Initialization of Recursive Mixture-based Clustering with Uniform Components

Evgenia Suzdaleva[1], Ivan Nagy[1,2], Pavla Pecherková[1,2] and Raissa Likhonina[1]
[1]Department of Signal Processing, The Institute of Information Theory and Automation of the Czech Academy of Sciences, Pod vodárenskou věží 4, 18208, Prague, Czech Republic
[2]Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 11000, Prague, Czech Republic

Keywords: Mixture-based Clustering, Recursive Mixture Estimation, Uniform Components, Bayesian Estimation.

Abstract: The paper deals with a task of initialization of the recursive mixture estimation for the case of uniform components. This task is significant as a part of mixture-based clustering, where data clusters are described by the uniform distributions. The issue is extensively explored for normal components. However, sometimes the assumption of normality is not suitable or limits potential application areas (e.g., in the case of data with fixed bounds). The use of uniform components can be beneficial for these cases. Initialization is always a critical task of the mixture estimation. Within the considered recursive estimation algorithm the key point of its initialization is a choice of initial statistics of components. The paper explores several initialization approaches and compares results of clustering with a theoretical counterpart. Experiments with real data are demonstrated.

1 INTRODUCTION

The use of mixture models is widespread in a range of applications working with multi-modal systems requiring to be described and identified (Hu et al., 2015; Bao and Shen, 2016), for example, industry, fault detection, transportation, marketing, medicine, etc. In the field of data analysis, mixtures are used for model-based clustering (Roy et al., 2017; Bouveyron and Brunet-Saumard, 2014; Scrucca, 2016), where clusters in the data space are described by distributions of mixture components.

Various distributions are intensively investigated for tasks of mixture-based clustering (Fernández et al., 2016; Suzdaleva et al., 2015; Browne and McNicholas, 2015; Morris and McNicholas, 2016). Gaussian mixtures are probably the most frequently met models, see, e.g., (Malsiner-Walli et al., 2016; Li et al., 2016; O'Hagan et al., 2016), etc.

This paper considers a clustering with uniform components of the mixture model. This is beneficial for applications producing specific measurements with fixed boundaries, where the assumption of normality or belongingness to the exponential family is not suitable. A focus of the paper is a task of the mixture initialization, which is known to be a critical part of the mixture estimation significant for starting an estimation algorithm.

Recent papers on the mixture initialization found in the literature (Scrucca and Raftery, 2015; Melnykov and Melnykov, 2012; Kwedlo, 2013; Shireman et al., 2015; Maitra, 2009) are mostly concerned with initialization of the expectation-maximization (EM) algorithm (Gupta and Chen, 2011) used in iterative approaches to mixture estimation. However, the approach discussed in the presented paper is based on the recursive Bayesian estimation avoiding iterative computations. It was considered for normal models in (Peterka, 1981) and for normal mixtures in (Kárný et al., 1998; Kárný et al., 2006; Nagy et al., 2011). Extension of the approach for uniform components is presented in (Nagy et al., 2016).

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 (Suzdaleva et al., 2016) based on (Kárný et al., 2003) is found, again devoted to the initialization with normal mixtures.

This paper explores several initialization approaches for estimation of the mixture of uniform 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 and on a combination of expert-based visualization techniques and well-known clustering methods applied to prior data.

The paper is organized in the following way. Sec-

Suzdaleva, E., Nagy, I., Pecherková, P. and Likhonina, R.
Initialization of Recursive Mixture-based Clustering with Uniform Components.
DOI: 10.5220/0006417104490458
In *Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2017) - Volume 1*, pages 449-458
ISBN: 978-989-758-263-9

tion 2 introduces models and gives basic facts about their individual estimation. Section 3 presents a brief summary of recursive Bayesian estimation of mixtures of uniform components. Section 4 specifies the initialization problem and considers four initialization approaches. Section 5 provides results of their experimental comparison. Conclusions and open problems are given in Section 6.

2 MODELS

A considered system generates the continuous data vector y_t at each discrete time instant $t = 1, 2, \ldots..$ 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 (Kárný et al., 1998).

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 $f(y_t|\Theta, c_t = i) = f(y_t|\Theta_i)$ for $c_t = i$.

The general component pdf (1) is specified as the uniform distribution. Under assumption of the independence of individual entries of the vector y_t (made in this paper) the pdf (1) takes the following form $\forall i \in \{1, 2, \ldots, m_c\}$

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

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.

