Comparison of Various Definitions of Proximity in Mixture Estimation

Ivan Nagy[1,2], Evgenia Suzdaleva[1] and Pavla Pecherková[1,2]
[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*
{*nagy, suzdalev, pecherkova*}*@utia.cas.cz*

Keywords: Classification, Recursive Mixture Estimation, Proximity, Bayesian Methods, Mixture based Clustering.

Abstract: Classification is one of the frequently demanded tasks in data analysis. There exists a series of approaches in this area. This paper is oriented towards classification using the mixture model estimation, which is based on detection of density clusters in the data space and fitting the component models to them. A chosen function of proximity of the actually measured data to individual mixture components and the component shape play a significant role in solving the mixture-based classification task. This paper considers definitions of the proximity for several types of distributions describing the mixture components and compares their properties with respect to speed and quality of the resulting estimation interpreted as a classification task. Normal, exponential and uniform distributions as the most important models used for describing both Gaussian and non-Gaussian data are considered. Illustrative experiments with results of the comparison are provided.

1 INTRODUCTION

Mixture models are a powerful class of models used for description of multi-modal systems, which work in different switching regimes. Detection of an active regime is a significant task in various application fields such as, for example, fault detection, diagnostics systems, medicine, big data related problems, etc., see, e.g., (Yu, 2012a; Yu, 2011; Yu, 2012b).

This task is often solved within (supervised) classification issues, which distribute data observed on the described system among some categories, using the training set of already classified data. A range of such methods is rather wide and includes, e.g., decision trees, Bayesian classifiers, rule-based methods, including fuzzy rules, neural networks, k-nearest-neighbor classifiers, genetic algorithms and other techniques. The detailed overview of these numerical techniques can be found, e.g., in sources (Larose, 2005; Han et al., 2011; Zaki and Meira, 2014; Calders and Verwer, 2010; Zhang, 2000; Ishibuchi et al., 2000), etc.

Unlike them, the unsupervised classification methods are directly based on clustering solutions looking for data clusters in the untrained data set, such as hierarchical and partitioning methods (centroid-, density-based) and many others, see an overview, in e.g., (Berkhin, 2006; Jain, 2010; Ester et al., 1996).

A separate group of such methods is created by model-based clustering and classification algorithms presented e.g., by (Bouveyron and Brunet-Saumard, 2014; Zeng and Cheung, 2014; Ng and McLachlan, 2014), etc. These methods are more complex and demanding.

Specific methods of this type of classification approaches are based on estimation of mixture models. The mixture model consists of components in the form of probability density functions (pdfs) describing individual system regimes and a model of their switching. The mixture-based classification starts from some pre-specified (mostly resulted from the initial data analysis) locations of components and performs a search for density clusters in the data space with the aim of fitting component models to data. The search for the data clusters is one of the most critical parts of the mixture estimation algorithms.

Fundamental approaches in this area mostly focus on: (i) the use of the EM algorithm (Gupta and Chen, 2011), see, e.g., (Boldea and Magnus, 2009; Wang et al., 2004); (ii) Variational Bayes methods (McGrory and Titterington, 2009; Šmídl and Quinn, 2006); (iii) Markov Chain Monte Carlo (MCMC) methods (Frühwirth-Schnatter, 2006; Doucet and Andrieu, 2001; Chen and Liu, 2000), often used for mixtures of state-space models; and (iv) recursive Bayesian estimation algorithms (Kárný et al., 1998;

Nagy, I., Suzdaleva, E. and Pecherková, P.
Comparison of Various Definitions of Proximity in Mixture Estimation.
In *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016) - Volume 1*, pages 527-534
ISBN: 978-989-758-198-4

Peterka, 1981; Kárný et al., 2006; Nagy et al., 2011; Suzdaleva et al., 2015), etc.

The present paper deals with the mixture-based classification using the last part of the enumerated approaches, which is directed at algebraic computing the statistics of the involved component distributions avoiding applying the numerical techniques. A significant role in this process is played by the chosen function of the proximity of the actually measured data item and individual mixture components.

