2016 IEEE 8th International Conference on Intelligent Systems

# Expert-based Initialization of Recursive Mixture Estimation

Evgenia Suzdaleva*, Ivan Nagy†*, Tereza Mlynářová*
*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
Email: suzdalev@utia.cas.cz
†Faculty of Transportation Sciences, Czech Technical University
Na Florenci 25, 11000 Prague, Czech Republic
Email: nagy@utia.cas.cz

***Abstract*—Initialization is an extremely important part of the mixture estimation process. There exists a series of initialization approaches in the literature concerning the mixture initialization. However, the majority of them is directed at initialization of the expectation-maximization algorithm widely used in this area. This paper focuses on the initialization of the mixture estimation with normal components based on the recursive statistics update of involved distributions, where the mentioned methods are not suitable. Its key part is the choice of the initial statistics. The paper describes several relatively simple initialization techniques primarily based on processing the prior data. The experimental part of the paper represents results of validation on real data.**

***Keywords*—mixture initialization, recursive estimation, component centers**

## I. INTRODUCTION

The initialization is one of the key problems of the mixture estimation. Mixture models are often used for description of multi-modal systems, whose behavior can switch among different working modes. Such modeling is demanded in a variety of application areas, including, e.g., fault detection (fault or non-fault mode), car diagnostics (eco-driving or sport mode, etc.), traffic flow control (the level of service), big data issues, etc., see, for instance, [1], [2], [3].

The mixture model consists of several components that describe the individual working modes of the observed system and of their switching model. The last is considered as the random Markov process called the pointer [4], [5], and its value at the corresponding time instant indicates the currently active component (i.e., the working mode). In reality, parameters of neither the components nor the pointer model are available. Thus the mixture estimation problem consists, in general, in estimation of the component and the pointer model parameters, and also in the pointer value estimation.

The mixture estimation approaches found in the literature are mainly based on (i) the iterative expectation-maximization (EM) algorithm [6], see, e.g., [7], [8]; (ii) the approximative Variational Bayes approach [9], [10]; (iii) sampling Markov Chain Monte Carlo techniques, e.g., [11], [12], [13]. Closely related tasks are also discussed in [14], [15].

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

A different non-numerical approach is given by the recursive Bayesian estimation theory for static mixtures [4], [5], individual normal components [16] and dynamic mixtures [17], which, unlike the above mentioned sources, represent on-line data-based estimation algorithms avoiding numerical iterative computations. The present research project supports their philosophy in developing the mixture estimation algorithms.

The mixture estimation algorithm should be initialized before starting. In the considered context the initialization primarily lies in specifying (i) distributions of components, (ii) the number of components, and (iii) the prior probability density functions (pdfs) describing parameters of components and of the pointer model. The present paper is limited by mixtures of normal components.

A series of papers was found in the area of the mixture initialization. For instance, the paper [18] proposes the initialization of the EM algorithm via a strategy defining mean vectors by choosing points with higher concentrations of neighbors. It uses a truncated normal distribution for the preliminary estimation of covariance matrices.

Another paper [19] describes a new method for random initialization of the EM algorithm based on selecting the feature vector from a set of candidate vectors, located farthest from already initialized components. The Mahalanobis distance is used. The paper [20] is devoted to simple and fast approaches of the initialization of the EM algorithm based on the well-known clustering algorithms. The paper [21] proposes the EM initialization method by partition of the training set to be modeled individually by single experts and the subsequent initialization of models on a partition subset. The paper [22] initializes a mixture via the EM algorithm using a product kernel estimate of pdfs and the gradient method for local extrema finding.

It is seen that the majority of studies is primarily oriented at application of the EM algorithm. Under the adopted theory [5], [4], [16], [17] not using the EM algorithm, the initialization

focuses on the number of components and the initial statistics of the Gauss-inverse-Wishart parameter pdfs. In this field the paper [23] is found, which (applied to the presented subproblem) leads to weighting the initial statistics of the parameter pdfs.

The present paper considers the initialization primarily based on detecting the initial centers of components via the visualization analysis of the prior or expert knowledge. In the case of static normal components this expert-based procedure is rather effective. For dynamic mixture components the task is more complicated. The paper considers several ways of initialization of dynamic components: (i) fixation of covariance matrices; (ii) imitation of the static case; (iii) repeated use of the data sample, see, e.g., [4]; and (iv) weighting the initial statistics [23], and validates them experimentally on real data. The paper demonstrates that a relatively small amount of prior data used for the mixture initialization contributes to a faster stabilization of parameter estimates during the on-line estimation.

