1

Clustering Non-gaussian Data Using Mixture Estimation with Uniform Components

Ivan Nagy[2,1], Evgenia Suzdaleva[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
suzdalev@utia.cas.cz
[2]Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 11000 Prague, Czech Republic
nagy@utia.cas.cz

Summary. This chapter considers the problem of clustering non-gaussian data with fixed bounds via recursive mixture estimation under the Bayesian methodology. Here a mixture of uniform distributions is taken, where individual clusters are described by mixture components. For the on-line detection of data clusters, the paper proposes a mixture estimation algorithm based on (i) the update of reproducible statistics of uniform components; (ii) the heuristic initialization via the method of moments; (iii) the non-trivial adaptive forgetting technique; (iv) the data-dependent dynamic pointer model. Results of validation experiments are presented.

1.1 INTRODUCTION

The cluster analysis is a powerful tool of data processing solved by a great number of methods (e.g., well-known centroid, density based methods, etc.), see e.g., [1, 2, 3]. One of the cluster analysis domain is the mixture-based clustering, which is discussed in this paper.

The mixture components describe individual clusters in the data space. This means that the location, size and shape of components are important in the task of covering the data clusters. The location and the size are given by the expectation and the covariance matrix of the component, while the shape is defined by its distribution. Gaussian components are traditionally successful

in detecting elliptic clusters [4, 5, 6]. However, clusters of a different shape require a solution with involved components of other distributions. The same situation occurs in the case of clustering non-negative, or somehow limited data (for instance, a vehicle speed under the speed limits). To take into account such a feature, components should have a limited support. Such a choice is e.g., the uniform distribution, which well covers clusters of the rectangle shape (for independent variables in the data space) or of the parallelogram shape in the case of dependent variables.

Estimation of data models with the bounded support including uniform ones was solved in various domains, e.g., clustering [7], individual state-space and regression models [8, 9] as well as mixture models [10], etc. In mixture-based clustering the bounded data the challenging task is the update of statistics of the uniform component parameters. Intuitively the prior chosen bounds of the uniform distribution are only expandable, but they are not floating in the data space to detect the centers of clusters. While estimating a uniform mixture, this feature is harmful and should be fixed. A solution to this task will allow to perform clustering on-line at each time instant using both the current and the previously available measurements. This task is highly desired in many application areas (fault detection, diagnostics, medicine, etc.).

This paper solves the problem based on the recursive Bayesian estimation algorithms proposed for normal regression components in [11, 12, 13] and applies them for derivation of the algorithm for the uniform mixture. The paper represents the extended version of the work [22]. The main contributions of the approach in addition to the recursive statistics update include also: (i) the initialization based on finding the initial centers of components with the help of the method of moments; (ii) the novel non-trivial adaptive technique of forgetting; (iii) using the categorical data-dependent dynamic model of switching, which assumes that the currently active component is modeled in dependence of the past active one and on discrete measurements too.

The proposed algorithm also enriches the clustering and classification tools developed within the current project under the adopted Bayesian recursive mixture estimation context. The systematic extension of the theory has already given algorithms for normal regression [14], state-space [15], mixed normal and categorical [16] and exponential components [17]. Here this line is continued by developing the systematic approach to uniform components and the data-dependent model of switching partially started in [16].

The present paper focuses on independent variables. Modeling dependent uniformly distributed variables with parallelogram-shaped clusters is definitely an important task that will be solved later. However, the main aim of this work is to cluster data with fixed lower and upper bounds, which is sufficiently covered by rectangle clusters provided by independent variables.

The paper is organized in the following way. Section 1.2 introduces models and formulates the problem. The preparative Section 1.3 recalls necessary basic facts about estimation of individual uniform and categorical models and explains the general idea of the approach. Section 1.4 presents a solution to

the formulated task and the structural algorithm. Section 1.5 provides results of the experimental validation of the proposed algorithm. Conclusions and open problems are given in Section 1.6.

1.2 MODELS AND PROBLEM FORMULATION

Let's consider a multi-modal system, which at each discrete time instant $t = 1, 2, ...$ generates continuous data y_t, whose values are bounded by minimal and maximal bounds different within each working mode, and discrete data z_t with the set of its possible values $\{1, 2, \ldots, m_z\}$. It is assumed that the observed system works in m_c working modes indicated by values of the unmeasured dynamic discrete variable $c_t \in \{1, 2, \ldots, m_c\}$, which is called the pointer [11], and each of the pointer values also depends on values of the measured variable z_t.

The system is described by a mixture of uniform distributions presented by the probability density functions (pdfs)

$$f(y_t|\Theta, c_t = i),\ i \in \{1, 2, \ldots, m_c\}, \tag{1.1}$$

where $\Theta = \{\Theta_i\}_{i=1}^{m_c}$ is a collection of unknown parameters of all components, and $\Theta_i = \{L_i, R_i\}$ (for $c_t = i$) are parameters of the i-th component, where L_i is the minimum bound of the data y_t, and R_i is the maximum bound.

