ESTIMATION OF PARAMETERS IN FINITE MIXTURES FROM CENSORED DATA

Ivana Malá

Abstract

In the contribution the problem of estimation of parameters in mixtures of probability distributions in case of the presence of right censored data is treated. Only models with known number of components and observed component membership (complete models) are studied in the text. 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. 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.

Key words: censored data, finite mixture, parameter estimation, simulation

JEL Code: C41, C63, C88

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 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 observations included in given subpopulation. The aim of the analysis is then an estimation of unknown parameters. The problem will be shortly described in the part 1. Then some results of simulations are shown in the part 2.

1 Methods

1.1 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 for given K components there are probability densities $f_j(y;\boldsymbol{\theta}_j)$ (j = 1,., K) depending on p dimensional (in general unknown) vector parameter $\boldsymbol{\theta}_j$. Furthermore, K weights π_j fulfil obvious constraints $\sum_{j=1}^{K}\pi_j = 1,\ 0 \le \pi_j \le 1,\ j = 1,.,K$. A density of the mixture of these probability distributions is defined as a weighted average of densities f_j with weights π_j as

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

The mixture density (1) depends on the vector parameter $\boldsymbol{\psi}$, $\boldsymbol{\psi} = (\pi_1,.,\pi_{K-1},\boldsymbol{\theta}_j,\ j = 1,.,K)$, 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. 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 distribution (1) is of the form

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

and in the case of lognormal distribution

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

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 used. It means that the likelihood function $L(\boldsymbol{\psi})$

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

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 contribution of the i-th observation to the function L in (4) is only $\pi_j f_j\left(y_i; \boldsymbol{\theta}_j\right)$ (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

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

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 (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 observed, the vectors $\mathbf{z}$ are random vectors and the formula (5) cannot be used. For the estimation so-called *EM* algorithm is frequently used (Titterington & al., 1985, McLachlan, Peel, 2000).

1.2 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 $\pi_j f_j\left(y_i; \boldsymbol{\theta}_j\right)$ (as above). If the observation is right censored at time y_i (it means that $Y_i > y_i$), its contribution to (4) is $\pi_j\left(1 - F_j\left(y_i; \boldsymbol{\theta}_j\right)\right)$. The formula (5) can be rewritten as

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

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

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 all parameters in the part 2 the package

Survival in 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 can be used. In this text known group 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 propositions about distributions (not used in this text that is a part of a wider simulation study).

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.

2 Simulation and results

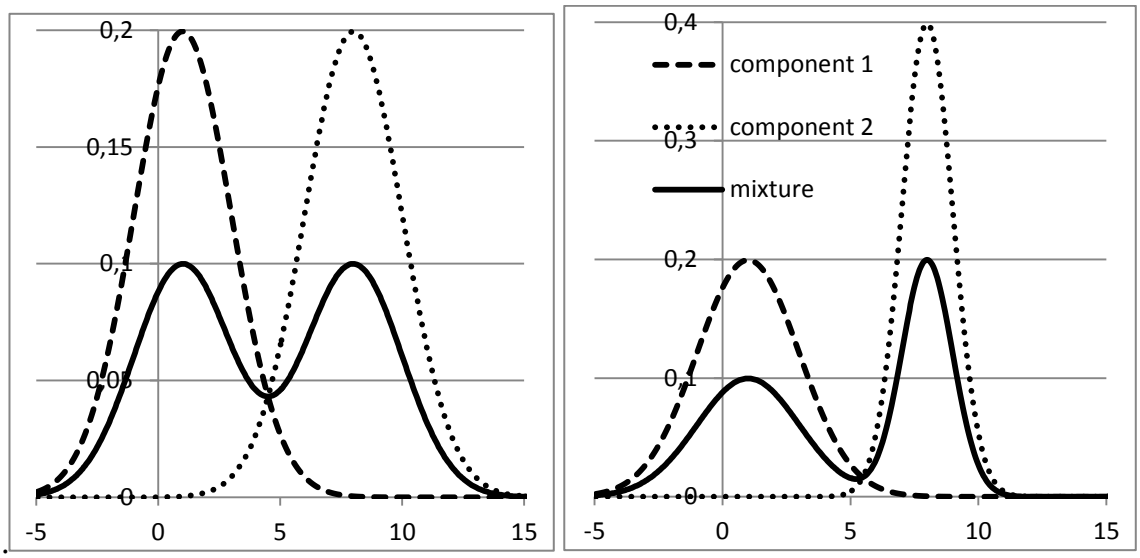
In order to illustrate estimation of the parameters and their properties a simulation was performed. 10,000 samples were generated from the mixtures of two components (with equal mixing proportions $\pi_1 = \pi_2 = 0.5$) from normal (symmetric distribution) and lognormal distributions (positively skewed distribution). In all simulations samples of 500 and 1,000 observations were generated with 10%, 30% and 50% of right censored observations. Note that under these assumptions there are 250 (500 observations) in each component. From all samples unknown parameters were estimated and then their minimum, mean, maximum and standard deviation were evaluated. These values are given in tables. The distribution of parameters is well described with the use of histograms.