The paper is organized in the following way. Section II introduces models. Section III gives necessary basic facts about the mixture estimation algorithm and specifies the initialization problem. Section IV describes the mentioned initialization approaches. Section V provides results of experiments. Conclusions and open problems are given in Section VI.

## II. 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 each 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 [5].

The observed system is supposed to be described by a mixture model, which (in this paper) consists of $m_c$ components. The components can be represented either by

$$\text{the static pdf} \quad f\left(y_t|\Theta, c_t = i\right), \ \forall i \in \{1, 2, \ldots, m_c\}, \quad (1)$$

$$\text{or by the dynamic pdf} \quad f\left(y_t|\psi_{t-1}, \Theta, c_t = i\right), \quad (2)$$

where $\Theta$ is a collection of parameters of all components, and $\Theta \equiv \{\Theta_i\}_{i=1}^{m_c}$, where $\Theta_i$ includes parameters of the $i$-th component in the sense that $f\left(y_t|\Theta, c_t = i\right) = f\left(y_t|\Theta_i\right)$ for $c_t = i$, and $\psi_{t-1} = [y_{t-1}, y_{t-2}, \ldots, y_{t-n}]'$ is the regression vector with the memory length $n$.

This paper focuses on using the pdfs (1) or (2) with the normally distributed white noise. In this case the pdfs are specified as follows.

### A. Static Components

The pdf (1) has the form $\forall i \in \{1, 2, \ldots, m_c\}$

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

where $N$ denotes a dimension of the vector $y_t$; $\theta_i$ 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 (i.e., in the case of the diagonal $r_i$ the component is round-shaped), and $\Theta_i \equiv \{\theta_i, r_i\}$.

### B. Dynamic Components

The pdf (2) is specified as

$$(2\pi)^{-N/2}|r_i|^{-1/2} \exp\left\{-\frac{1}{2}[y_t - \theta_i\psi_{t-1}]' r_i^{-1}[y_t - \theta_i\psi_{t-1}]\right\}, \quad (4)$$

where unlike (1) the parameter $\theta_i$ is a collection of regression coefficients of the $i$-th component, whose number corresponds to the memory length $n$ used for the regression vector $\psi_{t-1}$. A rest of notations are identical to the previous case.

### C. Dynamic Pointer Model

Switching the active components, either (1) or (2), is described by the dynamic model

$$f\left(c_t = i|c_{t-1} = j, \alpha\right), \ i, j \in \{1, 2, \ldots, m_c\}, \quad (5)$$

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 parameter $\alpha$ 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 $c_{t-1} = j$.

## III. RECURSIVE MIXTURE ESTIMATION

Formulation of the initialization problem requires a preliminary outline of the recursive approach to the Bayesian mixture estimation. The algorithm to be effectively initialized is based on the paper [5], which proposes the solution for normal mixtures with the static pointer model, and on [17] considered the problem for the dynamic pointer model. In the context of the introduced mixture of components (1) or (2) and of the pointer model (5), the estimation problem concerns the unknown parameters $\Theta$ and $\alpha$ and the pointer value $c_t$. Derivations are based on construction of the joint pdf of all variables to be estimated and application of the Bayes rule and of the chain rule, see e.g., [16]. Here they are outlined briefly to present the necessary theoretical background for static components (1) with a subsequent explanation of changes in the case of using (2).

Assuming that $\Theta$ and $\alpha$, and $y_t$ and $\alpha$, and $c_t$ and $\Theta$ are mutually independent, and denoting the data collection $y(t) = \{y_0, y_1, \ldots, y_t\}$, where $y_0$ stands for prior data,

the joint pdf of all variables to be estimated has the form $\forall i, j \in \{1, 2, \ldots, m_c\}$

$$\underbrace{f(\Theta, c_t = i, c_{t-1} = j, \alpha | y(t))}_{joint\ posterior\ pdf} \tag{6}$$

$$\propto \underbrace{f(y_t, \Theta, c_t = i, c_{t-1} = j, \alpha | y(t-1))}_{via\ chain\ rule\ and\ Bayes\ rule}$$

$$= \underbrace{f\left(y_t | \Theta, c_t = i\right)}_{(1)\ or\ (2)} \underbrace{f(\Theta | y(t-1))}_{prior\ pdf\ of\ \Theta}$$

$$\times \underbrace{f\left(c_t = i | c_{t-1} = j, \alpha\right)}_{(5)} \underbrace{f(\alpha | y(t-1))}_{prior\ pdf\ of\ \alpha} \underbrace{f(c_{t-1} = j | y(t-1))}_{prior\ pointer\ pdf}.$$

Recursive formulas for estimation of $c_t$, $\Theta$ and $\alpha$ via (6) are obtained using the marginalization of (6) firstly over the parameters $\Theta$ and $\alpha$. It results in the posterior pdf $f(c_t = i, c_{t-1} = j | y(t))$, which is joint for both $c_t$ and $c_{t-1}$. Further the resulted joint pdf should be again marginalized over the values of $c_{t-1}$ for obtaining the posterior pdf $f(c_t = i | y(t))$ of the current pointer.

