Practical Initialization of Recursive Mixture-based Clustering for Non-negative Data

Evženie Suzdaleva[1] and Ivan Nagy[1,2]

[1] Department of Signal Processing,
The Czech Academy of Sciences, Institute of Information Theory and Automation,
Pod vodárenskou věží 4, 18208 Prague, Czech Republic,
{nagy,suzdalev}@utia.cas.cz,
WWW home page: http://www.utia.cas.cz/people/nagy
[2] Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25,
11000 Prague, Czech Republic

Abstract. The paper provides a practical guide on initialization of the recursive mixture-based clustering of non-negative data. For modeling the non-negative data, mixtures of uniform, exponential, gamma and other distributions can be used. Initialization is known to be an important task for a start of the mixture estimation algorithm. Within the considered recursive approach, the key point of initialization is a choice of initial statistics of the involved prior distributions. The paper describes several initialization techniques for the mentioned types of components that can be beneficial primarily from a practical point of view.

Keywords: mixture-based clustering, recursive mixture estimation, different components, non-negative data, Bayesian estimation

1 Introduction

The Bayesian mixture estimation [1–4], which makes the basis for the clustering considered in this paper, performs the estimation online. It means that in the beginning of the estimation, where smaller data sets are used, the parameters can be determined very inaccurately and their point estimates have large variances. This can be unacceptable in some application domains with demands of quick and effective estimation (e.g., fault detection, online diagnostics, medicine, etc.). The model estimation is usually just a preparation for other tasks that use the estimated model. This may be, for example, the prediction of the system output or its control. Then at the beginning, a poorly estimated model may give either completely wrong predictions or faulty control values that can damage the controlled system. Within the clustering tasks, it leads to an unsuccessful search for data clusters. Therefore, it is very important to pre-set the task which includes the model estimation before its start, so that the model has already been roughly adjusted and the estimation has only "fine-tuned" it. This is exactly the main feature of the initialization problem.

The Bayesian approach to estimation allows us such a way. It means that the prior information can be used for this aim by means of preparing the prior

distribution statistics, so that the parameters from them are roughly matched with reality. However, the non-trivial question is how to convert the prior information about the system to the prior statistics. In the field of mixture-based clustering [5–7], the prior information should be transformed to the statistics of the mixture components used.

Different distributions are used within this task [8–11]. Gaussian mixtures are probably the most frequently met models, see, e.g., [12–14], etc.

This paper continues a line started in paper [15], which considered the initialization task for the clustering with uniform components of the mixture model. The uniform components are beneficial for applications where the analysis of data with fixed boundaries is required. This paper extends the study [15] for the domain of non-negative data, which means that other suitable distributions can be taken as the mixture components. The mixture initialization with them is not a trivial task.

The majority of studies, e.g., [16–20] are devoted to the initialization of the expectation-maximization (EM) algorithm [21] used in iterative approaches to mixture estimation. In this paper, similarly as in [15], the mixture-based clustering based on the recursive Bayesian estimation avoiding iterative computations is used. It was considered for normal models in [22] and for normal mixtures in [1–3]. A series of other components is discussed in [9, 23, 4]. Within the mentioned framework, the initialization is primarily concerned with a choice of (i) the number of components, (ii) the initial statistics of a model of switching the components and (iii) the initial statistics of components. In this area, paper [24] based on [25] is also found, again devoted to the initialization with normal mixtures.

This paper explores initialization approaches for the estimation of the mixture of uniform, Bernoulli, geometric, exponential and Gamma components. The main emphasis is on the choice of the initial statistics of components. The discussed methods are based on the use of prior data.

The paper is organized in the following way. Section 2 introduces a mixture model along with different types of its components as well as a model of their switching. Section 3 presents a brief summary of recursive Bayesian mixture estimation algorithm. Section 4 specifies the initialization problem and discusses the initialization techniques for all of the mentioned types of the components. Conclusions and open problems are given in Section 5.

2 Models

Let us consider a multi-modal system, which generates the continuous data vector y_t at each discrete time instant $t = 1, 2, ...$ The system is assumed to work in m_c working modes. Each of them is indicated at the time instant t by the value of the unmeasured dynamic discrete variable $c_t \in \{1, 2, \ldots, m_c\}$, which is called the pointer [1].

For description of such the multi-modal system a mixture model is used, which is here comprised of m_c components in the form of the following probability

density functions (pdfs)

$$f(y_t|\Theta, c_t = i),\ i \in \{1, 2, \ldots, m_c\}, \tag{1}$$

where $\Theta = \{\Theta_i\}_{i=1}^{m_c}$ is a collection of unknown parameters of all components, and Θ_i includes parameters of the i-th component in the sense that $f\left(y_t|\Theta, c_t = i\right) = f\left(y_t|\Theta_i\right)$ for $c_t = i$.

The general component pdf (1) is specified in dependence of the data nature and certain model assumptions. In this paper, several types of components are considered as follows.

2.1 Uniform Components

Under assumption of the independence of individual entries of the vector y_t, the uniform pdf (1) takes the following form $\forall i \in \{1, 2, \ldots, m_c\}$

$$f\left(y_t|L, R, c_t = i\right) = \begin{cases} \frac{1}{R_i - L_i} & \text{for } y_t \in (L_i, R_i)\,, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\{L_i, R_i\} \equiv \Theta_i$, and their entries $(L_l)_i$ and $(R_l)_i$ are minimal and maximal bounds of the l-th entry $y_{l;t}$ of the K-dimensional vector y_t within the i-th uniform component.

2.2 Bernoulli Components

The Bernoulli component (1) is a special case of the categorical components considered in [2, 4]. Here for the sake of simplicity, it is used in the form

$$f\left(y_t|\Theta, c_t = i\right) = \Theta_{i;1}^{y_t}\Theta_{i;0}^{1-y_t}, \tag{3}$$

where $y_t \in (0, 1)$ and $\Theta_i = [\Theta_{i;0}, \Theta_{i;1}]$ are parameters of the i-th component, $\Theta_{i;0}$ is the probability of $y_t = 0$ and $\Theta_{i;1}$ of the value of 1. It holds $\Theta_{i;0}, \Theta_{i;1} \geq 0$ and $\Theta_{i;0} + \Theta_{i;1} = 1$, which means $\Theta_{i;1} = 1 - \Theta_{i;0}$.

