

FINITE MIXTURES OF LOGNORMAL AND GAMMA DISTRIBUTIONS

Ivana Malá

Abstract

In the contribution 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. 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. The asymptotical properties of maximum likelihood estimates are discussed with respect to the asymptotic normal distribution and standard deviances of estimated parameters. Examples for uncensored data (incomes in the Czech Republic) and censored data (duration of unemployment in the Czech Republic) are given and particular problems are introduced. Sample deviations are used as the estimate of standard deviations of estimates in the Monte-Carlo simulation and Fisher information matrix was used for practical applications. Distributions with asymptotically independent estimates (lognormal distribution) and strongly dependent estimates (gamma distributions) were selected in the text. All calculations are made in the program R.

Key words: duration of unemployment, income distribution, censored data, mixtures of distributions

JEL Code: C41, C13, C15

Introduction

Finite mixture models are frequently used for the modelling of distributions of random variables defined on the population that is composed of subsets with non-homogenous distributions (McLachlan & Peel, 2000). Usually the same component distributions are used with the parameters depending on the subset. In this text the mixture models with two lognormal and gamma component distributions are studied and the unknown parameters are estimated with the use of maximum likelihood method. These models are applied to the simulated data, data about incomes and data dealing with the unemployment duration in the

Czech Republic in 2010. In all these situations models with known component membership can be applied (in the simulation the component membership is known, in the second and third application gender of a head of a household or an unemployed was selected as an explanatory variable). In the first part more general mixtures of K components are introduced. For the mentioned real problems a positively skewed component distribution should be applied as a model, both selected distributions (lognormal and gamma) are unimodal and have the mentioned properties. The mixtures of these distributions are not generally unimodal or positively skewed.

1 Methods

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

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

McLachlan & Peel, 2000. The weights fulfil constraints $\sum_{j=1}^K \pi_j = 1$, $0 \leq \pi_j \leq 1$, $j = 1, \dots, K$ and component densities depend on p dimensional (in general unknown) vector parameters $\boldsymbol{\theta}_j$. All unknown parameters are included in the vector parameter $\boldsymbol{\psi}$, where $\boldsymbol{\psi} = (\pi_1, \dots, \pi_{K-1}, \boldsymbol{\theta}_j, j = 1, \dots, K)$.

From (1) we obtain

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

where F_j are component distribution functions, $j = 1, \dots, K$. If X_j , $j = 1, \dots, K$ are random variables with densities f_j and X is a random variable with density (1), then

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

The choice of K is crucial for the proper model as well as the choice of densities f_j . In this text two-parameter distributions are used as component distributions. Lognormal distribution is given by the density ($\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2)$, $\mu \in R$, $\sigma_j^2 > 0$)

$$f_{LN}(x; \Psi) = \sum_{j=1}^K \frac{\pi_j}{\sqrt{2\pi}\sigma_j x} \exp\left(-\frac{(\ln x - \mu_j)^2}{2\sigma_j^2}\right), \quad (4)$$

with component expected values $e^{\mu_j + \sigma_j^2/2}$ and variances $e^{2\mu_j + \sigma_j^2} (e^{\sigma_j^2} - 1)$, $j = 1, \dots, K$.

The gamma distributions has a density ($\theta_j = (m_j, \delta_j)$, $m_j, \delta_j > 0$)

$$f_{\text{gamma}}(x; \Psi) = \sum_{j=1}^K \frac{\pi_j}{\Gamma(m_j) x^{m_j-1}} \exp\left(-\frac{x}{\delta_j}\right), \quad (5)$$

with component expected values $m_j \delta_j$ and component variances $m_j \delta_j^2$, $j = 1, \dots, K$.

For the estimation of unknown parameters (from a sample x_i , $i = 1, \dots, n$) the maximum likelihood estimation is used. From (1) it follows that the likelihood function $L(\Psi)$ is equal to

$$L(\Psi) = \prod_{i=1}^n \sum_{j=1}^K \pi_j f_j(x_i; \theta_j), \quad (6)$$

in case of complete (non-censored) data (simulation illustration and income data in this text).

If the data are right censored (in the value x_i) or interval censored (in the interval (l_i, u_i)),

formula (6) takes the form (we use (2) and Lawless, 2003)

$$L(\Psi) = \prod_{\substack{i: x_i \text{ right} \\ \text{censored}}} \left(1 - \sum_{j=1}^K \pi_j F_j(x_i; \theta_j)\right) \prod_{\substack{i: x_i \text{ interval} \\ \text{censored}}} \sum_{j=1}^K \pi_j [F_j(u_i; \theta_j) - F_j(l_i; \theta_j)]. \quad (7)$$

Right or interval censored data are treated in the unemployment example.

