ISSN 2045-3345

ISSN 2045-3345

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

THE USE OF FINITE MIXTURES OF LOGNORMAL AND GAMMA DISTRIBUTIONS

Ivana Malá[I]
University of Economics, Prague, Czech Republic

ABSTRACT

In the text the finite mixtures of distributions are studied with special emphasis to the models with known group membership. Component distributions are supposed to be two parameter lognormal and gamma distributions, both distributions are unimodal and positively skewed. The former distribution allows independent maximum likelihood estimates of parameters, the latter asymptotically strongly dependent estimates. Maximum likelihood method of estimation is used to estimate unknown parameters of the model - the parameters of component distributions and the component proportions from large samples. In the text large samples are treated, so the asymptotical properties of maximum likelihood estimates mentioned above are discussed with respect to the asymptotic normal bivariate distribution and standard deviances of estimated parameters. Models based on uncensored data (equivalized annual net incomes in the Czech Republic) and censored data (duration of unemployment in the Czech Republic) are analysed and particular problems connected to both estimation procedures are introduced. For both problems the model with lognormal components was found superior to the model with gamma components. All calculations are made in the program R.

JEL CLASSIFICATION & KEYWORDS

■ C41 ■ C13 ■ C15 ■ DURATION OF UNEMPLOYMENT ■ INCOME DISTRIBUTION ■ CENSORED DATA ■ MIXTURES OF DISTRIBUTIONS

INTRODUCTION

Finite mixture models are frequently used for the modelling of distributions of random variables defined on the non-homogenous population that is composed of subsets with homogenous distributions of analysed variable (McLachlan & Peel, 2000). The spectrum of possible applications is very large as can be seen from a huge literature can be found that is available to the topic. An extensive bibliography to the mixtures can be found in the mentioned monograph of McLachlan and Peel from 2000.

Frequently, same component distributions are used for all components with the parameters depending on the component. The same concept was selected in this text, where the mixture models with two-parametric lognormal and gamma component distributions were used. Mixture models are applied to two of very important characteristics of Czech economy - the equivalized net annual income and the distribution of the duration of unemployment. The models refer to years 2010 and 2011. The former problem is based on complete data about incomes, the latter problem includes (incomplete) censored data. The models with observed component membership are applied (gender and education of a head of a household (for income data) or of an unemployed (for unemployment data)), however also models with artificial components could be used. In our study the question whether the use of components given by gender or education can improve the fit of the overall distribution of income or unemployment duration. It can be well supposed that there is dependence between gender (education) and incomes (or unemployment duration). But we know, that the use of two parameter lognormal or gamma distribution is not sufficient for the modelling of the distribution of analysed variables. We will obtain description of components as a by-product of the analysis and we will show, that even the use of components (in case of gender) can improve the fit and provide the model, which can be accepted. The use of education with four levels (basic or no education, secondary education, secondary education with baccalaureate and tertiary education) and lognormal components gives the model superior in analysed models.

For both mentioned problems a positively skewed component distribution should be applied as a model, the selected distributions (lognormal and gamma) are unimodal and have the mentioned property. Both distributions are included into the set of income distributions (Kleiber & Kotz, 2013). Lognormal distribution (usually with three parameters) is frequently used for the modelling of incomes and wages, for example for the Czech Republic Bílková (2012) or Malá (2013). In the paper Chotikapanich & Griffiths (2008), the use of mixture of gamma distributions for the modelling of income distributions is discussed.

The mixtures of these distributions are not generally unimodal or positively skewed, but in the models analysed in this text all mixtures are as also unimodal and positively skewed, they are more gamma or lognormal distributed.

The text is organized in the following way. In the first part a mixture model is introduced together with properties that are useful for the analysis of data. Asymptotic properties of maximum likelihood estimates of parameters are discussed and differences between estimates in gamma and lognormal components are shown. In the second part a model for incomes is estimated and in the third part deals with the model for the unemployment duration.

