The Use Of Finite Mixtures Of Lognormal Distributions In The Modeling Of Incomes In The Czech Republic

Malá, Ivana

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

VOLUME 4, 2011

ISSN 2045-3345

ISSN 2045-3345

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

THE USE OF FINITE MIXTURES OF LOGNORMAL DISTRIBUTIONS IN THE MODELING OF INCOMES IN THE CZECH REPUBLIC

Ivana MALÁ
University of Economics, Prague, Czech Republic

ABSTRACT

Finite mixtures of probability distributions are used for the modeling of probability distributions of incomes. These distributions are typically non-homogenous, heavy tailed and positively skewed. In the text a net annual incomes per capita of the Czech households in 2004 and 2008 are analysed. The finite mixtures of lognormal distributions are fitted into data from the survey Results of the Living Conditions Survey (a national module of the European Union Statistics on Income and Living Conditions (EU-SILC)) that has been held by the Czech Statistical Office since 2005. Firstly, the components with known group membership are formed according to the education of a head of a household (factor with 5 levels) and number of children (2 levels factor children yes/no and more detailed 5 levels factor) in the household. Secondly, data are divided into groups with unknown group membership in order to obtain the best possible fit. In this case 1 to 5 components in the mixture are used. All models fitted into data are compared with the use of Akaike criterion.

JEL CLASSIFICATION & KEYWORDS

■ C13 ■ C51 ■ FINITE MIXTURE ■ LOGNORMAL DISTRIBUTION ■ INCOME MODELLING ■ CZECH REPUBLIC

INTRODUCTION

Studying and analysing incomes and wages is very important not only for experts in the field but also for general public. Characteristics of their levels (as values of the mean or median), characteristics of variability (standard deviation or coefficient of variation) and Gini index of inequality are frequently published and discussed from various points of view. In this article a method of mixtures is used for the estimation of distribution of annual income per capita in the Czech Republic and characteristics mentioned above are evaluated from these estimated distributions and compared with sample values. Lognormal distribution for components is used as it is known to be useful in the modelling of income or wage distributions (an overview of other ´income´ distributions as generalized gamma, beta or lambda distributions, Pareto or Weibull distributions can be found in McDonald (1984)). Many references of the topic are included in Kleiber, Kotz (2003)). The incomes in the Czech Republic with the use of lognormal distribution are analysed in Bartošová, Bína (2008), Bílková (2009) or Pavelka (2009). The last mentioned article by Pavelka shows the use of mixtures of lognormal distributions for wages in the Czech Republic, in Bartošová, Forbelská (2011) mixture of models for SILC data are used. The unknown parameters are estimated with the use of maximum likelihood method.

In the article data dealing with the Czech households for 2004 and 2008 are used. The set of all households is not homogenous, the households differ in structure (number of members, economically active members, pensioners, children etc.) as well as in economic activities or education of members. In the text lognormal mixture models with known component membership (complete data models) are fitted to incomes for subgroups given by education of a head of household and according to the existence of children or number of children in the household. Separate distributions can be found for these subgroups defined by explanatory variables as above and these distributions are mixed together in the overall distribution of the Czech households.

Moreover, mixture models with unknown component membership (incomplete data models) are constructed with 1 to 5 components. In this case group membership is not observable and observations are divided into subgroups in order to obtain the best fit. In this case lognormal distributions are used for components, distribution of a mixture is not any more lognormal. This approach to the analysis of distribution is similar to cluster analysis. For the clustering data into groups in this text no explanatory variables were used and probabilities of membership for data were not computed as the goal of the analysis is to identifie distributions in the mixture.

Methods

Finite mixtures of probability distributions

In this part the finite mixture of probability densities is defined and its properties that are used in this article are given (Titterington et al. (1985)). Suppose now, that for given K there are probability densities $f_j(y;\theta_j)$ (j = 1,., K) depending on p dimensional (in general unknown) vector parameter θ_j. Furthermore, K weights π_j fulfil obvious constraints $\sum_{j=1}^{K}\pi_j = 1,\ 0<\pi_j<1,\ j=1,..,K.$

A density of the mixture of these probability distributions is defined as a weighted average of densities f_j with weights (mixing proportions) π_j in the form

$$f(y;\psi) = \sum_{j=1}^{K} \pi_j f_j(y;\theta_j). \qquad (1)$$

The mixture density (1) depends on the vector parameter ψ, $\psi = (\pi_1,\ldots,\pi_{K-1},\theta_j,j=1,\ldots,K)$ with (K−1) parameters π_j and K_p parameters theta. If the probability distribution given by the formula (1) is used in a model, (K−1) + K_p unknown parameters are to be estimated. If all mixing proportions are supposed to be positive, all K components are present in the mixture. The choice of K is crucial for the proper model as well as probability densities f_j.