In this text the models with observable component membership are taken into account. Under this assumption the likelihood functions (6) and (7) 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.

With exception of estimates in lognormal components with complete data, no explicit formulas can be found. For this reason, all computations were made numerically with the use of the program R (RPROGRAM, 2012). For the fittings of the censored data the package Survival (RSURVIVAL, 2013), for complete data the package Fitdistrplus (RFITDISTR, 2013) were used.

2 Data and Results

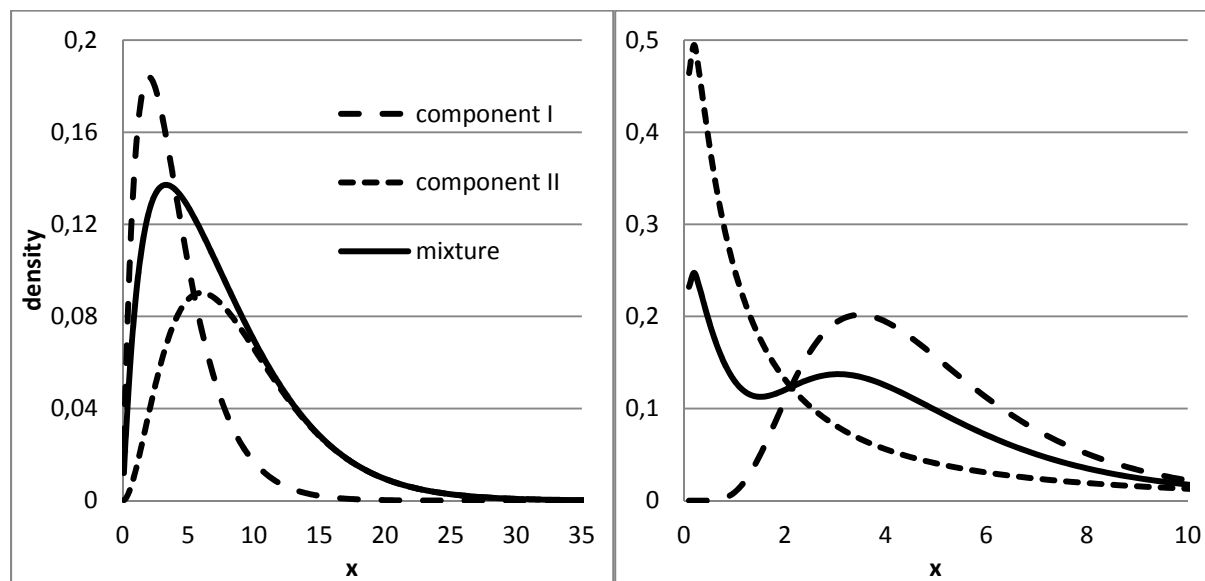
2.1 Simulation study

Simulation was made to illustrate the difference between both component distributions. The selected distributions were used in mixtures with two components. 1, 000 replications were made of samples with 1,000 observations composed of 500 observations from the first component and 500 from the second component.

Gamma distributions with the component densities (5) were used with parameters $m_1 = 2, \delta_1 = 2, m_2 = 3$ and $\delta_2 = 3$ (with the mixing proportion $\pi = 0.5$). In this case it follows $E(X_1) = 4, D(X_1) = 8, E(X_2) = 9$ and $D(X_2) = 27$. The mixture has an expected value 6.5 (3). In the Figure 1 the component densities and the mixture density is given in the left part.

There is not an explicit formula for the maximum likelihood estimates of the parameters of gamma distribution and the maximum of logarithmic likelihood was found numerically (with the use of the package Fitdistplus (RFITDISTR, 2013). Standard deviations of the estimates were evaluated as an standard deviation of 1,000 estimates in the study instead of the inverse Fisher information in (8).

Fig. 1: Probability densities used in the simulation, gamma component distributions (left), lognormal component distribution (right)