This paper considers several definitions of a function used as the proximity and compares their properties with respect to speed and quality of the resulting mixture estimation interpreted as a classification task. Three types of components are considered: the normal distribution as the most important type of components, and also exponential and uniform distributions, which are essential in modeling non-Gaussian data.

The paper is organized in the following way. The preparative Section 2 introduces the used models and recalls necessary basic facts about their individual recursive estimation under the Bayesian methodology. Section 3 is devoted to introducing the proximity within the mixture estimation algorithm and its chosen definitions. Section 3.4 specifies the algorithm in the unified form for all types of components and proximity definitions. Section 4 provides results of their experimental comparison. Conclusions and open problems are given in Section 5.

2 MODELS

Let's consider a multi-modal system, which at each discrete time instant $t = 1, 2, ...$ generates the continuous data vector y_t. It is assumed that the observed system works 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).

The observed system is supposed to be described by a mixture model, which (in this paper) consists of m_c components in the form of the following pdfs

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

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 component, which describes data generated by the system at the time instant t is said to be active.

2.1 Dynamic Pointer Model

Switching the active components (1) is described by the dynamic model of the pointer

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

which is 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 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 $c_{t-1} = j$.

According to (Kárný et al., 2006), the parameter α of the pointer model (2) is estimated using the conjugate prior Dirichlet pdf in the Bayes rule, recomputing its initially chosen statistics and its normalizing.

Depending on the nature of the measured data, the pdfs of components can be specified as follows.

2.2 Normal Components

Under assumption of normality of measurements, the pdf (1) can be specified as

$$(2\pi)^{-N/2}|r_i|^{-1/2}\exp\left\{-\frac{1}{2}[y_t - \theta_i]' r_i^{-1}[y_t - \theta_i]\right\}, \qquad (3)$$

where N denotes a dimension of the vector y_t; for each $i \in \{1, 2, \ldots, m_c\}$ the parameter θ_i represents the center of the i-th component; r_i is the covariance matrix of the involved normal noise, which defines the shape of the component, and here $\{\theta_i, r_i\} \equiv \Theta_i$.

The point estimates of parameters θ_i and r_i of individual components are used in computing the proximity (this will be explained later). Necessary basic facts about calculation of the point estimates are briefly given below.

According to (Peterka, 1981; Kárný et al., 1998), parameters θ_i and r_i of the individual i-th component (3) (omitting here for simplicity the subscript i) are estimated via the Bayes rule using the conjugate prior Gauss-inverse-Wishart pdf with the reproducible (initially chosen) statistics V_{t-1} and k_{t-1} of the appropriate dimensions. The statistics are updated as follows:

$$V_t = V_{t-1} + \begin{bmatrix} y_t \\ 1 \end{bmatrix}[y_t, 1], \qquad (4)$$

$$\kappa_t = \kappa_{t-1} + 1, \qquad (5)$$

and the point estimates of parameters are respectively

$$\hat{\theta}_t = V_1^{-1} V_y, \quad \hat{r}_t = \frac{V_{yy} - V_y' V_1^{-1} V_y}{\kappa_t} \tag{6}$$

with the help of partition $V_t = \begin{bmatrix} V_{yy} & V_y' \\ V_y & V_1 \end{bmatrix}$, (7)

where V_{yy} is the square matrix of the dimension $(N \times N)$, V_y' is N-dimensional column vector and V_1 is scalar, see details in (Peterka, 1981).

2.3 Exponential Components

The exponential distribution of components (1), which is often suitable for situations, where the assumption of normality brings a series of limitations (non-negative, bounded data, etc.) can be specified as

$$\left(\prod_{l=1}^{N} (a_l)_i \right) \exp\left\{ -a_i'(y_t - b_i) \right\}, \tag{8}$$

i.e., here $\{a_i, b_i\} \equiv \Theta_i$, and $(a_l)_i > 0$ and $(b_l)_i \in R$ are the l-th entries of the N-dimensional vectors a_i and b_i respectively with $l = \{1, 2 \ldots, N\}$. Currently the independence of entries of the vector y_t is assumed.