It follows immediately from (1) that a cumulative distribution function F of the mixture is defined as

$$F(y;\psi) = \sum_{j=1}^{K} \pi_j F_j(y;\theta_j), \qquad (2)$$

where $F_j(y;\theta_j)$ is a distribution function of the j-th distribution in the mixture. For an expected value of the mixture a formula similar to cumulative distribution function can be used and the expected value can be evaluated as a weighted average of the expected values of its components with weights π_j .

These simple formulas are not true for higher moments or for values of a quantile function. In the text standard deviation of the mixture is frequently used as well as quantiles. Let X_j is a random variable with density function fj, expected value $E(X_j)$ and finite variance $D(X_j)$, (j = 1,., K). Then formulas (3) and (4) valid

$$E(Y) = \sum_{j=1}^{K} \pi_j E(X_j), \quad (3)$$

$$D(Y) = \sum_{j=1}^{K} \pi_j E(X_j^2) - (E(Y))^2 = \sum_{j=1}^{K} \pi_j \left(D(X_j) + (E(X_j))^2\right) - (E(Y))^2. \quad (4)$$

The 100P% quantile y_P can be found as a solution of an equation

$$F(y_P;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j F(y_P;\boldsymbol{\theta}_j) = P,\ 0 < P < 1. \quad (5)$$

This equation should be solved numerically, as the weighted average of component 100P% quantiles is not equal to y_P. But it is a good first guess for numerical procedure. This approach was used in this text when selected quantiles were evaluated.

Likelihood function (from a sample y_i, i=1,.,n) can be written as

$$L(\psi) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j f_j(y_i;\boldsymbol{\theta}_j). \quad (6)$$

Suppose now, that the random sample arises from a population divided into K subpopulations and for each observation y_i the subpopulation j is observed together with the value. Data of this type are called complete. In this case i-th observation's contribution to the function L is only $\pi_j f_j(y;\theta_j)$ (if this observation comes from the j-th subpopulation). The likelihood function (6) can be then rewritten in the form (according to Titterington et al. (1985))

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{n} \prod_{j=1}^{K} \pi_j^{z_{ij}} f_j(y_i;\boldsymbol{\theta}_j)^{z_{ij}}, \quad (7)$$

where z_i are known 0/1 vectors with K components and z_{ij} is equal to 1 if i-th observation comes from the j-th density and 0 otherwise. The vector $\sum_{i=1}^{n} z_i$ contains subgroup frequencies (number of observations in each subgroup). Taking logarithm in (7), the logarithmic likelihood function l can be written in the form

$$l(\boldsymbol{\psi}) = \ln L(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \ln \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \ln f_j(y_i;\boldsymbol{\theta}_j). \quad (8)$$

The function l in (8) splits into two parts, the first part depends only on mixing proportions and the second one only on parameters of probability densities (values z_{ij} are known, as we suppose, that data are complete). Both parts in (8) can be maximized separately. Maximum likelihood estimates of proportions are sample relative frequencies of components and estimates of parameters of the component densities can be found as maximum likelihood estimates in each subgroup.

If the group membership is not known, the logarithm of (6) is equal to

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{K} \pi_j f_j(y_i;\boldsymbol{\theta}_j) \right).$$

In this case the logarithmic likelihood function cannot be split into parts as in (8) and the function is usually maximized with the use of EM logarithm (Pavelka (2009)). This is a numeric procedure that consists of two steps. First step is called Expectation (probabilities πj are estimated) and the second one Maximization, where estimated values from the first step are used in order to found new approximations of parameters theta. These two steps are repeated until a solution is found. Generally, EM algorithm doesn´t guarantee absolute maximum of the logarithmic likelihood function but only the local extreme (Titterington et al. (1985)). Moreover for higher number of components it can be time consuming and even after thousands of steps convergence may not occur. It is the case of poor initial approximations as it was in this text.