Source: own computations

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

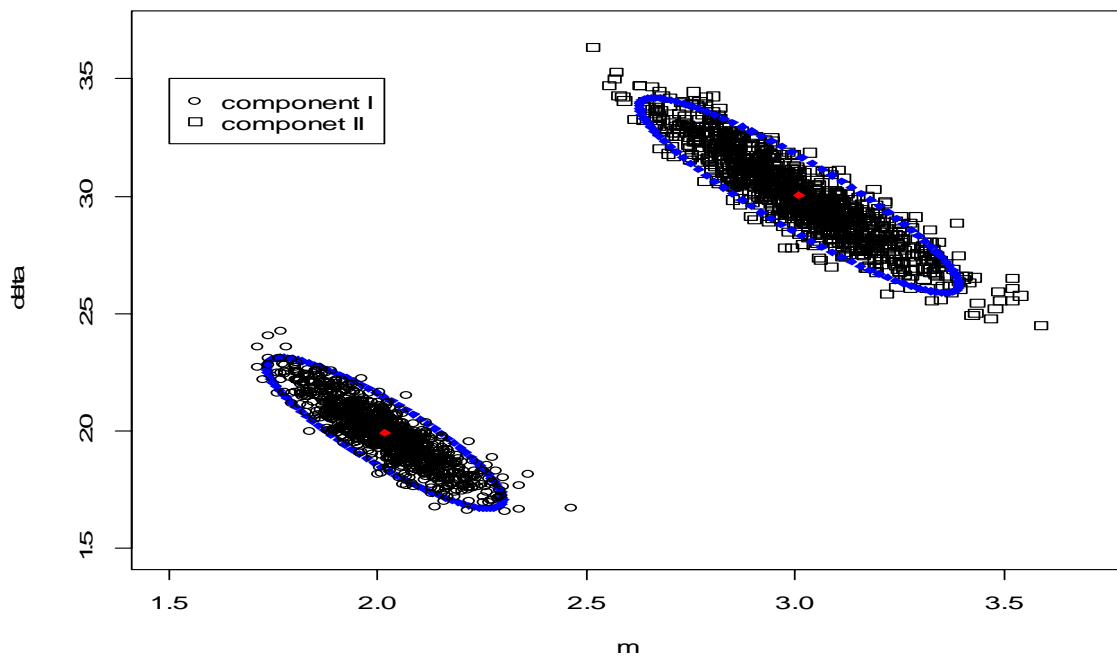
$$\mathbf{I}(m, \delta) = \begin{bmatrix} \frac{\Gamma''(m)\Gamma(m) - (\Gamma'(m))^2}{(\Gamma(m))^2} & \frac{1}{\delta} \\ \frac{1}{\delta} & \frac{m}{\delta^2} \end{bmatrix}, \quad (8)$$

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 $\ln(\Gamma(m))$, frequently called trigamma function. The correlation coefficient between maximum likelihood estimates of parameters does not depend on the parameter δ and it is equal to

$$\rho = -\frac{1}{\sqrt{\text{trigamma}(m)} \sqrt{m}}. \quad (9)$$

We can derive that the estimated parameters are dependent with negative correlation. In our simulation the maximum likelihood estimates of unknown parameters are highly correlated (the theoretical values given by (9) are -0.919 for the first component and -0.880 for the second one). The sample correlation coefficient from the 1,000 samples was found to be -0.88 for the first component and -0.92 for the second component. In the Figure 2 we can see two groups of 500 points and the 95% confidence region for the vector parameter (m, δ) based on the asymptotic normal distribution of maximum likelihood estimates. The ellipses have the centres in the points $(\hat{m}_j, \hat{\delta}_j)$, $j = 1, 2$, it is $(2.009, 0.998)$ and $(3.008, 3.002)$.

Fig. 2: Estimated parameters of gamma distribution of components, 95% asymptotic confidence region



Source: own computations

For the lognormal distribution the maximum likelihood estimates of the parameters are given as

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

From the Fisher information matrix

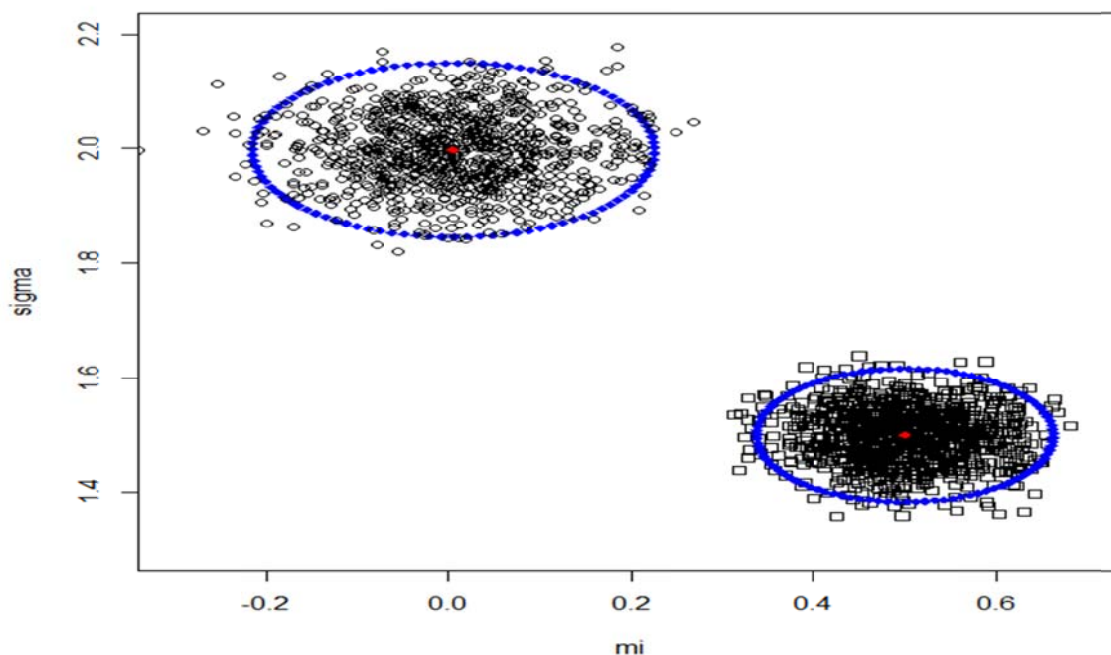
$$\mathbf{I}(\mu, \sigma^2) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-4} \end{bmatrix},$$