2.1 Normal distribution

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 distributions (μ_1, μ_2, σ) (we have $\pi_1 = \pi_2 = 0.5$). 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.

Suppose now, that parameters σ are not equal. In this case a separation of components according to (5) is used and the fit is performed separately in each component. In the Figure 1 (right part) the density of the mixture of two normal densities $\left(N(1,4), N(8,1)\right)$ with the mixing proportions 0.5 is shown.

Fig. 1: Normal mixture with two components with equal scale parameters (left) and unequal scale parameters (right)



Source: own computations

In the Table 1 sample characteristics mean, minimum, maximum evaluated from 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 shift to the right from given values of parameters is obvious. Its value decreases with sample size and the percentage of censored data. The sample with a half of observations is heavy-censored but we frequently

Tab. 1: Sample characteristics of estimates: mean (minimum-maximum)

	n=1,000 10 percent	*n*=1,000 30 percent	*n*=1,000 50 percent
$\hat{\mu}_1$	1.19 (0.80-1.53)	1.67 (1.35-2.04)	2.36 (2.00-2.72)
$\hat{\mu}_2$	8.18 (7.84-8.51)	8.67 (8.32-9.00)	9.36 (8.95-9.75)
$\hat{\sigma}$	2.06 (1.85-2.26)	2.21 (1.99-2.45)	2.43 (2.14-2.71)
	n=500 10 percent	*n*=500 30 percent	*n*=50050 percent
$\hat{\mu}_1$	1.66 (1.11-2.17)	1.18 (0.70-1.70)	2.35 (1.84-2.85)
$\hat{\mu}_2$	8.66 (8.15-9.14)	8.18 (7.69-8.68)	9.35 (8.81-9.87)
$\hat{\sigma}$	2.21 (1.87-2.56)	2.06 (1.78-2.30)	2.42 (1.99-2.81)

Source: own computations

meet such situation in applications. Sample variances of both estimated parameters are similar and approximately 0.05 (σ) and 0.06 (μ_i) for sample size 500 and 0.04 for sigma and 0.05 for estimates of μ for sample size 1000. The range is also similar for all selected percentages of censored data. In the case of unequal 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. In this case means are similar for smaller and larger samples. Range of values is greater for small sample (and it is true also for standard deviation).

Tab. 2: Sample characteristics of estimates: mean (minimum-maximum)

	n=1,000 10 percent	*n*=1,000 30 percent	*n*=1,000 50 percent
$\hat{\mu}_1$	1.19 (0.83-1.52)	1.67 (1.31-2.02)	2.36 (2.02-2.73)
$\hat{\sigma}_1$	2.06 (1.82-2.33)	2.21 (1.93-2.52)	2.43 (2,03-2.81)
$\hat{\mu}_2$	8.09 (7.92-8.27)	8.35 (8.15 – 8.50)	8.68 (8.46-8.82)
$\hat{\sigma}_2$	1.03 (0.91-1.16)	1.11 (0.96-1.25)	1.21 (1.00-1.39)
	n=500 10 percent	*n*=500 30 percent	*n*=500 50 percent
$\hat{\mu}_1$	1.18 (0.72-1.72)	1.67 (1.18-2.12)	2.34 (1.80-2.84)
$\hat{\sigma}_1$	2.05 (1.73-2.44)	2.21 (1.80-2.62)	2.42 (1,84-3.07)
$\hat{\mu}_2$	8.09 (7.83-8.34)	8.33 (8.09-8.55)	8.68 (8.38-8.92)
$\hat{\sigma}_2$	1.02 (0.84-1.22)	1.10 (0.87 -1.13)	1.21 (0.95-1.52)

