Kybernetika

Ivan Nagy; Evgenia Suzdaleva; Miroslav Kárný

Bayesian estimation of mixtures with dynamic transitions and known component parameters

Kybernetika, Vol. 47 (2011), No. 4, 572--594

Persistent URL: http://dml.cz/dmlcz/141660

KYBERNETIKA — VOLUME 47 (2011), NUMBER 4, PAGES 572–594

BAYESIAN ESTIMATION OF MIXTURES WITH DYNAMIC TRANSITIONS AND KNOWN COMPONENT PARAMETERS

Ivan Nagy, Evgenia Suzdaleva and Miroslav Kárný

Probabilistic mixtures provide flexible "universal" approximation of probability density functions. Their wide use is enabled by the availability of a range of efficient estimation algorithms. Among them, quasi-Bayesian estimation plays a prominent role as it runs "naturally" in one-pass mode. This is important in on-line applications and/or extensive databases. It even copes with dynamic nature of components forming the mixture. However, the quasi-Bayesian estimation relies on mixing via constant component weights. Thus, mixtures with dynamic components and dynamic transitions between them are not supported. The present paper fills this gap. For the sake of simplicity and to give a better insight into the task, the paper considers mixtures with known components. A general case with unknown components will be presented soon.

Keywords: mixture model, Bayesian estimation, approximation, clustering, classification

Classification: 93E12, 68T05

1. INTRODUCTION

When dealing with practical applications, one can often meet a situation, when a system to be monitored exists in several different modes of behavior, see e. g. [16]. Transitions between individual modes need not be sharp and they really are rather diffused in practice. The model of such a system is composed of individual components (local models) corresponding to the particular modes. A special "pointer variable" distinguishes the mode (or combination of modes), in which the system actually works. The described model is called a mixture model [22]. The ability of the mixture model to approximate the probability density functions (pdfs) "universally" [7] enhances its importance far beyond modeling of systems with several modes. It found application in a wide variety of areas.

The application width has motivated development of a range of sophisticated algorithms for estimation of mixture parameters, e. g. [12], especially those used in the area of data mining, e. g. [8, 21, 23, 24]. Majority of approaches, represented by expectation-maximization algorithm [5], are oriented towards point estimation. Other well known algorithms are Variational Bayes [19] and Regime Switching Models [6]. However, these approaches are not fully on-line, rely on completely numerical solutions and consequently they are not ready for large industrial applications.

Quasi-Bayesian estimation for static systems generating independently distributed data was proposed in [22]. It was generalized to mixtures with dynamic components allowing dependence among the modeled data [11]. The generalization even allows recursive implementation of the quasi-Bayesian estimation for components in a dynamic exponential family [10]. The proposed approximate estimation replaces the unobserved pointer to the active component by its expectation conditioned on all data measured up to the current time instant. However, it relies on a static model of generating the pointer values. This significantly restricts the expressive power of such semi-dynamic mixtures. The present paper removes this drawback. It considers estimation of a mixture model with several components and a dynamic pointer to the active component. It means that the value of the pointer depends on its realization at the previous time instant. The estimation of this dynamic mixture essentially uses the same basis as the discussed quasi-Bayesian estimation. It also preserves its key features: the evolved posterior pdfs are approximately self-reproducing and determined by a finite-dimensional statistics. The computational complexity is fixed during this one-pass Bayesian estimation. This is important in on-line applications as well as in handling extensive databases. Unlike the quasi-Bayes algorithm, the approximation used here follows a universal approach, as it is based on the Kullback–Leibler divergence [13]. This fact is crucial for further development of the presented algorithms.

In order to stress the methodological nature of the paper, the presentation is made as straightforward as possible (Problem formulation, Solution, Algorithmic Summary, Illustrative Example and Concluding Remarks). The inevitable technical details are to be found in the Appendix.