Basic facts about the recursive estimation of the parameters a_i and b_i, which are necessary to obtain the proximity (explained later) are recalled below. With the help of the Bayes rule and the prior exponential pdf, the parameters of the individual i-th exponential component (8) are obtained, using the algebraic update of the initially chosen statistics S_{t-1}, K_{t-1} and B_{t-1}. It is done at each time instant t as follows (omitting here the subscript i for simplicity):

$$S_t = S_{t-1} + y_t, \tag{9}$$
$$K_t = K_{t-1} + 1, \tag{10}$$
$$\delta_l = B_{l;t-1} - y_{l;t}, \tag{11}$$

where, denoting δ_l for the l-th entry $y_{l;t}$ of y_t,

$$\text{if } \delta_l > 0, \; B_{l;t} = B_{l;t-1} - \delta_l; \tag{12}$$
$$\text{else } B_{l;t} = B_{l;t} + \varepsilon, \tag{13}$$

where $B_{l;t}$ is the l-th entry of the statistics B_t, and ε can be taken as 0.1. The point estimates of parameters are obtained from

$$\hat{a}_{l;t} = \frac{K_t}{S_{l;t} - K_t B_{l;t}}, \tag{14}$$
$$\hat{b}_{l;t} = B_{l;t}, \tag{15}$$

where $\hat{a}_{l;t}$ and $\hat{b}_{l;t}$ are entries of vectors $\hat{a}_t$ and $\hat{b}_t$ respectively, see, e.g., (Yang et al., 2013), and $S_{l;t}$ is the l-th entry of the statistics S_t.

2.4 Uniform Components

The pdf (1) with the uniform distribution often serves as an appropriate tool for description of bounded data. In this paper the independence of entries of the vector y_t is assumed, and therefore, $\forall i \in \{1, 2, \ldots, m_c\}$ the pdf (1) takes the form

$$f(y_t | L, R, c_t = i) = \begin{cases} \frac{1}{R_i - L_i} & \text{for } y_t \in (L_i, R_i), \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

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

The estimation of parameters of the individual i-th uniform component (16) 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 $l \in \{1, \ldots, N\}$ in the following form, see, e.g., (Casella and Berger, 2001):

$$\text{if } y_{l;t} < \mathcal{L}_{l;t-1}, \quad \text{then } \mathcal{L}_{l;t} = y_{l;t}, \tag{17}$$
$$\text{if } y_{l;t} > \mathcal{R}_{l;t-1}, \quad \text{then } \mathcal{R}_{l;t} = y_{l;t}, \tag{18}$$

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

$$\hat{L}_t = \mathcal{L}_t, \quad \hat{R}_t = \mathcal{R}_t. \tag{19}$$

3 PROXIMITY DEFINITIONS

3.1 General Estimation Algorithm

The mixture estimation algorithm is derived using the joint pdf for all variables to be estimated, i.e., Θ, α and c_t, and its gradual marginalization over all of them. In general, based on (Kárný et al., 1998; Kárný et al., 2006; Peterka, 1981; Nagy et al., 2011), after initialization, the recursive mixture estimation algorithm includes the following steps at each time instant t (here they are given only to help a reader to be oriented in the discussed field):

1. Measure new data.
2. Compute the proximity of the current data item to each component. It is done by substituting the measured data item and the point parameter estimates from the previous time instant into corresponding components.
3. Construct the weighting vector containing the probabilities of the activity of components at the actual time instant. It is obtained by using the obtained proximities, the prior pointer pdf and the previous point estimate of α.