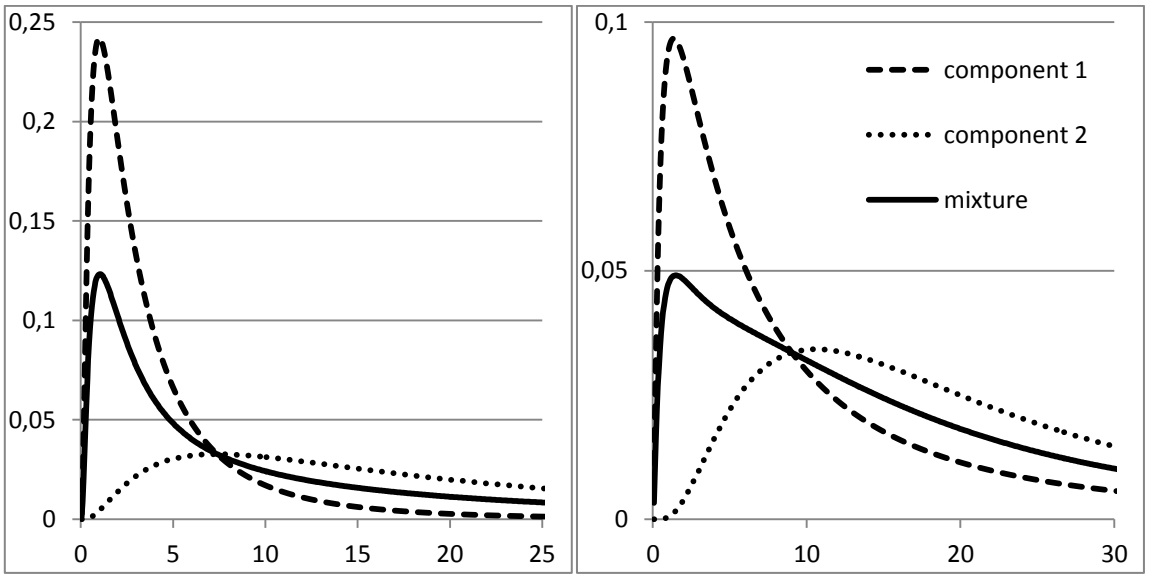
Source: own computations

2.2 Lognormal distribution

For the lognormal distribution component densities are not well separated (Figure 2) as it was for 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 have again two components with equal mixing proportions. Figure 2 (left part) and Table 3 refers to the components with equal scale parameters sigma $\left(LN(1;1), LN(3;1)\right)$. Figure 2 (right part) and Table 4 deals with component probability distributions $\left(LN(1;1), LN(3;0.64)\right)$. Note now, that in both examples components have unequal variances as parameters σ are standard deviances of logarithms of *Y* and variance of the lognormal distribution depends not only on σ but also on the parameter μ. Results were 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.

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



Source: own computations

In the Tables 3 and 4 shifts to the right are again obvious. The higher the percentage of censored data, the more shifted the mean of estimates to the right occurs. Ranges for estimates of sigma in the Table 2 are smaller than for estimates of μ_i (values of μ_i are estimated separately in each component from one half of observations).

Tab. 3: Sample characteristics of parameters: mean (minimum-maximum)

	n=1,000 10 percent	n=1,000 30 percent	n=1,000 50 percent
$\hat{\mu}_1$	1.09 (0.87-1.32)	1.31 (1.13-1.48)	1.63 (1.76-1.81)
$\hat{\mu}_2$	3.08 (2.89-3.27)	3.30 (3.13-3.44)	3.60 (3.56-3.76)
$\hat{\sigma}$	0.93 (0.80-1.07)	1.00 (0.90-1.11)	1.10 (0.98-1.23)
	n=500 10 percent	n=500 30 percent	n=500 50 percent
$\hat{\mu}_1$	1.09 (0.92-1.26)	1,31 (1.08-1.56)	1.62 (1.36-1,88)
$\hat{\mu}_2$	3.08 (2.95-3.21)	3,29 (3.07-3.51)	3.59 (3.38-3.82)
$\hat{\sigma}$	0.93 (0.86-1.01)	1,00 (0.87-1.16)	1.09 (0.90-1.29)

Source: own computations

In the Figures 3-5 histograms of estimates for samples sizes 500 and 30 % of censored observations for both equal and unequal values of sigma are shown. In all the figures we can see approximately normal distribution of estimates, however all histograms show slightly

