ISSN 2045-3345

ISSN 2045-3345

ESTIMATION OF PARAMETERS IN FINITE MIXTURES OF DISTRIBUTIONS FROM RIGHT CENSORED DATA

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

ABSTRACT

In the text the problem of estimation of parameters in finite mixtures of probability distributions in case of the presence of right censored data is treated. Only models with known number of components and observable component membership (complete mixture models) are studied. Possible estimation method of unknown parameters (parameters of components and mixing proportions) is described. Specific theoretical and numerical problems associated with this this type of the modelling are discussed with respect to the mixture models with complete data. A simulation study is presented to illustrate properties of estimates and sensitivity of results on the proportion of censored data and the sample size. For the simulation 10,000 samples with 500 and 1,000 observations from the mixtures with two components of normal (symmetric distribution) and lognormal (asymmetric distribution) distributions are generated. Results are given in the tables and selected histograms of estimates of parameters are shown. All computations are made in the package R.

JEL CLASSIFICATION & KEYWORDS

■ C41 ■ C63 ■ C88 ■ CENSORED DATA ■ FINITE MIXTURE ■ PARAMETER ESTIMATION ■ SIMULATION

INTRODUCTION

The models with censored data are often used in many statistical modelling problems. Censored data are met in medical applications, demography, economics, insurance, technics and a lot of other fields of study. All these applications naturally sometimes treat time-to-event variables as time to death or time of unemployment. In this case complete data (the time of the occurrence of the event of interest is observed) or incomplete data (the event didn´t occur by the end of the study) are included in analysed datasets. In this text only complete and right censored data (for censored data we have information that the event occurs after observed time) are treated. Suppose now, that a based population consists of K subpopulations and a probability distribution of analysed time to event is described by a known (chosen) probability density in each subpopulation. In this problem the probability distribution of time-to-event variable can be described with the use of mixture of K distributions in subsets with mixing proportions that describes percentage of members of the component in the whole population. The aim of the analysis is then an estimation of unknown parameters of component densities and these percentages. The problem will be shortly described and then some results of simulations are shown below.

Finite mixtures of probability distributions

In this part the finite mixture of probability density is defined and its properties that are used in this article are given (Titterington & al. (1985), McLachlan, Peel (2000)). Suppose now, that Y is a time to event random variable with continuous distribution. For given number of components 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

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

A density of the mixture distribution of Y is defined as a weighted average of densities f_j with weights π_j as

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

The mixture density (1) depends on the vector parameter $\boldsymbol{\psi}$, $\boldsymbol{\psi}=(\pi_1,\ldots,\pi_{K-1};\theta_j, j=1,\ldots,K)$ 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. 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.

In this text two-parametric distributions are used as component distributions. For the normal component distribution formula (1) is of the form

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

In case of lognormal component distribution we have

$$f(y;\boldsymbol{\psi}) = \sum_{j=1}^{K} \frac{\pi_j}{\sqrt{2\pi}\sigma_j y} \exp\left(-\frac{(\ln y-\mu_j)^2}{2\sigma_j^2}\right). \tag{3}$$

The vector of parameters $\boldsymbol{\psi}$ for both models (2) − (3) has generally $3K-1$ unknown parameters. For the estimation of unknown parameters (from a sample y_i, i =1,..,n) the maximum likelihood estimation is frequently used. It means that the likelihood function $L(\boldsymbol{\psi})$

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j f_j(y_i;\boldsymbol{\theta}_j). \tag{4}$$

is maximised.

Suppose, that the random sample arises from a population divided into K subpopulations and for each observation y_i the component j is observed together with the value (complete data problem). In this case the contribution of the i-th observation to the function L in (4) is only $\pi_j f_j(y_i;\theta_j)$ (if this observation comes from the j-th component). If no censored observations are present in the dataset, logarithmic likelihood function l can be written as

$$l(\boldsymbol{\psi}) = \ln(L(\boldsymbol{\psi})) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij} \ln \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij} \ln f_j(y_i;\boldsymbol{\theta}_j), \tag{5}$$

where $\mathbf{z}_i$ 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. Both parts in (5) can be maximized separately. Maximum likelihood estimates of proportions are sample relative frequencies of components and estimates

[I] malai@vse.cz

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

of parameters of the component densities can be found as maximum likelihood estimates in each subgroup.

