THE USE OF FINITE MIXTURES OF LOGNORMAL DISTRIBUTIONS IN THE MODELLING OF INCOMES OF THE CZECH HOUSEHOLDS

Ivana Malá

Abstract

Finite mixtures of probability distributions may be successfully used in the modelling of probability distributions of incomes. These distributions are typically 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.

Key words: finite mixture, income distribution, lognormal distribution, maximum likelihood estimate, EM algorithm

JEL Code: C13, C51

Introduction

Studying and analyzing 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 ones. 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 in McDonald, 1984). 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. The unknown parameters are estimated with the use of maximum likelihood method.

In the article data dealing with the Czech households for years 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 complete data are fitted to incomes for groups given by education of a head of a household and according to the existence 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, data are divided into groups with unknown group membership for 1 to 5 components.

1. Methods

1.1. 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 K probability densities $f_j(y;\boldsymbol{\theta}_j)$ (j = 1,..,K) depend on p dimensional (in general unknown) vector parameter $\boldsymbol{\theta}_j$. Furthermore, K weights π_j fulfil obvious constraints $\sum_{j=1}^{K}\pi_j = 1,\ 0 \le \pi_j \le 1,\ \ j = 1,.,K.$

The 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

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

The mixture density (1) depends on the vector parameter $\boldsymbol{\psi}$, $\psi = (\pi_1,.,\pi_{K-1},\theta_j,\ j = 1,.,K)$, with ($K$−1) parameters π_j and Kp parameters theta. If the probability distribution given by the formula (1) is used in a model, (K−1) + Kp unknown parameters are to be estimated. It follows immediately from (1) that a cumulative distribution function F of the mixture is defined as $F(y;\psi)=\sum_{j=1}^{K}\pi_j F_j(y;\theta_j)$, 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. If X_j is a random variable with density function f_j, expected value $E(X_j)$ and finite variance $D(X_j)$, (j = 1,., K), variance of Y with probability distribution defined by (1) can be computed as

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

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

$$F\left(y_P;\boldsymbol{\psi}\right)=\sum_{j=1}^{K}\pi_j F\left(y_P;\boldsymbol{\theta}_j\right)=P,\ 0<P<1. \qquad (3)$$

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

$$L(\psi)=\prod_{i=1}^{n}f\left(y_i;\psi\right)=\prod_{i=1}^{n}\sum_{j=1}^{K}\pi_j f_j\left(y_i;\psi\right). \qquad (4)$$

Suppose that the random sample arises from the mixture of K subpopulations and for each observation y_i the subpopulation j is observed together with its value. Data of this type are called complete. In this case i-th observation's contribution to the function L is only $\pi_j f_j\left(y_i;\theta_j\right)$ (if this observation comes from the j-th subpopulation). The likelihood function (4) can be then rewritten in the form (according to Titterington,1985)

$$L(\boldsymbol{\psi})=\prod_{i=1}^{n}\prod_{j=1}^{K}\pi_j^{z_{ij}}\ f_j\left(y_i;\boldsymbol{\theta}_j\right)^{z_{ij}}, \qquad (5)$$

where $\boldsymbol{z}_i$ 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 $\sum_{i=1}^{n}\boldsymbol{z}_i$ contains subgroup frequencies (number of observations in each subgroup). Taking logarithm in (5) the logarithmic likelihood function l can be written in the form

$$l(\boldsymbol{\psi})=\ln L(\boldsymbol{\psi})=\sum_{i=1}^{n}\sum_{j=1}^{K}z_{ij}\ln\pi_j+\sum_{i=1}^{n}\sum_{j=1}^{K}z_{ij}\ln f_j\left(y_i;\boldsymbol{\theta}_j\right). \qquad (6)$$

The function l in (6) 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 (6) 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 (4) 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 (6) 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 *E*xpectation (probabilities π_j are estimated) and the second one *M*aximization, 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).

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

$$\text{AIC}=-2*l(\boldsymbol{\psi}) + 2*\text{number of parameters} \qquad (7)$$

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

1.2. Lognormal distribution

For the modelling of distribution of incomes, the lognormal distribution is frequently used with satisfactory results. In this paper two-parametric lognormal distribution is used for densities f_j. Suppose that a random variable Y with distribution from (1) has 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 $\boldsymbol{\psi}$ has $(K-1) + 2K$ components $(\pi_j, \mu_j, \sigma_j^2, j = 1,.,K)$.
The estimates $\hat{\pi}, \hat{\mu}, \hat{\sigma}^2$ of unknown parameters in (6) can be evaluated as (j =1,.,K)

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

For the incomplete data, a package *flexmix* (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.