Switching the components describing the data is described by the following data-dependent dynamic pointer model:

$$f\left(c_t = i|\alpha, c_{t-1} = j, z_t = k\right) = \tag{1.2}$$

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

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 z_t. 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 $z_t = k$.

The task is to cluster the data on-line at each time instant t according to the determined active component based on the available data collection and newly arriving data. Under Bayesian methodology adopted in [11, 12, 13] it leads to looking for a recursive algebraic computation of statistics of the involved distributions, which is obtained by substituting the prior pdf to be propagated into the Bayes rule [18, 19]:

$$f(\Theta|\Delta(t)) \propto f\left(y_t|\Theta\right) f(\Theta|\Delta(t-1)), \tag{1.3}$$

where the denotation $\Delta(t) = \{\Delta_0, \Delta_1, \ldots, \Delta_t\}$ represents the collection of data available up to the time instant t; Δ_0 denotes the prior knowledge; the data item Δ_t includes the pair $\{y_t, z_t\}$; and $f(\Theta|\Delta(t-1))$ is the prior pdf.

Within the considered context the clustering problem is specified as the recursive estimation of

- all component parameters Θ;
- the pointer model parameter α;
- the value of the pointer c_t expressing the active component at each time instant t.

1.3 PRELIMINARIES

1.3.1 Individual Uniform Model Estimation

As it is known [20], description of the individual uniform pdf can be presented twofold: via minimal and maximal bounds or using the mid-point and the mid-range. The multivariate uniform pdf for independent variables will be the product of univariate marginal pdfs, and the distribution will have generally the rectangle support. The assumed independence of variables leads to a straightforward extension of the univariate case up to the multivariate one, which means that the whole estimation is performed independently over individual dimensions. Here, for simplicity omitting c_t from the condition of (1.1), the uniform pdf can be presented as

$$f(y_t|\Theta) = f(y_t|L,R) = \begin{cases} \frac{1}{R-L} & \text{for } y_t \in (L,R), \\ 0 & \text{otherwise,} \end{cases} \tag{1.4}$$

$$= f(y_t|S,h)$$

$$= \begin{cases} \frac{1}{2^K \prod_{l=1}^{K} h_l} & \text{for } y_t \in (S-h,\, S+h) \\ 0 & \text{otherwise,} \end{cases} \tag{1.5}$$

where K denotes the dimension of the vector y_t, $S = [S_1, \ldots, S_K]'$ is the vector of mid-points of the distribution support, and $h = [h_1, \ldots, h_K]'$ is the vector of mid-ranges.

For description (1.4) the maximum likelihood (ML) estimation leads to using the K-dimensional statistics $\mathcal{L}_t$ and $\mathcal{R}_t$ with their update for the new measurement y_t at time t in the following form [20] for each $l \in \{1, \ldots, K\}$

$$\text{if } y_{l;t} < \mathcal{L}_{l;t-1}, \text{ then } \mathcal{L}_{l;t} = y_{l;t}, \tag{1.6}$$

$$\text{if } y_{l;t} > \mathcal{R}_{l;t-1}, \text{ then } \mathcal{R}_{l;t} = y_{l;t}. \tag{1.7}$$

The point estimates of parameters L and R at time t are then obtained as

$$\hat{L}_t = \mathcal{L}_t, \quad \hat{R}_t = \mathcal{R}_t. \tag{1.8}$$

Notice here that in dependence of the initially chosen statistics $\mathcal{L}_0$ a $\mathcal{R}_0$, the point estimate $\hat{L}_t$ can be located to the left from $\mathcal{L}_0$, and then $\hat{R}_t$ to the right from $\mathcal{R}_0$. It means that the prior statistics can be only extended, which is problematic in the case of several components.

For description (1.5) the statistics for the parameter estimation are chosen based on the method of moments (MM) [21] as the sum $s_t = [s_{1;t}, \ldots, s_{K;t}]'$ and the sum of squares q_t, which is the matrix with the diagonal $[q_{1;t}, \ldots, q_{K;t}]$. Starting with the chosen initial statistics, their update for the actual data item y_t measured at the time instant t is

$$s_t = s_{t-1} + y_t, \tag{1.9}$$

$$q_t = q_{t-1} + y_t y_t'. \tag{1.10}$$

The point estimates of parameters S and h are obtained via expressing the covariance matrix of the uniform distribution with the help of the statistics s_t and q_t

$$D_t = \left(q_t - s_t s_t'/t\right)/t, \tag{1.11}$$

which gives

$$\hat{S}_t = s_t/t, \tag{1.12}$$

$$\hat{h}_t = \sqrt{3\,\mathrm{diag}(D_t)}, \tag{1.13}$$

where $\sqrt{3\,\mathrm{diag}(D_t)}$ denotes the square roots of entries of the vector $\mathrm{diag}(D_t)$, which follows from the variance of the univariate uniform distribution.