If the group membership is not observed, the vectors **z** are random vectors with multinomial distribution and the formula (5) cannot be used. For the estimation so-called EM algorithm is frequently used (Titterington & al. (1985), McLachlan, Peel (2000)).

Right censored data

We will suppose now, that the analysed dataset includes right censored data. If a complete (non-censored) *i*-th observation comes from the *j*-th density, it contributes to *L* as $\pi_j f_j(y_i; \theta_j)$ (as above). If the observation is right censored at time y_i (it means that $Y_i > y_i$ its contribution to (4) is $\pi_j(1 - F_j(y_i; \theta_j))$. The formula (5) can be rewritten as

$$l(\psi) = \ln\left(\prod_{i:y_i complete} \sum_{j=1}^{K} \pi_j f_j(y_i;\theta_j) \prod_{i:y_i censored} \sum_{j=1}^{K} \pi_j (1 - F(y_i;\theta_j)) \right) =$$

$$= \sum_{i=1}^{n} z_{ij} \ln \pi_j + \left(\sum_{i:y_i complete} z_{ij} \ln f_j(y_i;\theta_j) + \sum_{i:y_i censored} z_{ij} \ln\left(1 - F_j(y_i;\theta_j)\right) \right).$$

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. For the estimation of unknown parameters the package Survival (Therneau (2012)) in the program R v. 2.3.1 was used. For the parametric approach to the estimation (described above) a function SurvReg was found to be well applicable. With the use of this function so-called survival distributions as normal, lognormal, Weibull, exponential, logistic and loglogistic may be fitted into censored data of different type and moreover explanatory variables (covariates) can be used. In this text the known component membership was applied as an explanatory variable. If we are interested only in survival function 1–*F*, the nonparametric Kaplan-Meier procedure is applicable (built in the function SurvFit) and it can be used without any assumption about distributions (not used in this text).

The properties of estimates for unknown component membership (obtained by EM algorithm) – consistency and asymptotic normality – are given in Svensson, Sjostedt-de Luna (2010). In the literature more variations of this algorithm can be found for censored data (for example Pilla, Lindsay (2001)). The article Svensson & al. (2006) includes the technical application of these methods. The EM algorithm for the fitting of normal distributions implemented in R was modified to obtain estimates in this case.

Simulation and results

In order to illustrate estimation of the parameters and their properties a simulation was performed. 10,000 samples were generated from the mixtures of two components (with equal mixing proportions $\pi_1 = \pi_2 = 0.05$) from normal (symmetric distribution) and lognormal distributions (positively skewed distribution). In all simulations samples of 500 and 1,000 observations were generated with approximately 10 %, 30 % and 50 % of right censored observations. The censoring was done with the use of exponential random variables, the expected value was different for both components and it was evaluated by the numeric method to maintain selected percent of censored data (approximately) in both components. Both values were independently generated in random and the observation was taken as a minimum value of these two values. If smaller value was the exponential one, the observation was supposed to be censored.