2. PROBLEM FORMULATION

The considered system operates in n different modes, which can change gradually and randomly one into other. The system modes are supposed so different that each of them has to have its special model. These models are called components. In real time, the modes of the system vary so that at each time instant $t \in \{1, 2, \ldots, N\} = t^*$ the system works in single, so called actual, working point.

For this system we consider a set of all data from time $t = 0$ to $t = N$ denoted by $d\,(0 : N) = \{d_0, d_1, d_2, \ldots, d_N\} = d(N)$, where $d(0) = d_0$ is prior information (preliminary measured data or expert knowledge) supposed to be known. The data samples $d\,(1 : N) = \{d_1, d_2, \ldots, d_N\}$, i. e., for $t > 0$, are measured on the system, i.e., $d(N) = \{d(0), d(1 : N)\}$.

At each time instant, the overall system can be described by one component properly chosen from the set of all components. This component is called the actual one and its label (order number) at time instant t is indicated by the value of the random variable $c_t \in \{1, 2, \ldots, n\} = c^*$. The random variables c_t form a stochastic process $c\,(1 : N) = \{c_1, c_2, \ldots, c_N\}$ which is called a pointer. Similarly, its prior value c_0 is available and formally denoted by $c(0)$. Description of the pointer uses a parameter α. Both, the pointer $c\,(1 : N)$ and the parameter α are unknown and have to be estimated.

We suppose that a global model of such system can be expressed as a joint probability density function (pdf) of all measured or estimated variables: $d\,(1:N)$ (all measured data), $c\,(1:N)$ (all estimated pointer entries) and α (parameter of the pointer model)

$$f\left(d\,(1:N)\,,c\,(1:N)\,,\alpha|d(0),c(0)\right), \tag{1}$$

where $f(\cdot|\cdot)$ denotes a conditional pdf. The subscript at f denoting the random variable is omitted from brevity reasons and the random variable is given by its realization in the argument of the pdf.

Our goal is to derive on-line estimation algorithms. That is why we need models working in real time. They can be obtained by applying the well known chain rule to the global model (1) and taking some assumptions. We obtain

$$f\left(d\,(1:N)\,,c\,(1:N)\,,\alpha|d(0),c(0)\right)=\prod_{t=1}^{N} f\left(d_t,c_t|d\,(t-1)\,,c\,(t-1)\,,\alpha\right)$$

$$=\quad\prod_{t=1}^{N} f\left(d_t|c_t,d\,(t-1)\,,c\,(t-1)\,,\alpha\right)f\left(c_t|d\,(t-1)\,,c\,(t-1)\,,\alpha\right). \tag{2}$$

Generally, the pdfs in the factorization (2) represent models in real time. The first pdfs after the product sign are descriptions of components, the second ones are descriptions of the pointer. To obtain the final form of the models, we accept the following assumptions

- components depend only on the actual entry c_t of the pointer and they do not depend on the pointer parameter α, so it holds

$$f\left(d_t|c_t,d\,(t-1)\,,c\,(t-1)\,,\alpha\right)=f\left(d_t|c_t,d\,(t-1)\right),$$

- pointer model does not depend on data and it depends only on the last pointer, not the older ones (Markov property). The pointer model thus takes the form

$$f\left(c_t|d\,(t-1)\,,c\,(t-1)\,,\alpha\right)=f\left(c_t|c_{t-1},\alpha\right).$$

Under these assumptions, the global model (1) reads

$$f\left(d\,(1:N)\,,c\,(1:N)\,,\alpha|d(0),c(0)\right)=\prod_{t=1}^{N} f\left(d_t|c_t,d\,(t-1)\right)f\left(c_t|c_{t-1},\alpha\right).$$

From the above factorization of the global model (1) the needed real time models (component and pointer ones) naturally follow.

Component model

For any pointer value, the operating mode of the system is described by a component model specified by the conditional probability density function

$$f\left(d_t|c,d\,(t-1)\right). \tag{3}$$

These component models are supposed fully known. They have a specific form (for instance, normal pdf of scalar d_t with known expectation and variance). In general, they can be described by any known time and data dependent pdf.