Methods

Suppose X to be a positive value random variable with continuous distribution. The density function f is given as a weighted average of K component densities $f_j(x,\theta_j)$ (j = 1, ..., K) with weights (mixing proportions) π_j

$$f\left(x;\boldsymbol{\psi}\right)=\sum_{j=1}^{K}\pi_j f_j\left(x;\boldsymbol{\theta}_j\right), \qquad (1)$$

McLachlan & Peel, 2000. We suppose that the weights fulfil obvious constraints

$$\sum_{j=1}^{K}\pi_j=1,\ 0\le\pi_j\le 1,\ j=1,...,K$$

and component densities depend on p-dimensional (in general unknown) vector parameters θ_j, j = 1,…,K. All

[I] malai@vse.cz

unknown parameters are included in the vector of unknown parameter to be estimated $\boldsymbol{\psi}$, where $\boldsymbol{\psi} = (\pi_1,\ldots, \pi_{K-1}, \theta_j, j = 1,\ldots,K)$.

From (1) we obtain

$$F(x;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j F_j(x;\boldsymbol{\theta}_j), \tag{2}$$

where F_j are component distribution functions, $j = 1,\ldots,K$. If X_j, $j = 1, \ldots, K$ are random variables with densities f_j and X is a random variable with density (1), then it follows for the expected value of X

$$E(X) = \sum_{j=1}^{K} \pi_j E(X_j). \tag{3}$$

The formula for the variances $D(X)$ is more complicated, but it can be written as

$$D(X) = \sum_{j=1}^{K} \pi_j \left(D(X_j) + \left[D(X_j) \right]^2 \right) - \left[E(X) \right]^2. \tag{4}$$

Percentiles x_P, $0 < P < 1$, of the mixtures were found from the definition (with the use of (2)) by numeric solving the equation

$$P = F(x_P;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j F_j(x_P;\boldsymbol{\theta}_j). \tag{5}$$

Median was evaluated for $P = 0.5$, quartile deviation was found from quartiles ($P = 0.25$ and 0.75) as it is defined as $0.5(x_{0.75} - x_{0.25})$.

The choice of number of components K is crucial for the proper model as well as the choice of component densities f_j. In this text only models with observable component membership, the number K is given as number of values in factor independent variable ($K = 2$ for gender and $K = 4$ for education). Two-parameter lognormal distribution is given by the density $\left(\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2), \mu_j \in R, \sigma_j^2 > 0\right)$

$$f_{LN}(x;\boldsymbol{\psi}) = \sum_{j=1}^{K} \frac{\pi_j}{\sqrt{2\pi}\sigma_j x} \exp\left(-\frac{(\ln x - \mu_j)^2}{2\sigma_j^2} \right), \tag{6}$$

with component expected values and variances

$$E(X_j) = e^{\mu_j + \sigma_j^2/2},\ D(X_j) = e^{2\mu_j + \sigma_j^2}\left(e^{\sigma_j^2} - 1\right),\ j = 1, \ldots, K.$$

Gamma distribution has a density

$$f_{gamma}(x;\boldsymbol{\psi}) = \sum_{j=1}^{K} \frac{\pi_j}{\Gamma(m_j) x^{m-1}} \exp\left(-\frac{x}{\delta} \right), \tag{7}$$

with component expected values and variances

$$E(X_j) = m_j \delta_j,\ D(X_j) = m_j \delta_j^2,\ j = 1, \ldots, K.$$

As mentioned above, for the estimation of unknown parameters (from a random sample x_i, $i = 1,\ldots,n$) the maximum likelihood estimation is usually used in order to obtain maximum likelihood estimate $\hat{\boldsymbol{\psi}}$ of the parameter $\boldsymbol{\psi}$. From (1) it follows that the likelihood function $L(\boldsymbol{\psi})$ is equal to

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j f_j(x_i;\boldsymbol{\theta}_j), \tag{8}$$

in case of complete (non-censored) data (income data in this text). If the data are right censored (in the value x_i) or interval censored (in the interval $\langle l_i, u_i \rangle$), formula (8) takes the form (we use (2) and Lawless, 2003)

$$L(\boldsymbol{\psi}) = \prod_{\substack{i:x_i \text{ right} \\ \text{censored}}} \left(1 - \sum_{j=1}^{K} \pi_j F_j(x_i;\boldsymbol{\theta}_j) \right) \prod_{\substack{i:x_i \text{ interval} \\ \text{censored}}} \sum_{j=1}^{K} \pi_j \left[F_j(u_i;\boldsymbol{\theta}_j) - F_j(l_i;\boldsymbol{\theta}_j) \right]. \tag{9}$$

Right and interval censored data are treated in the unemployment model.