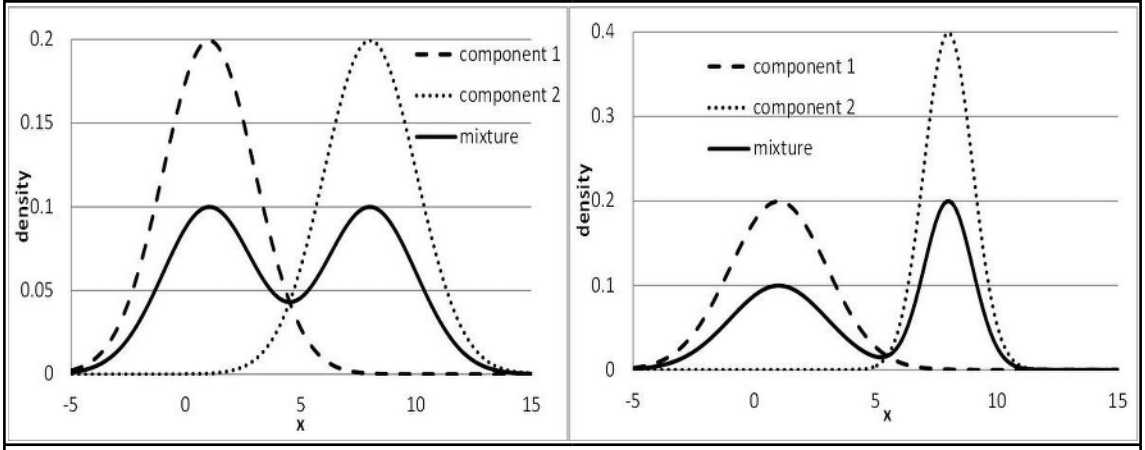
Note that under these assumptions there are 250 (500 observations) in each component. From all samples unknown parameters were estimated and then the mean and standard deviation were evaluated. These values are given in the Tables 1–4. Distributions of estimated parameters are given in the Figures 3–8 as histograms. Only samples of 1,000 observations and 10 % of censored observations are shown in the figures.

Normal distribution mixture

In the case of two components with the same parameter σ ($\sigma_1 = \sigma_2 = \sigma$), 5 parameters in the mixture are to be estimated. We will concentrate on three parameters of component distributions (μ_1, μ_2, σ^2) (we suppose $\pi_1 = \pi_2 = 0.5$). In the Figure 1 (left part) the density of the mixture of two chosen normal densities ($N(1,4)$, $N(8,4)$) with the mixing proportions 0.5 is given. If the sigma parameters are not equal, a separation of components according to (5) is used and the fit is performed separately in each component; 4 unknown parameters (μ_1, μ_2, σ_1^2, σ_2^2) are to be estimated. In the Figure 1 (right part) the density of the mixture of two normal densities ($N(1,4)$, $N(8,1)$) with the mixing proportions 0.5 is shown.

In the Table 1 sample characteristics of estimated parameters mean and standard deviation evaluated from

Figure 1: Mixture of two normal components with equal scale parameters (left) and unequal scale parameters (right)



Source: own computations

10,000 generated samples are shown for equal standard deviations. For the estimation of σ all values in the sample were used, for estimates of μ_i only 50 % of them. The mean of estimates coincides with the selected value and it is obvious that only sample standard deviations depend on the sample size and percentage of censored data. The values of standard deviations decrease with sample size and increase with the percentage of censored data. The sample with a half of observations is heavy-censored but we frequently meet such situation in applications.

In case of unequal unknown standard deviations a separation of components from (5) is used and the fit is performed separately in each component. In the Table 2 characteristics of estimated parameters based on generated samples are given without the assumption of equal variances.

Table 1: Sample characteristics of estimates, normal components: mean (standard deviation)

parameter	***n*=1,000 10 %**	***n*=1,000 30 %**	***n*=1,000 50 %**
$\hat{\mu}_1$	0.999 (0.046)	0.999 (0.048)	1.001 (0.052)
$\hat{\mu}_2$	8,001 (0.024)	8.000 (0.055)	8.002 (0.060)
$\hat{\sigma}$	0.997 (0.024)	0.998 (0.027)	0.999 (0.031)
	***n*=500 10 %**	***n*=500 30 %**	***n*=500 50 %**
$\hat{\mu}_1$	1.001 (0.035)	1.001 (0.068)	0.995 (0.072)
$\hat{\mu}_2$	7.997 (0.068)	7.998 (0.078)	7.997 (0.083)
$\hat{\sigma}$	0.996 (0.033)	0.996 (0.039)	0.997 (0.044)