Furthermore characteristics of the mixture (expected value $E(Y)$, median and standard deviation) were evaluated as it was discussed in the part 1.1 with the use of known properties of the lognormal distribution.

2. Data and results

In this part of the article the concept of mixtures of lognormal distributions from previous part is used to the modelling 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 part 1. 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 a 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)). In this text only the impact of education of the head of the household is analysed without taking into account education of other members (especially of the partner of the head of the household). 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.

Moreover mixtures of one to five components with unknown group membership (incomplete data models) 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.

In the Table 1 quality of fits is compared for all 8 models mentioned above. 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 is 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 a household is for both analysed years better even in comparison with subgroups with unknown group membership. In 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 was necessary to pay attention to the choice of initial approximation of the parameters. The combination of random group membership (provided by *flexmix* 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.

Tab. 1: Quality of fits in 2004 and 2008

	2004		2008			2004		2008	
mixture	$-l$	*AIC*	$-l$	*AIC*	mixture	$-l$	*AIC*	$-l$	*AIC*
children 2	55,169	110,349	133,473	259,606	children 5	56,503	113,033	129,789	260,956
education	49,727	99,481	115,080	230,186	K=5	52,502	105,032	121,520	243,159
K=1	52,785	105,575	122,297	244,598	K=2	52,534	105,078	121,630	243,669
K=3	52,508	105,031	121,526	243,067	K=4	52,502	105,026	121,509	243,040

Source: own computations

In the Tables 2-4 the estimated characteristics of the level and variability of subgroups are given in order to analyse and compare them. In the Tables 2 and 3 results obtained from complete data are given, in the Table 4 these characteristics are shown for incomplete data. In the Table 2 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 real 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 the households with the head with magister education multiplied by inflation gives 181,552 CZK. The real value (Table 2) is 199,691 CZK and it means more than 11 percent of real increase.

Tab. 2: Estimated characteristics of the level and variability of income distribution. The complete data, groups divided according to education

	Expected value					Median				
	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
	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

Source: own computations

In the Table 3 the negative impact of number of children in the household on incomes is obvious. This fact could be reduced in case of the use of equalized incomes (CZSO) instead of incomes per capita.

Tab. 3: Estimated characteristics of the level and variability of mixture components (CZK) for complete data divided 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
	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

Source: own computations

Components in the Table 4 are arranged according to estimated values of the parameter μ_j. The expected value of the lognormal distribution depends also on σ^2 and the expected values of components in the table are not always ordered from the lowest to the highest. Relative variability (relative to the expected value) is smaller for groups of households with low incomes then for high income households with coefficient of variance greater than 100 percent, in 2008 for the four components model the standard deviation is 140 percent of the expected value for the group of the highest incomes per capita.

In the Table 5 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 are fitted into same data and the estimated values in the Table 5 can be compared to sample values: sample means 111,024 CZK in 2004 and 145,277 CZK in 2008, sample medians 97,050 and 126,596 CZK and standard deviations 77,676 in 2004 and 93,397 CZK in 2008. 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 deviation. Standard deviations of all fits underestimate (some of them remarkably) sample standard deviations.

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

	Expected value					Median				
	K=2		K=3			K=2		K=3		
year	j=1	j=2	j=1	j=2	j=3	j=1	j=2	j=1	j=2	j=3
2004	96,967	118,081	95,613	109,866	145,136	95,703	101,114	94,845	99,509	105,979
2008	128,551	171,787	119,535	146,527	197,689	124,991	143,057	118,302	136,216	140,084
	K=4					K=4				
	j=1	j=2	j=3	j=4		j=1	j=2	j=3	j=4	
2004	95,100	113,336	110,616	378,488		94,372	95,511	102,950	254,485	
2008	118,064	141,862	157,710	268,866		117,008	134,996	135,944	155,905	
	Standard deviation					Coefficient of variation				
	K=2		K=3			K=2		K=3		
year	j=1	j=2	j=1	j=2	j=3	j=1	j=2	j=1	j=2	j=3
2004	15,812	71,218	12,192	51,413	135,797	0.16	0.60	0.13	0.47	0.94
2008	30,900	114,208	17,303	58,079	196,849	0.24	0.66	0.14	0.40	1.00
	K=4					K=4				
	j=1	j=2	j=3	j=4		j=1	j=2	j=3	j=4	
2004	11,838	72,400	43,475	416,675		0.12	0.64	0.39	1.10	
2008	15,892	45,818	92,747	377,762		0.13	0.32	0.59	1.41	

Source: own computations

Tab. 5: Estimated characteristics of the level and variability of income distribution (CZK) for the complete data (first part) and incomplete data for K=2, 3, 4 (second part)