The following remarks can be given regarding (1.9)–(1.10):

- The obtained statistics are similar to those used for the recursive estimation of parameters of the normal regression model [12]. This juncture between the uniform and the normal distribution gives a chance to apply the similar systematic approach.
- The first moment, which is the basis of the obtained statistics, has a property of "floating" in the data space. For instance, in the case of starting at zero and continuing at one hundred, the zero value will be forgotten and the pdf will be located around the hundred. It does not hold for the statistics $\mathcal{L}_t$ and $\mathcal{R}_t$, where the pdf would be expanded from zero to the hundred value. This difference is significant and can be used for detecting centers of components.
- However, the drawback of the statistics is a lack of the characteristics of limiting the data by the support. As a consequence, the component support can move to the prohibited data area. To avoid this, further restrictions should be used.

1.3.2 Individual Categorical Model Estimation

The individual categorical model (1.2) in the case of the measured values of c_t, c_{t-1} and z_t is estimated via (1.3) using the conjugate prior Dirichlet pdf according to [13] with the recomputable statistics $(v_{t-1})_k$, which is here the square m_c-dimensional matrix, existing for each value of z_t. Its entries for $c_t = i$, $c_{t-1} = j$ and $z_t = k$ would be updated for $i, j \in \{1, \ldots, m_c\}$, $k \in \{1, \ldots, m_z\}$ in the following way:

$$(v_{i|j;t})_k = (v_{i|j;t-1})_k + \delta(i, j, k; c_t, c_{t-1}, z_t), \tag{1.14}$$

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

$$(\hat{\alpha}_{i|j;t})_k = \frac{(v_{i|j;t})_k}{\sum_{l=1}^{m_c} (v_{l|j;t})_k}, \tag{1.15}$$

The value of the pointer c_t at time t points to the active component. However, values of c_t and c_{t-1} are unavailable and should be estimated.

1.4 MIXTURE ESTIMATION WITH UNIFORM COMPONENTS

The discussed mixture-based clustering the bounded data is based on estimation of a mixture of uniform components and determination of the currently active one. Generally one of the key problems during the mixture estimation is initialization of the algorithm, i.e., the initial location of components in the data space. Difficulties with initialization of the mixture estimation algorithm still grow in the case of uniform components.

In order to estimate the mixture under the adopted theory [11, 12] (see the state of the art in Section 1.1), it is necessary to obtain the algebraic recursion of the update of the individual component and the pointer statistics in the form

$$\begin{aligned} \text{actual statistics} &= \text{previous statistics} \\ &+ \text{weighted data-based statistics increment.} \end{aligned}$$

Another demand is the use of the statistics for effective computing the point estimates of parameters. If these two requirements are met, it is possible to use the following estimation scheme, see, e.g., [17]:

- 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 and the past activity, where the maximal probability declares the currently active component;
- Updating the statistics of all components and the pointer model;
- Re-computing the point estimates of parameters necessary for calculating the proximity.

Uniform components do not belong to the exponential family, and extension of the estimation approach to this class of components is not entirely trivial. The way how to utilize advantages of both the types of statistics during the recursive estimation and to keep the disjoint components is the algorithm of their combination along with the forgetting technique, which is proposed in this paper. This way is rather suitable, if positions of components are known before (at least, one data item from each component). Thus, it is suitable to initialize the estimation by using the statistics (1.9)–(1.10) for detecting the initial centers of components with the help of prior data, and then actualize (1.6)–(1.7) by new data for specifying the bounds.

1.4.1 Proximity as the Approximated Likelihood

The derivations of individual points of the above scheme are based on construction of the joint pdf of all variables to be estimated and application of the Bayes rule, which takes the form (under assumption of the mutual independence of Θ and α, and Δ_t and α, and c_t and Θ):

$$\begin{aligned}
&\underbrace{f(\Theta, c_t = i, c_{t-1} = j, \alpha | \Delta(t))}_{joint\ posterior\ pdf} \\
&\propto \underbrace{f(y_t, \Theta, c_t = i, z_t = k, c_{t-1} = j, \alpha | \Delta(t-1))}_{via\ chain\ rule\ and\ Bayes\ rule} \\
&= \underbrace{f\left(y_t | \Theta, c_t = i\right)}_{(1.1)} \underbrace{f(\Theta | \Delta(t-1))}_{prior\ pdf\ of\ \Theta} \\
&\times \underbrace{f\left(c_t = i | \alpha, c_{t-1} = j, z_t = k\right)}_{(1.2)} \underbrace{f(\alpha | \Delta(t-1))}_{prior\ pdf\ of\ \alpha} \\
&\times \underbrace{f(c_{t-1} = j | \Delta(t-1)),}_{prior\ pointer\ pdf}
\end{aligned} \tag{1.16}$$

$\forall i, j \in \{1, 2, \ldots, m_c\}$ and for $k \in \{1, 2, \ldots, m_z\}$. To obtain recursive formulas for estimation of c_t, Θ and α with the help of (1.16), it is necessary to marginalize it firstly over the parameters Θ and α.