4. Update the statistics of all components using the weighting vector, and statistics of the pointer model, using the weighting matrix joint for the current pointer c_t and the previous one c_{t-1}.
5. Recompute the point estimates of all parameters and then use them as the initial ones in the first step of the on-line algorithm.

More detailed information can be also found in (Suzdaleva et al., 2015).

3.2 Proximity as the Approximated Likelihood

The proximity appeared in Step 2 of the above algorithm is explained in this section. Originally, according to (Kárný et al., 2006) the proximity has been introduced for the normal distribution as the likelihood for different variants of the model – as the so-called v-likelihood. In the considered context the variants of the normal distribution are given by normal components (3) labeled by the pointer value $c_t = i$. Generally the proximity is derived based on the following scheme. For each $i \in \{1,2,\ldots,m_c\}$ under assumption of the mutual independence of Θ and α, and y_t and α, and c_t and Θ it takes the form

$$\begin{aligned}
&\underbrace{f(\Theta, c_t = i, c_{t-1} = j, \alpha | y(t))}_{joint\ posterior\ pdf} \\
&\propto \underbrace{f(y_t, \Theta, c_t = i, c_{t-1} = j, \alpha | y(t-1))}_{via\ the\ chain\ rule\ and\ Bayes\ rule} \\
&= \underbrace{f(y_t | \Theta, c_t = i)}_{(1)} \underbrace{f(\Theta | y(t-1))}_{prior\ pdf\ of\ \Theta} \\
&\times \underbrace{f(c_t = i | \alpha, c_{t-1} = j)}_{(2)} \underbrace{f(\alpha | y(t-1))}_{prior\ pdf\ of\ \alpha} \\
&\times \underbrace{f(c_{t-1} = j | y(t-1))}_{prior\ pointer\ pdf}, \qquad (20)
\end{aligned}$$

where $y(t) = \{y_0, y_1, \ldots, y_t\}$ represents the collection of data available up to the time instant t, and y_0 denotes the prior knowledge. With the help of integrals of (20) over Θ, α and summation over c_{t-1}, the v-likelihood denoted by L_v for the pointer value $c_t = i$ is obtained as

$$\begin{aligned}
L_v(c_t = i | y(t-1)) = &\sum_{c_{t-1}=1}^{m_c} \int_{\Theta^*} \int_{\alpha^*} \underbrace{f(y_t | \Theta, c_t = i)}_{(1)} \\
&\times \underbrace{f(\Theta | y(t-1))}_{prior\ pdf\ of\ \Theta} \underbrace{f(c_t = i | \alpha, c_{t-1} = j)}_{(2)} \\
&\times \underbrace{f(\alpha | y(t-1))}_{prior\ pdf\ of\ \alpha} \underbrace{f(c_{t-1} = j | y(t-1))}_{prior\ pointer\ pdf} d\Theta\, d\alpha. \qquad (21)
\end{aligned}$$

The considered approximation via the Dirac delta function consists in substituting the point estimates of parameters Θ_i and α into (21), which gives

$$\begin{aligned}
L_v(c_t = i | y(t-1)) \doteq &f\left(y_t | \hat{\Theta}_{i;t-1}, c_t = i\right) \\
&\times f\left(c_t = i | \hat{\alpha}_{i|j;t-1}, c_{t-1} = j\right), \qquad (22)
\end{aligned}$$

where $\hat{\Theta}_{i;t-1}$ is the point estimate of Θ from the previous time instant $t-1$, and $\hat{\alpha}_{i|j;t-1}$ is the entry of the point estimate $\hat{\alpha}_{t-1}$ at time $t-1$. Thus, the result (22) is the predictive pdf existing for each $i, j \in \{1,2,\ldots,m_c\}$ with the substituted current data item y_t and the available parameter point estimates. This pdf is defined as the proximity of the actual data item and the i-th component.

For normal components (3) the predictive pdf has the Student distribution. It is chosen as one of definitions of the proximity to be used during the recursive mixture estimation.