All estimates in the text are maximum likelihood estimates and in order to compare different fits, Akaike criterion was used in the form

AIC=−2*l(ψ) + 2*number of parameters (9)

If different models are compared, the smaller the value of AIC the better fit.

All estimated characteristics of distributions based on maximum likelihood estimates of unknown parameters are also maximum likelihood estimates and have all theoretical properties of such estimates.

Lognormal distribution

For the modeling of distribution of incomes, the lognormal distribution is frequently used with satisfactory results. In this paper two-parametric lognormal distribution is used for densities fj and components differ only in the parameters. Under these assumptions the random variable Y has a distribution given by a mixture density

$$f(y;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j f_j(y;\mu_j,\sigma_j^2) = \sum_{j=1}^{K} \frac{\pi_j}{\sqrt{2\pi}\sigma_j y} \exp\left(-\frac{(\ln y - \mu_j)^2}{2\sigma_j^2}\right).$$

The vector of parameters ψ has (K−1) + 2K components (π_j, μ_j, σ^2_j,j =1,…K).

Known formulas for distribution function, quantile function and moments of lognormal distribution (Johnson et al. (1994)) are used in (2) - (5) in order to evaluate characteristics of the mixture.

The estimates $\hat{\pi}, \hat{\mu}, \hat{\sigma}^2$ of unknown parameters in (8) can be evaluated as (j =1,.,K)

$$\hat{\boldsymbol{\pi}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_j, \hat{\mu}_j = \frac{1}{n}\sum_{i:z_{ij}=1} \ln y_i, \sigma_j^2 = \frac{1}{n}\sum_{i:z_{ij}=1} (\ln y_i - \mu_j)^2.$$

For the incomplete data, a package flex mix (Grün, Leisch (2008)) in program R 2.13.1 was used for the maximization of the logarithmic likelihood function l. The package estimates parameters for mixtures of normal distributions (mixing proportions, expected values and standard deviations of normal distributions). This program was used for the logarithms of analysed incomes. Values of logarithmic likelihood function were evaluated in original incomes, not in their logarithms.

Data and results

In this part of the article the concept of mixtures of lognormal distributions from previous part is used to the modeling of incomes of the Czech households. Data from EU-SILC (European Union – Statistics on Income and Living Conditions) survey from two years 2005 and 2009 were used. The survey has been held by the Czech Statistical Office yearly since 2005, the survey EU-SILC 2005 refers to the incomes from 2004 and EU-SILC 2009 to 2008. The aim of the survey is to gather representative data on income distribution for the whole population and for various household types. For each household in the sample an annual income per capita (in CZK) was evaluated as a ratio

of a total of all incomes (net) and a total of members of the household. All incomes in the text are in CZK, average rates were 1Euro=31.90 CZK in 2004 and 24.94 CZK in 2008. Suppose that the income of a household per capita is the random variable Y with mixture distribution discussed in the previous parts. The survey from 2005 consists of 4,341 households, in 2008 there were 9,911 households included in the sample. In this text the households are divided into subgroups according to education of a head of household (5 levels – the head with primary (or without any education) (B), secondary and vocational (without leaving exam) (S), complete secondary (CS), tertiary up to baccalaureate (BS), university education with the magister or PhD titles (MS)). This model is referred as education. In this text only the impact of education of the head of household is analysed without taking into account education of other members (especially of the partner of the head of the household). The head of household for couples with or without children is always the male, regardless of his economic activity. In lone-parent families or non-family households the first criterion for determining of head of household was economic activity and the secondary criterion was income of household members (CZSO).

Number of children in the household is used as a second explanatory variable. Two models are constructed: one model with only two components (households with children and without children) and more detailed division with 5 components (number of children 0-3 and more than 3). One can expect these groups to be suitable for improving the fit. Data are complete in all these models and estimation of unknown parameters was performed with the use of formulas given above. These two models are referred as childer2 and children 5.

Moreover mixtures of one to five components with unknown group membership (models with incomplete data) were fitted into the sample. In this text the estimated values of unknown parameters are not given. We will concentrate on the quality of fits and the analysis of given or estimated subgroups. These models are referred K=1 to K=5.

In the Table 1 quality of fits is compared for all 8 models mentioned above. In the Table values of logarithmic likelihood function (for estimates in the model) and Akaike criterion are given for all models. In order to obtain comparable results (in each analysed year separately) all values are evaluated from the original sample, not from logarithms of income. It means that values in Table 1 are not those given by flexmix program.