### A. Component Parameters

The integral of (6) over $\Theta$ is evaluated by substituting the point estimates of $\theta_i$ and $r_i$ available from the previous time instant $t-1$ and the currently measured $y_t$ into the corresponding $i$-th normal component, either (1) or (2). The mentioned point estimates of parameters of the $i$-th component are computed based on using the conjugate prior Gauss-inverse-Wishart pdf with the recomputable (initially chosen) statistics $(V_{t-1})_i$ and $k_{i;t-1}$ in the Bayes rule, which according to [16], [5] gives the algebraic recursion for static components

$$(V_t)_i = (V_{t-1})_i + w_{i;t} \begin{bmatrix} y_t \\ 1 \end{bmatrix} [y_t,\, 1]\,, \tag{7}$$

for dynamic components

$$(V_t)_i = (V_{t-1})_i + w_{i;t} \begin{bmatrix} y_t \\ \psi_{t-1} \end{bmatrix} [y_t,\, \psi_{t-1}]\,, \tag{8}$$

and valid for both of them

$$\kappa_{i;t} = \kappa_{i;t-1} + w_{i;t}, \tag{9}$$

where $w_{i;t}$ will be explained later. The needed point estimates are computed at time $t$ for each component as follows [16]:

$$(\hat{\theta}_t)_i = V_1^{-1} V_y, \quad (\hat{r}_t)_i = \frac{V_{yy} - V_y^{'} V_1^{-1} V_y}{\kappa_{i;t}}, \tag{10}$$

where $(V_t)_i$ is partitioned (for simplicity with the omitted subscript $i$)

$$(V_t)_i = \begin{bmatrix} V_{yy} & V_y^{'} \\ V_y & V_1 \end{bmatrix}, \tag{11}$$

so that in the static case $V_{yy}$ is the square matrix of the dimension $N$ of the vector $y_t$, $V_y'$ is $N$-dimensional column vector and $V_1$ is scalar. For dynamic components (2), the partition changes according to the memory length $n$ used in the regression vector $\psi_{t-1}$, i.e., $V_y$ and $V_1$ become matrices of appropriate dimensions. The substitution of (10) and $y_t$ into the corresponding $i$-th normal pdf provides the proximity of each component to the current data item.

### B. Pointer Parameters

Similarly, the integral of (6) over $\alpha$ provides the computation of its point estimate using the previous-time statistics denoted by $\vartheta_{t-1}$ of the conjugate prior Dirichlet pdf according to [4]. Here the mentioned statistics is the square $m_c$-dimensional matrix, whose entries for $c_t = i$ and $c_{t-1} = j$ are recursively computed in the following way:

$$\vartheta_{i|j;t} = \vartheta_{i|j;t-1} + W_{i,j;t}, \tag{12}$$

where $W_{i,j;t}$ will be explained a bit later, and which was introduced by [17] with the approximation based on the Kerridge inaccuracy [24]. However, here, for simplicity, it is updated similarly to [5], but modified for the dynamic pointer model. The point estimate of $\alpha$ is then obtained by simple normalizing the updated statistics

$$\hat{\alpha}_{i|j;t} = \frac{\vartheta_{i|j;t}}{\sum_{k=1}^{m_c} \vartheta_{k|j;t}}. \tag{13}$$

### C. Component Weights

Here the above denotations $w_{i;t}$ and $W_{i,j;t}$ are explained. After the described marginalization the posterior pdf $f(c_t = i, c_{t-1} = j | y(t))$ is obtained by entry-wise multiplying the proximity obtained from each component, the previous-time point estimate of $\alpha$ (13) and the prior pointer pdf $(c_{t-1} = j | y(t-1))$. The last is the weight of the components at the previous time instant, and it is denoted by $w_{j;t-1}$ and expresses the (initially chosen and then actualized) probability of the activity of the $j$-th component at time $t-1$.

For all $i, j \in \{1, 2, \ldots, m_c\}$, the posterior pdfs $f(c_t = i, c_{t-1} = j | y(t))$ are entries denotes by $W_{i,j;t}$ of the square $m_c$-dimensional matrix, which is normalized and summed up over rows to obtain the posterior pdf $f(c_t = i | y(t))$. The last provides the updated weight $w_{i;t}$ of each $i$-th component at time $t$. The maximal weight $w_{i;t}$ defines the currently active component, i.e., the point estimate of the pointer $c_t$ at time $t$.