Pointer model

Thus, the pointer model is assumed dynamic, data independent and parameterized by the unknown parameter α in the following form

$$f\left(c_t|c_{t-1}, d\left(t-1\right), \alpha\right) = f\left(c_t|c_{t-1}, \alpha\right) = \alpha_{c_t|c_{t-1}}, \tag{4}$$

where α is an unknown matrix parameter, whose entry $\alpha_{i|j}$ denotes a probability of the ith currently active component, when the jth one was active previously. For α it holds:

$$\alpha \in \alpha^* = \left\{ \alpha_{i|j} \geq 0, \, \forall i, j \in c^*; \, \sum_{i \in c^*} \alpha_{i|j} = 1, \, \forall j \in c^* \right\}.$$

The addressed task

The paper proposes an approximate *recursive* Bayesian estimation of the pointer c_t to the active component at each time $t \in t^*$ generally described by its posterior pdf

$$f\left(c_t|d\left(t\right)\right), \tag{5}$$

which is represented by a vector of probabilities for each component to be active at time t. It holds $f\left(c_t|d\left(t\right)\right) \geq 0$, $\forall c_t \in c^*$, and $\sum_{c_t \in c^*} f\left(c_t|d\left(t\right)\right) = 1$, for all $t \in t^*$.

Its computation is bound to the recursive evolution of the posterior pdf of α for which exact evaluations are not feasible. Thus, the paper proposes its on-line implementable approximation.

3. PROBLEM SOLUTION

The main task, as written in the previous paragraph, is to estimate the currently active component(s), i.e., to provide the pdf (5) at each time t of the time interval t^*. Due to the form of the models (3) and (4), the mentioned pointer pdf (5) cannot be derived separately. It is closely connected to the parameter α that has to be recursively estimated together with the pointer variable c_t. Thus, the basic task is to find a recursion for the evolution of the posterior pdf of unknown variables $[c_t, \alpha]$, conditioned on the past data $d\left(t-1\right)$ for $t \in t^*$. Nevertheless, a problem is encountered. The computations during the evolution of the posterior pdf are not feasible. The form of the posterior pdf changes during the estimation and its complexity grows with running time. It means that an approximation of the posterior pdf is necessary at each step of the estimation. This approximation restores the original form of the prior pdf (posterior from the last estimation step). The desired pointer pdf (5) is then derived from the evolved posterior pdf.

Joint pdf

The estimation is based on the joint pdf of d_t, c_t, c_{t-1}, α, conditioned on observed data $d\left(t-1\right)$. With the help of decomposition of the joint pdf into a product of the conditional ones, and according to (3) and (4), this pdf can be expressed as follows

$$\begin{aligned} & f\left(d_t, c_t, c_{t-1}, \alpha | d\left(t-1\right)\right) \\ &= f\left(d_t | c_t, c_{t-1}, \alpha, d\left(t-1\right)\right) f\left(c_t | c_{t-1}, \alpha, d\left(t-1\right)\right) f\left(c_{t-1}, \alpha | d\left(t-1\right)\right) \\ &= f\left(d_t | c_t, d\left(t-1\right)\right) \alpha_{c_t | c_{t-1}} f\left(c_{t-1}, \alpha | d\left(t-1\right)\right), \end{aligned} \tag{6}$$

where

- $f\left(d_t | c_t, c_{t-1}, \alpha, d\left(t-1\right)\right) = f\left(d_t | c_t, d\left(t-1\right)\right)$ is the model of the c_tth component (3) , which is supposed to be known and to meet the conditional of independency on α and c_{t-1}
- $f\left(c_t | c_{t-1}, \alpha, d\left(t-1\right)\right) = \alpha_{c_t | c_{t-1}}$ is the model of the pointer (4) that is independent of data $d\left(t-1\right)$ and whose form is known but the specific values of α are unknown.