3.3 Proximity as Decreasing Functions

Distinguishing the individual components is a complicated part of the mixture estimation. The used proximity can unambiguously assign the measured data item (if possible) to one of the components. Since components in the data space can be of a different shape and located on a different distance from each other, and also with a variously decreasing density of data items on their edges, it is important to choose a suitable distance measure. This is determined by the chosen definition of the proximity.

The main requirement for the proximity function is to be at its maximum at the center of a component and to decrease rapidly with the increasing distance. The speed of decreasing should be fast enough. That's is why it is advantageous to apply the proximity function with a curve of the normal distribution.

Other functions also have similar properties and thus can be explored as the proximity. Generally a power function can decrease in the beginning of its course faster than an exponential function, but at a larger distance the exponential function is closer to zero. Comparison of these functions as the proximity is the main task of the presented research.

Based on this idea, different definitions of the proximity can be chosen using a rapidly decreasing function depending on

$$\Delta_t = y_t - E[y_t], \qquad (23)$$

where the expectation $E[y_t]$ is computed for the corresponding type of components.

For the i-th normal component (3) the expectation is given by the point estimate of θ_i according to (6).

For the i-th exponential component the expectation can be obtained as

$$\frac{1}{(\hat{a}_{l;t-1})_i} + (\hat{b}_{l;t-1})_i \tag{24}$$

using the point estimates (14)–(15) for the l-th entries of the vector y_t.

For the i-th uniform component the expectation is a simple average of the component bounds L_i and R_i for corresponding entries of the vector y_t.

Using Δ_t from (23) the following rapidly decreasing functions (as combinations of the mentioned above) are considered:

1. The polynomial of the form

$$1/\Delta_t^5. \tag{25}$$

2. The normal approximation of other distributions, optimal in the sense of the Kullback-Leibler divergence, see, e.g., (Kárný et al., 2006), which is the pdf (3) with the substituted expectation of the corresponding type of the components instead of θ_i. The covariance matrices can be either used from individual distributions or chosen as the diagonal ones in case only the expectations should be estimated.

3. The polynomial in the form

$$\exp\{-(2\Delta_t)^5\}. \tag{26}$$

4. The Student distribution based relation

$$\left(1+\frac{\Delta_t^2}{t}\right)^{-\frac{t}{2}}. \tag{27}$$

3.4 Algorithm

The mixture estimation algorithm tailored to the above proximity definitions is summarized as follows.

Initialization (for t=1)

- Set the number of components m_c.
- Set the initial statistics for the corresponding type of components, i.e., either for (3) or (8), or (16), see Section 2.
- Set the initial Dirichlet statistics of the pointer model (2) according to (Kárný et al., 2006) as the $(m_c \times m_c)$-dimensional matrix denoted by γ_0.
- Using the initial statistics, compute the initial point estimates of parameters of all components using either (6) or (14)–(15), or (19).
- Compute the initial point estimate of α by normalizing the Dirichlet statistics γ_0.
- Set the initial m_c-dimensional weighting vector w_0.

On-line part (for t=2,...)

1. Measure the new data item y_t.
2. Compute the expectations $E[y_t]$ of all components using the corresponding point estimates.
3. For all components, obtain (23) by substituting y_t and the expectations $E[y_t]$.
4. For all components, substitute the obtained Δ_t into the chosen type of the proximity, i.e., either (25), (3), (26) or (27).
5. In each case the result is the m_c-dimensional vector of proximities, which has one entry for each component. These entries are proportional to the inverse distance of the data item to the individual components. The higher the number is, the closer the data item is to the component.
6. Multiply entry-wise the resulted vector from the previous step, the previous weighting vector w_{t-1} and the previous point estimate of α. The result of the multiplication is the matrix of weights denoted by $W_{i,j;t}$ joint for c_t and c_{t-1}.
7. Perform the summation of the normalized matrix from the previous step over rows and obtain the updated weighting vector w_t. The maximal entry of the weighting vector gives the point estimate of the pointer c_t, i.e., the currently active component.
8. For all components, update the statistics with the help of multiplying the data-dependent increment in each statistics by the corresponding entry from the actualized weighting vector w_t, see (Kárný et al., 1998; Kárný et al., 2006; Peterka, 1981; Nagy et al., 2011; Suzdaleva et al., 2015; Nagy et al., 2016).
9. Update the statistics of the pointer model as