### D. Initialization Problem Specification

The outlined relations are summarized as the following algorithm steps performed on-line for $t = 2, \ldots$:

1) Measure the new data $y_t$.
2) Obtain proximities of all components, using the previous-time point estimates (10).
3) Multiply entry-wise the proximities, the prior weighting vector $w_{t-1}$ and the previous-time point estimate $\hat{\alpha}_{t-1}$.
4) The result of this entry-wise multiplication is the matrix with entries $W_{i,j;t}$. Normalize this matrix.
5) Perform the summation of the normalized matrix over rows and obtain the updated vector $w_t$ with entries $w_{i;t}$.
6) Update all statistics, using $w_{i;t}$ and $W_{i,j;t}$ according to (7) or (8), (9), and (12).

7) Recompute the point estimates of all parameters according to (10) and (13) and go to Step 1.

Thus, the initialization of this on-line part of the algorithm lies in setting at time $t = 1$:

- the number of components $m_c$,
- the initial statistics of all components $(V_0)_i$, $\kappa_{i;0}$ and the pointer statistics $\vartheta_0$ (the initial estimates in Steps 2 and 3 are computed from them),
- the initial $m_c$-dimensional weighting vector $w_0$,

where $m_c$ and $(V_0)_i$ are the key ones and they will be the focus of the subsequent sections. The rest of statistics can be initialized either uniformly or randomly in combination with their updating by prior data.

## IV. EXPERT-BASED MIXTURE INITIALIZATION

The proposed initialization is based on convincing that in the beginning of the mixture estimation (as well as generally description of the multi-modal system) in a specific domain some type of prior or expert knowledge is always available. Such a kind of the knowledge can be in the form of specially previously measured data, realistic simulations (e.g., from Aimsun (www.aimsun.com) in the traffic flow control area) or, at least, the expert information about the expected number of components (disease symptoms in medicine, types of failures in car diagnostics, success in elections, etc.).

Anyway the start of the estimation is always critical due to a risk of dominance of a single active component resulted from the temporary non-activeness of others as well as noisy data. This can lead to joining other components and finally the failure of the estimation. To avoid the mentioned dominance the following expert-based procedures can be performed:

- fixing the covariance matrices of components as diagonal ones with entries $0.1$ and running their estimation later, which is very simple and effective way;
- detection of the initial component centers by the visual analysis;
- repeated use of the prior data sample inspired by [4], [23].
- suppressing the influence of the first measured data on the estimation to support the initial estimates obtained from the initial statistics to produce proper weights of components based on [23].

These ways of processing the prior data to extract the information necessary for a successful initialization is described below. Thus, in this section the time instant $t$ corresponds to prior data items. The implementation is prepared in the open source programming environment Scilab (www.scilab.org).

To determine the area of the interest in the data-parameter space it is suitable to work with the normalized data with zero expectations and the unit covariance matrices. This is reached by extracting the mean value from each prior data item and division by the standard deviation. However, it is not a necessary condition.

### A. Static Component Initialization

For the initialization of static components (1) it is extremely important to detect the initial centers of clusters in the data space. This task covers both the determination of the number of components and of the initial statistics. Covariance matrices for the normalized data could be used as diagonal ones with entries $0.1$.

For this aim the prior data sample is processed as follows. Individual entries of the multidimensional vector $y_t$ are visualized by pairs against each other. The analysis of the visualization gives a possibility to distinguish the number of plotted components and get their centers. Here for demonstration, the real data sample measured on a driven vehicle is taken, where the vector $y_t$ contains the following entries: (i) $y_{1;t}$ is the instantaneous fuel consumption $[\mu l]$, (ii) $y_{2;t}$ is the vehicle speed [km/h], (iii) $y_{3;t}$ is pressing the gas pedal [%], (iv) $y_{4;t}$ is the engine speed [rpm]. The sampling period is 1 second. The number of prior data is 400.

Two-dimensional clusters of each variable are shown in Figure 1. The visualization represents the upper triangular matrix of figures, where each row corresponds to the entry of the vector $y_t$ from $y_{1;t}$ to $y_{4;t}$ plotted firstly against itself and then against the rest of entries. The normalized data with zero expectations and unit variances are used, which means that values on axes do not express real ranges of data items. Individual figures are denoted by numbers $l, k \in 1, \ldots, N$ corresponding to the entries indices. Under assumption that the processed data are of a multi-modal character, clusters are clearly visible. Here three clusters are seen, thus $m_c = 3$. For detection of initial centers of components, figures 1-2, 2-3 and 3-4 located above the diagonal are of the main interest.