In this text only models with observable component membership are taken into account. Under this assumption the likelihood functions (8) and (9) can be split into K components, where maximum likelihood estimates are evaluated (McLachlan & Peel, 2000). Maximum likelihood estimates of the mixing proportions can be found as relative frequencies of observations from the components in the whole sample.

There exist close form for maximum likelihood estimates of the parameters of lognormal distribution (based on complete data) and these estimates can be evaluated as (Kleiber & Kotz, 2003)

$$\hat{\mu} = \overline{\ln(x_i)} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\ln(x_i) - \overline{\ln(x_i)} \right)^2. \tag{10}$$

The Fisher information matrix for lognormal distribution is given as

$$\mathbf{I}(\mu, \sigma^2) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \left(2\sigma^4\right)^{-1} \end{bmatrix},$$

the inverse matrix $\mathbf{I}^{-1}$ is then

$$\mathbf{I}^{-1}(\mu, \sigma^2) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}. \tag{11}$$

From the theory of asymptotic properties of maximum likelihood estimates it follows that the vectors of estimated parameters are asymptotically independent with the covariance matrix estimated (maximum likelihood estimate) as

$$\frac{1}{n} \hat{\mathbf{I}}^{-1}(\mu, \sigma^2) = \frac{1}{n} \mathbf{I}^{-1}(\hat{\mu}, \hat{\sigma}^2),$$

with $\mathbf{I}^{-1}$ given in (11).

The Fisher information matrix $\mathbf{I}(m, \delta)$ of the gamma distribution is equal to (Miura, 2011)

$$\mathbf{I}(m, \delta) = \begin{bmatrix} \dfrac{\Gamma''(m)\Gamma(m) - \left(\Gamma'(m)\right)^2}{\left(\Gamma(m)\right)^2} & \dfrac{1}{\delta} \\ \dfrac{1}{\delta} & \dfrac{m}{\delta^2} \end{bmatrix}, \tag{12}$$

where Γ' and Γ'' are the first and the second derivatives of gamma function. The term on the first row and first column is the second derivative of the function $\ln(\Gamma(m))$, frequently called trigamma function. The asymptotic covariance matrix of the estimated parameters of gamma distribution $(\hat{m}, \hat{\delta})$ is equal to $\frac{1}{n}\mathbf{I}^{-1}(m, \delta)$ and the matrix can be estimated as $\frac{1}{n}\mathbf{I}^{-1}(\hat{m}, \hat{\delta})$, where $\mathbf{I}$ is given in (12).

It follows from (11) that estimates of μ and σ^2 of lognormal distribution are independent with standard deviations

$$\frac{\sigma}{\sqrt{n}} \left(\frac{\sqrt{2}\sigma^2}{\sqrt{n}}, \text{respectively} \right).$$

From (12) we obtain for the estimates of parameters in gamma distribution (asymptotic) standard deviations

$$\sqrt{D(\hat{m})} = \frac{1}{\sqrt{n}} \sqrt{\frac{m}{m\ \text{trigamma}(m) - 1}} \tag{13}$$

and

$$\sqrt{D(\hat{\delta})} = \frac{\delta}{\sqrt{n}} \sqrt{\frac{\text{trigamma}(m)}{m\ \text{trigamma}(m) - 1}}. \tag{14}$$

Moreover, the correlation coefficient between maximum likelihood estimates of parameters does not depend on the scale parameter and it is equal to

$$\rho_{\hat{m},\hat{\delta}} = -\frac{1}{\sqrt{\text{trigamma}(m)}\ \sqrt{m}}. \tag{15}$$

With exception of estimates in lognormal components with complete data (formula (10)), no explicit formulas for parameter estimation can be found. For gamma distributed

components in the model with complete data maximum of logarithmic numeric maximization procedure was used in order to find values of unknown parameters (package Fitdistrplus in the program R (RPROGRAM, 2013, RFITDISTR, 2013)). For the fittings of the censored data the R package Survival (RSURVIVAL, 2013) was used for lognormal components and package Fitdistrplus for gamma components.