2.3 Geometric Components

The geometric distribution with a large number of possible values can be used for modeling the non-negative data as well. Here, for the sake of simplicity, it is considered for the case $y_t \in \{0, 1\}$. Based on the Bernoulli distribution, the geometric component (1) has the form

$$f\left(y_t|\Theta, c_t = i\right) = \Theta_i\left(1 - \Theta_i\right)^{1-y_t},\ \text{where } \Theta_i \in (0, 1)\,,\ y_t \in \{0, 1\}. \tag{4}$$

2.4 Exponential Components

The exponential distribution is well suited for modeling the non-negative data. The exponential component (1) is the pdf

$$f\left(y_t|\Theta, c_t = i\right) = \Theta_i \exp\left\{-\Theta_i y_t\right\},\ \text{where } \Theta_i > 0,\ y_t \geq 0. \tag{5}$$

2.5 Gamma Components

The Gamma pdf is the generalization of the exponential distribution. It is suitable for modeling the non-negative data, whose maximum frequency does not lie at zero or near zero. The probability density decreases exponentially with increasing argument. The Gamma component (1) has the form

$$f\left(y_t|\gamma,\beta,c_t=i\right)=\frac{\beta_i^{\gamma_i}}{\Gamma\left(\gamma_i\right)}y_t^{\gamma_i-1}\exp\left\{-\beta_i y_t\right\}, \tag{6}$$

where $\Theta_i=\left\{\gamma_i,\beta_i\right\},\gamma_i>0,\beta_i>0,y_t\geq 0$.

2.6 Pointer Model

A component, which describes data generated by the system at the time instant t is said to be active. Switching the active components is described by a model of the pointer c_t as follows:

$$f\left(c_t=i|\alpha,c_{t-1}=j,z_t=k\right)= \tag{7}$$

	$c_t=1$	$c_t=2$	$\cdots$	$c_t=m_c$
$c_{t-1}=1$	$(\alpha_{1\vert 1})_k$	$(\alpha_{2\vert 1})_k$	$\cdots$	$(\alpha_{m_c\vert 1})_k$
$c_{t-1}=2$	$(\alpha_{1\vert 2})_k$		$\cdots$	
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$c_{t-1}=m_c$	$(\alpha_{1\vert m_c})_k$		$\cdots$	$(\alpha_{m_c\vert m_c)_k}$

where the unknown parameter α is the $(m_c\times m_c)$-dimensional matrix, which exists for each value $k\in\{1,2,\ldots,m_z\}$ of the discrete variable z_t obtained from y_t by its discretization. Its entries $(\alpha_{i|j})_k$ are non-negative probabilities of the pointer $c_t=i$ under condition that the previous pointer $c_{t-1}=j$ with $i,j\in\{1,2,\ldots,m_c\}$ and the variable $z_t=k$.

3 Mixture-Based Clustering Summary

The clustering considered in this paper is based on recursive mixture estimation algorithms, most of those are described in literature, e.g., [1, 22, 2–4]. The key point of the recursive clustering is to estimate parameters of components and the pointer model and determine which component is active at time t, i.e., currently generates data.

The following algorithmic scheme of the clustering summarizes its main steps at each time instant:

1. Measuring the new data item;
2. Computing the proximity of the data item to individual components, see [26];
3. Computing the probability of the activity of components (i.e., weights) using the proximity, the point estimate of the pointer model and the past activity, where the maximal probability declares the currently active component, see [1–3];

4. Classifying data according to the declared active component;
5. Updating the statistics of all components and the pointer model, see [1, 2, 4, 15], etc.;
6. Re-computing the point estimates of parameters necessary for calculating the proximity, see [22, 2, 27], etc.
7. Go to Step 1.

These steps belong to the online part of the estimation, which should be initialized before a start.

4 Mixture Initialization

The initialization task is specified for the above recursive algorithm in the following way. For time $t = 0$, it is necessary to set:

- the number of components m_c,
- the initial weighting vector,
- the initial statistics of the pointer model and the components.

The number of components can be determined offline using prior data, for example, by their visualization, e.g., [24] or with the help of well-known clustering methods such as, e.g., k-means [28], etc.

As regards the initial weighting vector and the pointer statistics, it is sufficient to initialize them either uniformly or randomly in combination with their updating by prior data.

The choice of the initial statistics of the components is the key point. It is explained by computing the proximity value, which depends on the parameter point estimates and, therefore, on the component statistics. With the accurately chosen number of components and the pointer statistics the proximity with wrong initial component statistics leads to the unsuccessful clustering. The subsequent sections are devoted to this part of the initialization task.

4.1 Initialization with Uniform Components

The mixture initialization for the case of uniform components (2) was described in details in [15]. This section summarizes this initialization approach within the bounds of the task of modeling the non-negative data.

The initial statistics of the uniform components can be chosen according to the following four techniques.

Component Centers via Mid-point Update One of the approaches is to find centers of components instead of the left and right bounds for initial detection of components [23]. In this case the statistics $(s_{l;0})_i$, $(q_{l;0})_i$ should be used, which are l-th entries of the K-dimensional vectors s_t and q_t, where the last comprises a diagonal of a matrix [15]. Starting from random values, they

are updated using a set of prior data $\forall i \in \{1, 2, \ldots, m_c\}$ and $\forall l = \{1, \ldots, K\}$ in the following way.

$$(s_{l;t})_i = (s_{l;t-1})_i + w_{i;t} y_{l;t}, \tag{8}$$

$$(q_{l;t})_i = (q_{l;t-1})_i + w_{i;t} y_{l;t}^2, \tag{9}$$

where $w_{i;t}$ is a weight of the i-th component, see for details [23, 15], etc. After updating they are used to compute the point estimates of the mid-point and mid-range vectors of each component $(S_t)_i$ and $(h_t)_i$ respectively as follows (similarly as for normal components).

$$(\hat{S}_t)_i = (s_t)_i / t, \tag{10}$$

$$(D_t)_i = \left((q_t)_i - (s_t)_i (s_t')_i / t\right) / t, \tag{11}$$

$$(\hat{h}_t)_i = \sqrt{3\,\mathrm{diag}((D_t)_i)}, \tag{12}$$

where $(D_t)_i$ is the covariance matrix of the uniform pdf, and $\sqrt{3\,\mathrm{diag}((D_t)_i)}$ denotes the square roots of entries of the vector $\mathrm{diag}((D_t)_i)$. (10) and (11) from the previous time instant are placed instead of the expectation and the covariance matrix into the proximity, see [15]. In the end of updating by prior data the mid-point $(\hat{S}_{l;t})_i$ is the center of the i-th component for the l-th data entry. The point estimates of the minimum and maximum bounds are then obtained as

$$(\hat{L}_{l;t})_i = (\hat{S}_{l;t})_i - \varepsilon, \tag{13}$$

$$(\hat{R}_{l;t})_i = (\hat{S}_{l;t})_i + \varepsilon, \tag{14}$$

with small ε, and they are used during the on-line estimation.