Figure 1-2 exhibits three clusters at positions $[0, -0.5]$, $[0.5, 1]$ and $[1, -1]$, which indicate three positions of clusters of the variable $y_{2;t}$: i.e., -0.5, 1 and -1. These values are explored in the second figure 2-3 on the x axis, where the variable $y_{2;t}$ is shown. Figure 2-3 gives the coordinates $[-0.5, 0.5]$, $[-1, -1.2]$ and $[1, 0.5]$, which provide positions 0.5 a -1.2 for the entry $y_{3;t}$. Using them in figure 3-4 the centers of components between entries $y_{3;t}$ and $y_{4;t}$ are detected as $[0.5, -0.3]$, $[0.5, 0.7]$ and $[-1.2, 0]$.

Based on this visual analysis, positions of the four-dimensional initial centers denoted by $s_i$ $\forall i\{1, \ldots, m_c\}$ are summarized in Table I.

TABLE I
INITIAL CENTERS OF STATIC COMPONENTS

| Data entry | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| $y_{1;t}$ | 0 | 0.5 | 1 |
| $y_{2;t}$ | -0.5 | 1 | -1 |
| $y_{3;t}$ | 0.5 | -1.2 | 0.5 |
| $y_{4;t}$ | -0.3 | 0.7 | 0 |

The $i$-th initial center is substituted into the initial statistics

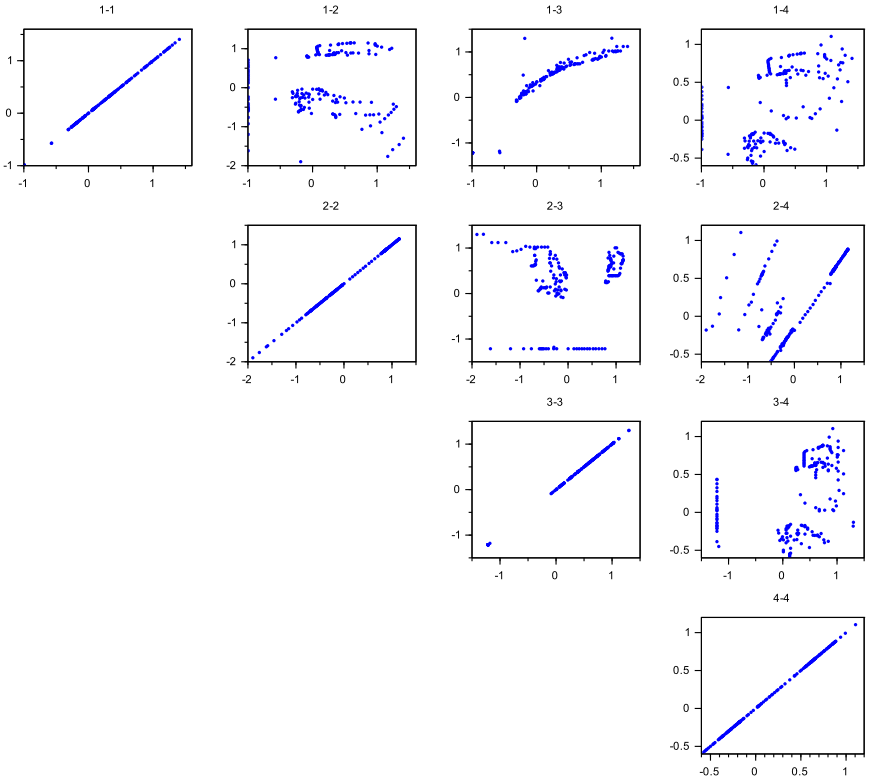


Fig. 1. Visualization of pairs of the normalized data vector against each other. Notice visible clusters plotted in figures denoted by 1-2, 2-3 and 3-4.

$(V_0)_i$ of the $i$-component as follows:

$$(V_0)_i = \begin{bmatrix} 1 & 0 & 0 & 0 & (s_1)_i \\ 0 & 1 & 0 & 0 & (s_2)_i \\ 0 & 0 & 1 & 0 & (s_3)_i \\ 0 & 0 & 0 & 1 & (s_4)_i \\ (s_1)_i & (s_2)_i & (s_3)_i & (s_4)_i & 1 \end{bmatrix} \quad (14)$$

where $(s_l)_i$ $\forall l\{1,\ldots,N\}$ is the $l$-th entry of the vector $s_i$.

Thus, the expert-based initialization procedure for static components includes the steps: (i) normalize data (optionally); (ii) plot all data entries against each other, (iii) find subsequently positions of clusters in corresponding figures (here above the diagonal). The constructed initial statistics is used in the on-line part of the estimation algorithm. Validation of the approach is discussed in Section V.

