by both data-driven methods and economic theory.

Having identified stability as an important statistical feature of the household incomes, we define a continuum of indices for its assessment free of any references to models. Variations of this definition may fit a particular purpose much more closely.

After extracting the relevant data, all the computing was carried out in R. The software developed (functions) can be obtained from the first author on request.

Acknowledgements

Research for this paper was conducted at CEPS/INSTEAD in Differdange, Luxembourg, where the authors were visitors under the Access to Research Infrastructure (ARI) Programme of the European Commission (Contract HPRI–CT–2001–00128). Philippe Van Kerm's assistance with background information and orientation in the ECHP database is acknowledged.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

Aitkin, M., and Wilson, G. T. (1980). Mixture models, outliers and the EM algorithm. *Technometrics* **22**, 325–331.

Bianchi, M. (1997). Testing for convergence: Evidence from nonparametric multimodality tests. *Journal of Applied Econometrics* **12**, 393–409.

Burkhauser, R. V., Cutts, A. C., Daly, M. C., and Jenkins, S. P. (1999). Testing the significance of income distribution changes over the 1980's business cycle: A cross-national comparison. *Journal of Applied Econometrics* **14**, 253–272.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society* Ser. B **39**, 1–38.

European Commission (1996a). European Community Household Panel (ECHP): Survey methodology and implementation, Volume 1. Eurostat, Luxembourg.

European Commission (1996b). European Community Household Panel (ECHP): Methods, Volume 1. Survey questionnaires: Waves 1–3. Eurostat, Luxembourg.

European Commission (2002a). ECHP UDB Manual. European Community Household Panel Longitudinal Users' Database. Waves 1 to 6, survey years 1994 to 1999. Document PAN 168/2002–12. Eurostat, Luxembourg.

European Commission (2002b). Imputation of income in the ECHP. Document PAN 164/2002–12. Eurostat, Luxembourg.

Flachaire, E., and Nuñez, O. (2002). Estimation of income distribution and detection of subpopulations: an explanatory model. Working Paper No. WS 030201, Departamento de Estadística y Econometria, University of Carlos III, Madrid.

Gelman, A., Carlin, B. P., Stern, H., and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall, London.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley and Sons, New York.

Marron, J. S., and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* **2**, 712–736.

McLachlan, G., and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley and Sons, New York.

McLachlan, G., and Peel, D. (2000). *Finite Mixture Models.* Wiley and Sons, New York.

Meng, X.-L., and van Dyk, D. A. (1997). The EM algorithm — an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society* Ser. B **59**, 511–567.

Paap, R., and van Dijk, H. K. (1998). Distribution and mobility of wealth of nations. *European Economic Review* **42**, 1269–1293.

Peracchi, F. (2002). The European Community Household Panel: A review. *Empirical Economics* **27**, 63–90.

Pittau, M. G. (2003). Fitting regional income distribution in Europe via finite mixture models. Submitted.

Quah, D. (1996). Empirics for economic growth and convergence. *European Economic Review* **40**, 1353–1375.

Richardson, S., and Green, P. J. (1997). On Bayesian analysis of mixtures with unknown number of components. *Journal of the Royal Statistical Society* Ser. B **59**, 731–792.

Scott, D. W., and Szewczyk, W. F. (2001). From kernels to mixtures. *Technometrics* **43**, 323–335.

IRISS-C/I is currently supported by the European Community under the *Transnational Access to Major Research Infrastructures* action of the *Improving the Human Research Potential and the Socio-Economic Knowledge Base* programme (5th framework programme)

[contract HPRI-CT-2001-00128]

Please refer to this document as

IRISS Working Paper 2003-08, CEPS/INSTEAD, Differdange, G.-D. Luxembourg

(CEPS/INSTEAD internal doc. #07-03-0027-E)