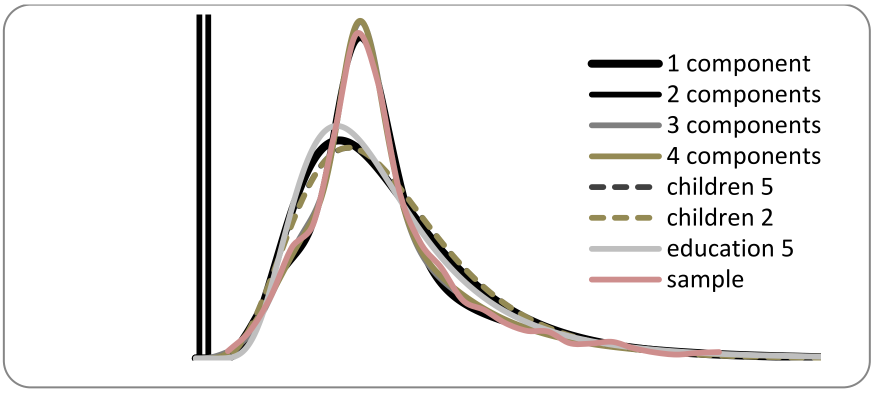
	Education (5 levels)			Children (2 levels)			Children (5 levels)		
year	$E(Y)$	$y_{0.5}$	$\sqrt{D(Y)}$	$E(Y)$	$y_{0.5}$	$\sqrt{D(Y)}$	$E(Y)$	$y_{0.5}$	$\sqrt{D(Y)}$

2004	110,238	97,390	56,671	109,556	100,953	49,873	109,572	97,959	49,971
2008	144,113	129,487	68,340	143,354	132,969	61,095	143,267	142,091	61,305
	K=2			*K*=3			*K*=4		
2004	110,269	97,463	58,239	110,583	97,101	64,649	111,041	97,143	75,442
2008	144,808	128,246	77,063	144,834	126,806	83,550	145,263	126,814	94,711

Source: own computations

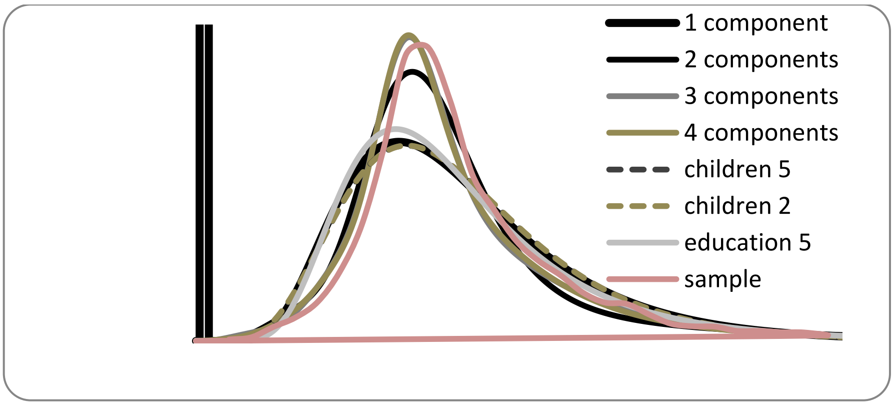
In the Figures 1 and 2 estimated mixture densities are shown for 2004 (Figure 1) and 2008 (Figure 2). For both years the estimated density from the fit with incomplete data is reasonably closed to sample one even for only 2 components. The fits from complete data are similar to the density obtained from single lognormal distribution.

Fig. 1: Estimated mixture densities in 2004



Source: own computations

Fig. 2: Estimated mixture densities in 2008



Source: own computations

Conclusions

In the paper the use of the mixtures of lognormal distributions is proposed as a suitable model for the incomes in the Czech Republic. The expected as well as strange properties of the models are described and quantified.

The concept of mixture distributions is well applicable to income data, as these values form usually very non-homogenous set. If data are divided into subgroups according to a known explanatory variable, we have 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. This fact was quite apparent in the models that took into account number of children in the household.

In case of incomplete data, the algorithm search for more homogenous groups and the fit is improved with every new component. For too many components there are many parameters in the model and Akaike criterion increases. Moreover there could be numeric problems and the approximation could become time consuming. It is sometimes difficult to clearly interpret subgroups in such models.

Acknowledgment

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

References

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

Bílková,D. 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 E., Nunez O. Estimation of the Income Distribution and Detection of Subpopulations: an Explanatory Model. *Computational Statistics & Data* Analysis. 2007.

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

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

Pavelka, R. 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., Smith, A.F., Makov, U.E. Statistical analysis of finite mixture distributions, Wiley, 1985.

Contact

Ivana Malá
University of Economics, Prague
nám. W. Churchilla 4, Praha, Czech Republic
malai@vse.cz