Data and results

Mixture models for equivalised net yearly income in the Czech Republic

Data from EU-SILC Survey (a national module of the European Union Statistics on Income and Living Conditions 2011, (CZSO, 2012)) performed by the Czech Statistical Office is used for the modelling of the net yearly equivalised incomes (in CZK) of the Czech households in 2010. The survey is based on households, each household is included in the survey for five years (yearly 20 % of households leave the survey and new households enter). The annual net equivalised income of each household (in CZK) was evaluated as a ratio of annual net income of the household and number of units (equivalent adults) that reflects number of members and the structure of the household. The number of units evaluated according to European Union methodology was used. It assigns the weight 1 to the first adult, other adult members of household have weight 0.5 and each child has weight 0.3. Total number of units in the household is lower than number of members, the equality is obtained only for single member households. It follows that income per capita is lower (or equal to for single member households) than equivalised income.

The mean annual exchange rate of CZK in 2010 was 25.29 CZK/EUR and 29.485 CZK/GBP (CNB, 2013). It follows that average yearly income was 8,090 EUR or 6,939 GBP, sample median 7,077 EUR or 6,070 GBP (Table 1).

In the Table 1 8,866 observations from the survey is divided into components according to gender (man, woman) or education (no education or basic education, secondary, complete secondary and tertiary) of the head of the household. Sample characteristics of the location and variability are given for the whole sample and all subsets defined above. Percentages of subsets in the whole sample will be used as maximum likelihood estimates of the mixing proportions in (1).

In the Tables 2 and 3 (Table 2 describes the model with two components (gender) and Table 3 the four component model (education)). All estimates of parameters of components are given with standard deviations. Standard deviations are provided by packages in R, only for gamma distributions standard deviation of the parameter $\hat{\delta}$ was approximated from the standard deviations of the estimate of the intensity (called rate) $1/\hat{\delta}$. The estimate of the scale parameter δ was evaluated as the inverse value of the rate and the standard deviation was estimated from the standard deviation of the estimate of the rate with the application of Taylor approximation (up to second order of the polynomial). Moment estimates m^+ and δ^+ in gamma distribution

$$\delta^+ = m_2 / \bar{x}$$

$$m^+ = \bar{x}^2 / m_2$$

where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is a sample second order central moment evaluated from a sample, were used as good initial estimates for the numeric approximation.

From the estimated component parameters the values of maximum likelihood estimates component of expected values and standard deviations were evaluated by substituting estimates of parameters into the theoretical formulas. These expected values are then mixed to obtain the maximum likelihood estimate of the expected value of equivalised income (formula (3)). Estimated standard deviation of the mixture was evaluated from (4).

From the estimates of parameters maximum likelihood estimates of various quantities can be evaluated; in the Tables 2 and 3 expected values and standard deviations. But it could be possible to find an estimate of any characteristics of interest. From the Tables 2 and 3 we can see that estimated characteristics of location copy well sample characteristics for all models. It is not true for the variance, all models underestimate the sample variability (with the difference 20,000 CZK for both distributions for gender, 2,000 CZK for lognormal and 16,000 CZK for gamma distribution and education models). If four component model is used for lognormal distribution, the mixture is able to model variability well. From this point gamma distribution is not skewed enough to model well heavy tail of sample distribution. The variability of incomes for households with a man as a head are greater than this variability for women. To compare variance coefficients, we obtain for sample characteristics 52.8 % and 54.6 %, for estimates only 42.1 % and 41.1 % for gamma components (resp. 42.9 % and 40.6 % for lognormal components). For the model with components given by education, the higher the education, the larger spectrum of possible jobs brings higher variance in incomes.

The model provides not only information about the whole set of the Czech households, but also about each component separately. Lower level of income for households headed by woman in comparison with households headed by a man is visible.