### B. Dynamic Component Initialization

A character of dynamic components (2) requires both to support the dynamics of models and to avoid a preliminary dominance of any of components due to noisy data. The following initialization procedures can be considered (notice that they can be also combined).

*1) Static Case Imitation:* The first one is to imitate the static case described above and to detect both the number of components and their initial centers, using the visual analysis of prior data. The initial statistics $(V_0)_i$ in this case is constructed with the help of substituting a matrix of the form (14) with the detected initial centers instead of its part $V_1$ in (11). The rest of corresponding matrix entries are zero values. Such the initialization can be in many cases efficient, i.e., for data with the rather slow dynamics.

*2) Initial Centers with Support of Dynamics:* Another option is to combine the above approach with diagonal matrices, entries of which represent the chosen initial model dynamics. In this case (using the diagonal noise covariance matrix too) the component is decomposed into independent equations (in dependence on a dimension of the vector $y_t$). This allows to use the stabilized positions of centers for the initial statistics.

Construction of the initial statistics $(V_0)_i$ is based on the fact that the initial centers detected for static components are their constant expectations. Thus for the dynamic model (here for simplicity for the first order component with $n = 1$) the constant in (2), or precisely (4), is determined from

$$y_{l;t} - (a_{l|l})_i y_{l;t-1} - (s_l)_i, \quad (15)$$

where $\{a_i, s_i\} \in \theta_i$ of the $i$-th component, and $(a_{l|l})_i$ is the diagonal entry of the matrix of regression coefficients $a_i$ with $l \in \{1,\ldots,N\}$, and $(s_l)_i$ is the entry of the vector $s_i$. Then the diagonal entry $(a_{l|l})_i$ expressing the dynamics (a small value about 0.1 brings more dynamics, and a value approaching 1 corresponds to slow dynamics) can be used for constructing the initial statistics. For the previous example,

the initial statistics of the $i$-th strongly dynamic component is constructed by substituting

$$V_y = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \\ (s_1)_i & (s_2)_i & (s_3)_i & (s_4)_i \end{bmatrix} \tag{16}$$

into (11). It also defines $V_y'$, and the rest of submatrices are the unit matrices.

*3) Repeated Use of the Data Sample:* Another expert-based procedure, which is rather helpful in the initialization mostly under condition of the lack of data is the repeated use of the available prior data sample [4], [23]. Firstly the estimation starts according to the algorithm from Section III-D with small diagonal initial statistics $(V_0)_i$. The actualized statistics after the course of the estimation with the whole sample of prior data are used as the new initial one, and the estimation algorithm starts again. The resulted updated statistics are used as initial for the on-line estimation.

This way of initialization can be also combined with weighting the initial statistics to suppress the influence of data in the beginning of the algorithm running, which is described below.

*4) Weighting the Initial Statistics:* This initialization approach is primarily based on [23], which in the considered context takes the following form.

The prior or expert given data are firstly substituted in the extended regression vectors $[y_t,\ \psi_{t-1}]'$ used then in the statistics update (8). The amount of the used extended regression vectors should correspond to the number of parameters (regression coefficients) to be estimated. The statistics $\kappa_{i;0}$ expresses the number of the used data.

The Bayesian estimation is strengthened with gradually measuring new data, which means that a weight of the new data item is inversely proportional to the statistics, and therefore the possible disturbance in data takes the inversely proportional effect on the estimation. Thus the same prior regression vectors multiplied by the chosen weight are used again for the initial statistics as follows:

$$(V_0)_i = \mu(V_0)_i,\ \ \kappa_{i;0} = \mu, \tag{17}$$

where $\mu$ expresses the number of the used prior data items, which means that the first newly measured data item takes the effect $\frac{1}{\mu}$ on the estimation.

Improvements brought by the mentioned initialization methods appear primarily in the speed of finding the stabilized estimates of regression coefficients during the on-line mixture estimation. Validation of the enumerated approaches is presented below.

## V. EXPERIMENTS

The initialization of the mixture estimation algorithm can be validated in accordance with the following criteria.