it follows that the estimates in (10) are asymptotically independent.

The parameters for the simulation were chosen to be $\mu_1 = 0.5, \sigma_1 = 0.5, \mu_2 = 1.5$ and $\sigma_2 = 0.5$. The moment characteristics are than

$$E(X_1) = E(X_2) = 5.08, \quad D(X_1) = 7.3 \quad \text{and} \quad D(X_2) = 219.$$

Fig. 3: Estimated parameters of lognormal distributions of components, 95 % asymptotic confidence region



Source: own computations

Both components have the same expected values 5.08 and they differ in the shape (Figure 1, right part). The ellipses of the 95% asymptotic confidence region (Figure 3) has main axes parallel with horizontal and vertical axes, as the estimates are independent and the correlation coefficient is equal to 0. The centres of the ellipses have coordinates $(\hat{\mu}_j, \hat{\sigma}_j)$, $j = 1, 2$, it means $(1.500, 0.499)$ and $(0.502, 1.497)$.

For both models, the estimated parameters, standard deviations and estimated component expected values are given in the Table 1. Maximum likelihood estimates (MLE) of $E(X_j)$, $j = 1, 2$ were found from the estimated parameters and formulas for theoretical expected values of lognormal and gamma distributions and the MLE estimates of $E(X)$ were evaluated as a weighted average of estimated component expected values with weights 0.5 (according to (3)).

Tab. 1: Estimated parameters and expected values for both models (Fig. 1 and Fig. 2). Standard errors of estimates are given in brackets.

parameter	\hat{m}	$\hat{\delta}$	expected value X_j	expected value X
-----------	-----------	----------------	----------------------	--------------------

I	2.009 (0.121)	1.998 (0.181)	4.014	6.525
II	3.021 (0.134)	2.991 (0.195)	9.036	
	$\hat{\mu}$	$\hat{\sigma}$		
I	1.500 (0.022)	0.499 (0.066)	5.076	5.073
II	0.500 (0.016)	1.499 (0.047)	5.071	

Source: own computations

For the gamma distribution standard deviations in the Table 1 are greater for δ than for m . It follows from (8), that

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

and

$$\sqrt{D(\hat{\delta})} = \frac{\delta}{\sqrt{n}} \sqrt{\frac{\operatorname{trigamma}(m)}{m \operatorname{trigamma}(m) - 1}}$$

For the components we obtain theoretical asymptotic values of standard deviations $\sqrt{D(\hat{m})} = 0.117$ and $\sqrt{D(\hat{\delta})} = 0.133$ for the first component and $\sqrt{D(\hat{m})} = 0.180$ and $\sqrt{D(\hat{\delta})} = 0.196$ for the second component.

2.2 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, CZSO, 2012) performed by the Czech Statistical Office is used for the modelling of the net yearly equivalised incomes of the Czech households in 2010. An 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.

Lognormal distribution is one of so called income distributions and it is frequently used for the modelling of incomes and wages, for example Bílková, 2012. In the paper Chotikapanich & Griffiths, 2008 the use of mixture of gamma distributions for the modelling

of income distributions is discussed.

Tab. 2: Results of the modelling of the net yearly income of the Czech household in 2010

($\hat{\pi}_1 = 0.73$, $\hat{\pi}_2 = 0.27$, $\bar{X}_1 = 220,165$ CZK, $\bar{X}_2 = 162,331$ CZK, $\bar{X} = 204,607$ CZK).