The positive impact of the high education of the head is obvious, the estimated expected value for basic education

Table 1: Descriptive statistics of the net annual income of the Czech household in 2010

parameter / gender	n_i	percentage (%)	average (CZK)	median (CZK)	standard deviation (CZK)	quartile deviation (CZK)
man	6,482	73.11	220,165	192,013	116,255	47,054
woman	2,385	26.90	162,331	140,48	88,699	30,456
basic	978	11.3	145,386	136,294	48,153	22,014
secondary	3,824	43.13	182,524	169,6	70,897	36,685
secondary	2,748	30.99	216,138	191,617	118,121	49,715
tertiary	1,316	14.84	288,708	242,51	168,619	81,426
sample	8,866	100	204,607	178,969	112,484	46,177

Source: Author

Table 2: Results of the modelling of the net annual income of the Czech household in 2010. Standard errors of estimates in brackets ($\hat{\pi}_1 = 0.731, \hat{\pi}_2 = 0.269$)

parameter / gender	$\hat{m}$	$\hat{\delta}$	expected value (CZK)	standard deviation (CZK)	expected value (CZK)	standard deviation (CZK)
man	5.638 (0.045)	39,05 (263)	220,164	92,722	204,609	90,209
woman	5.930 (0.098)	27,376 (420)	162,34	66,665		
	$\hat{\mu}$	$\hat{\sigma}$				
man	12.211 (0.005)	0.411 (0.003)	218,701	93,819	203,105	90,792
woman	11.911 (0.008)	0.391 (0.004)	160,723	65,323		
Source: Author						

Table 3: Results of the modelling of the net yearly income of the Czech household in 2010. Standard errors of estimates in brackets ($\hat{\pi}_1 = 0.111, \hat{\pi}_2 = 0.431, \hat{\pi}_3 = 0.310, \hat{\pi}_4 = 0.148$)

parameter / education	$\hat{m}$	$\hat{\delta}$	expected value (CZK)	standard deviation (CZK)	expected value (CZK)	standard deviation (CZK)
basic	9.198 (0.309)	15,873 (536)	146, 000	48,14	204,675	96,335
secondary	7.995 (0.134)	22,831 (384)	182,534	64,556		
complete secondary	5.711 (0.057)	37,846 (270)	216,139	90,443		
tertiary	4.106 (0.077)	70,307 (1,082)	288,681	142,465		
	$\hat{\mu}$	$\hat{\sigma}$				
basic	11.836 (0.010)	0.324 (0.004)	145,606	51,119	203,884	110,454
secondary	12.051 (0.006)	0.359 (0.003)	182,609	72,141		
complete secondary	12.194 (0.008)	0.405 (0.004)	214,417	98,333		
tertiary	12.447 (0.014)	0.492 (0.009)	287,022	169,326		
Source: Author						

is approximately one half of the value for tertiary education. But recall now, that we are modelling equivalised income of a household, not income of a person.

In the Figure 1 estimated probability densities according to (1) of all models (maximum likelihood estimates) are shown and compared to the histogram of the sample and kernel density (with normal kernel function). The best fit provides kernel density, than model with components given by education and with lognormal component distribution.

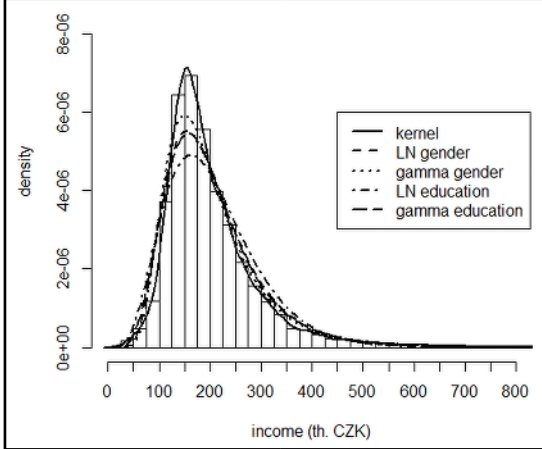
The models with observable components are not constructed to fit the data as well as possible. The Akaike criterion finds as the best the model with four lognormal components.

The estimation of parameters for the mixture of lognormal distributions was fast and there were no problems to obtain solution of maximization. More iterations were needed to find the solution for the mixture of gamma distributions. The repeated estimations with various initial values of parameters were used in order to find global extreme of logarithmic likelihood. Moreover some of the processes tended to converge to the unlimited parameters, from this reason the values of both parameters were set bounded in the numerical procedures.