Centers Based on *K*-means Another way is to use the centers of clusters initially detected by the k-means method [28] from prior data and put them into (13) and (14) to be used during the on-line estimation.

Centers as Averages The average values from individual prior data entries with small deviations can be taken as initial centers of components and then substituted into (13) and (14).

Bounds as Minimum and Maximum Here the minimum and maximum values of corresponding entries of the data vector y_t are used directly as the component statistics denoted by $(\mathcal{L}_{l;0})_i$ and $(\mathcal{R}_{l;0})_i$ respectively.

Finally, the main results of the first three techniques above are $(\hat{S}_{l;T})_i$, which is the center of the i-th component for the l-th entry of y_t and T is the number of prior data items. With the help of the last technique, the initial bounds of components are obtained.

For the on-line (i.e., for $t = T+1, T+2, \ldots$) estimation of the component bounds and classification of data among components according to the actual maximum weight, the algorithm summarized in Section 3 is applied. For the three first initialization techniques, relations (13) and (14) should be used before measuring the first data item y_t.

Results A set of anonymized medical hematological prior data is used for demonstration of the initialization approach for the uniform components. The following specific variables comprise the 8-dimensional vector y_t:

- $y_{1;t}$ – precollection number of leucocytes, $[10^9/l]$;
- $y_{2;t}$ – precollection number of HTK, [%];
- $y_{3;t}$ – precollection number of Hemoglobin (Hbg), $[g/dl]$;
- $y_{4;t}$ – precollection number of platelet count (PLT), $[10^9/l]$;
- $y_{5;t}$ – precollection number of CD34+, $[\mu l]$;
- $y_{6;t}$ – precollection number of CD34+ in total blood volume (TBV), $[10^6]$,
- $y_{7;t}$ – concentration of mono-nuclear cells (MNC), [%];
- $y_{8;t}$ – concentration of CD34+/kg, $[10^6]$.

The number of components is initialized as 3. The verification of the initialization techniques is performed according to the following three criteria.

Evolution of component weights Evolution of component weights, which express the activity of components, is observed during the on-line estimation. The rare activity of some component or its absence indicates that the number of components is incorrectly initialized and probably too high. The regular activity of all components validates the correct choice of the number of components.

A fragment of the evolution of the component weights with the statistics initialized via the mid-point update is demonstrated in Figure 1. It can be seen that all three components are regularly active. The k-means based initial statistics give the similar activity, see Figure 2.

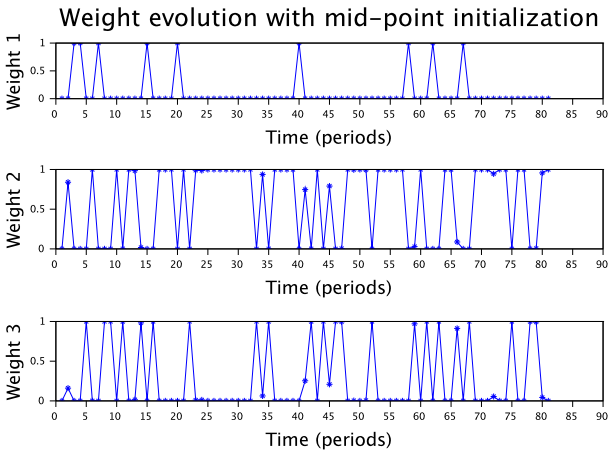

Fig. 1. The weight evolution with the initialization via the mid-point update [15].

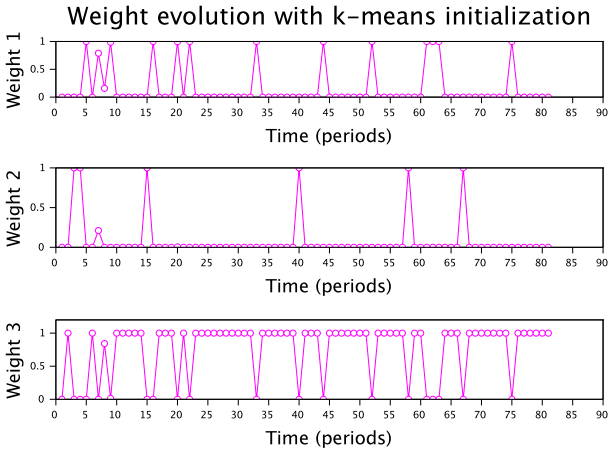

Fig. 2. Evolution of component weights with the initialization based on k-means [15].

The initialization via centers as averages is shown in Figure 3. It produces a bit more probabilities close to 0.5. However, in general, the result is similar to two first methods.

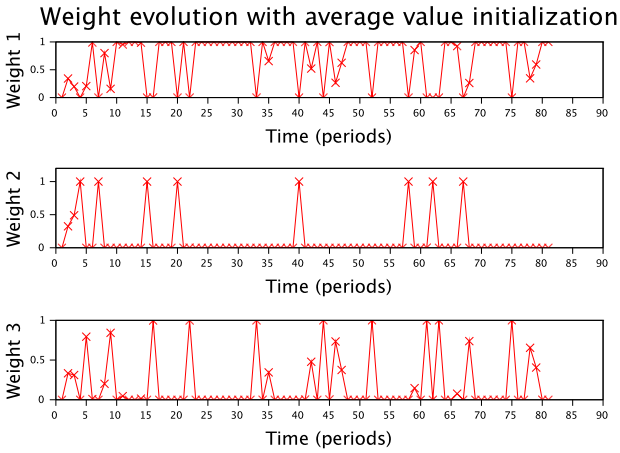

Fig. 3. Evolution of component weights with the initialized centers as averages [15].