Source: own computations

Table 2: Sample characteristics of estimates, normal components: mean (standard deviation)

parameter	***n*=1,000 10 %**	***n*=1,000 30 %**	***n*=1,000 50 %**
$\hat{\mu}_1$	0.999 (0.090)	0.995 (0.096)	1.004 (0.109)
$\hat{\sigma}_1$	1.997 (0.067)	1.997 (0.077)	1.995 (0.090)
$\hat{\mu}_2$	8.001 (0.047)	7.997 (0.052)	7.995 (0.060)
$\hat{\sigma}_2$	0.998 (0.033)	0.997 (0.037)	1.001 (0.050)
parameter	***n*=500 10 %**	***n*=500 30 %**	***n*=500 50 %**
$\hat{\mu}_1$	0.999 (0.128)	1.003 (0.135)	1.005 (0.151)
$\hat{\sigma}_1$	1.992 (0.094)	1.996 (0.109)	1.997 (0.111)
$\hat{\mu}_2$	8.001 (0.067)	8.001 (0.073)	8.002 (0.075)
$\hat{\sigma}_2$	0.996 (0.046)	0.996 (0.052)	0.994 (0.058)

Source: own computations

Mixture of lognormal distributions

For the lognormal distribution the component densities are not well separated (Figure 2) as it was in the previous part concerning normal distribution (Figure 1). In general, it is more complicated to identify it. This problem doesn't occur if the component membership is known as it is supposed in this text. We will treat again a mixture model with two components with equal (and known) mixing proportions. The Figure 2 (left part) and the Table 3 refer to the components with equal scale parameters sigma (*LN*(1,1), *LN*(3,1)). The Figure 2 (right part) and the Table 4 deal with component probability distributions with unequal variances (*LN*(1,1), *LN*(3,0.64)). Note now, that in both examples the components have unequal variances as parameters σ are standard deviances of the logarithms of time to event variable *Y* and the variance of lognormal distribution *LN*(μ, σ^2) depends not only on σ but also on the parameter μ. The results are not different for this distribution to compare with the normal distribution. In fact both problems mean estimation of parameters in normal distributions – for values *Y* in the first problem and ln *Y* in the second one.

In the Table 3 estimated characteristics of location (mean) and variability (standard deviation) are given for estimated parameters in the model of components with equal parameter sigma. As in the Tables 1 and 2 the sample standard deviations are smaller for estimates of sigma then for estimates of the parameters mi. In the Table 4 sample characteristics are given for the model with different sigmas. In this table estimates are made from smaller samples, because the components are separated and estimates are evaluated only from the half of observations. It is reflected in higher standard deviations of estimates of parameters sigma comparing to the Table 3. The mean value of the sample of estimates copy well the known values, in the majority of estimates the means are less than estimated parameters. Sample standard deviations are smaller for larger samples and again they increase with the percentage of censored data.

Table 3: Sample characteristics of estimates: mean (standard deviation)

parameter	***n*=1,000 10 %**	***n*=1,000 30 %**	***n*=1,000 50 %**
$\hat{\mu}_1$	0.990 (0.046)	0.999 (0.049)	1.000 (0.058)
$\hat{\mu}_2$	2.990 (0.045)	2.999 (0.050)	3.001 (0.055)
$\hat{\sigma}$	0.998 (0.024)	0.998 (0.027)	0.998 (0.033)
parameter	***n*=500 10 %**	***n*=500 30 %**	***n*=500 50 %**
$\hat{\mu}_1$	1.002 (0.050)	1.009 (0.069)	1.001 (0.081)
$\hat{\mu}_2$	2.997 (0.039)	2.997 (0.069)	3.002 (0.077)
$\hat{\sigma}$	0.997 (0.029)	0.997 (0.038)	0.996 (0.046)

Source: own computations

Table 4: Sample characteristics of estimates: mean (standard deviation)

parameter	***n*=1,000 10 %**	***n*=1,000 30 %**	***n*=1,000 50 %**
$\hat{\mu}_1$	0.998 (0.046)	1.001 (0.050)	1.003 (0.057)
$\hat{\sigma}_1$	0.998 (0.034)	0.998 (0.039)	0.998 (0.045)
$\hat{\mu}_2$	3.001 (0.030)	2.997 (0.045)	3.001 (0.037)
$\hat{\sigma}_2$	0.639 (0.021)	0.637 (0.034)	0.639 (0.029)
parameter	***n*=500 10 %**	***n*=500 30 %**	***n*=500 50 %**
$\hat{\mu}_1$	0.999 (0.065)	0.999 (0.071)	1.001 (0.080)
$\hat{\sigma}_1$	0.996 (0.048)	0.997 (0.055)	0.996 (0.065)
$\hat{\mu}_2$	3.001 (0.042)	2.997 (0.045)	3.001 (0.053)
$\hat{\sigma}_2$	0.644 (0.030)	0.637 (0.034)	0.632 (0.040)