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:

$$f(c_t = i|c_{t-1} = j, \alpha),\ i, j \in \{1, 2, \ldots, m_c\}, \tag{3}$$

represented by the transition table

	$c_t = 1$	$c_t = 2$	$\cdots$	$c_t = m_c$
$c_{t-1} = 1$	$\alpha_{1\|1}$	$\alpha_{2\|1}$	$\cdots$	$\alpha_{m_c\|1}$
$c_{t-1} = 2$	$\alpha_{1\|2}$		$\cdots$	
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$c_{t-1} = m_c$	$\alpha_{1\|m_c}$		$\cdots$	$\alpha_{m_c\|m_c}$

where a current value of the pointer corresponds to the active component, and the unknown parameter α is the $(m_c \times m_c)$-dimensional matrix, and its entries $\alpha_{i|j}$ are non-negative probabilities of the pointer $c_t = i$ (expressing that the i-th component is active at time t) under condition that the previous pointer $c_{t-1} = j$.

2.1 Individual Model Estimation

The estimation of parameters of the individual i-th uniform component (2) in the case of independent data entries is performed using the initially chosen statistics $\mathcal{L}_{t-1}$ and $\mathcal{R}_{t-1}$ with the update of their l-th entries for each $l \in \{1, \ldots, K\}$ in the following form, see, e.g., (Casella and Berger, 2001):

$$\text{if } y_{l;t} < \mathcal{L}_{l;t-1}, \qquad \text{then } \mathcal{L}_{l;t} = y_{l;t}, \tag{4}$$

$$\text{if } y_{l;t} > \mathcal{R}_{l;t-1}, \qquad \text{then } \mathcal{R}_{l;t} = y_{l;t}, \tag{5}$$

where the subscript i is omitted for simplicity. The point estimates of parameters are computed via

$$\hat{L}_t = \mathcal{L}_t,\ \ \hat{R}_t = \mathcal{R}_t. \tag{6}$$

According to (Kárný et al., 2006), parameter α of the pointer model (3) is estimated using the conjugate prior Dirichlet pdf in the Bayes rule, recomputing its initially chosen statistics and its normalizing. The mentioned statistics is denoted by ν_{t-1}, which is here the square m_c-dimensional matrix. Its entries in the case of available values $c_t = i$ and $c_{t-1} = j$ are updated for $i, j \in \{1, \ldots, m_c\}$ in the following way:

$$\nu_{i|j;t} = \nu_{i|j;t-1} + \delta(i, j; c_t, c_{t-1}), \tag{7}$$

where $\delta(i, j; c_t, c_{t-1})$ is the Kronecker delta function, which is equal to 1, if $c_t = i$ and $c_{t-1} = j$, and it is 0 otherwise. The point estimate of α is then obtained by

$$\hat{\alpha}_{i|j;t} = \frac{\nu_{i|j;t}}{\sum_{k=1}^{m_c} \nu_{k|j;t}}. \tag{8}$$

However, values of c_t and c_{t-1} are unavailable and should be estimated. It means that generally for the aim of the mixture-based clustering with the introduced models it is necessary to estimate parameters Θ and α and the pointer values.

3 UNIFORM MIXTURE ESTIMATION

To specify a task of the mixture initialization, a necessary theoretical background on recursive mixture estimation with uniform components should be given.

The uniform distribution does not belong to the exponential family. Thus, extension of the general approach to recursive estimation (Kárný et al., 1998; Peterka, 1981; Kárný et al., 2006; Nagy et al., 2011) for this class of components is not straightforward and might need the use of specific techniques of forgetting (Nagy et al., 2016).

Generally, the estimation algorithm is based on using the joint pdf of unknown variables to be estimated and the Bayes and the chain rule (Peterka, 1981). The unknown parameters Θ and α (assumed to be mutually independent) and the pointer values c_t and c_{t-1} enter the joint pdf as follows:

$$\begin{aligned}
&\underbrace{f(\Theta, c_t = i, c_{t-1} = j, \alpha | y(t))}_{joint\ posterior\ pdf}\\
&\propto \underbrace{f(y_t, \Theta, c_t = i, c_{t-1} = j, \alpha | y(t-1))}_{via\ chain\ rule\ and\ Bayes\ rule}\\
&= \underbrace{f(y_t|\Theta, c_t = i)}_{(1)} \underbrace{f(\Theta|y(t-1))}_{prior\ pdf\ of\ \Theta}\\
&\times \underbrace{f(c_t = i|\alpha, c_{t-1} = j)}_{(3)} \underbrace{f(\alpha|y(t-1))}_{prior\ pdf\ of\ \alpha}\\
&\times \underbrace{f(c_{t-1} = j|y(t-1)),}_{prior\ pointer\ pdf} \qquad (9)
\end{aligned}$$

$\forall i, j \in \{1, 2, \dots, m_c\}$, where denotation $y(t) = \{y_0, y_1, \dots, y_t\}$ stands for the data collection up to the time instant t, and y_0 denotes the prior information.