A new object appears in (6)

$$f\left(c_{t-1}, \alpha | d\left(t-1\right)\right),$$

which is a prior pdf for time t of the estimated pointer c_{t-1} and parameter α. The variables c_{t-1} and α are assumed to be conditionally independent

$$f\left(c_{t-1}, \alpha | d\left(t-1\right)\right) = f\left(c_{t-1} | d\left(t-1\right)\right) f\left(\alpha | d\left(t-1\right)\right). \tag{7}$$

The intuitively plausible property, that this joint pdf is given as a product of marginal pdfs, can be assumed a priori. In each step of the estimation the form of the prior pdf is lost and must be restored using an approximation. The form of the prior pdf is important for feasibility of computations.

Prior pdf for estimation of c_t

In the relation (7), the pdf

$$f\left(c_{t-1} | d\left(t-1\right)\right)$$

is a vector of probabilities similarly as in (5) for $f\left(c_t | d\left(t\right)\right)$, but with the time index t replaced by $t-1$. For the time instant t it is the prior pdf for estimation of the pointer variable c_t.

Prior pdf for estimation of α

The second pdf in (7) $f\left(\alpha | d\left(t-1\right)\right)$ describes the unknown transition probabilities $\alpha_{i|j}$. It is chosen as Dirichlet pdf of α with the matrix statistic ν_{t-1}, $\nu_{i|j;t-1} > 0$,

$$\mathcal{D}_\alpha\left(\nu_{t-1}\right) = \frac{1}{B\left(\nu_{t-1}\right)} \prod_{i \in c*} \prod_{j \in c^*} \alpha_{i|j}^{\nu_{i|j;t-1}-1}, \tag{8}$$

which is, in more detail, defined in Appendix 7.1. This pdf is conjugated (self-reproducing) to the model (4), see [10].

The prior pdfs are evolved in time. This evolution starts with prior pdfs from the beginning of estimation, i. e., from the (prior) time instant $t = 0$. The prior pdfs can be constructed from two sources; *(i)* from a priori measured data, *(ii)* from expert information. The prior data that are measured on the system before the estimation can be elaborated by the same algorithm which is used in the subsequent estimation, thus producing the prior pdf. The incorporation of the expert knowledge into the prior statistics is not an easy task. However, some general recommendations exist [2]. E.g., the prior statistics for the pointer can be set according to the prior belief in the probability of individual transitions. The normalized values in the rows of the prior statistics ν_0 express the probabilities of transitions; the absolute values of these statistics express the belief in the prior guess.

Modification of the joint pdf

After substituting the introduced pdf forms, notation and independency assumption (7), the joint pdf (6) gets the form

$$f\left(d_t, c_t, c_{t-1}, \alpha | d\left(t-1\right)\right) = f\left(d_t | c_t, d\left(t-1\right)\right) \alpha_{c_t|c_{t-1}} f\left(c_{t-1} | d\left(t-1\right)\right) \mathcal{D}_\alpha\left(\nu_{t-1}\right). \quad (9)$$

Nevertheless, this form of the joint pdf is not final. Using the basic equality

$$\alpha_{c_t|c_{t-1}} \mathcal{D}_\alpha\left(\nu_{t-1}\right) = \hat{\alpha}_{c_t|c_{t-1};t-1} \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right), \quad (10)$$

which is proved in Appendix 7.1, it can be given the form

$$f\left(d_t, c_t, c_{t-1}, \alpha | d\left(t-1\right)\right) = f\left(d_t | c_t, d\left(t-1\right)\right) \hat{\alpha}_{c_t|c_{t-1};t-1} f\left(c_{t-1} | d\left(t-1\right)\right) \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right), \quad (11)$$

where

- $\hat{\alpha}_{c_t|c_{t-1};t-1}$ is quadratically optimal point estimate of $\alpha_{c_t|c_{t-1}}$ given by the relation

$$\hat{\alpha}_{c_t|c_{t-1};t-1} = \frac{\nu_{c_t|c_{t-1};t-1}}{\sum_{i\in c^*} \nu_{i|c_{t-1};t-1}}, \quad (12)$$

derivation of which can be found in Appendix 7.1

- and

$$\nu_{t-1}^{c_t|c_{t-1}} = \nu_{t-1} + \delta_{c_t|c_{t-1}}, \quad (13)$$

  where $\delta_{c_t|c_{t-1}}$ is a zero matrix of compatible dimensions with the number one on the position $c_t|c_{t-1}$ (matrix of products of the Kronecker delta functions $\delta\left(j, c_t\right) \delta\left(i, c_{t-1}\right)$), see Appendix 7.1.