Standard errors of estimates in brackets.

parameter	\hat{m}	$\hat{\delta}$	expected value	expected value
gender			X_j (CZK)	X (CZK)
man	5.638 (0.045)	39,050 (263)	220,163	204,550
woman	5.930 (0.098)	27,376 (420)	162,340	
	$\hat{\mu}$	$\hat{\sigma}$		
man	12.211 (0.005)	0.411 (0.003)	218,701	203,047
woman	11.911 (0.008)	0.391 (0.004)	160,723	

Source: own computations

The households were divided into two components according to the gender of the head of the household. From 8,866 households in the sample there were 6,481 households headed by a man and 2,385 by a woman. In the Table 2 the estimated values of the parameters of both components are given. Standard deviations of maximum likelihood estimates are provided by the package *fitdistrplus* for gamma distribution and they were evaluated from the Fisher matrix in the solution for lognormal distributions.

In the program R the package *Fitdistrplus* (RFITDISTR, 2013) estimates the shape parameter m and the rate $1/\delta$ of gamma distribution. 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.

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. From the estimates of parameters maximum likelihood estimates of various quantities can be evaluated (moment characteristics, quantile characteristics, probabilities of intervals). The estimates of the level and variability were lower for gamma model than for lognormal model.

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.

2.3 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 2010 are 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 (with observable or unknown component membership) seems to be a suitable approach to the modelling of the distribution of the duration of unemployment. Positively skewed component distributions, examined in this text, have expected properties of the suitable model distribution for the unemployment duration, unimodal component variables were selected. Both distributions have hazard function with one extreme – global maximum (Lawless, 2003). 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. We will use gender of the unemployed as an indicator of the component. There were 4,753 unemployed in the data set, 2,352 men and 2,401 women.

If the censored data are included in the estimation, there are explicit formulas for maximum likelihood estimates for neither of component distributions. The likelihood (7) has to be maximized with the use of numeric methods. The package Survival (RSURVIVAL, 2013) was used for the computations.

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. The repeated estimations with various initial values of estimates were used to find the global extreme of the likelihood.

Tab. 3: Results of the modelling of the duration of unemployment (in months) in the Czech Republic in 2010 ($\hat{\pi}_1 = 0.495, \hat{\pi}_2 = 0.505$)

parameter	\hat{m}	$\hat{\delta}$	expected value (months)	expected value (months)
man	1.961 (0.090)	8.157 (0.532)	16.0	16.9
woman	1.987 (0.095)	8.913 (0.635)	17.7	
	$\hat{\mu}$	$\hat{\sigma}$		
man	2.588 (0.029)	0.937 (0.020)	20,6	21.9
woman	2.703 (0.030)	0.937 (0.020)	23,2	

Source: own computations

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). According to the Akaike criterion the model with lognormal components provides better fit to data than gamma distributed components.

Conclusion

In this text finite mixtures of distributions with observed component membership were treated. Three illustrations of two component mixtures were given and comment approaches and differences were discussed. All samples of interest were large (the components have 500 observations in the simulation illustration and at least 2,300 observations in income or duration of unemployment data). The asymptotical properties of maximum likelihood estimates (unbiased estimates with normal distribution with the variance given by Fisher information matrix) were used.

Both 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.

All simulations and computations were performed in the program R. This program seems to be the useful tool for estimation in such models with both uncensored and censored data as well for the simulations.

Acknowledgment

The article was supported 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.
- CZSO. (2012) *Household income and living conditions 2011*. Retrieved from <http://www.czso.cz/csu/2012edicniplan.nsf/engp/3012-12>
- 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.
- Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. Hoboken: Wiley series in Probability and Mathematical Statistics.
- LFSS (2012). Labour market in the Czech Republic. Retrieved from <http://www.czso.cz/csu/2012edicniplan.nsf/engp/3104-12>
- Loster, T., & Langhamrova, J. (2011). *Analysis of long-term unemployment in the czech republic*. In Loster Tomas, Pavelka Tomas (Eds.), *International Days of Statistics and Economics* (pp. 307-316). ISBN 978-80-86175-77-5.
- Malá, I. (2012). *Estimation of parameters in finite mixtures from censored data*. 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
- McLachlan, G. J. & Peel, D. (2000). *Finite mixture models*. New York: Wiley series in Probability and Mathematical Statistics.
- Miura, K. (2011). An introduction to maximum likelihood estimation and information geometry. *Interdisciplinary Information Sciences*, 17(3), 155-174.
- 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>.

RFITDISTR. 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>.

Contact

Ivana Malá

University of Economics, Prague

W. Churchill Sq. 4, Prague 3

malai@vse.cz