The last method based on minimum and maximum prior values provides only two detected components. Figure 4 shows at the y-axis that the weights of the

first component in the top plot are too low, and this component is never declared to be active.

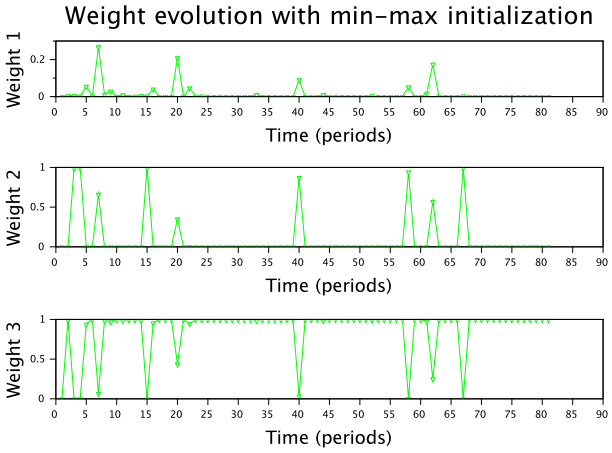

Fig. 4. The weight evolution with the initialization based on minimum and maximum prior values [15].

Evolution of Bounds Evolution of the point estimates of component parameters (i.e., bounds) is monitored at the beginning of the on-line estimation. Fast locating the stabilized values of the point estimates means that the initialization is successful. Comparing the evolution of the minimum and maximum bounds of individual entries within each component, it can be noticed that a speed of localization of stabilized estimate values is similar for the first three methods, i.e., the bounds of the components detect their final values relatively quickly, see Figure 5.

The initialization according to minimum and maximum prior values provides a worse stabilization in search of the values of the bounds, see an example of the left bound evolution for the third component in Figure 6, where the evolution of the left (minimum) bounds of individual data entries is presented.

Clusters The shape and the location of final clusters detected in the data space by starting the estimation algorithm with the mentioned initialization techniques are compared. Comparison with k-means clustering is also demonstrated. Clusters of the most interesting pair of data entries from the practical (hematological) point of view are presented here. The entries $y_{5;t}$, which is the precollection number of CD34+, and $y_{8;t}$, which is the concentration of CD34+/kg, are chosen. Their clusters detected according to the estimated pointer value can be seen in Figures 7 and 8, where the comparison of the results initialized according to all

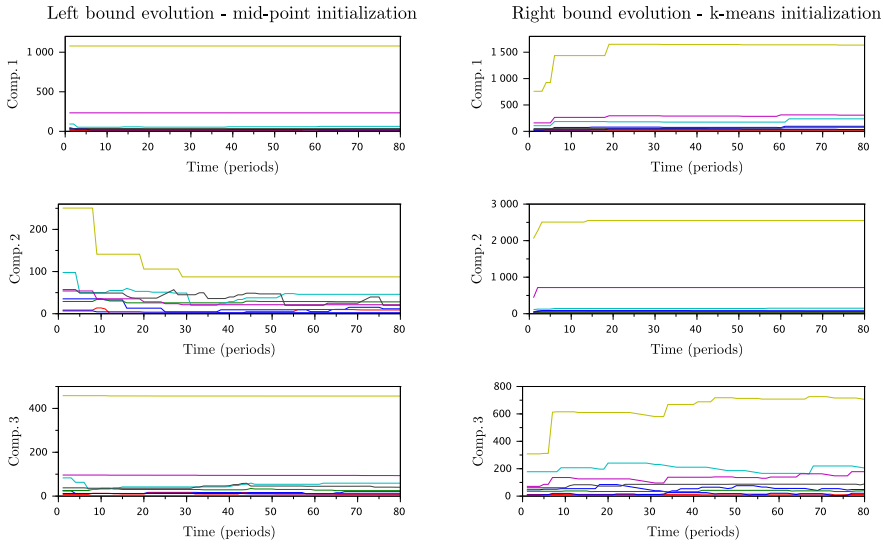

Fig. 5. Example of the bound evolution with the initialization via the mid-point update (left) and k-means (right) [15].

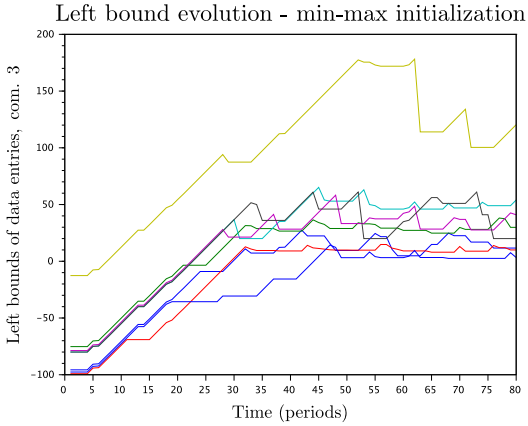

Fig. 6. Evolution of the left bounds of the third component with the initialization according to minimum and maximum prior values [15].

of the discussed methods is demonstrated. The colors of the clusters in the figure are chosen randomly in all of the plots. The clusters are enumerated according to the order in which they have been detected and plotted. The shapes and the location of the detected clusters should be compared.