The marginalization of (1.16) over parameters Θ provides the proximity, i.e., the closeness of the current data item y_t to individual components at each time instant t. It is evaluated in the same way as for the mixture estimation

with normal regression components, i.e., as the approximated likelihood. It is the value of the normal pdf, which is the normal approximation of the uniform component, optimal in the sense of the Kullback-Leibler divergence, see, e.g., [13]. The proximity is obtained by substituting the point estimates of the expectation $(E_{t-1})_i$ and the covariance matrix $(D_{t-1})_i$ of each i-th uniform component from the previous time instant $t-1$ and the currently measured y_t into the pdf

$$m_i = (2\pi)^{-K/2}|(D_{t-1})_i|^{-1/2} \times \exp\left\{-\frac{1}{2}\left(y_t - (E_{t-1})_i\right)' (D_{t-1}^{-1})_i \left(y_t - (E_{t-1})_i\right)\right\}, \tag{1.17}$$

where $(E_{t-1})_i$ and $(D_{t-1})_i$ are either (1.12) and (1.11) respectively obtained for the i-th component via the statistics (1.9)–(1.10), or the expectation $(E_{t-1})_i$ is the K-dimensional vector, each l-th entry of which is

$$(E_{l;t-1})_i = \frac{1}{2}((\hat{L}_{l;t-1})_i + (\hat{R}_{l;t-1})_i), \tag{1.18}$$

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

$$(D_{l;t-1})_i = \frac{1}{12}((\hat{R}_{l;t-1})_i - (\hat{L}_{l;t-1})_i)^2 \tag{1.19}$$

obtained via (1.8). The proximities from all m_c components form the m_c-dimensional vector m.

Similarly, the integral of (1.16) over α provides the computation of its point estimate (1.15) using the previous-time statistics $(v_{t-1})_k$ for the actual value k of z_t.

1.4.2 Component Weights

In order to obtain the i-th component weight (a probability that the component is currently active) the proximities (1.17) are multiplied entry-wise by the previous-time point estimate of the parameter α (1.15) and the prior weighting m_c-dimensional vector w_{t-1}, whose entries are the prior (initially chosen) pointer pdfs $(c_{t-1} = j|\Delta(t-1))$, i.e.,

$$W_t \propto (w_{t-1}m') \,.*\, (\hat{\alpha}_{t-1})_k \tag{1.20}$$

where W_t denotes the square m_c-dimensional matrix comprised from pdfs $f(c_t = i, c_{t-1} = j|\Delta(t))$ 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.

1.4.3 The Component Statistics Update

Using the obtained weights $w_{i;t}$ at time t, the component statistics are updated as follows. The updates (1.9)–(1.10) for the i-th component takes the form

$$(s_{l;t})_i = (s_{l;t-1})_i + w_{i;t} y_{l;t}, \tag{1.21}$$

$$(q_{l;t})_i = (q_{l;t-1})_i + w_{i;t} y_{l;t}^2, \tag{1.22}$$

$\forall i \in \{1, 2, \ldots, m_c\}$ and $\forall l = \{1, \ldots, K\}$.

The update of the statistics (1.6)–(1.7) is performed with the adaptive forgetting technique. The principle of this forgetting is as follows. If the larger value of the entry $y_{l;t}$ is measured, the corresponding l-th maximum bound $(\mathcal{R}_{l;t})_i$ of the i-th component moves on it. Otherwise it is a bit narrowed. If the larger values of $y_{l;t}$ do not arrive for some number of time instants, the maximum bound will decrease until it meets the active area where the data are measured. The identical principle is for the minimum bound on the opposite direction. The question is a reasonable start of forgetting. The idea is very simple: forgetting should not be performed as soon as the bound does not move, but after some period of time. Therefore, it is necessary to estimate how often the bounds should be updated, and start forgetting when they do not move too long.

If the maximum statistics $(\mathcal{R}_{l;t-1})_i$ is assumed to lie in e.g., 80% of the interval of the i-th uniform component (from the minimum to the maximum), the probability of measuring the value of $y_{l;t} > (\mathcal{R}_{l;t-1})_i$, which leads to its update (1.7), is 0.2. The number n of the time instants, when $(\mathcal{R}_{l;t-1})_i$ was not updated (due to $y_{l;t} < (\mathcal{R}_{l;t-1})_i$)) is described by the geometrical distribution with the distribution function

$$F(n) = 1 - (1 - p)^n, \tag{1.23}$$

here with $p = 0.2$. Taking the confidence level, for instance, 0.05, it is possible to compute the number n, for which the following relation holds:

$$1 - F(n) = 0.05. \tag{1.24}$$

It is

$$n = \frac{\ln(0.05)}{\ln(0.8)} \doteq 13. \tag{1.25}$$

It means that if the statistics was not updated during $n = 13$ time instants, and the update follows, then with the probability 0.95 the current point estimate (according to (1.8)) lies in 20% from the population maximum border. This can occur either (i) if it really lies in 20% from the right bound – then shifting the bound to the left leads to the higher frequency of updating, or (ii) the point estimate of the right bound is caused by some outlier – then the shifting remains until the bound estimate reaches the corresponding data cluster. This will enable to get rid of inaccuracies brought by the prior statistics.