Recursive formulas for estimation of c_t, Θ and α are derived by marginalization of (9) over Θ, α and c_{t-1}. In the first case, the marginalization over parameters Θ gives a closeness of the current data item y_t to individual components at each time instant t, which is called the proximity, see, e.g., (Nagy et al., 2016). Here, the normal approximation of the uniform component, optimal in the sense of the Kullback-Leibler divergence, see (Kárný et al., 2006), is taken. It means that the proximity m_i of the i-th component is the value of the normal pdf obtained by putting the point estimates of the expectation and the covariance matrix of this uniform component from the previous time instant $t-1$ and the currently measured y_t into

$$\begin{aligned}
m_i &= (2\pi)^{-K/2} |(D_{t-1})_i|^{-1/2}\\
&\times \exp\left\{ -\frac{1}{2} (y_t - (E_{t-1})_i)' (D_{t-1}^{-1})_i (y_t - (E_{t-1})_i) \right\},
\end{aligned} \qquad (10)$$

where K is a dimension of the vector y_t, $(E_{t-1})_i$ is the K-dimensional expectation vector of the i-th component, each l-th entry of which is obtained via (6) as follows:

$$(E_{l;t-1})_i = \frac{1}{2}((\hat{L}_{l;t-1})_i + (\hat{R}_{l;t-1})_i) \qquad (11)$$

and $(D_{t-1})_i$ is the covariance matrix containing on the diagonal

$$(D_{l;t-1})_i = \frac{1}{12}((\hat{R}_{l;t-1})_i - (\hat{L}_{l;t-1})_i)^2. \qquad (12)$$

The proximities from all m_c components comprise the m_c-dimensional vector m.

Similarly, the integral of (9) over α provides the computation of its point estimate (8) using the previous-time statistics v_{t-1}.

However, the general purpose of the estimation is to obtain the component weights (i.e., probabilities that the components are currently active). For this aim, the proximities (10) are multiplied entry-wise by the previous-time point estimate of the parameter α (8) and the prior weighting m_c-dimensional vector w_{t-1}, whose entries are the prior (initially chosen) pointer pdfs $(c_{t-1} = j|y(t-1))$, i.e.,

$$W_t \propto \left(w_{t-1} m'\right) .* \hat{\alpha}_{t-1} \qquad (13)$$

where W_t denotes the square m_c-dimensional matrix containing pdfs $f(c_t = i, c_{t-1} = j|y(t))$ joint for c_t and c_{t-1}, and $.*$ is a "dot product" that multiplies the matrices entry by entry. The matrix W_t is normalized so that the overall sum of all its entries is equal to 1, and subsequently it is summed up over rows, which allows to obtain the vector w_t with updated component weights $w_{i;t}$ for all components.

The maximal $w_{i;t}$ defines the currently active component, i.e., the point estimate of the pointer c_t at time t. This point estimate is subsequently used for data clustering.

3.1 The Statistics Updates

The above theoretical background leads to the following relations for updating the component statistics $(\mathcal{L}_{l;t-1})_i$ and $(\mathcal{R}_{l;t-1})_i$ with the help of the obtained weights $w_{i;t}$ at time t (Nagy et al., 2016). A specific feature of the uniform component statistics is their moving depending on a newly arrived data item. The number of non-updates of each statistics is described by the geometrical distribution. When the statistics is not updated for a relatively long time, it is forgotten. A scheme of forgetting is as follows. For the minimum bound statistics $(\mathcal{L}_{l;t-1})_i$ of the l-th entry of the i-th component, the counter of its non-updates is set as 0, i.e.,

$$(\lambda^L_{l;t-1})_i = 0 \qquad (14)$$

and then the update with forgetting takes the form

$$\begin{aligned}
\delta_L &= y_{l;t} - (\mathcal{L}_{l;t-1})_i, \quad &(15)\\
\text{if } \delta_L < 0,\ (\mathcal{L}_{l;t})_i &= (\mathcal{L}_{l;t-1})_i - w_{i;t}\delta_L, &(16)\\
(\lambda^L_{l;t})_i &= 0, &(17)\\
\text{else } (\lambda^L_{l;t})_i &= (\lambda^L_{l;t-1})_i + 1, &(18)\\
\text{if } (\lambda^L_{l;t})_i > n,\ (\mathcal{L}_{l;t})_i &= (\mathcal{L}_{l;t-1})_i + \phi w_{i;t}, &(19)
\end{aligned}$$

where n is the allowed number of non-updates computed from the distribution function of the geometrical distribution depending on the used confidence interval and assumption of the statistics location, see (Nagy et al., 2016), and ϕ is a small forgetting factor.