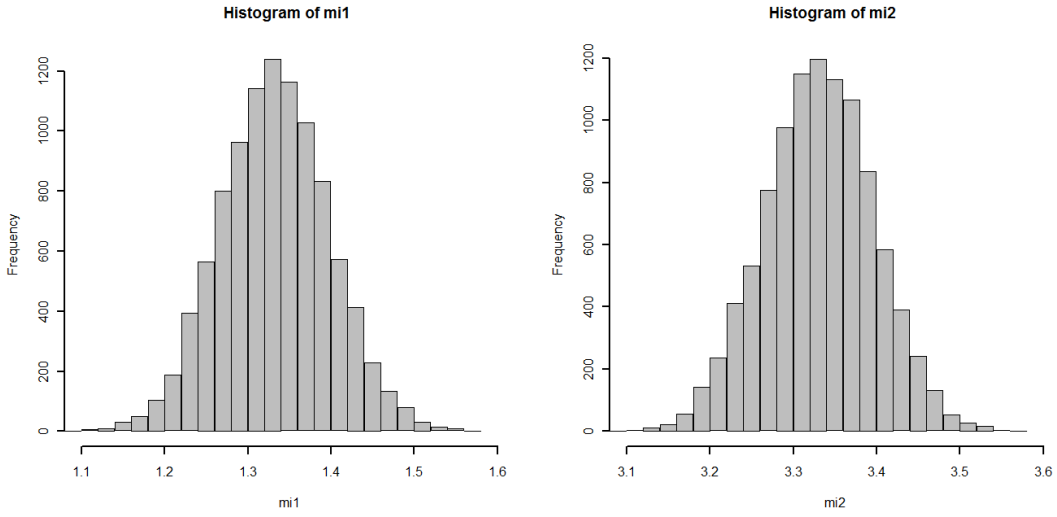
negatively skewed estimates of parameters. These pictures illustrate central limit theorem and asymptotic normality of estimates.

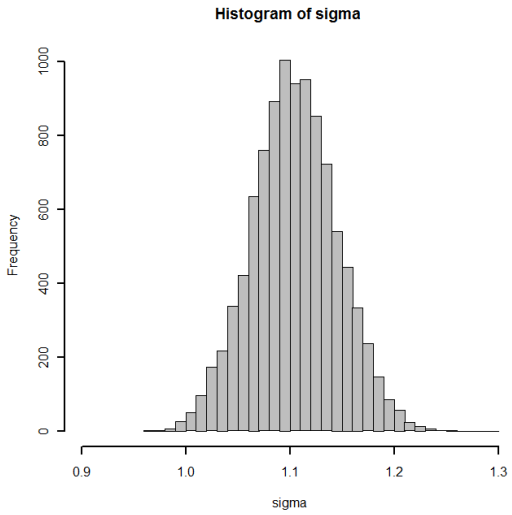
Tab. 4: Sample characteristics of parameters: mean (minimum-maximum)

	n=1,000 10 percent	*n*=1,000 30 percent	*n*=1,000 50 percent
$\hat{\mu}_1$	1.02 (0.89-1.36)	1.43 (0.21-1.64)	1.89 (0.66-2.12)
$\hat{\sigma}_1$	1.34 (1.18-1.50)	1.44 (1.23-1.63)	1.58 (1,32-1.83)
$\hat{\mu}_2$	3.08 (2.93-3.22)	3.27 (3.12-3.42)	3.54 (3.37-3.71)
$\hat{\sigma}_2$	0.82 (0.72-0.92)	0.89 (0.77-1.00)	0.97 (0.80-1.11)
	n=500 10 percent	*n*=500 30 percent	*n*=500 50 percent
$\hat{\mu}_1$	1.12 (0.75-1.45)	1.43 (0.11-1.76)	1.88 (1.52-2.19)
$\hat{\sigma}_1$	1.33 (1.09-1.57)	1.43 (1.15-1.72)	1.57 (1.19-1.99)
$\hat{\mu}_2$	3.07 (2.86-3.27)	3.27 (3.07-3.46)	3.54 (3.31-3.74)
$\hat{\sigma}_2$	0.82 (0.68-0.98)	0.88 (0.70-1.05)	0.97 (0.76-1.22)