In this way, the update (1.6)–(1.7) takes the following form $\forall i \in \{1, 2, \ldots, m_c\}$ and $\forall l = \{1, \ldots, K\}$. For the minimum bound, the counter of non-updates is set as

$$(\lambda^L_{l;t-1})_i = 0, \tag{1.26}$$

and then

$$\delta_L = y_{l;t} - (\mathcal{L}_{l;t-1})_i, \tag{1.27}$$

$$\text{if } \delta_L < 0, \quad (\mathcal{L}_{l;t})_i = (\mathcal{L}_{l;t-1})_i - w_{i;t}\delta_L, \tag{1.28}$$

$$(\lambda^L_{l;t})_i = 0, \tag{1.29}$$

$$\text{else} \quad (\lambda^L_{l;t})_i = (\lambda^L_{l;t-1})_i + 1, \tag{1.30}$$

$$\text{if } (\lambda^L_{l;t})_i > \underbrace{n}_{(1.25)}, \quad (\mathcal{L}_{l;t})_i = (\mathcal{L}_{l;t-1})_i + \phi w_{i;t}, \tag{1.31}$$

where ϕ is the forgetting factor, often set as 0.01. Similarly the update is performed for the maximum bound with the counter of non-updates set as $(\lambda^R_{l;t-1})_i = 0$, i.e., and

$$\delta_R = y_{l;t} - (\mathcal{R}_{l;t-1})_i, \tag{1.32}$$

$$\text{if } \delta_R > 0, \quad (\mathcal{R}_{l;t})_i = (\mathcal{R}_{l;t-1})_i + w_{i;t}\delta_R, \tag{1.33}$$

$$(\lambda^R_{l;t})_i = 0, \tag{1.34}$$

$$\text{else} \quad (\lambda^R_{l;t})_i = (\lambda^R_{l;t-1})_i + 1, \tag{1.35}$$

$$\text{if } (\lambda^R_{l;t})_i > n, \quad (\mathcal{R}_{l;t})_i = (\mathcal{R}_{l;t-1})_i - \phi w_{i;t}. \tag{1.36}$$

1.4.4 The Pointer Update

The statistics of the pointer model is updated similarly to the update of the individual categorical model and based on [13, 11], however, with the joint weights $W_{i,j;t}$ [17] from the matrix (1.20), where the row j corresponds to the value of c_{t-1}, and the column i to the current pointer c_t

$$(v_{i|j;t})_k = (v_{i|j;t-1})_k + \delta(k; z_t)W_{j,i;t}, \tag{1.37}$$

and the Kronecker delta function $\delta(k; z_t) = 1$ for $z_t = k$ and 0 otherwise.

1.4.5 Algorithm

The following algorithm specifies the estimation scheme with the above relations.
Initialization (for $t = 0$)

1. Set the number of components m_c.

2. Set the initial (expert-based or random) values of all component statistics $(s_{l;0})_i$, $(q_{l;0})_i$ and the pointer statistics $(v_0)_k$ $\forall k \in \{1, 2, \ldots, m_z\}$.
3. Using the initial statistics, compute the point estimates (1.12), (1.13), (1.11) and (1.15).
4. Set the initial weighting vector w_0.

Initialization of component centers (for $t = 1, \ldots, T$)

1. Load the prior data item y_t, z_t.
2. Substitute (1.12), (1.11) and y_t into (1.17) to obtain the proximities.
3. Using (1.15) for the actual value k of z_t, compute the weighting vector w_t via (1.20), its normalization and summation over rows.
4. Update the statistics (1.21), (1.22) and (1.37).
5. Re-compute the point estimates (1.12), (1.13), (1.11) and (1.15) and go to Step 1 of the initialization of component centers.
6. For $t = T$ the result is $(\hat{S}_{l;T})_i$, which is the center of the i-th component for the l-th entry of y_t.

On-line bound estimation (for t=T+1, T+2, ...)

1. Use the obtained centers to set the initial bounds $(\hat{L}_{l;T})_i = (\hat{S}_{l;T})_i - \varepsilon$, and $(\hat{R}_{l;T})_i = (\hat{S}_{l;T})_i + \varepsilon$ with small ε.
2. Measure the data item y_t, z_t.
3. For all components, compute the expectations and the covariance matrices via (1.18) and (1.19).
4. Substitute (1.18), (1.19) and the current y_t into (1.17).
5. Using (1.15) for the actual value k of z_t, compute the weighting vector w_t via (1.20), its normalization and summation over rows.
6. Declare the active component according the biggest entry of the vector w_t, which is the point estimate of the pointer c_t at time t.
7. Update the component statistics according to (1.26)–(1.36) for both the bounds.
8. Update the pointer statistics (1.37).
9. Re-compute the point estimates (1.18), (1.19) and (1.15) and go to Step 2 of the on-line bound estimation.