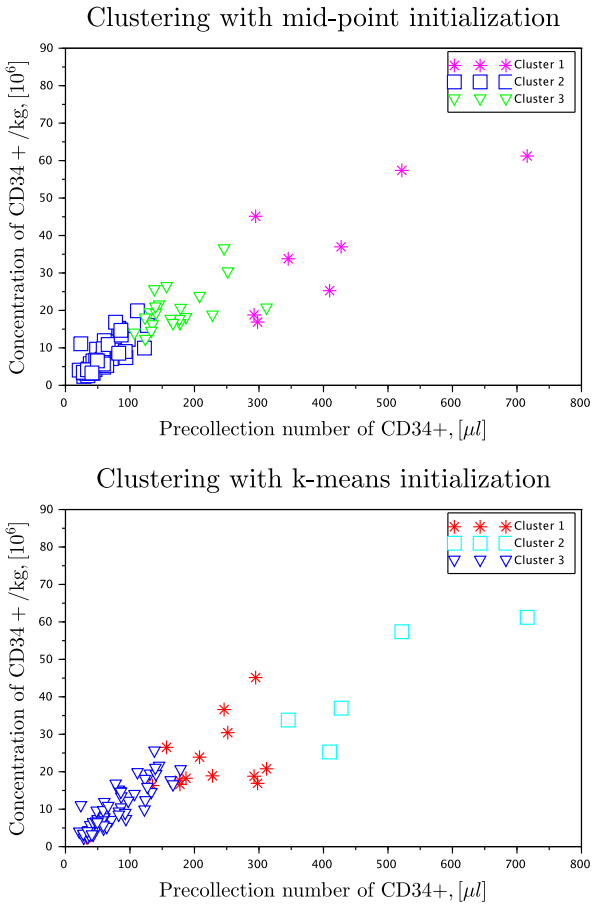

Fig. 7. Comparison of clusters of $y_{5;t}$ and $y_{8;t}$ with the mid-point (top) and k-means based (bottom) initialization techniques [15].

The insignificant difference in the location of two upper clusters can be seen in Figures 7 and 8 (top), while in Figure 8 (bottom) the clustering practically fails. Only two data items are classified as belonging to the first cluster, i.e., two clusters are detected instead of three.

4.2 Initialization with Bernoulli Components

For the Bernoulli components (3) as well as other distributions considered in subsequent sections, the approach used for the uniform components is not suitable. The reason is as follows. The proximity function [26], (i.e., the approximation of the posterior pdf by the Dirac delta function and substitution of the pa-

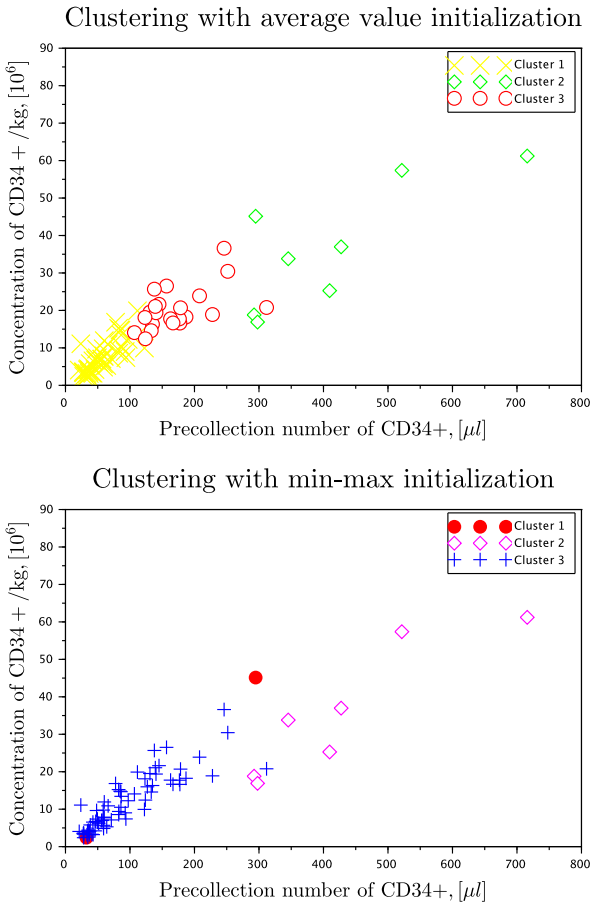

Fig. 8. Comparison of clusters of $y_{5;t}$ and $y_{8;t}$ with the initialization techniques based the average (top) and minimum and maximum prior values (bottom) [15].

rameter point estimates into the components [4]) cannot be used, because for non-negative data the model has a different form than the likelihood function. The approach based on the likelihood function derivation can be used instead.

For the initialization of the statistics of the Bernoulli component (3), it is advantageous to use the following construction of the likelihood function. For instance, for the first two data items y_1 and y_2, the product of pdfs (3) takes the form (here omitting the subscript i for simplicity)

$$\Theta_1^{y_1}\Theta_0^{1-y_1}\Theta_1^{y_2}\Theta_0^{1-y_2} = \Theta_1^{y_1+y_2}\Theta_0^{2-y_1+y_2}, \tag{15}$$

which means that for t data items, the likelihood is

$$L_t(\Theta) = \Theta_1^{\sum_{\tau=1}^{t} y_\tau}\Theta_0^{t-\sum_{\tau=1}^{t} y_\tau} = \Theta_1^{S_t}\Theta_0^{t-S_t}, \tag{16}$$

where $S_t = \sum_{\tau=1}^{t} y_\tau$. The data items $y_\tau \in \{0, 1\}$, therefore S_t counts the values 1 and $t - S_t$ – the values 0.

The posterior pdf via the Bayes rule [22] is the product of the likelihood function and the prior pdf. It means that it is suitable to choose the prior pdf in the same form [2], i.e.,

$$f\left(\Theta|y\left(0\right)\right) \propto \Theta_1^{S_0}\Theta_0^{t_0-S_0}, \tag{17}$$

where $y(0)$ denotes a collection of prior data, S_0 is the number of the prior values of 1, t_0 is the number of prior data and $t_0 - S_0$ is the number of the prior values of 0.

The number of prior data can be either real when working with some prior data set or fictitious in the case of using, e.g., expert knowledge. To utilize the prior knowledge, which says that both the values 1 and 0 are measured in the same ratio, it is suitable to choose an arbitrary value of t_0 and S_0 will be a half of it. However, if t_0 is high, it means that this information was extracted from a large data set and can dominate. If it is small, the influence of the prior knowledge is also weak.

The posterior pdf takes the form

$$f\left(\Theta|y\left(t\right)\right) \propto \Theta_1^{S_t}\Theta_0^{\kappa_t-S_t}, \tag{18}$$

where for the statistics it holds

$$S_t = \sum_{\tau=1}^{t} y_\tau + S_0, \tag{19}$$

$$\kappa_t = t + t_0, \tag{20}$$

which highlights the previous remark: low values of S_0 and t_0 will not influence the statistics. The higher values, the more influence.