For the maximum bound statistics $(\mathcal{R}_{l;t-1})_i$ the update is performed similarly, i.e., $(\lambda^R_{l;t-1})_i = 0$,

$$\begin{aligned}
\delta_R &= y_{l;t} - (\mathcal{R}_{l;t-1})_i, \quad &(20)\\
\text{if } \delta_R > 0,\ (\mathcal{R}_{l;t})_i &= (\mathcal{R}_{l;t-1})_i + w_{i;t}\delta_R, &(21)\\
(\lambda^R_{l;t})_i &= 0, &(22)\\
\text{else } (\lambda^R_{l;t})_i &= (\lambda^R_{l;t-1})_i + 1, &(23)\\
\text{if } (\lambda^R_{l;t})_i > n,\ (\mathcal{R}_{l;t})_i &= (\mathcal{R}_{l;t-1})_i - \phi w_{i;t}. &(24)
\end{aligned}$$

3.2 The Pointer Update

The statistics of the pointer model is updated similarly to the update of the individual categorical model and based on (Kárný et al., 2006; Kárný et al., 1998), but with the joint weights $W_{i,j;t}$ from the matrix (13), where the row j corresponds to the value of c_{t-1}, and the column i to the current pointer c_t

$$v_{i|j;t} = v_{i|j;t-1} + W_{j,i;t}. \quad (25)$$

3.3 Algorithmic Summary

The briefly summarized above relations comprise the following algorithmic scheme of clustering at each time instant:

- Measuring the new data item;
- Computing the proximity of the data item to individual components;
- 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;
- Classifying data according to the declared active component;
- Updating the statistics of all components and the pointer model;
- Re-computing the point estimates of parameters necessary for calculating the proximity.

4 MIXTURE INITIALIZATION

The main feature of the discussed recursive clustering is its on-line performance and updating with each new measurement. It is dangerous from the point of view of the unsuccessful start of the algorithm, as it can lead to dominance of one of the components. However, prior data sets, which are usually available in most application areas (e.g., previous measurements, realistic simulations, etc.) can be analyzed off-line for the initialization purposes using a combination of relatively simple expert-based techniques, e.g., (Suzdaleva et al., 2016) and well-known clustering methods such as, e.g., k-means (Jain, 2010), etc.

The initialization task is specified for the above recursive algorithm in the following way. For time $t = 0$, $\forall i, j \in \{1, 2, \ldots, m_c\}$ and for $l \in \{1, 2, \ldots, K\}$ it is necessary to set:

- the number of components m_c,
- the initial statistics of the pointer model $v_{i|j;0}$ and the initial weighting vector w_0,
- the initial components statistics $(\mathcal{L}_{l;0})_i$ and $(\mathcal{R}_{l;0})_i$.

The last point is the key one. 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.

4.1 Choice of Number of Components

Here a set of anonymized medical hematological prior data is used for demonstration of the data visualization with the aim of determining the number of 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, $[10^6]$.

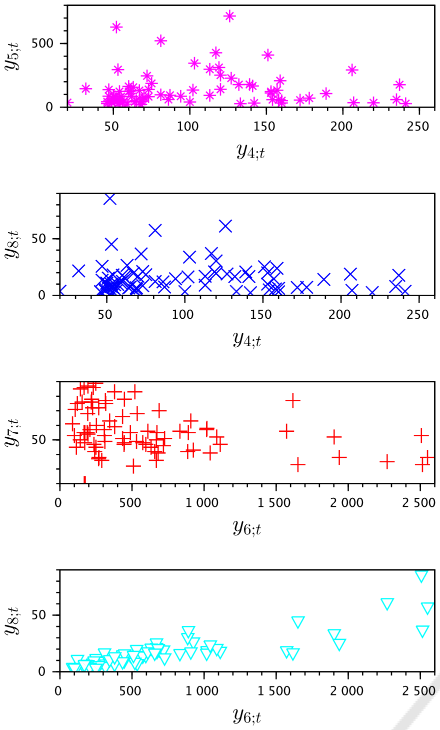

Figure 1: Visualization of selected prior data entries.

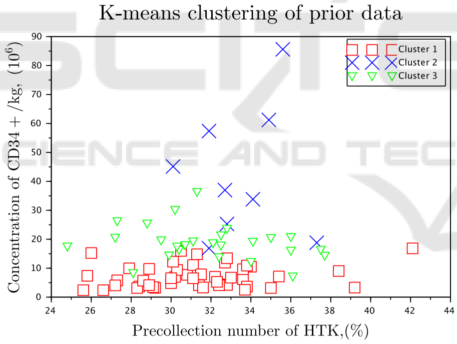

Figure 2: Three clusters detected by k-means in selected prior data entries.