Expression (11) is the final form of the joint pdf. However, this result is just preparatory and it forms a basis for further derivations.

Time evolution of the posterior pdf

The crucial point of the whole estimation algorithm is the time evolution of the pdf $f\left(c_{t-1},\alpha|d\left(t-1\right)\right)$, describing the prior pdf for the unknown variables c_t and α to the posterior one $f\left(c_t,\alpha|d\left(t\right)\right)$. As this evolution is recursive using the Bayes rule, an important question is, if these pdfs are self-reproducing, i. e., whether the complexity of the algorithm does not grow in time.

Using the Bayes rule saying that $f\left(A|B,C\right)\propto f\left(A,B|C\right)$, the posterior pdf $f\left(c_t,\alpha|d\left(t\right)\right)$ becomes

$$f\left(c_t,\alpha|d\left(t\right)\right)\propto f\left(d_t,c_t,\alpha|d\left(t-1\right)\right),$$

where $A=\{c_t,\alpha\}$, $B=d_t$ and $C=d\left(t-1\right)$

Now, adding and immediately marginalizing the past pointer variable c_{t-1}, the posterior pdf takes the form

$$f\left(c_t,\alpha|d\left(t\right)\right)\propto\sum_{c_{t-1}\in c^*}f\left(d_t,c_t,c_{t-1},\alpha|d\left(t-1\right)\right)$$

Substituting the joint pdf from (11), the posterior pdf takes its final form

$$f\left(c_t,\alpha|d\left(t\right)\right)\propto f\left(d_t|c_t,d\left(t-1\right)\right)\sum_{c_{t-1}\in c^*}f\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}\mathcal{D}_{\alpha}\left(\nu_{t-1}^{c_t|c_{t-1}}\right). \tag{14}$$

The above relation (14) shows that the exact evolution from the prior to the posterior pdf is not feasible. Recursive use of (14) gives the posterior pdf as a product of sums. This destroys the form of the prior pdfs. The demands for memory and computing time necessary for their exact evaluation grow unacceptably. Thus, an approximation restoring the form of the posterior pdf is necessary.

Estimation of the pointer c_t

A recursion for the pointer c_t, i. e. the recomputation of the prior $f\left(c_t|d\left(t-1\right)\right)$ to the posterior $f\left(c_t|d\left(t\right)\right)$ can be obtained from the joint posterior (14) by integration over α

$$\begin{aligned} & f\left(c_t|d\left(t\right)\right)\propto\int_{\alpha^*}f\left(c_t,\alpha|d\left(t\right)\right)\mathrm{d}\alpha \\ = & \int_{\alpha^*}f\left(d_t|c_t,d\left(t-1\right)\right)\sum_{c_{t-1}\in c^*}f\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}\mathcal{D}_{\alpha}\left(\nu_{t-1}^{c_t|c_{t-1}}\right)\mathrm{d}\alpha \\ = & f\left(d_t|c_t,d\left(t-1\right)\right)\sum_{c_{t-1}\in c^*}f\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}. \end{aligned} \tag{15}$$

As the posterior $f\left(c_t|d\left(t\right)\right)$ is a mere vector, its summation form does not matter and its form is preserved. Thus, the recursive formula for estimation of c_t is ready. However, it needs the term $\hat{\alpha}_{c_t|c_{t-1};t-1}$, which follows from estimation of α.