The fit of two parametric lognormal distribution into data sets can be seen for incomplete data and K=1. This fit is supposed to be really unsatisfactory. In the case of complete data we obtain information about the distribution of different groups but as it can be seen in the Table 1 the resulting mixture density provides not generally better fit to data than the two-parametric lognormal distribution. For the division of households according to number of children the resulting fit is worst (in comparison by AIC) than two parametric lognormal distribution. The division given by the education of a head of household is for both analysed years better even in comparison with subgroups with unknown group membership. For both years the best fit from incomplete data was met with the choice K=4. In case of 5 components the numeric procedure took really a lot of steps to obtain maximum likelihood estimates of (4+10)=14 unknown parameters and it were necessary to pay attention to the choice of initial approximation of the parameters. The combination of random group membership (provided by flex mix package) and the membership guessed from order values of incomes was used and the numeric procedure was performed from more initial guess, the higher number of components K, the greater number of fits and iterations and so the longer time to perform the analysis.

Table 1: Quality of fits in 2004 and 2008

	2004		2008	
Model	−l	AIC	−l	AIC
children2	55,169	110,349	133,473	259,606
education	49,727	99,481	115,08	230,186
K=1	52,785	105,575	122,297	244,598
K=3	52,508	105,031	121,526	243,067
Model	−l	AIC	−l	AIC
children5	56,503	113,033	129,789	260,956
K=5	52,502	105,032	121,52	243,159
K=2	52,534	105,078	121,63	243,669
K=4	52,502	105,026	121,509	243,04

Source: own computations

In Table 2 estimated mixing proportions are given for all models. For the models with complete data these values are relative frequencies of subgroups in the sample. For models with incomplete data estimated values of component probabilities are shown. For the model children5 proportions of the last group of households with more than 3 children (0.005) are not shown in the Table. This group is very small even in the large samples in this analysis. For incomplete data subgroups are sorted in increasing order according to estimated parameters μ. Expected value of the lognormal distribution depends not only on this value but also on σ^2 and the expected values of components in the table are not always ordered from the lowest to the highest. In the case of 3 artificial groups for incomplete data we can see similar structure of subgroups, for 4 components there is an increase in the second group (+7%) and in the group of highest incomes (+2.1%). Similar values are for groups according to education and even for models based on number of children.

Table 2: Estimated mixing proportions for mixtures

	K=2		K=3			K=4			
Year	j=1	j=2	j=1	j=2	j=3	j=1	j=2	j=3	j=4
2004	0,37	0,63	0,259	0,616	0,125	0,245	0,372	0,371	0,012
2008	0,624	0,376	0,275	0,613	0,112	0,231	0,439	0,297	0,033

	Education					Children				
Year	B	S	CS	BS	MS	No	Yes	1	2	3
2004	0,127	0,451	0,293	0,015	0,114	0,674	0,326	0,148	0,148	0,025
2008	0,125	0,452	0,297	0,017	0,109	0,689	0,311	0,143	0,137	0,026

Source: own computations

In the Tables 3-5 the estimated characteristics of the level (mean and median, first part of the table) and variability (standard deviation and coefficient of variation, second part of the table) of all subgroups are given in order to analyse and compare them. In the Tables 3 and 4 results obtained from complete data are given, in the Table 4 these characteristics are shown for incomplete data. In the Table 3 we can see that it is worth studying or at least to live in a household with a head with high education. All results are in nominal values of incomes. The inflation rate from 2004 to 2008 was (CZSO) 1.1413. For example the estimated expected value (year 2004) of income per capita for households with the head with magister education multiplied by inflation gives 181,552 CZK, the nominal value (Table 3) is 199,691 CZK. In addition to estimated values of characteristics an increase (in % of 2004 value) in all characteristics (except for coefficient of variation) is given in the Tables 3 and 4.