With the help of substitution of the distributions into the Bayes rule [2, 22], the recursive update of the statistics is obtained in the form

$$S_\tau = S_{\tau-1} + y_\tau, \tag{21}$$

$$\kappa_\tau = \kappa_{\tau-1} + 1 \tag{22}$$

for $\tau = 1, 2, \cdots, t$, which starts for the given S_0 a κ_0.

The point estimates of the model parameters can be obtained via MAP (Maximum A Posteriori) method [29] as the argument of a maximum of the posterior pdf.

The maximum is obtained by setting the derivation of the posterior pdf equal to zero and computing the parameter estimate. Let's denote in (18) $\Theta_1 = \Theta$ and $\Theta_2 = 1 - \Theta$, then

$$S_t\left(1-\Theta\right)^{\kappa_t-S_t} - \Theta^{S_t}\left(\kappa_t - S_t\right)\left(1-\Theta\right)^{\kappa_t-S_t-1} = 0, \tag{23}$$

and therefore

$$\hat{\Theta}_t = \frac{S_t}{\kappa_t}, \tag{24}$$

i.e., the number of the values of 1 is divided by the number of data, which is the result already mentioned above.

From relation (24), it can be seen that the point estimate of the parameter Θ is the average of the measured output. This can be used for the construction of the prior statistics as follows. The average of the prior data is denoted by y_0 and it is set as the initial value of the $\hat{\Theta}_0$, which means that according to (24)

$$y_0 = \frac{S_0}{\kappa_0}. \tag{25}$$

Let's define

$$S_0 = n_0 y_0, \tag{26}$$

$$\kappa_0 = n_0, \tag{27}$$

where n_0 is chosen according to the emphasis with which the prior knowledge is intended to be used. After substitution, it is obtained

$$\frac{S_0}{\kappa_0} = \frac{n_0 y_0}{n_0} = y_0 = \hat{\Theta}_0. \tag{28}$$

This is directly the desired result, which does not depend on n_0.

Figures 9 – 11 demonstrate the evolution of the point estimates of the parameter Θ. It can be seen how the prior knowledge can influence the stabilization of the values of the point estimates. In dependence on the strength and the correctness of the prior information the values of the point estimates are approaching to the true values with a different speed.

4.3 Initialization with Geometric Components

For the geometric component pdf (4) [30], similarly as in the previous case, the model product for the first two data items y_1 and y_2 gives (omitting the subscript i for the sake of simplicity)

$$\Theta\left(1-\Theta\right)^{1-y_1}\Theta\left(1-\Theta\right)^{1-y_2} = \Theta^2\left(1-\Theta\right)^{2-(y_1+y_2)}. \tag{29}$$

Hence, the likelihood function again takes the form

$$L_t\left(\Theta\right) = \Theta^{\kappa_t}\left(1-\Theta\right)^{\kappa_t - S_t}, \tag{30}$$

where

$$S_t = \sum_{\tau=1}^{t} y_\tau + S_0 \text{ and } \quad \kappa_t = t + t_0. \tag{31}$$

The likelihood derivation equal to zero gives the similar result as in the case of the Bernoulli pdf, i.e.,

$$\Theta = \frac{\kappa_t}{S_t + \kappa_t} = \frac{1}{\bar{y} + 1}, \tag{32}$$

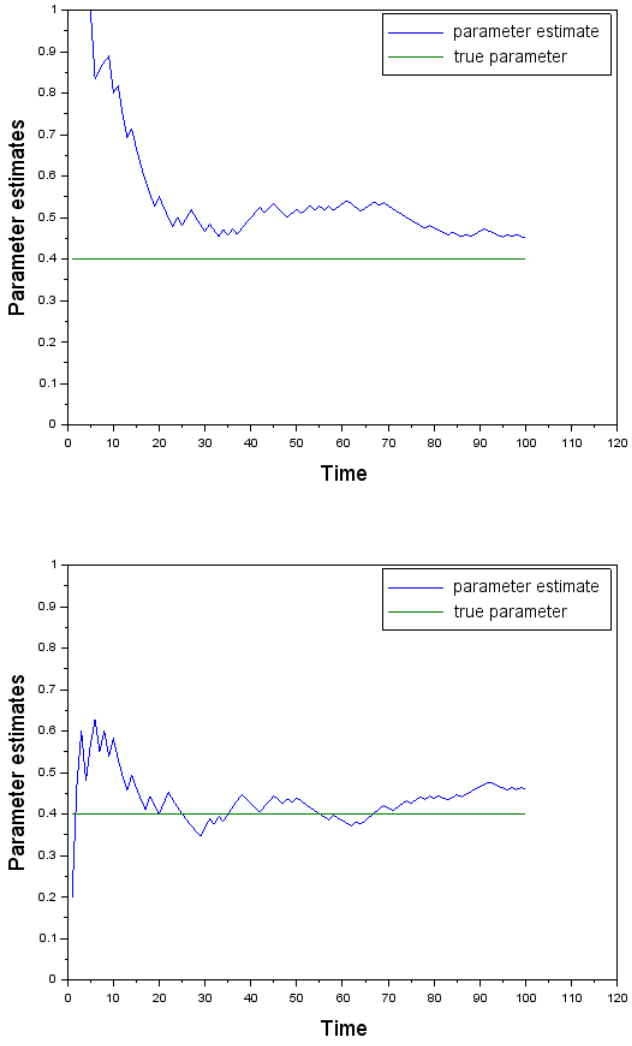

Fig. 9. The evolution of the point estimates without prior knowledge (top) and with a weak correct prior knowledge (bottom)

where

$$\bar{y} = \frac{S_t}{\kappa_t} \tag{33}$$

is the average output. It allows us to use the initialization technique described above for the Bernoulli components.