$$\gamma_{i|j;t} = \gamma_{i|j;t-1} + W_{i,j;t} \tag{28}$$

   based on (Kárný et al., 1998; Nagy et al., 2011).
10. Recompute the point estimates of all parameters and use them for Step 1 of the on-line part.

4 EXPERIMENTS

This section provides the experimental comparison of evolution of the proximities (25), (3), (26) and (27) for all three discussed types of components according to the above algorithm. The simulated two-dimensional data were used, which is quite sufficient for adequate judging the proximity evolution. Three components of each type were simulated. A series of experiments with the various number of data was performed. Here typical results are demonstrated.

The quality of classification as the detection of the active component at each time instant according to the maximal weights was compared. Firstly, easily distinguishable components as a simple case were tested. Results for them are provided in Table 1. Then, a case with two components closer to each other and the third at a larger distance was tested, see Table 2. Finally, the most complicated case with overlapped components was tested, see results in Table 3.

Table 1: Average number of incorrect pointer point estimates (%) for easily distinguishable components.

Proximities / components	(3)	(8)	(16)
$1/\Delta_t^5$	1.2	0	3.4
Approximation (3)	0	65.5	0.4
$\exp\{-(2\Delta_t)^5\}$	0	0	0.4
$\left(1+\frac{\Delta_t^2}{t}\right)^{-\frac{t}{2}}$	26.4	0	0.4

Table 2: Average number of incorrect pointer point estimates (%) for variously located components.

Proximities / components	(3)	(8)	(16)
$1/\Delta_t^5$	7.4	6.6	15.8
Approximation (3)	27.2	67.4	11.2
$\exp\{-(2\Delta_t)^5\}$	15.8	3.6	11.2
$\left(1+\frac{\Delta_t^2}{t}\right)^{-\frac{t}{2}}$	21.4	4.6	12

Table 3: Average number of incorrect pointer point estimates (%) for overlapped components.

Proximities / components	(3)	(8)	(16)
$1/\Delta_t^5$	39.4	39.2	19.2
Approximation (3)	37.8	67.4	15.4
$\exp\{-(2\Delta_t)^5\}$	39.6	33.4	15.4
$\left(1+\frac{\Delta_t^2}{t}\right)^{-\frac{t}{2}}$	39.2	37.2	16

These tables briefly (due to a lack of space) present results of comparison. Nevertheless, to summarize results obtained for variously noised data (that determines distances among components) the following trend is observed.

For easily distinguishable components all the proximity definitions are similarly successful with the insignificant difference, excepting the Student distribution based function (27) for normal components, and the normal approximation of the exponential distributions.

Variously located components, where two components are closer to each other and the third is at a larger distance, is a more challenging task. In this case the trend in success of classification of exponential and uniform components is preserved. However,

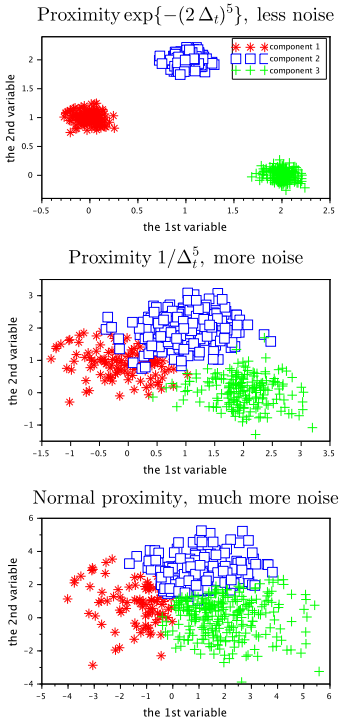

Figure 1: The most successful classification results with normal components from Tables 1, 2 and 3.

significant worsening can be seen for the proximity (3) for normal components. The proximity $1/\Delta_t^5$ gives the best results.