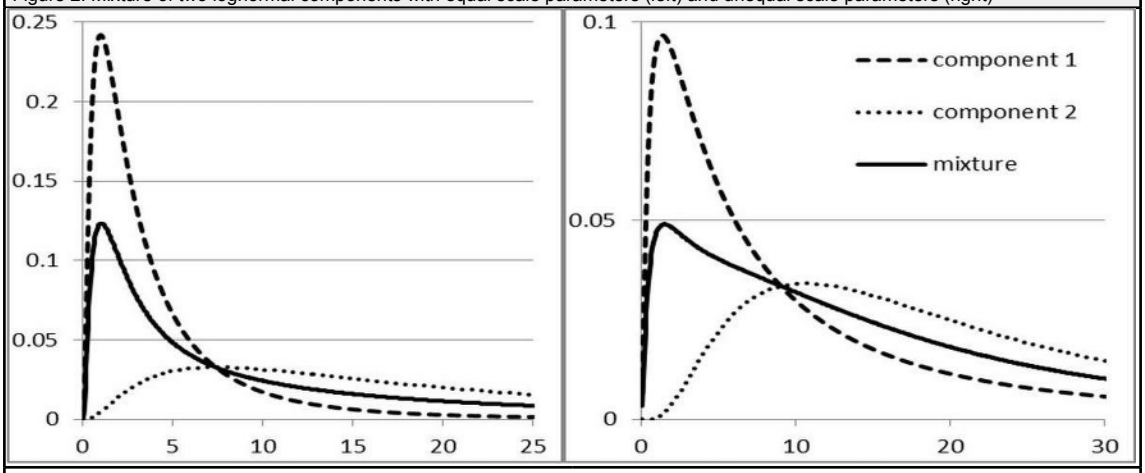
Source: own computations

From the values in the Tables 1-4 the asymptotic confidence intervals can be constructed for estimated parameters. All these intervals include selected values of parameters. In the text covariance matrices are not shown, however from the program SurvFit this information can be obtained.

In the previous tables only mean and sample standard deviations are given, the whole empirical distribution of parameters is shown in the Figures 3-8. Histograms of estimates for samples sizes 1,000 observations and 10 % of censored observations for both equal and unequal values of sigma and both mixtures are shown. In all the figures we can see approximately normal distribution of estimates, however all histograms show slightly negatively skewed estimates of parameters. All these pictures illustrate central limit theorem and asymptotic normality of estimates. On the horizontal axis ranges of values are visible and possible strange results relatively distant from the estimated (known) values of parameters of mixtures are noticeable. In the vertical axis of all histograms values of estimated density of estimates (not frequencies that are in thousands of observation due to the 10,000 fits of each model) are given.

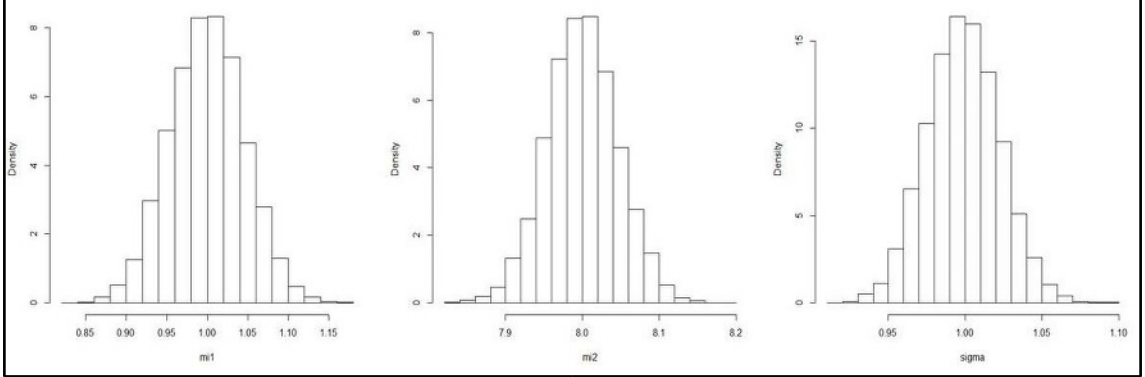
In the Figures 3 and 4 the histograms of three estimated parameters (μ_1, μ_2, σ^2) are shown for both mixtures