Table 3: Estimated characteristics of the level and variability of components (CZK). The complete data, households divided according to education of a head of household

	Expected value					Median				
Year	B	S	CS	BS	MS	B	S	CS	BS	MS
2004	89,457	99,113	116,285	131,421	159,075	84,288	91,309	104,611	114,921	139,246
2008	119,826	130,207	152,848	183,481	199,691	112,308	121,905	139,944	159,692	175,606
%	34	31	31	39	25	33	34	34	39	26
	Standard deviation					Coefficient of variation				
2004	31,804	41,844	56,450	72,909	87,862	0.372	0.375	0.439	0.566	0.541
2008	44,574	48,866	67,134	103,813	108,111	0.391	0.453	0.412	0.452	0.348
%	40	17	19	42	23	-	-	-	-	-

Source: own computations

The subgroup referred as children No is the subgroup in both models dealing with children. It is a subgroup in the model chidren2 as a group of household without children and in the model children5 a subgroup with number of children 0.

In the Table 4 negative impact of number of children in the household on incomes is obvious. This fact could be reduced in case of the use of equivalised incomes (CZSO) instead of incomes per capita. In the methodology of European Union the head of household has weight 1, other adult members of household weight 0.5 and children 0.3. It means, that a complete family with two children has total of weights 2.1=1+0.5+2 . 0.3 instead of 4. For a complete family with four children the difference between EU weight and number of members is even higher with 2.7 instead of 6. Than total income of this household is divided by 2.7 to obtain equivalised income and by 6 to obtain per capita income. Second problem is very small number of observations of households with 3 and more than 3 children. This fact makes results for these groups of households very imprecise.

Table 4: Estimated characteristics of the level and variability of mixture components (CZK) for complete data, components according to number of children

	Expected value						Median					
Year	No	yes	1	2	3	≥4	no	Yes	1	2	3	≥4
2004	120,625	86,670	97,968	81,195	58,858	56,641	111,748	77,497	87,641	73,865	53,637	53,423
2008	154,518	118,620	136,123	107,625	89,759	65,064	143,918	107,581	123,995	99,509	81,797	61,451
%	28	37	39	33	53	15	29	39	41	35	56	15
	Standard deviation						Coefficient of variation					
2004	49,026	43,398	48,940	37,059	26,593	19,952	0.41	0.50	0.50	0.46	0.45	0.35
2008	60,386	55,098	61,658	44,346	40,554	22,635	0.39	0.46	0.45	0.41	0.45	0.35
%	23	23	23	20	52	13	-	-	-	-	-	-

Source: own computations

In the Table 5 estimated values of location (first part of the table) and variability (second part of the table) characteristics are shown for artificial components in incomplete data problem. The components are ordered from the lowest expected value to the highest expected value. The interpretation of these groups for K=2 divide households into subgroups with lower and higher per capita income, for K=3 groups with low, medium and high per capita income and in case of K=4 the group of households with very high per capita income appears. The estimates of probabilities of components are given in the Table 2 and the structure of groups according to the level of per capita income is obvious from the table. Expected value of the low income group for 3 components increased by 25 per cent (from 2004 to 2008), the group of medium incomes by 33% and the group of high incomes by 36 per cent. Standard deviation increase was higher, in these groups it was 42, 13 and 45 per cent. Standard deviations in particular groups increase with the expected value. Relative variability (relative to the expected value) is smaller for groups of households with lower incomes then for higher income households with coefficient of variance greater than 100 per cent, in 2008 for the four components model the standard deviation is 140 per cent of the expected value for the group of the highest incomes per capita. The only decline in expected value or median occurs in 4 components model in the group of households with very high incomes. Absolute variability decreased, coefficient of variation increased from 110% to 141%.

Table 5: Estimated characteristics of the level and variability of mixture components (CZK) for incomplete data, K=2, 3, 4

	Expected value				
	K=2		K=3		
Year	j=1	j=2	j=1	j=2	j=3
2004	96,967	118,081	95,613	109,866	145,136
2008	128,551	171,787	119,535	146,527	197,689
	K=4				
	j=1	j=2	j=3	j=4	
2004	95,1	113,336	110,616	378,488	
2008	118,064	141,862	157,71	268,866	
	Standard deviation				
	K=2		K=3		
year	j=1	j=2	j=1	j=2	j=3
2004	15,812	71,218	12,192	51,413	135,797
2008	30,9	114,208	17,303	58,079	196,849
	K=4				
	j=1	j=2	j=3	j=4	
2004	11,838	72,4	43,475	416,675	
2008	15,892	45,818	92,747	377,762	
	Median				
	K=2		K=3		
Year	j=1	j=2	j=1	j=2	j=3
2004	95,703	101,114	94,845	99,509	105,979
2008	124,991	143,057	118,302	136,216	140,084
	K=4				
	j=1	j=2	j=3	j=4	
2004	94,372	95,511	102,95	254,485	
2008	117,008	134,996	135,944	155,905	
	Coefficient of variation				
	K=2		K=3		
year	j=1	j=2	j=1	j=2	j=3
2004	0,16	0,6	0,13	0,47	0,94
2008	0,24	0,66	0,14	0,4	1
	K=4				
	j=1	j=2	j=3	j=4	
2004	0,12	0,64	0,39	1,1	
2008	0,13	0,32	0,59	1,41	