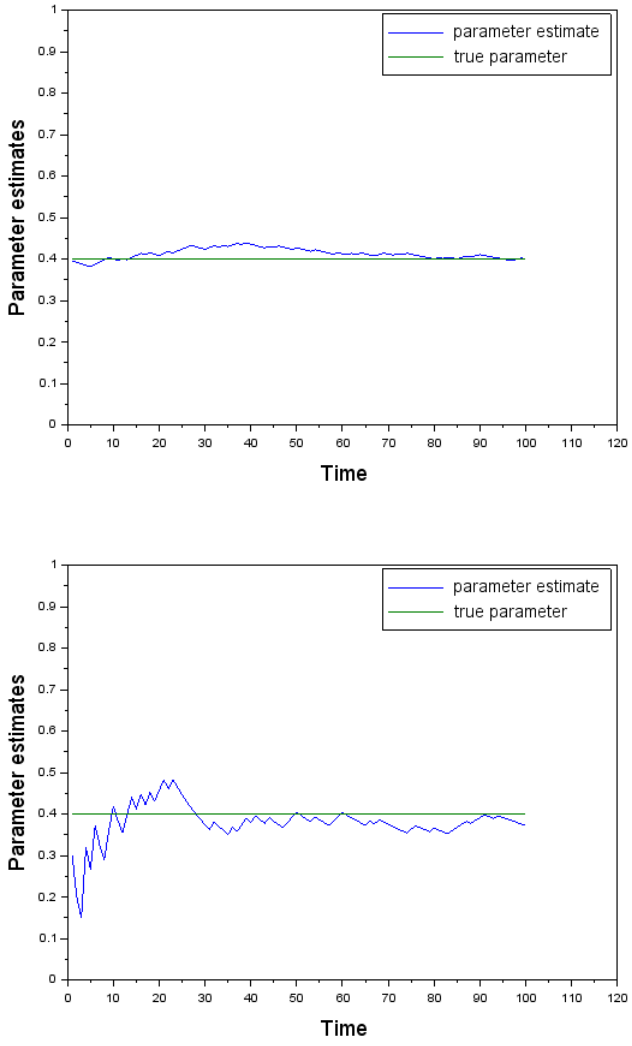

Fig. 10. The evolution of the point estimates with a strong correct prior knowledge (top) and with a weak incorrect prior knowledge (bottom)

4.4 Initialization with Exponential Components

For the exponential component (5) [31, 32], the product of the models (omitting the subscript i for the sake of simplicity) is

$$\Theta \exp\{-\Theta y_1\}\, \Theta \exp\{-\Theta y_2\} = \Theta^2 \exp\{-\Theta\,(y_1 + y_2)\} \tag{34}$$

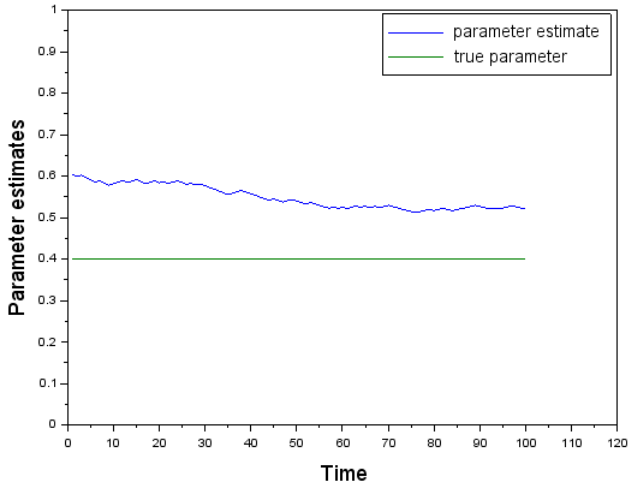

Fig. 11. The evolution of the point estimates with a strong incorrect prior knowledge

and the likelihood function is therefore

$$L_t(\Theta) = \Theta^{\kappa_t} \exp\{-\Theta S_t\}, \tag{35}$$

where again

$$S_t = \sum_{\tau=1}^{t} y_\tau + S_0 \quad \text{and} \quad \kappa_t = t + t_0. \tag{36}$$

After its derivation, the point estimate of the parameter Θ is computed as

$$\hat{\Theta}_t = \frac{\kappa_t}{S_t}, \tag{37}$$

which is the inverse average output from the used data. Thus, for the initialization, the technique described for the Bernoulli components can be used again.

4.5 Initialization with Gamma Components

For the Gamma pdf (6), the model product for the first two data items y_1 and y_2 (omitting the subscript i for the sake of simplicity) is

$$\begin{aligned} &\frac{\beta^\gamma}{\Gamma(\gamma)} y_1^{\gamma-1} \exp\{-\beta y_1\} \frac{\beta^\gamma}{\Gamma(\gamma)} y_2^{\gamma-1} \exp\{-\beta y_2\} \\ &= \left(\frac{\beta^\gamma}{\Gamma(\gamma)}\right)^2 (y_1 y_2)^{\gamma-1} \exp\{-\beta(y_1 + y_2)\}. \end{aligned} \tag{38}$$

According to this relation, the likelihood function takes the form

$$L_t(\Theta) = \left(\frac{\beta^\gamma}{\Gamma(\gamma)}\right)^{\kappa_t} P_t^{\gamma-1} \exp\{-\beta S_t\}, \tag{39}$$

where

$$\kappa_t = t + t_0, \tag{40}$$

$$S_t = \sum_{\tau-1}^{t} y_\tau + S_0, \tag{41}$$

$$P_t = \prod_{\tau=1}^{t} y_\tau \cdot P_0. \tag{42}$$

Here, the computation of the point estimates of the parameter Θ is not so straightforward as in the previous cases. They can be derived as follows, e.g., [33]. Let's denote

$$s = \ln\left(\frac{S_t}{\kappa_t}\right) - \frac{1}{\kappa_t}\ln(P_t) \tag{43}$$

and then

$$\gamma \doteq \frac{3 - s + \sqrt{(s-3)^2 + 24s}}{12s}, \tag{44}$$

$$\beta = \frac{\gamma \kappa_t}{S_t}. \tag{45}$$

It is known, e.g., [33] (and it can be easily derived) that for the Gamma distribution, it holds

$$\text{the average of the output } \bar{y} = \frac{\gamma}{\beta}, \tag{46}$$

$$\text{the mode of the output } \hat{y} = \frac{\gamma - 1}{\beta}, \tag{47}$$

which enables us to obtain

$$\gamma = \frac{\bar{y}}{\bar{y} - \hat{y}} \quad \text{and} \quad \beta = \frac{1}{\bar{y} - \hat{y}}. \tag{48}$$

For the initialization purposes, it is again assumed that the average $\bar{y}_0$ and the mode $\hat{y}$ are available from the prior data.