Figure 2: Mixture of two lognormal components with equal scale parameters (left) and unequal scale parameters (right)



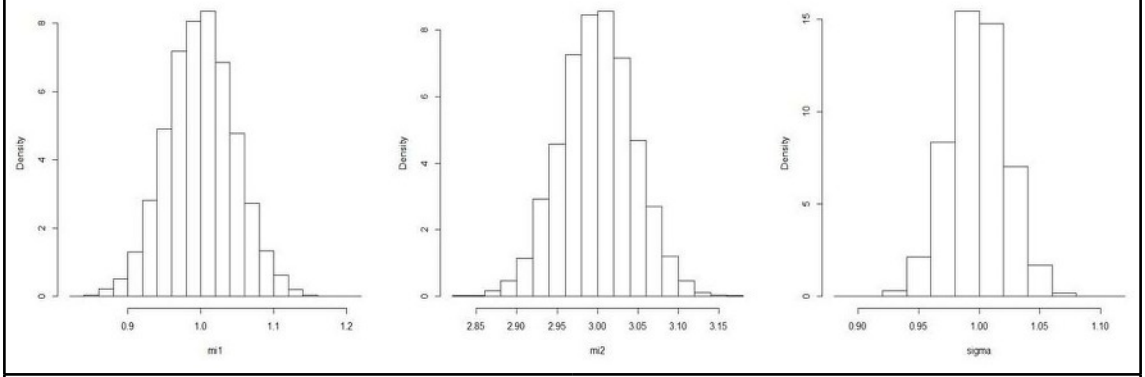
Source: own computations

Figure 3: Distribution of parameters of normal distributions, equal variances: *n*=1,000, 10 % of censored data (Table 1)



Source: own computations

Figure 4: Distribution of parameters lognormal distributions, equal variances: *n*=1,000, 10 % of censored data (Table 3)

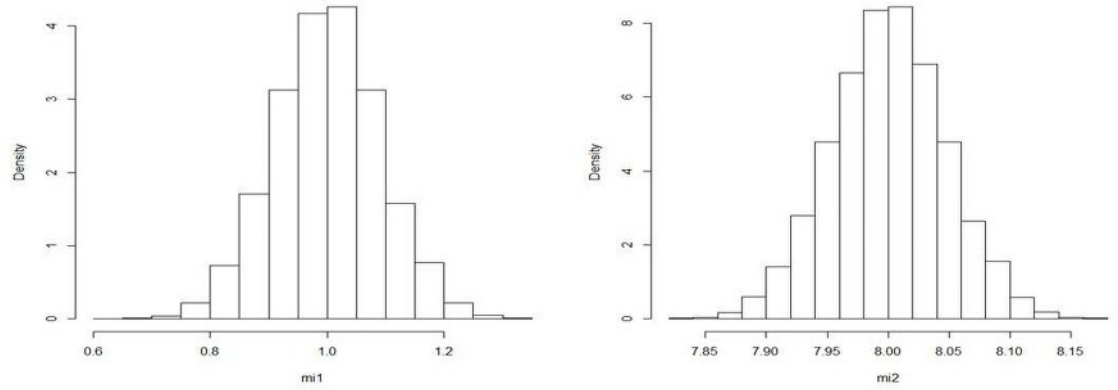


Source: own computations

(Figure 3 for normal and Figure 4 for lognormal component densities). On the horizontal axis estimates of these parameters μ_1, μ_2 and sigma are labelled.