All the data entries are plotted against each other in the form of upper triangular matrix of figures to detect a number of visible clusters. If the visual analysis is successful, and data clusters are distinguishable, their number can be validated by using the k-means method. To save space, selected data entries plotted against each other are demonstrated in Figure 1. Two bottom plots indicate that three clusters can be detected, and they are also slightly distinguishable in the two top plots. To verify the number of components, the k-means is used, see Figure 2.

The initialized number of components can be also validated by evolution of components weights during the on-line estimation. This is demonstrated in Section 5.

4.2 Initial Pointer Settings

The initial statistics of the pointer model $v_{i|j;0}$ and the initial weighting vector w_0 are initialized either uniformly or randomly in combination with their updating by prior data.

4.3 Initial Components Statistics

Here four approaches to setting the initial statistics of components are explored.

4.3.1 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 (Nagy et al., 2016). In this case additional statistics should be used. They are $(s_{l;0})_i$, $(q_{l;0})_i$, which are l-th entries of the K-dimensional vectors s_t and q_t, where the last comprises a diagonal of a matrix. Starting from random values, they are updated by a small set of prior data $\forall i \in \{1,2,\ldots,m_c\}$ and $\forall l = \{1,\ldots,K\}$ in the following way.

$$(s_{l;t})_i \quad = \quad (s_{l;t-1})_i + w_{i;t} y_{l;t}, \qquad (26)$$

$$(q_{l;t})_i \quad = \quad (q_{l;t-1})_i + w_{i;t} y_{l;t}^2, \qquad (27)$$

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.

$$(\hat{S}_t)_i \quad = \quad (s_t)_i/t, \qquad (28)$$

$$(D_t)_i \quad = \quad \left((q_t)_i - (s_t)_i(s_t')_i/t\right)/t, \qquad (29)$$

$$(\hat{h}_t)_i \quad = \quad \sqrt{3\,\mathrm{diag}((D_t)_i)}, \qquad (30)$$

where $(D_t)_i$ is the covariance matrix of the uniform distribution, and $\sqrt{3\,\mathrm{diag}((D_t)_i)}$ denotes the square roots of entries of the vector $\mathrm{diag}((D_t)_i)$. (28) and (29) from the previous time instant are placed instead of the expectation and the covariance matrix into the proximity (10). 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

$$(\hat{L}_{l;t})_i \quad = \quad (\hat{S}_{l;t})_i - \varepsilon, \qquad (31)$$

$$(\hat{R}_{l;t})_i \quad = \quad (\hat{S}_{l;t})_i + \varepsilon, \qquad (32)$$

with small ε, and they are used in (11) and (12) during the on-line estimation according to Section 3.

4.3.2 Centers based on K-means

Another way is to use the centers of clusters initially detected by the k-means method from prior data and put them into (31) and (32) to be used during the on-line estimation according to Section 3.

4.3.3 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 (31) and (32).

4.3.4 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 $(\mathcal{L}_{l;0})_i$ and $(\mathcal{R}_{l;0})_i$ respectively. Then via (6) they enter (11) and (12).

4.3.5 Initialization Algorithm

This section presents Algorithm 1 tailored to the discussed initialization approaches. It is supposed to run only with the available prior data set up to the time instant $t = T$ (where T is whole number of prior data) before the on-line time loop of the clustering.

Finally, in the case of using Section 4.3.1, results of Algorithm 1 are $(\hat{S}_{l;T})_i$, which is the center of the i-th component for the l-th entry of y_t, and the component weights, both recursively updated by all prior data. Results obtained according to Sections 4.3.2 and 4.3.3 are the component centers computed offline and the component weights. With the help of the last technique from Section 4.3.4, the initial bounds of components are obtained along with the weighting vector.

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 3.3 is applied. For the three first initialization techniques, relations (31) and (32) should be used before measuring the first data item y_t.