However, it is necessary to have the initial statistics, not the initial values of parameters. The derivation of the statistics from the parameters is a relatively complicated procedure. That is why a good choice is not to try to derive them as it was done in the previous sections, but to simulate some data using the parameter estimates and use them as the prior data for the computation of the initial statistics.

5 Conclusions

The paper summarizes practical approaches to the initialization of mixture components for a task of recursive mixture-based clustering under the Bayesian methodology. The investigated approaches are based on processing the prior data set with the aim of setting the initial statistics of several types of components. The potential application of the discussed techniques can be beneficial for areas of the data analysis of non-negative variables.

Acknowledgements

The paper was supported by project GAČR GA15-03564S.

References

1. Kárný, M., Kadlec, J., Sutanto, E.L. (1998). Quasi-Bayes estimation applied to normal mixture, In: *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing (eds. J. Rojíček, M. Valečková, M. Kárný, K. Warwick)*, CMP'98 /3./, Prague, CZ, p. 77–82.
2. Kárný, M., Böhm, J., Guy, T. V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L. (2006). *Optimized Bayesian Dynamic Advising: Theory and Algorithms*, Springer-Verlag London.
3. Nagy, I., Suzdaleva, E., Kárný, M., Mlynářová, T. (2011). Bayesian estimation of dynamic finite mixtures. *Int. Journal of Adaptive Control and Signal Processing*, vol. 25, 9, p. 765–787.
4. Nagy, I., Suzdaleva, E., (2017). *Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components*, SpringerBriefs in Statistics. Springer International Publishing.
5. Roy, A., Pal, A., Garain, U. (2017). JCLMM: A Finite Mixture Model for Clustering of Circular-Linear data and its application to Psoriatic Plaque Segmentation, *Pattern Recognition*, doi 10.1016/j.patcog.2016.12.016.
6. Bouveyron, C., Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis.* 71(0), p. 52–78.
7. Scrucca, L., (2016). Genetic algorithms for subset selection in model-based clustering. *Unsupervised Learning Algorithms*, p. 55–70, Springer International Publishing.
8. Fernández, D., Arnold, R., Pledger, S., (2016). Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis*, 93, p.46–75.
9. Suzdaleva, E., Nagy, I., Mlynářová, T. (2015). Recursive Estimation of Mixtures of Exponential and Normal Distributions. In: *Proceedings of the 8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Warsaw, Poland, September 24–26, p.137–142.
10. Browne, R.P. and McNicholas, P.D., (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2), p.176–198.
11. Morris, K. and McNicholas, P.D., (2016). Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. *Computational Statistics & Data Analysis*, 97, p.133–150.

12. Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B., (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and computing*, 26(1–2), p.303–324.
13. Li, R., Wang, Z., Gu, C., Li, F., Wu, H., (2016). A novel time-of-use tariff design based on Gaussian Mixture Model. *Applied Energy*, 162, p.1530–1536.
14. O'Hagan, A., Murphy, T.B., Gormley, I.C., McNicholas, P.D., Karlis, D., (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics & Data Analysis*, 93, p.18–30.
15. Suzdaleva, E., Nagy, I., Pecherková, P., Likhonina, R., (2017). Initialization of Recursive Mixture-based Clustering with Uniform Components. In: Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2017), Madrid, Spain, July, 26 – 28, 2017, p. 449–458.
16. Scrucca, L. and Raftery, A.E., (2015). Improved initialisation of model-based clustering using Gaussian hierarchical partitions. *Advances in data analysis and classification*, 9(4), p.447–460.
17. Melnykov, V., Melnykov, I. (2012). Initializing the EM algorithm in Gaussian mixture models with an unknown number of components, *Computational Statistics & Data Analysis*, 56(6), p.1381–1395.
18. Kwedlo, W. (2013). A new method for random initialization of the EM algorithm for multivariate Gaussian mixture learning, In: *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*, (eds. R. Burduk, K. Jackowski, M. Kurzynski, M. Wozniak, A. Zolnierek), Springer International Publishing, Heidelberg, p. 81–90.
19. Shireman, E., Steinley, D. and Brusco, M.J., (2015). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior research methods*, p.1–12.
20. Maitra, R., (2009). Initializing partition-optimization algorithms. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(1), p.144–157.
21. Gupta, M. R. , Chen, Y. (2011). Theory and use of the EM method. In: *Foundations and Trends in Signal Processing*, vol. 4, 3, p. 223–296.
22. Peterka, V. (1981). Bayesian system identification. In: *Trends and Progress in System Identification (ed. P. Eykhoff)*, Oxford, Pergamon Press, 1981, p. 239–304.
23. Nagy, I., Suzdaleva, E., Mlynářová, T. (2016). Mixture-based clustering non-gaussian data with fixed bounds. In: *Proceedings of the IEEE International conference Intelligent systems IS'16*, p. 265–271.
24. Suzdaleva, E., Nagy, I., Mlynářová, T. (2016). Expert-based initialization of recursive mixture estimation. In: *Proceedings of the IEEE International conference Intelligent systems IS'16*, p. 308–315.
25. Kárný, M., Nedoma, P., Khailova, N., Pavelková, L., (2003). Prior information in structure estimation. In: *IEE Proceedings, Control Theory and Applications*, 150(6), pp. 643–653.
26. Nagy, I., Suzdaleva, E., Pecherková, P. (2016). Comparison of Various Definitions of Proximity in Mixture Estimation. In: *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, p. 527–534
27. Casella, G., Berger R.L. (2001). *Statistical Inference, 2nd ed.*, Duxbury Press.
28. Jain, A. K., (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp. 651–666.
29. DeGroot, M., (1970). *Optimal Statistical Decisions*, McGraw-Hill,.
30. Spiegel, M. R., (1992), Theory and Problems of Probability and Statistics. New York: McGraw-Hill.

31. Johnson, R. A., Wichern, D. W., (2007). *Applied Multivariate Statistical Analysis.* Pearson Prentice Hall.
32. Elfessi, A., Reineke, D. M., (2001). A Bayesian look at classical estimation: the exponential distribution. Journal of Statistics Education, 9 (1).
33. Minka, T. P. (2002). Estimating a Gamma distribution, doi: 10.1.1.142.7635.