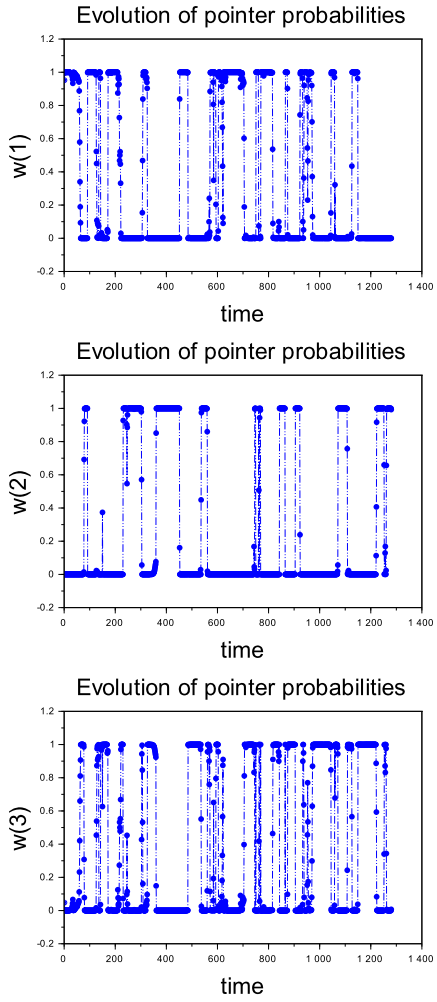


Fig. 2. The evolution of the activity of three components. Notice that values of the weights are approaching 0 or 1.

### A. Weight Evolution

The first one concerns with the initialized number of components, which is identical both for the static and dynamic components. It is verified by the evolution of the component weights during the on-line part of the estimation algorithm using 6400 data. For better visibility, fragments with 1200 data items are shown. The evolution should demonstrate a reasonable way of switching the components. In that case it confirms that the model is correctly established. For the prior data used in Section IV-A the evolution of the corresponding entries of the weighting vector $w_t$ of three detected components is shown in Figure 2.

It can be seen that (i) the components switch in a reasonable way, (ii) the plotted values of probabilities are mostly approaching 1 or 0, which means the unambiguous decision for the currently active component.

### B. Parameter Estimate Evolution

The evolution of the component centers for static components (1) and of the estimates of regression coefficients for dynamic ones (2) is a sufficient indicator of the successful initialization.

Stabilization of positions of component centers in the data space after their initial search indicates that the estimation is correct. Otherwise, if some resulting centers are identical or very close one to another, this mostly signalizes that too many components are chosen, and their number should be reduced.

The evolution of the component centers can be seen in Figure 3 in the parameter space for the normalized entries $y_{1;t}$ and $y_{2;t}$, and $y_{3;t}$ and $y_{4;t}$ plotted against each other.

The evolution of the estimation of regression coefficients of two of the components (to save space) is shown in Figure 4. The stabilization of the estimation can be seen, where after the initial search the steady-state is reached.

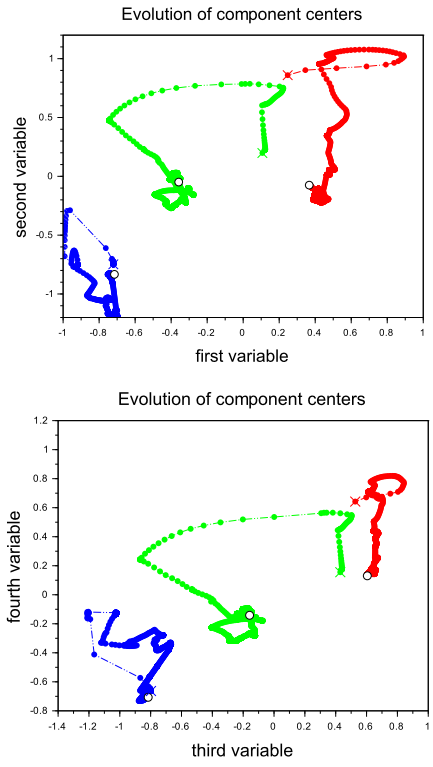


Fig. 3. Evolution of three component centers. The start position is denoted by 'x', and the end of the search is marked by 'o'. The density of points corresponds to the speed of movement.

### C. Validation via Data Prediction

The graphically represented comparison of the predicted data items obtained from components with the substituted estimates and the real data is shown in Figure 5. For the lack of space, only two normalized entries of the data vector $y_t$ are shown. Graphs demonstrate the coincidence between predictions and real data items.

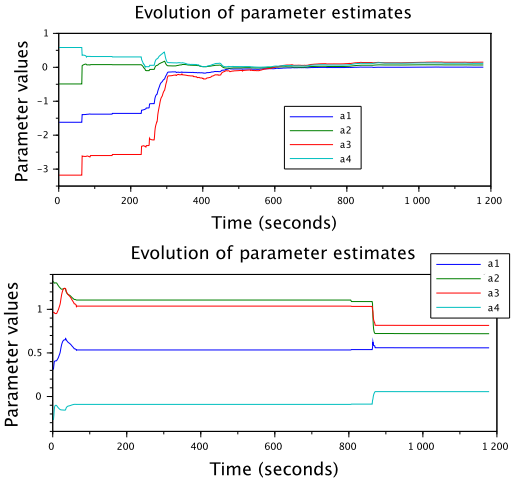


Fig. 4. Evolution of regression coefficients. Notice that after the initial search the stabilized state is reached. In the bottom figure the initialization has given the already stabilized values.