Algorithm 1.

```
{Preliminary initialization (for $t = 0$)}
Set the number of components $m_c$.
for all $i, j \in \{1, 2, \ldots, m_c\}$ and $l \in \{1, 2 \ldots, K\}$ do
    Set the initial random values of the component mid-point and mid-range statistics $(s_{l;0})_i$, $(q_{l;0})_i$ and the pointer statistics $v_0$ according to (26), (27) and (25).
    Using these statistics, compute the point estimates (28), (29), (30) and (8).
end for
for all $i \in \{1, 2, \ldots, m_c\}$ do
    Set the initial (random or uniform) weighting vector $w_0$.
end for
{Initialization with prior data set (for $t = 1, \ldots, T$)}
for $t = 1, 2, \ldots, T$ do
    Load the prior data item $y_t$.
    for all $i, j \in \{1, \ldots, m_c\}$ and $l \in \{1, \ldots, K\}$ do
        Obtain the proximities (10) as follows:
        if According to Section 4.3.1, then
            Use (28) and (29) in (10).
        else if According to Section 4.3.2, then
            Put the $k$-means centers in (31) and (32).
            Compute (11), (12) and (10).
        else if According to Section 4.3.3, then
            Use means of data entries as $(\hat{S}_{l;t})_i$ in (31) and (32).
            Compute (11), (12) and (10).
        else if According to Section 4.3.4, then
            Set $(\mathcal{L}_{l;t})_i$ and $(\mathcal{R}_{l;t})_i$ as minimum and maximum values of prior data entries.
            Using (6), compute (11), (12) and (10).
        end if
        Using (10) and (8), compute the weighting vector $w_t$ via (13), its normalization and summation over rows.
        Update the pointer statistics (25).
        Re-compute its point estimate (8).
        if According to Section 4.3.1, then
            Update the statistics (26), (27).
            Re-compute the point estimates (28), (30) and (29) and go to the first step of the initialization with prior data.
        end if
    end for
end for
```

5 EXPERIMENTS

This section provides the experimental comparison of the described initialization approaches with the help of real data introduced in Section 4.1. The validation of approaches was performed according to three following criteria:

- 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.
- 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.
- The shape and the location of final clusters de-

tected in the data space by starting the estimation algorithm with the mentioned initialization techniques are compared. Comparison with *k*-means clustering is also demonstrated.

5.1 Evolution of Component Weights

A fragment of the evolution of the component weights with the statistics initialized according to Section 4.3.1 is demonstrated in Figure 3. It can be seen that all three components are regularly active. The plotted weights are approaching to 0 or 1 that unambiguously expresses the activity of components. The *k*-means based initial statistics give the similar activity, see Figure 4. The initialization via methods from Sections 4.3.3 is shown in Figure 5. It produces a bit more probabilities close to 0.5. However, in general, the result is similar to two first methods.

The last method based on minimum and maximum prior values according to Section 4.3.4 provides only two detected components. Figure 6 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.

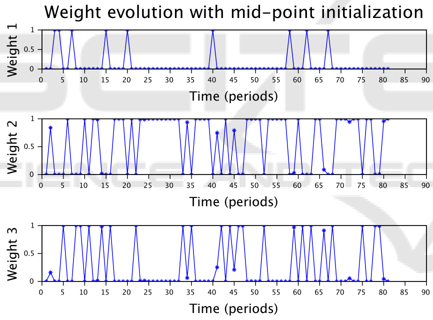

Figure 3: Evolution of component weights with initial statistics according to Section 4.3.1.

5.2 Evolution of Bounds

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., with exception of Section 4.3.4. The bounds of one of the components are stabilized a bit slower than the others: component 2 in the case of Section 4.3.1, component 3 with Section 4.3.2 and component 1 with Sections 4.3.3. The bounds of the rest of the components detect their final values relatively quickly. To save space, an example of the left bound evolution is shown in Figure 7 for initialization based on Section 4.3.1, where the difference between the second component and the rest of them should be noticed. The right bound evolution with the *k*-means initialization due to Section 4.3.2 is demonstrated in Figure 8, where the same can be said about the third component.

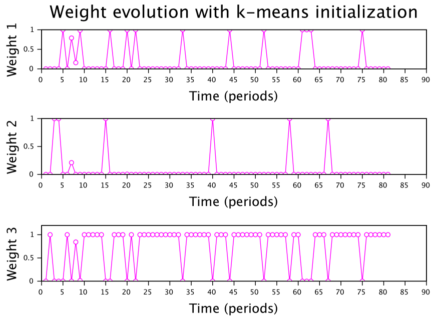

Figure 4: Evolution of component weights with initialization according to Section 4.3.2.

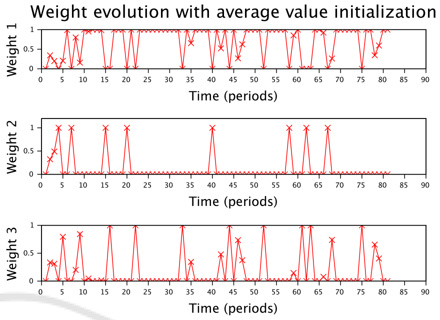

Figure 5: Evolution of component weights with initialization according to Section 4.3.3.