Mixture models for duration of unemployment in the Czech Republic

In this part data dealing with the duration of unemployment in the Czech Republic (in months) in 2010 and 2011 are

Figure 1: Histogram of income and estimated densities (Tables 2 and 3)



Source: Author

analysed. It is known, that the rate of unemployment and its duration depends on various factors. For example men and highly educated people have shorter duration of unemployment, lower rate of unemployment and higher chance to find a new job. The mixtures of probability distributions treated in this text seems to be a suitable approach to the modelling of the distribution of duration of

unemployment. Positively skewed unimodal component distributions lognormal and gamma distributions have expected properties of the suitable model distribution for the unemployment duration. In the survival analysis (dealing with censored data) in addition to the cumulative distribution function $F(x)$ the survival function $S(x)$ is frequently used. From (2) we obtain

$$S(x;\boldsymbol{\psi}) = 1 - F(x;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j S_j\left(x;\boldsymbol{\theta}_j\right), \qquad (16)$$

where

$$S_j\left(x;\boldsymbol{\theta}_j\right) = 1 - F\left(x;\boldsymbol{\theta}_j\right).$$

The hazard function is given by the formula (from (2) and (16))

$$h(x;\boldsymbol{\psi}) = \lim_{\Delta x \to \infty} \frac{P\left(x \le X < x + \Delta x \middle| X \ge x\right)}{\Delta x} = \frac{f(x;\boldsymbol{\psi})}{S(x;\boldsymbol{\psi})} = \frac{\sum_{j=1}^{K} \pi_j f_j\left(x;\boldsymbol{\theta}_j\right)}{\sum_{j=1}^{K} \pi_j S_j\left(x;\boldsymbol{\theta}_j\right)}, x > 0.$$

Both component distributions have hazard function with one extreme – global maximum (Lawless, 2003). It means that there exists a value of time x, where an intensity of finding of a new job is highest. It increases to this time x and decreases from this time point. The longer is a duration of unemployment, the smaller is hazard function. More information about the duration of unemployment in the Czech Republic in the analysed period can be found in Čabla, 2012, where nonparametric estimation was used for those, who found a job. In Löster & Langhamrová, 2011 the authors deal with the development of long-term unemployment (unemployment longer than one year), in this text we use data for the unemployed with the duration of unemployment shorter than two years.

The Labour force sample survey (LFSS, 2012) is organized quarterly by the Czech Statistical Office. As in the previous part and EU-SILC data, this survey is based on the Czech households. The households form a rotating panel, each household is followed for five quarters, it means more than one year. Data on unemployment duration are either right or interval-censored, no exact values are recorded. The duration is reported in intervals 0-3 months, 3-6 months, from 6 months to one year, from one to two years and more than two years. In the data all the unemployed of the age 15-65 years with unemployment duration up to two years from the LFSS from the first quarter of 2010 to the first quarter of 2011 were included. If the unemployed found a job in the period in the study, the observation is interval censored in a finite interval. Observations for those, who did not find a new job, are right censored.

We will use again gender of the unemployed as an indicator of the component. There were 4,753 unemployed included in the data set, 2,352 men and 2,401 women. The components defined by education have number of cases 713 for the unemployed without education or with basic education, 2,246 for secondary education, 1,447 for secondary education with baccalaureate and 347 unemployed with tertiary education. The relative counts based on these values were used as estimates of mixing proportions and they are given in Tables 4 and 5.

If the censored data are included in the estimation, there are no explicit formulas for maximum likelihood estimates for any of component distributions. The likelihood (9) has to be maximized with the use of numeric methods in each component. The package Survival (RSURVIVAL, 2013) was used for the estimation of lognormal model and Fitdistrplus (RFITDISTRPLUS, 2013) for fitting gamma distributed component distributions.

The estimation of parameters for the mixture of lognormal distributions was quick and there were no problems to obtain solution of maximization. More iterations were needed to find the solution for the mixture of gamma distributions. Repeated estimations with various initial values of estimates were used to find the global extreme of the likelihood.