Estimation of the parameter α

Similarly as for c_t, the recursion for estimating of α can be obtained from the joint posterior pdf (14) by marginalization over c_t. We obtain

$$\tilde{f}\left(\alpha|d\left(t\right)\right) \propto \sum_{c_t\in c^*} f\left(d_t|c_t, d\left(t-1\right)\right) \sum_{c_{t-1}\in c^*} f\left(c_{t-1}|d\left(t-1\right)\right) \hat{\alpha}_{c_t|c_{t-1};t-1} \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right) \tag{16}$$

where $\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)$ is the prior pdf for α updated for a specific combination of values of c_t and c_{t-1}.

Remark. The resulting "updated posterior pdf" is denoted by $\tilde{f}\left(\alpha|d\left(t\right)\right)$ to stress, that it has not the final form. It is a mixture of Dirichlet pdfs and must be further approximated to a single Dirichlet pdf. This one will be denoted by $f\left(\alpha|d\left(t\right)\right)$ and will be called the "approximated posterior pdf".

Here, the situation is more complicated. The resulting posterior is constructed as a weighted sum of updated priors (with Dirichlet distributions) and thus the result does not have the Dirichlet distribution. To preserve feasibility of the resulting algorithm, its form must be approximately restored.

Approximation

As mentioned, the posterior pdf $\tilde{f}\left(\alpha|d\left(t\right)\right)$ is a mixture of Dirichlet prior pdfs and must be approximated to a single Dirichlet pdf which we denote $f\left(\alpha|d\left(t\right)\right) = \mathcal{D}_\alpha\left(\nu_t\right)$. The pdf f is computed so that it minimizes the Kullback–Leibler divergence KL defined as

$$KL = \int_{\alpha^*} f\left(\alpha|d\left(t\right)\right) \ln \frac{f\left(\alpha|d\left(t\right)\right)}{\tilde{f}\left(\alpha|d\left(t\right)\right)} \mathrm{d}\alpha. \tag{17}$$

The Kullback–Leibler divergence in this form is shown [1] to be best compatible with the Bayesian approach to estimation.

Solution to the approximation task

The solution is leads to a solution of the following system of equations with the $n \times n$-matrix statistics ν_t

$$G_{t-1}\,\Xi\left(\nu_{i|j;t}\right) - H_{i|j;t-1} = 0,\ i,\ j\ \in c^*, \tag{18}$$

where

$$G_{t-1} \equiv \sum_{c_t\in c^*} f\left(d_t|c_t, d\left(t-1\right)\right) \sum_{c_{t-1}\in c^*} f\left(c_{t-1}|d\left(t-1\right)\right) \hat{\alpha}_{c_t|c_{t-1};t-1}, \tag{19}$$

$$H_{i|j;t-1} \equiv \sum_{c_t\in c^*} f\left(d_t|c_t, d\left(t-1\right)\right) \sum_{c_{t-1}\in c^*} f\left(c_{t-1}|d\left(t-1\right)\right) \hat{\alpha}_{c_t|c_{t-1};t-1} \Xi\left(\nu_{i|j;t-1}^{c_t|c_{t-1}}\right), \tag{20}$$

for $i, j \in c^*$, with

$$\Xi\left(\nu_{i|j;t}\right) = \Psi\left(\nu_{i|j}\right) - \Psi\left(\sum_{i \in c^*} \nu_{i|j}\right), \quad i, j \in c^*, \tag{21}$$

and Ψ function is

$$\Psi(z) = \frac{\mathrm{d}}{\mathrm{d}z} \ln \Gamma(z).$$

The matrix elements $\nu_{i|j;t-1}^{c_t|c_{t-1}}$ are introduced in (13). The full derivation can be found in Appendix 7.2. The numerical solution to the equation (18) must be performed at each step of the estimation. However, it is quick and without problems due to the fact, that the minimized function KL is a convex function of ν. Thus, the search for the extreme is straightforward and the extreme found is always the global minimum. For a proof of this assertion, see Appendix 7.3.