1.5 RESULTS

1.5.1 Illustrative example

A simple example of the algorithm application can be done by using simulated data with three uniform components. For programming the open-source software Scilab (www.scilab.org) was used.

Advantages of the update (1.26)–(1.36) with forgetting can be demonstrated as follows. 500 simulated data items of two-dimensional vector y_t are

taken. Figure 1.1 (top) shows clusters detected by using the update with forgetting and compares them with clustering, where the update was taken without forgetting (bottom). It can be seen that even in the simple case of well-distinguishable components the clustering without forgetting in the bottom plot is not successful.

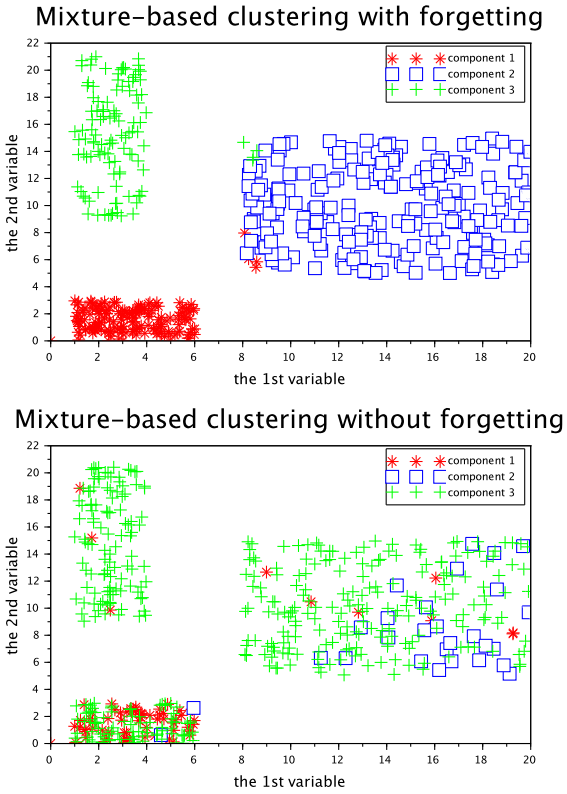

Fig. 1.1. Comparison of clustering with and without forgetting with well-distinguishable components.

The top figure compares results of clustering with forgetting (top) and without forgetting (bottom). Notice that three well-distinguishable components are correctly detected in the top plot, but they are incorrectly determined in the bottom figure.

Figure 1.2 compares results for a more complicated situation, where components are located close to each other. Mixture-based clustering with forgetting coped with this task, which is demonstrated in Figure 1.2 (top). Clustering without forgetting gave similar results as in Figure 1.1 (bottom).

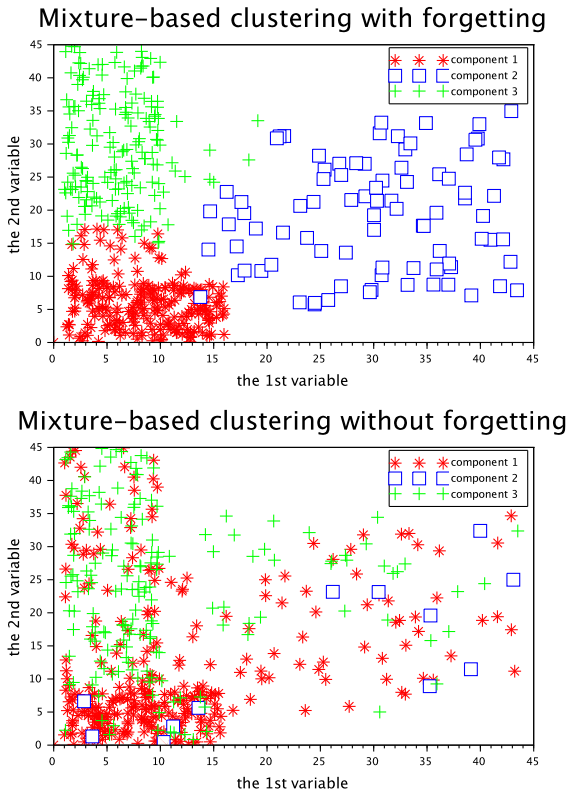

Fig. 1.2. Comparison of clustering with and without forgetting with closely located components.

Notice that three components are sharply visible in the top figure, and they are incorrectly detected in the bottom figure.

These results only verify correctness of programming. The next section provides validation experiments with realistic traffic simulations, which is a much more complicated task.