For strongly overlapped components (which is the most complicated and also realistic case) the situation changes. For normal components all results are similar, but now the proximity definition (3) is the most successful.

It seems that a decision about a choice of the proximity definition is suitable to make in accordance to the noise covariance matrix estimation. For the less noised data the distance-based proximities $1/\Delta_t^5$ and $\exp\{-(2\Delta_t)^5\}$ can be used. The normal approximation (3) is not suitable for exponential components, but it can be used for uniformed components. For more noised data the normal approximation (3) can be used both for normal and uniform components.

The graphical presentation of the classification results is given for normal components in Figure 1, where (due to a lack of space) the most successful combinations of the proximity definition and a distance among components are shown. In all figures three detected components are sharply visible.

Selected results for exponential components can be seen in Figure 2, where the evolution of the component centers during the on-line estimation is plotted for the proximity (26) (top) and (27) (bottom). The

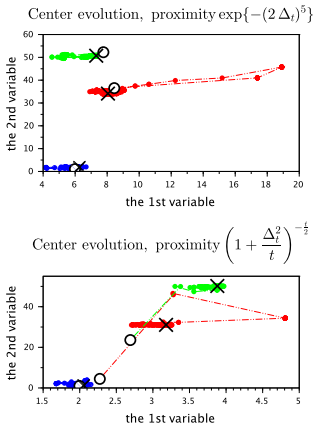

Figure 2: Evolution of exponential component centers during the estimation with the proximity $\exp\{-(2\,\Delta_t)^5\}$ (top) and the Student based proximity (bottom).

initial centers are denoted by the circle, the final – by 'x'. The figures show that although the success of the estimation according to Table 1 were the same for the proximity (26) and (27), the stabilization of component centers took a different time. In the top figure the centers were correctly located just in the beginning of estimation, while in the bottom figure the search for correct locations was performed a bit longer.

For uniform components the classification results were close to each other for all tested types of location of components. It is worth noting that the normal approximation (3) and the distance-based proximity (26) were similarly successful. Results with one of them for variously located components are demonstrated in Figure 3 (top).

The proximity evolution during the chosen time interval is shown in Figure 3 (bottom). The distance-based proximity $1/\Delta_t^5$ demonstrates the strongest sharpness. The normal approximation has a smoother course. The third proximity from Table 1 is influenced by the constant, which allows to move its course as necessary. The Student distribution based proximity is close to flat.

5 CONCLUSIONS

The aim of the described recursive classification is the analysis of real data in real time, which suggests that the detected clusters can be of various shapes. Thus solutions with mixtures of different components are highly desired.

Using the point estimates directly in the exponential and uniform models as the proximity gives unsuccessful results during the mixture estimation, which can be explained by asymmetric distributions. One

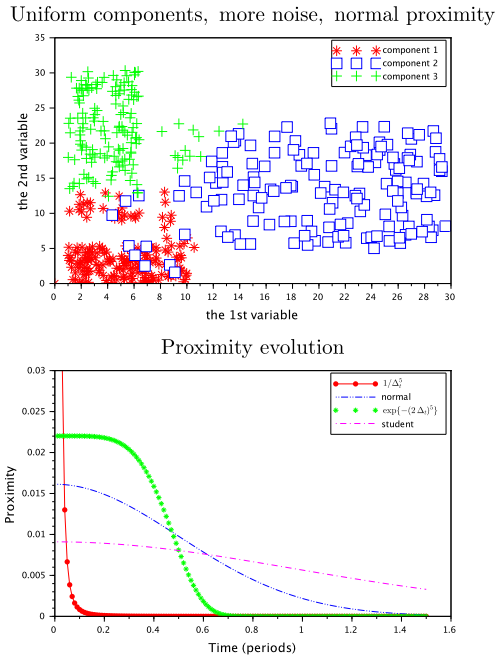

Figure 3: Detected uniform components with the proximity (3) (top) and the proximity evolution for a selected interval (bottom).