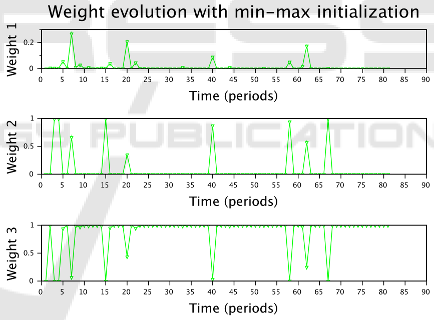

Figure 6: Evolution of component weights with initialization according to Section 4.3.4.

Initialization according to Section 4.3.4 provides a worse stabilization in search of stabilized values of the bounds, see an example of the left bound evolution for the third component in Figure 9, where the evolution of the left (minimum) bounds of individual data entries is presented.

5.3 Clusters

Clusters of the most interesting pair of data entries from the practical (hematological) point of view are

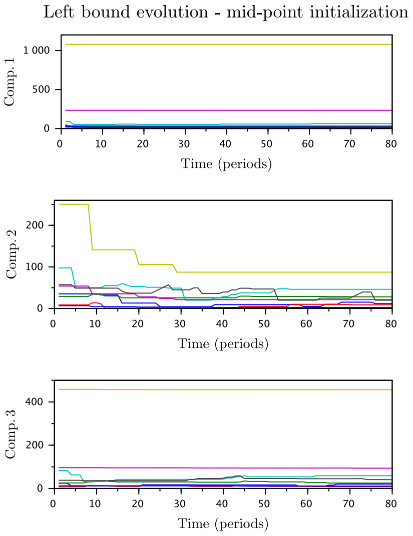

Figure 7: Evolution of the left bounds with initialization according to Section 4.3.1.

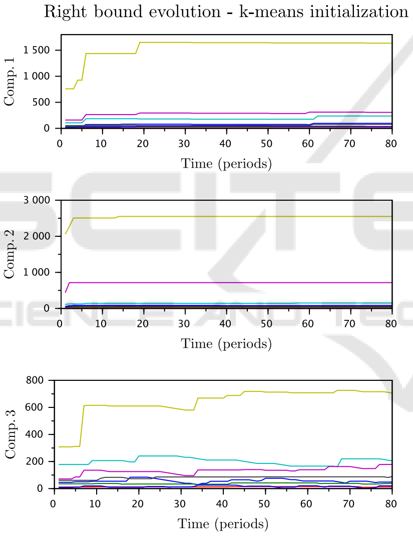

Figure 8: Evolution of the right bounds with initialization according to Section 4.3.2.

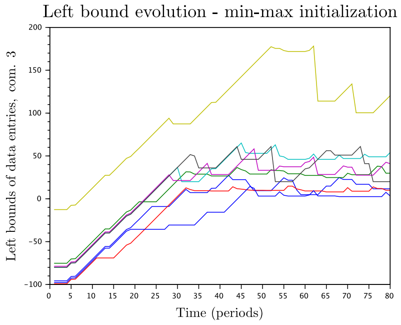

Figure 9: Evolution of the left bounds of the third component with initialization according to Section 4.3.4.

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 Figure 10, where comparison of results initialized according to Sections 4.3.1, Section 4.3.2, 4.3.3 and 4.3.4 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.

The insignificant difference in the location of two upper clusters can be seen in three first figures, while in the bottom figure the clustering practically fails. Only two data items are classified as belonging to the first cluster, i.e., two clusters are detected.

For validation of clustering, the *k*-means algorithm was run with whole data set. Results of this clustering are given in Figure 11. It can be seen that the shape and location of clusters are very similar in Figure 11 and in the first and the third top plots in Figure 10. The second top plot differs a bit.

5.4 Discussion

To summarize the experimental part of the work, it can be stated that the obtained results of the recursive clustering are validated by such the well-known theoretical counterpart as *k*-means. It should not be forgotten that *k*-means works with whole available data set off-line, while the recursive clustering is based on a completely different philosophy of on-line estimation. The use of the normal approximation as the proximity function for uniform components is also successfully validated.

Among the discussed initialization techniques the last method concerned with using the minimum and maximum bound statistics has the worst results. This indicates that the initialization via centers of uniform components is a reasonable way of starting the estimation algorithm. The initialization with randomly chosen centers (which is not shown here to save space) mostly leads to a dominance of one components.

The described initialization still need an expert's intervention, namely, for a choice of the component number. Surely, automatization of this process would be preferable. This will be one of the tasks within the current project.