1.5.2 Validation experiments

Realistic simulations from the transportation microscopic simulator Aimsun (www.aimsun.com) were used for testing the proposed algorithm. A series of validation experiments was performed. Here typical results are shown.

The aim of clustering was to detect different working modes in the traffic flow data.

Data

The data vector y_t contained:

- $y_{1;t}$ – the non-negative traffic flow intensity [vehicle/period] bounded by the saturated flow on the considered intersection on the maximal value 32;
- $y_{2;t}$ – the non-negative occupancy of the measuring detector [%] with the maximal value 100.

The discrete variable z_t was the measured indicator of the vehicle queue existence such that $z_t \in \{0, 1\}$, where 1 denotes that the queue is observed, and 0 – there is no queue. The data were measured each 90 seconds.

Initialization

400 prior data items were used for the initial detection of component centers. The assumed number of components expressing the traffic flow modes was $m_c = 3$. The initial centers of components obtained according to the corresponding part of the algorithm are given in Table 1.1.

Table 1.1. Initial centers of three components

	$(\hat{S}_{T=400})_i$
$i = 1$	$[16.1 \quad 85.4]'$
$i = 2$	$[16.3 \quad 33.5]'$
$i = 3$	$[8.6 \quad 2.4]'$

Results

1000 data items were used for the bound estimation. Figure 1.3 demonstrates results of the proposed mixture-based clustering (top) and compares them with the k-means clustering [2] (bottom).

The proposed algorithm detected three clusters on-line by actualizing the statistics by each arriving data item, see Figure 1.3 (top).

The upper cluster with the center $[16.1 \quad 85.4]'$ (according to $i = 1$ in Table 1.1) corresponds to the approaching unstable traffic flow. This explains why the intensity is almost the same as for the middle cluster (corresponding to $i = 2$ in Table 1.1, i.e., with the initial value 16.3), however, the detector occupancy reports a high degree of workload for the upper cluster. The middle cluster can be interpreted as the stable flow. The bottom cluster in Figure 1.3 (top) with the lower intensity and lower occupancy and the initial centers $[8.6 \;\; 2.4]'$ corresponds to the free traffic flow.

Figure 1.3 (bottom) shows three clusters detected by iterative processing of the whole data sample by the k-means algorithm. The results differ from

those obtained in Figure 1.3 (top): the cluster with the free flow is partitioned in two clusters with two centers. The middle cluster is not found. The upper cluster is the same as in Figure 1.3 (top).

1.5.3 Discussion

The obtained results look promising. Surely, the question can arise of application of the approach in the case of lack of prior data. The estimation with both the types of statistics can be performed also independently. When using the algorithm only with the statistics (1.9)–(1.10), the bounds can be probably exceeded, which can be fixed by additional restrictions. For the algorithm only with the statistics (1.6)–(1.7) the initialization with random centers can be applied in the absence of prior data. In this case setting the adaptive forgetting technique is expected to considerably help in finding the bounds. Analyzing the series of performed experiments, it can be said that the most suitable application of the approach is the combination of statistics given in Algorithm 1.4.5.

One of the advantages of the proposed algorithm is also a possibility of its on-line running with real-time measurements and the gradual update of component bounds, while iterative algorithms in this area (e.g., k-means) focus on off-line processing the whole available data sample at once. The unified systematic Bayesian approach used for other types of components (see Section 1.1), where each of models requires a specific update of statistics, is here presented for uniform components. A combination with other distributions for covering different shapes of clusters can be also a further extension of the approach.

1.6 CONCLUSION

This paper proposes the algorithm of mixture-based clustering the non-Gaussian data with fixed bounds. The specific solutions include (i) the recursive Bayesian estimation of uniform components and the switching model; (ii) the initialization via the moment method; (iii) the forgetting technique. The results of testing the algorithm are provided.

However, there still exists a series of open problems in the discussed area, where the first of them is modeling dependent uniformly distributed variables with parallelogram-shaped clusters. Further, extension of the clustering and classification tools for other distributions is planned within the future work on the present project.