of the attempts was the use of the normal approximation optimal from the point of view of the Kullback-Leibler divergence. The important property of the proximity is its rapidly decreasing course. This is satisfied by the normal approximation, but other functions can be also relevant for the considered task. The results of the comparison show that for different types of components a choice of the proximity is not unambiguous and influences the classification results. Thus, the significance of the proximity choice is confirmed.

However, there is still a series of open problems in this area, including (i) the solution for dependent entries of the data vector for exponential and uniform distributions, (ii) extension of the recursive mixture estimation theory for other distributions describing non-Gaussian data, and (iii) classification with a mixture of different distributions. This will be the plan of the future work within the present project.

ACKNOWLEDGEMENTS

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

REFERENCES

Yu, J. (2012). A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes, *Chemical Engineering Science*, vol. 68, 1, p. 506–519.

Yu, J. (2012). A particle filter driven dynamic Gaussian mixture model approach for complex process monitoring and fault diagnosis, *Journal of Process Control*, vol. 22, 4 , p. 778–788.

Yu, Jianbo. (2011). Fault detection using principal components-based Gaussian mixture model for semiconductor manufacturing processes, *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, 3, p. 432–444.

Larose, D. T. (2005). *Discovering Knowledge in Data. An Introduction to Data Mining*. Willey.

Han, J., Kamber, M., Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd ed. (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

Zaki, M.J., Meira Jr.. W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Calders, T., Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*. 21(2), p. 277–292.

Zhang, G. P. (2000). Neural Networks for Classification: A Survey. In: *IEEE Transactions on System, Man, and Cybernetics – Part C: Applications and Reviews*. 30(4), November, p. 451–462.

Ishibuchi, H., Nakashima, T., Nii, M. (2000). Fuzzy If-Then Rules for Pattern Classification. In: *The Springer International Series in Engineering and Computer Science*. 553, p. 267–295.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In: *Grouping Multidimensional Data.* Eds.: J. Kogan, C. Nicholas, M. Teboulle. Springer Berlin Heidelberg, p.25–71.

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

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases. In: *Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, Portland, OR, August, p. 226–231.

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

Zeng, H., Cheung, Y. (2014). Learning a mixture model for clustering with the completed likelihood minimum message length criterion. *Pattern Recognition*. 47(5), p. 2011–2030.

Ng, S.K., McLachlan, G.J. (2014). Mixture models for clustering multilevel growth trajectories. *Computational Statistics & Data Analysis.* 71(0), p. 43–51.

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.

Boldea, O., Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model, *Journal of The American Statistical Association*, vol. 104, 488, p. 1539–1549.

Wang, H.X., Luo, B., Zhang, Q. B., Wei, S. (2004). Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm, *Pattern Recognition Letters*, vol. 25, 16, p. 1799–1809.

McGrory, C. A., Titterington, D. M. (2009). Variational Bayesian analysis for hidden Markov models, *Australian & New Zealand Journal of Statistics*, vol. 51, p. 227–244.

Šmídl, V., Quinn, A. (2006). *The Variational Bayes Method in Signal Processing*, Springer-Verlag Berlin Heidelberg.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer-Verlag New York.

Doucet, A., Andrieu, C. (2001). Iterative algorithms for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing*, vol. 49, 6, p. 1216–1227.

Chen, R., Liu, J.S. (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, p. 493–508.

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.

Yang, L., Zhou, H., Yuan, H. (2013). Bayes Estimation of Parameter of Exponential Distribution under a Bounded Loss Function. *Research Journal of Mathematics and Statistics*, vol.5, 4, p.28–31.

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*, Sofia, Bulgaria, September 4–6, accepted.