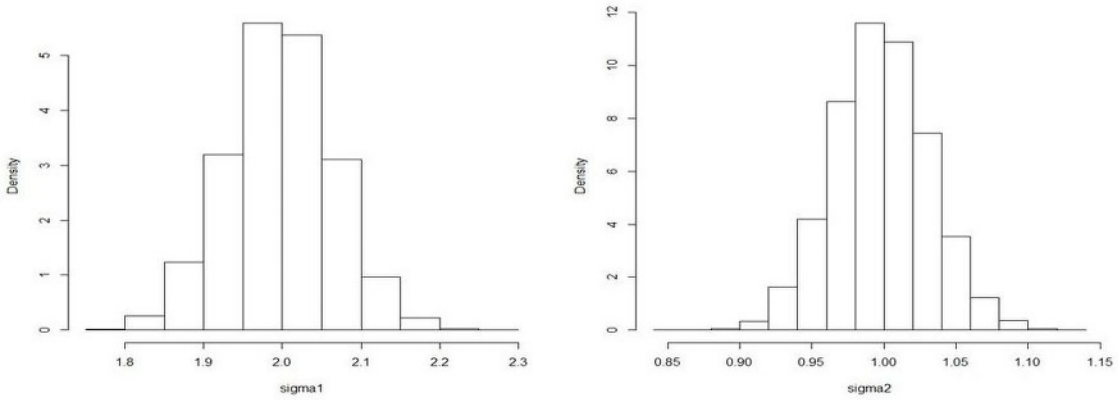
The fits for components with different scale parameters are shown in two figures. In the Figure 5 (normal components) and the Figure 7 (lognormal components) histograms of estimates of parameters μ are shown (labelled mi1 for the first component and μ_2 for the second component). The "true" values are 1 and 8 for the normal components in the Figure 5 and 1 and 3 for the lognormal components in the Figure 7. Estimated scale parameters are given in the Figure 6 (normal component distribution) and Figure 8 (lognormal component distribution) and labelled sigma1 and sigma2 on the horizontal axis. The "true" values in these figures are 2 and 1 for the normal components in the Figure 6 and 1 and 0.64 for the lognormal components in the Figure 8.

Figure 5: Distribution of location parameters of normal distribution, unequal variances, *n*=1,000, 10 % of censored data (Table 2)



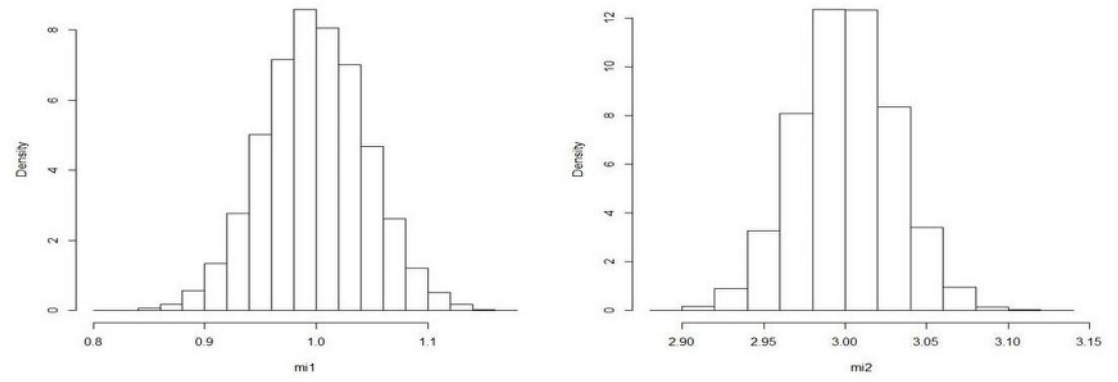
Source: own computations

Figure 6: Distribution of scale parameters of normal distribution, unequal variances, *n*=1,000, 10 % of censored data (Table 2)