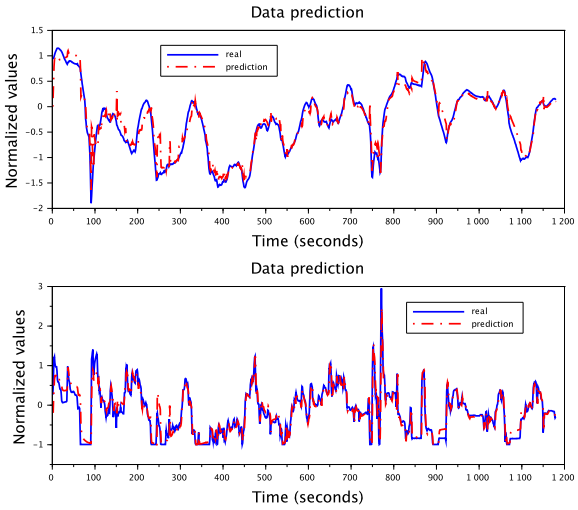


Fig. 5. Selected results of data prediction. Notice that predicted values correspond to real data items.

The presented results are shown for the combination of the visual analysis with the dynamics support initialization, the repeated use of the prior sample and weighting the initial statistics, which gives the minimal prediction error in comparison with other combinations.

## VI. CONCLUSION

The presented approach is based on the availability of prior or expert data, which is always the case in real application fields. Thus the intervention of an expert in processing the prior data is realistic and, as it can be seen, advantageous for such a critical task as the mixture initialization. This paper focuses on initialization of mixtures of normal components.

However, the present research project aims at the recursive estimation of mixtures of different distributions (namely, categorical, exponential, uniform components), which all require specific initialization approaches. This will be part of the future project work.

## ACKNOWLEDGMENT

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

## REFERENCES

[1] Park, B.-J., Zhang, Y., Lord, D. (2010). Bayesian mixture modeling approach to account for heterogeneity in speed data, *Transportation Research Part B: Methodological*, vol. 44, 5, p. 662–673.

[2] Yoshigoe, K., Dai, W., Abramson, M., Jacobs, A. (2015). Overcoming invasion of privacy in smart home environment with synthetic packet injection, In: *TRON Symposium (TRONSHOW)*, Tokyo, Japan, 2014, pp. 1-7.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Šmídl, V., Quinn, A. (2006). The Variational Bayes method in signal processing, Springer-Verlag Berlin Heidelberg.

[11] Frühwirth-Schnatter, S. (2006). Finite mixture and Markov switching models, Springer-Verlag New York, 2006.

[12] 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.

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

[14] Aggarwal, C. Ch. (2015). Outlier Analysis. In: *Data Mining: The Textbook*, Springer International Publishing, p. 237–263.

[15] Fukunaga, K. (2013). Introduction to Statistical Pattern Recognition, series Computer science and scientific computing, Elsevier Science.

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

[17] 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.

[18] Melnykov, V., Melnykov, I. (2012). Initializing the EM algorithm in Gaussian mixture models with an unknown number of components, *Computational Statistics & Data Analysis*, vol. 56, 6, p. 1381–1395.

[19] 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.

[20] Blömer, J., Bujna, K. (2013). Simple methods for initializing the EM algorithm for Gaussian mixture models, *CoRR*, vol. abs/1312.5946, http://arxiv.org/abs/1312.5946.

[21] Ning, H., Yuxiao Hu, Y., Huang, T. (2008). Efficient initialization of mixtures of experts for human pose estimation, in: *Proceedings of the IEEE International Conference on Image Processing*.

[22] Paclik, P., Novovičová, J. (2001). Number of components and initialization in Gaussian mixture model for pattern recognition, in: *Proceedings of the 5th International Conference on Artificial Neural Networks and Genetic Algorithms, ICANNGA 2001*, Prague, Czech Republic, p. 406–409.

[23] Kárný, M., Nedoma, P., Khailova, N, Pavelková L. (2003). Prior information in structure estimation, *IEE Proceedings, Control Theory and Applications*, vol.150, 6, p. 643–653.

[24] Kerridge, D. (1961). Inaccuracy and inference, *Journal of Royal Statistical Society B*, vol. 23, p. 284–294.