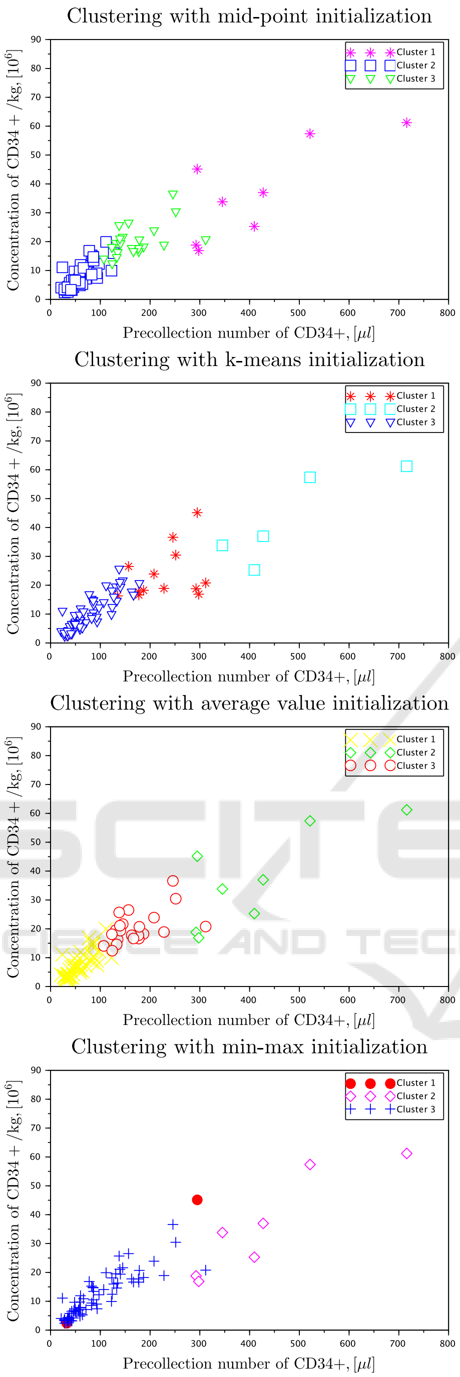

Figure 10: Comparison of clusters of $y_{5;t}$ and $y_{8;t}$ with different initialization techniques obtained via recursive mixture-based clustering.

6 CONCLUSIONS

The paper explores four approaches to a task of initialization of recursive mixture-based clustering with the uniform components under the Bayesian methodology. The investigated approaches are based on processing the prior data set with the aim of setting the initial statistics of uniform components. The comparison with the theoretical counterpart shows that the presented results are promising.

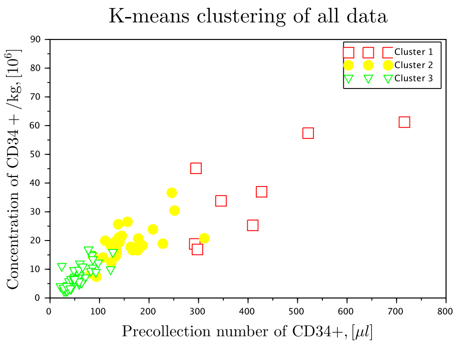

Figure 11: Clusters of $y_{5;t}$ and $y_{8;t}$ detected by k-means.

However, there still exists a series of open problems in the discussed area, e.g., a start of forgetting both the bounds (it must not be the same). Further, modeling dependent uniformly distributed variables with parallelogram-shaped clusters is still not solved. This is also a subject of the planned research work.

ACKNOWLEDGEMENTS

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

REFERENCES

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.

Hu, X., Munkin, M.K. and Trivedi, P.K., (2015). Estimating Incentive and Selection Effects in the Medigap Insurance Market: An Application with Dirichlet Process Mixture Model. , *Journal of Applied Econometrics*, 30(7), p.1115-1143.

Bao, L. and Shen, X., (2016). Improved Gaussian mixture model and application in speaker recognition. *In 2nd IEEE International Conference on Control, Automation and Robotics (ICCAR)*, p. 387–390.

Bouveyron, C., Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis.* 71(0), p. 52–78.

Scrucca, L., (2016). Genetic algorithms for subset selection in model-based clustering. *Unsupervised Learning Algorithms*, p. 55–70, Springer International Publishing.

Fernández, D., Arnold, R., Pledger, S., (2016). Mixture-based clustering for the ordered stereotype model.

Computational Statistics & Data Analysis, 93, p.46–75.

Browne, R.P. and McNicholas, P.D., (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2), p.176–198.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Maitra, R., (2009). Initializing partition-optimization algorithms. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(1), p.144–157.

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.

Peterka, V. (1981). Bayesian system identification. In: *Trends and Progress in System Identification (ed. P. Eykhoff)*, Oxford, Pergamon Press, 1981, p. 239–304.

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.

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.

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.

Casella, G., Berger R.L. (2001). *Statistical Inference, 2nd ed.*, Duxbury Press.

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.

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.

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.

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

Jain, A. K., (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp. 651–666.