Median of a mixture was evaluated by the numerical solution of (5) for $P = 0.5$.

It follows from both models that women are unemployed longer than men. There is relatively large difference between both models (we can compare it to the very similar models for incomes). In the Table 5 positive impact of education is visible. Characteristics obtained from both models in this table are comparable with exception of the component including the unemployed with basic education. The estimated lognormal distribution has higher location and it is more skewed than estimated gamma component distribution. Medians of the duration of unemployment are greater than one year with exception of the highest (tertiary) education.

According to the Akaike criterion the model with lognormal components provides better fit to data than gamma distributed components. It seems that gamma distribution doesn´t allow the good fit of sample distribution with so heavy tailed empirical distribution.

In the Figure 2 estimated densities from all models (given in Tables 4 and 5) are shown. For both component distributions there is not any difference in densities for models with two and four components.

Conclusion

In this text finite mixtures of distributions with observed component membership were treated. The models for two positively skewed variables defined on non-homogenous

Table 4: Results of the modelling of the duration of unemployment (in months) in the Czech Republic in 2010 and 2011 ($\hat{\pi}_1 = 0.495, \hat{\pi}_2 = 0.505$)

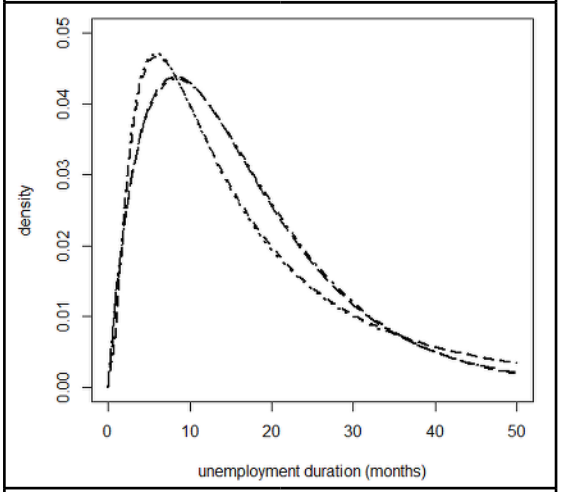
parameter / gender	$\hat{m}$	$\hat{\delta}$	expected value (months)	median (months)	expected value (months)	median (months)
man	1.961 (0.090)	8.157 (0.532)	16.0	13.4	16.5	14.0
woman	1.987 (0.095)	8.913 (0.635)	17.7	14.8		
	$\hat{\mu}$	$\hat{\sigma}$				
man	2.588 (0.029)	0.937 (0.020)	20.6	13.3	21.3	14.2
woman	2.703 (0.030)	0.937 (0.020)	23.2	14.9		
Source: Author						

Table 5: Results of the modelling of the duration of unemployment (in months) in the Czech Republic in 2010 and 2011 ($\hat{\pi}_1 = 0.150, \hat{\pi}_2 = 0.473, \hat{\pi}_3 = 0.304, \hat{\pi}_4 = 0.073$)

parameter / education	$\hat{m}$	$\hat{\delta}$	expected value	median	expected value	median
basic	1.980 (0.198)	12.608 (0.010)	25.0	20.9	17.1	14.9
secondary	2.08 (0.099)	7.932 (0.012)	16.5	13.9		
complete secondary	1.993 (0.115)	7.547 (0.014)	15.0	12.6		
tertiary	1.802 (0.0209)	7.478 (0.020)	13.5	11.1		
	$\hat{\mu}$	$\hat{\sigma}$				
basic	3.165 (0.040)	1.061 (0.060)	41.6	23.7	23.0	15.1
secondary	2.628 (0.019)	0.902 (0.024)	20.8	13.8		
complete secondary	2.511 (0.024)	0.907 (0.031)	18.6	12.3		
tertiary	2.371 (0.051)	0.958 (0.070)	16.9	10.7		

Source: Author

Figure 2: Histogram of income and estimated densities (Tables 2 and 3)



Source: Author

set of the Czech households annual eqvivalised income and duration of unemployment are constructed, unknown parameters are estimated with the use of maximum likelihood estimation method and results are given in the tables and figures. All subsamples in components of interest were large (from 347 observations for the component consisted of tertiary educated unemployed people to 6,482 households headed by a man in income data). For this reason the asymptotical properties of maximum likelihood estimates (unbiased estimates with normal distribution with the variance given by Fisher information matrix) were used to derive additional characteristics of component distributions.