ACKNOWLEDGMENT

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

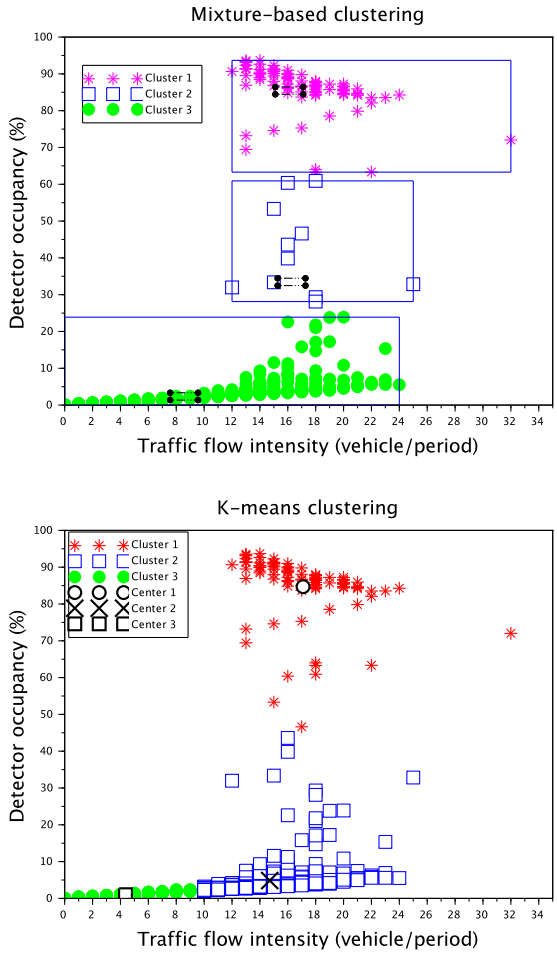

Fig. 1.3. Comparison of mixture-based (MB) (top) and k-means (KM) (bottom) clustering the traffic flow data.

The top figure shows three MB clusters: the upper one corresponds to the almost unstable traffic flow, the middle one – to the stable flow, and the bottom cluster is the free traffic flow. Initial bounds of clusters are shown as internal rectangles indicated by the dashed line. The estimated bounds of clusters are plotted as external rectangles. The bottom figure plots three KM clusters and their final centers. The upper cluster is similar to the MB results, however, the clusters of stable and free traffic flow are different.

References

1. Berkhin, P. A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data.* Eds.: J. Kogan, C. Nicholas, M. Teboulle. Springer Berlin Heidelberg, 2006, p. 25–71.
2. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, vol.31, 8 (2010), p. 651–666.
3. Bouveyron, C., Hammer B., Villmann T. Recent developments in clustering algorithms. In: *Verleysen M*, ed. ESANN 2012, p. 447–458.
4. Fraley, C., Raftery, A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Computer Journal*, vol.41, 8 (1998), p. 578-588.
5. Fraley, C., Raftery, A. E. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Journal of Classification*, vol.24, 2 (2007), p.155–181.
6. Bouveyron, C., Brunet-Saumard, C. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, vol.71 (2014), pages 52–78.
7. Banfield, J. D., Raftery, A. E. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, vol.49, 3 (1993), p.803–821.
8. Pavelková, L., Kárný, M. State and parameter estimation of state-space model with entry-wise correlated uniform noise. *International Journal of Adaptive Control and Signal Processing*, vol.28, 11 (2014), p. 1189-1205.
9. Jirsa, L., Pavelková, L. Estimation of Uniform Static Regression Model with Abruptly Varying Parameters. In: *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics ICINCO 2015*, Eds: Filipe J., Madani K., Gusikhin O., Sasiadek J., Colmar, France, July 21–23, 2015, p. 603–607.
10. McLachlan, G., Peel, D. Finite Mixture Models. Wiley-Interscience, 2000.
11. Kárný, M., Kadlec, J., Sutanto, E.L. 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, 07.09.1998–09.09.1998, p. 77–82.
12. Peterka, V. Bayesian system identification. In: Trends and Progress in System Identification (ed. P. Eykhoff), Oxford, Pergamon Press, 1981, p. 239–304.

13. Kárný, M., Böhm, J., Guy, T. V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L. Optimized Bayesian Dynamic Advising: Theory and Algorithms, Springer-Verlag London, 2006.
14. Nagy, I., Suzdaleva, E., Kárný, M., Mlynářová, T. Bayesian estimation of dynamic finite mixtures. *Int. Journal of Adaptive Control and Signal Processing*, vol.25, 9 (2011), p. 765–787.
15. Nagy, I., Suzdaleva, E. Mixture estimation with state-space components and Markov model of switching. *Applied Mathematical Modelling*, vol. 37, 24 (2013), p. 9970–9984.
16. Suzdaleva, E., Nagy, I. Recursive Mixture Estimation with Data Dependent Dynamic Model of Switching. *Journal of Computational and Graphical Statistics*, submitted.
17. Suzdaleva, E., Nagy, I., Mlynářová, T. 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, 2015, p.137–142.
18. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. Bayesian Data Analysis (Chapman & Hall/CRC Texts in Statistical Science), 3rd ed., Chapman and Hall/CRC, 2013.
19. Lee, P. M. Bayesian Statistics: An Introduction, 4th ed., Wiley, 2012.
20. Casella, G., Berger R.L. Statistical Inference, 2nd ed., Duxbury Press, 2001.
21. Bowman, K. O., Shenton, L. R. Estimator: Method of Moments. in: Encyclopedia of statistical sciences, Wiley, 1998, p. 2092-2098.
22. Nagy, I., Suzdaleva, E., Mlynářová, T. Mixture-based clustering non-gaussian data with fixed bounds. In Proceedings of 2016 IEEE 8th International Conference on Intelligent Systems IS'2016, p. 265-271, Sofia, Bulgaria, September 4-6 2016.