Source: own computations

In the Table 7 the estimated characteristics of the level and variability of corresponding mixture distributions are shown for 6 fits (results are given only for incomplete data with two to four components). All the models (in each year) are fitted into same data and the estimated values in the Table 7 can be compared to sample characteristics given in Table 6. From the table we can see that expected values evaluated from all fits are very similar and characterise well the sample values. The same is true for the medians, but it is not the case of standard deviations. Standard deviations of all fits underestimate (some of them remarkably) sample standard deviations. The best agreement between sample and estimated standard deviations is for the 4 components incomplete data model. The model with subgroups defined

with the use of education of the head of household is good according to the value of AIC criterion, but the model doesn´t express well standard deviations.

In both tables 6 and 7 an increase of all characteristics is given (in % of the value from 2004). The percentages for location characteristics are similar (approximately 30 per cent, with one exception of 45% for the model defined by number of children in the household).

Table 6: Sample characteristics of location and variability (CZK)

Year	Mean	Median	Standard deviation
2004	111,024	97,05	77,676
2008	145,277	126,595	93,397
%	31	30	20

Source: own computations

Table 7: Estimated characteristics of the level and variability of per capita income (CZK) for the complete data (first part) and incomplete data for K=2, 3, 4 (second part)

	Education (5 levels)			Children (2 levels)		
year	E(Y)	y0.5	$\sqrt{D(Y)}$	E(Y)	y0.5	$\sqrt{D(Y)}$
2004	110,238	97,39	56,671	109,556	100,953	49,873
2008	144,113	129,487	68,34	143,354	132,969	61,095
%	31	33	18	31	31	23
	K=2			K=3		
2004	110,269	97,463	58,239	110,583	97,101	64,649
2008	144,808	128,246	77,063	144,834	126,806	83,55
%	31	32	32	31	31	29
	Children (5 levels)					
year	E(Y)	y0.5	$\sqrt{D(Y)}$			
2004	109,572	97,959	49,971			
2008	143,267	142,091	61,305			
%	31	45	23			
	K=4					
2004	111,041	97,143	75,442			
2008	145,263	126,814	94,711			
%	31	31	26			

Source: own computations

In the Figure 1 the estimated probability densities for selected models (children5, education5, artificial components K=2-4) are shown for 2004 (left part of the figure) and 2008 (right part of the figure). For both years the estimated density from the fit with incomplete data is closed to sample one even for only 2 components. The fits from complete data are similar to the density obtained from single lognormal distribution (not shown in the figure). In the figure also the modification of 2004 curves into 2008 curves can be assessed.

Conclusions

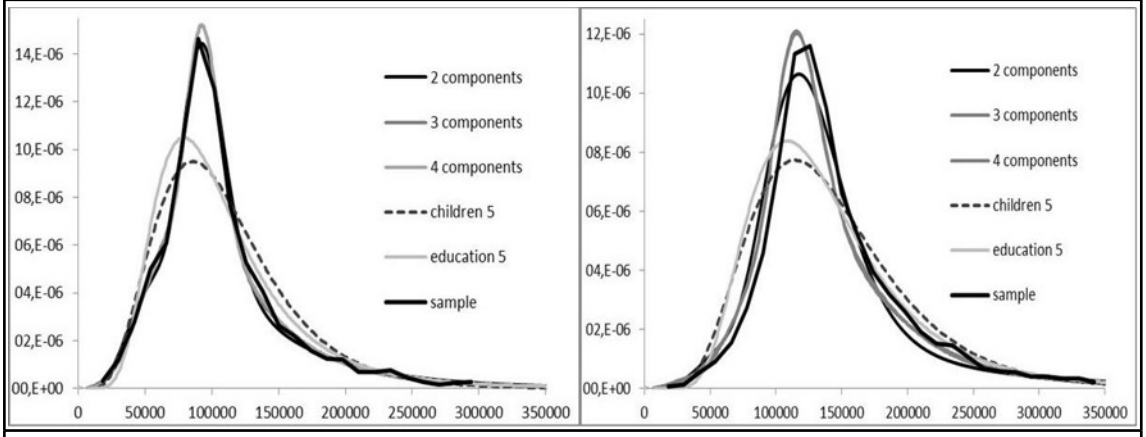
In the paper the use of the mixtures of lognormal distributions is proposed as a suitable model for annual per capita incomes in the Czech Republic. The expected as well as strange properties of the models are described and quantified in the text. The goal of the text is a description of component distributions. Differences in the use of mixture models with known and unknown group membership are illustrated.