Source: own computations

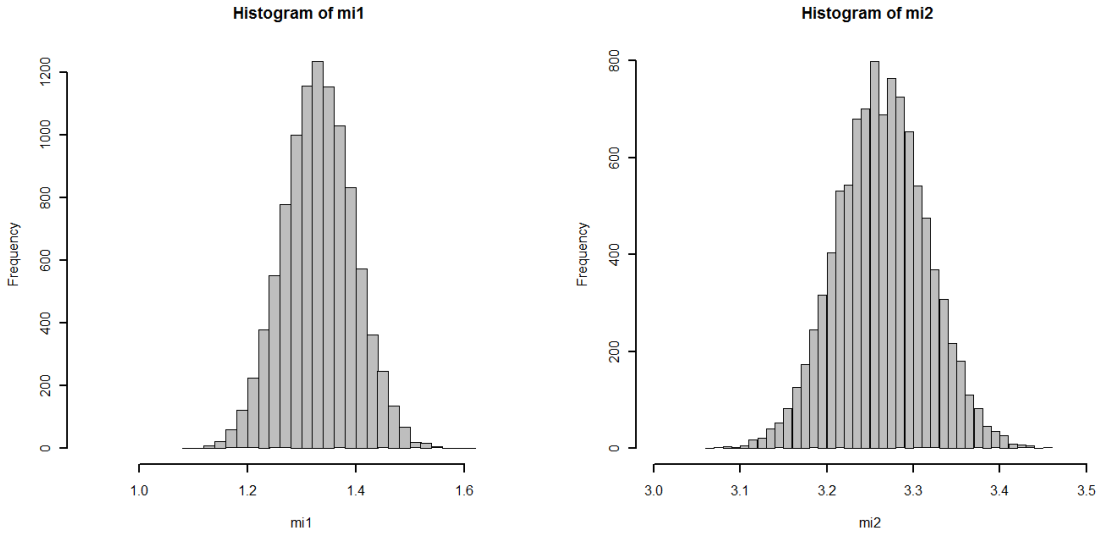
Fig. 3: Distribution of location parameters (Table 3)





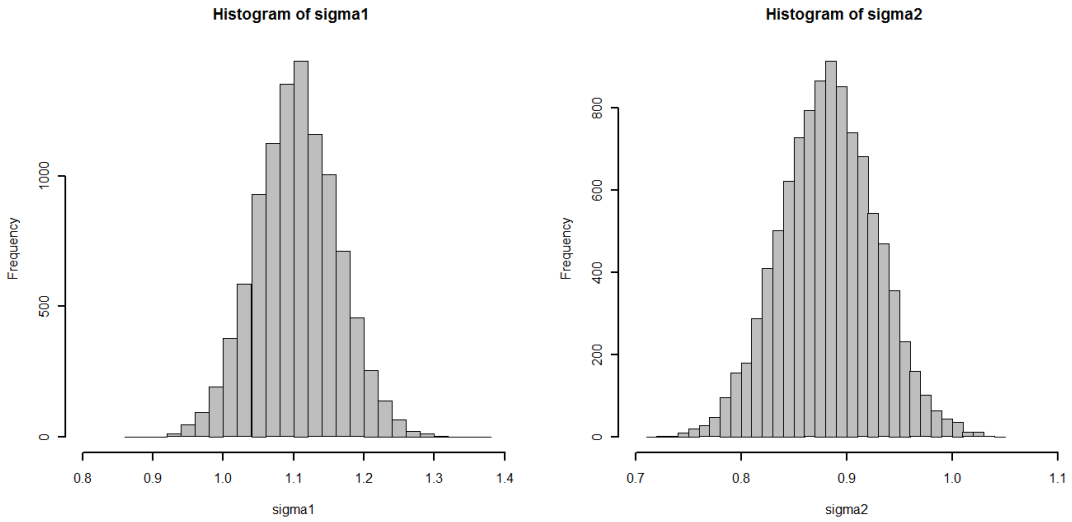
Source: own computations

Fig. 4: Distribution of location parameters (Table 4)



Source: own computations

Fig. 5: Distribution of scale parameters σ (Table 4)



Source: own computations

Conclusions

In this text the problem of estimation of parameters of the mixtures with censored data was treated. It was shown (with the use of the simulation) that even in the case of large samples and known component membership the estimates are shifted to the right (due to right censored data). In the figures asymptotic normality of estimates was illustrated. All simulations and computations were performed in the program R. This program was introduced as the useful tool to be used for such estimation problems.

References

McLachlan, G., Peel, D. *Finite Mixture Models*. Wiley Series in Probability and Statistics, John Wiley. 2000.

Pilla, R.S., Lindsay, B.G. Alternative EM Methods for Nonparametric Finite Mixture Models, *Biometrika*, Vol. 88, No. 2, 535-550. 2001.

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.

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.

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.

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

Contact

Ivana Malá
University of Economics in Prague
W.Churchilla pl. 4, Prague 3
malai@vse.cz