Source: own computations

Figure 7: Distribution of location parameters of lognormal distribution, unequal variances, *n*=1,000, 10 % of censored data (Table 4)



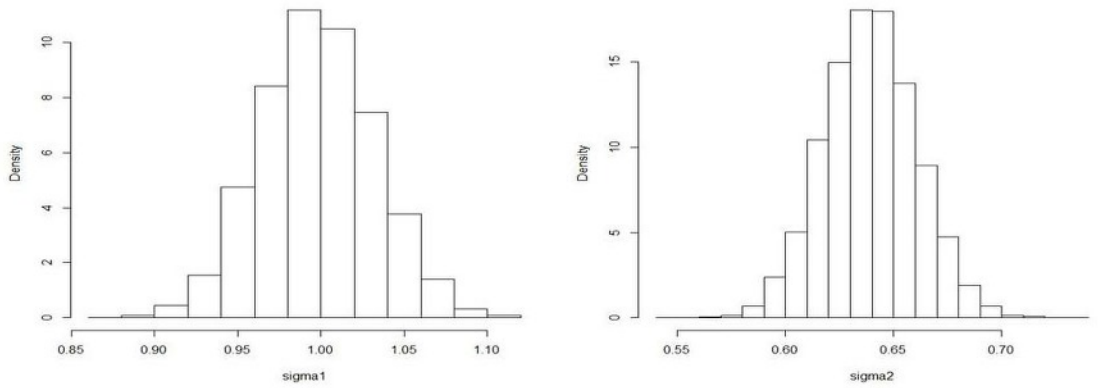
Source: own computations

Conclusion

In this text the problem of estimation of parameters of the finite mixtures of distributions with right censored data was treated. The problem with known (or observable) component membership was discussed; possible methods for models without this information (with artificial components constructed to obtain the best model to fit data) are mentioned together with citations.

It was shown that for relatively large samples (with the use of the simulation) and the percentage of censored data under 50 % the maximum likelihood estimates well approximate unknown values of parameters. The numeric procedure that enables the use of covariances was used to incorporate prior information about parameters (equality of scale parameters) and observed component membership.

Figure 8: Distribution of scale parameters of lognormal distribution, unequal variances, *n*=1,000, 10 % of censored data (Table 2)



Source: own computations

In the histograms based on 10,000 of fits the asymptotic normality of estimates was illustrated.

All simulations and numeric computations were performed in the program R. This program was introduced as the useful tool to be used for estimation problems with various so-called survival distributions (these distributions include normal and lognormal distributions that are used in the text). Moreover, user defined probability distributions can be used in the package Survival through the user defined probability density function called by the procedure.

Acknowledgement

The research was supported by the grant IG-410062 from the University of Economics in Prague, Czech Republic.

References

McLachlan, G., Peel, D. Finite Mixture Models. Wiley Series in Probability and Statistics, John Wiley. 2000. http://dx.doi.org/10.1002/0471721182

Pilla, R.S., Lindsay, B.G. Alternative EM Methods for Nonparametric Finite Mixture Models, Biometrika, Vol. 88, No. 2, 535-550. 2001. http://dx.doi.org/10.1093/biomet/88.2.535

Scallan, A.J. Fitting a mixture distribution to complex censored survival data using generalized linear model. Journal of Applied Statistics, Vol. 69, No. 6, 747-753. 1999. http://dx.doi.org/10.1080/02664769922188

Svensson, I, Sjostedt-de Luna, S. Asymptotic properties of a stochastic EM algorithm for mixtures with censored data. Journal of Statistical Planning and Inference. Vol. 140, 111–127. 2010. http://dx.doi.org/10.1016/j.jspi.2009.06.014

Svensson, I., Sjäostedt-de Luna, S. and Bondesson, L. Estimation of wood fibre length distributions from censored data through an EM algorithm. Scand. J. Statist. Vol.33, 503-522. 2006. http://dx.doi.org/10.1111/j.1467-9469.2006.00501.x

Therneau, T. A Package for Survival Analysis in S. R package version 2.37-2. (2012). http://CRAN.R-project.org/package=survival

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