The concept of mixture distributions is well applicable to income data, as these values form usually very non-homogenous data sets. In some cases we have information about a person´s (or a household) characteristics as education, age or location. In this case it is useful to take into account only the distribution of incomes in the particular subgroup, because such information frequently determines well personal income.

If data are divided into subgroups according to a known explanatory variable, we obtain information about subgroups and additionally these distributions can be weighted into a distribution for the whole sample. This model doesn´t ensure better fit even in case of subgroups with rather different shapes of distributions. In the text this fact was quite apparent in the models that use number of children in the household.

In case of incomplete data, where the group membership is not known and observations are clustered into artificial groups, the numerical algorithm searches for a chosen number of homogenous subgroups in order to obtain optimal model. These subgroups are artificial and there is no easy interpretation of the model. For the values in the sample we don´t have group membership even after the proces of estimation, only probabilities for the groups can be estimated. This concept is closed to cluster analysis, but we are interested in distributions in the clusters (and their size) more than in group membership of observations. The fitted model improves with every new component and the choice of the proper number of components is very important as well as the choice of suitable probability distribution for components. For too many components there are too many parameters in the model and numerical problems can

Figure 1: Estimated mixture densities in 2004 (left) and 2008 (right)



Source: own computations

occure, too few components don´t provide acceptable fit. In the case of too many components the procedure could become time consuming and even after many observations the solution is not obtained (the procedure doesn´t converge). It is sometimes difficult to clearly interpret subgroups in such complicated models.

In this text all component densities come from the same family of distributions (lognormal distributions) and differ only in parameters. In some applications the mixtures tha consists of components from different distributions can be succesfully used.

Acknowledgment

The paper was supported by a project IGS 24/2010 from the University of Economics in Prague.

References

Bartošová, Jitka and Vladislav Bína. Modelling of Income Distribution of Czech Households in Years 1996 – 2005. Acta Oeconomica Pragensia. Vol. 17. Iss. 4. 3 − 18. 2009.

Bartošová, Jitka and Marie Forbelská. 2011. Differentiation and dynamics of household incomes in the Czech EU-SILC survey in the years 2005-2008. In: Kováčová, M. (ed.), 10th International Conference APLIMAT 2011, Bratislava, February 1–4, 2011, Proceedings. Slovak University of Technology, Bratislava, 2011, s. 1451-1460.

Bílková, Diana. Application of Lognormal Curves in Modeling of Wage Distributions. Journal of Applied Mathematics. Vol. 1. Iss. 2. 341 − 352. 2008.

CZSO, Czech Statistical Office. www.czso.cz.

CNB, Czech National bank. www.cnb.cz.

Flachaire, Emmanuel and Olivier Nunez. Estimation of the Income Distribution and Detection of Subpopulations: an Explanatory Model. Computational Statistics & Data Analysis. 2007.

Grün, Bettina and Friedrich Leisch. Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. Journal of Statistical Software, 28(4):1-35, 2008.

Johnson, N. L., Narayanaswamy Balakrishnan and Samuel Kotz. Continuous Univariate Distributions. Vol. 1. New York: John Wiley & Sons, 1994.

McDonald, J.B. Some Generalized Functions for the Size Distribution of Income. Econometrica, Vol. 52, No. 3, 647-665, 1984.

Pavelka, Roman. Application of density mixture in the probability model construction of wage distributions, Applications of Mathematics and Statistics in Economy: AMSE 2009, Uherské hradiště, 2009, 341-350, 2009.

Titterington, D.M., A.F. Smith and U.E. Makov. Statistical analysis of finite mixture distributions, Wiley, 1985.