The models give information about probability distributions in components, these distributions are then mixed to the mixture distribution in the whole population. From simple, widely implemented component distributions (lognormal and gamma) with known properties a more complicated distribution is constructed. The component distributions enable to quantify intuitive and thoroughly theoretically and empirically proved relations and impacts of gender or education on incomes and duration of unemployment. The mixture distributions provides the model for the distribution of equivalised income or duration of unemployment.

Both selected distributions can well fit similar all empirical distributions. In the text the problem of the asymptotic correlation of estimates was shown, estimates in the gamma distribution (in the parameterisation shape and scale) are strongly correlated in comparison with independent estimates in the lognormal distribution. The numerical estimation was easier in lognormal than in gamma distribution, where it was complicated to obtain extremes of the log-likelihood. The strong linear dependence of estimates of parameters m and causes numerical problems, as a shift in one parameter can be compensated by a change in other one.

All computations were performed in the program R. This program seems to be the useful tool for estimation in models with both uncensored and censored data. The problem of finite mixture models is well cover in R packages, which provide user with large spectrum of methods and procedures.

Acknowledgements

The article was support by the grant IGA 410062 from the University of Economics, Prague.

References

Bílková, D. (2012). Recent Development of the Wage and Income Distribution in the Czech Republic. Prague Economic Papers, 21(2), 233-250.

Čabla, A. (2012). Unemployment duration in the Czech Republic. In Löster, T., Pavelka T. (Eds.), THE 6TH INTERNATIONAL DAYS OF STATISTICS AND ECONOMICS, Conference Proceedings. Retrieved from http://msed.vse.cz/msed_2012/en/front

Chotikapanich, D. & Griffiths, W. E. (2008). Estimating income distributions using a mixture of gamma densities. In D. Chotikapanich (Ed.), Modeling Income Distributions and Lorenz Curves (Vol. 5, pp. 285-302). Springer New York.

Kleiber, C., Kotz, S. (2003). Statistical Size Distributions in Economics and Actuarial Sciences, Wiley-Interscience, New York.

Lawless, J. F. (2003). Statistical models and methods for lifetime data. Hoboken: Wiley series in Probability and Mathematical Statistics.

Löster, T. & Langhamrová, J. (2011). Analysis of Long-term Unemployment in the Czech Republic. In Löster, T., Pavelka T. (Eds.), THE 5TH INTERNATIONAL DAYS OF STATISTICS AND ECONOMICS, Conference Proceedings. Retrieved from http://msed.vse.cz/msed_2011/en/front

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

Malá, I. (2013). Finite Mixtures of Lognormal and Gamma Distributions, In Löster, T., Pavelka T. (Eds.), THE 7TH INTERNATIONAL DAYS OF STATISTICS AND ECONOMICS, Conference Proceedings. Retrieved from http://msed.vse.cz/msed_2013/en/front

Miura, K. (2011). An introduction to maximum likelihood estimation and information geometry. Interdisciplinary Information Sciences, 17(3), 155-174.

CNB. (2013). Czech National Bank. http://www.cnb.cz. 10.10.2013.

CZSO. (2011). Household income and living conditions 2011. 10.10.2013. Retrieved from http://www.czso.cz/csu/2012edicniplan.nsf/engp/3012-12

CZSO. (2012). LFSS (2012). Labour market in the Czech Republic. 10.10.2013. Retrieved from http://www.czso.cz/csu/2012edicniplan.nsf/engp/3104-12

RPROGRAM. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

RSURVIVAL. Therneau, T. (2013). A Package for Survival Analysis in S. R package version 2.37-4. Retrieved from http://CRAN.R-project.org/package=survival.

RFITDISTRPLUS. Delignette-Muller, M. L., Pouillot, R., Denis, J.-B. & Dutang, C. (2013). Fitdistrplus: help to fit of a parametric distribution to non-censored or censored data. Retrieved from http://CRAN.R-project.org/package=fitdistrplus.