The resulting approximated posterior pdf

$$f(\alpha|d(t)) = \mathcal{D}_\alpha(\nu_t)$$

with the statistics ν_t determined by the solution of (18).

4. ALGORITHM

The obtained results can now be summarized in the form of an algorithm.

Initial part (start of the algorithm)

- Specify the number of components n and their pdfs $f(d_t|c, d(t-1)), \ c \in c^*$.
- Choose the initial values of the statistics for α-estimation, i. e., the numbers $\nu_{0;i|j} > 0, \ i, j \in c^*$.
- Set $t = 0$ and evaluate $\hat{\alpha}_t$ according to (12).
- Set the initial values of the statistics for estimation of c_t as a vector of probabilities $f(c_0|d(0)), \ c_t \in c^*$.

On-line part (time run of the algorithm) for $t = 1$ up to N do

1. Acquire the current data item d_t.
2. Compute values of the component pdfs for the measured d_t and all $c \in c^*$
$$f(d_t|c, d(t-1)).$$
3. Update the statistics for estimation of c_t, $\forall c_t \in c^*$,
$$f(c_t|d(t)) = f(d_t|c_t, d(t-1)) \sum_{c_{t-1} \in c^*} f(c_{t-1}|d(t-1))\, \hat{\alpha}_{c_t|c_{t-1};t-1},$$
where $\hat{\alpha}_{c_t|c_{t-1};t-1} = \nu_{c_t|c_{t-1};t-1} / \sum_{c_t \in c^*} \nu_{c_t|c_{t-1};t-1}$.

4. Compute the matrix

$$\Xi\left(\nu_{i|j;t-1}\right)=\Psi\left(\nu_{i|j;t-1}\right)-\Psi\left(\sum_{i\in c^*}\nu_{i|j;t-1}\right),\quad i,j\in c^*,$$

   where Ψ is the psi function, see (21).

5. Construct G and H from (18)

$$\begin{aligned} G_{t-1} &= \sum_{c_t\in c^*} f\left(d_t|c_t,d\left(t-1\right)\right)\sum_{c_{t-1}\in c^*} f\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1} \\ H_{i|j;t-1} &= \sum_{c_t\in c^*} f\left(d_t|c_t,d\left(t-1\right)\right) \\ &\quad\times\sum_{c_{t-1}\in c^*} f\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}\Xi\left(\nu_{i|j;t-1}^{c_t|c_{t-1}}\right), \end{aligned}$$

   for $i,j\in c^*$ with

$$\nu_{i|j;t-1}^{c_t|c_{t-1}}=\nu_{i|j;t-1}+\delta_{c_t|c_{t-1}}$$

   according to (19) and (20).

6. Use a numerical method to solve the equation

$$G_{t-1}\,\Xi\left(\nu_{i|j;t}\right)-H_{i|j;t-1}=0,\quad i,j\ \in c^*$$

   to obtain the optimal statistics $\nu_{i|j;t},\ i,j\in c^*$.

7. Compute the approximated posterior pdf for α

$$f\left(\alpha|d\left(t\right)\right)=\mathcal{D}_\alpha\left(\nu_t\right)$$

   with the optimal statistics ν_t, i. e., such one that possesses the Dirichlet form and approximates the updated posterior $\tilde{f}\left(\alpha|d\left(t\right)\right)$.

8. Guess the active component(s), e. g., on the basis of the point estimate, for instance, as the most probable component.

End of the loop for t

5. ILLUSTRATIVE EXPERIMENTS

The presented experiments demonstrate a contribution of the proposed mixture estimation with the dynamic pointer. For comparison, a simple, heuristic, mixture estimator is taken. The reason of the experiments is to demonstrate the main principles of the proposed estimation algorithm.

Simple mixture estimator

The simple mixture estimator uses a heuristic way of computation the weights of individual components for the current data item (i. e., the probabilities $f\left(c_t|d\left(t\right)\right)$). It is done via substituting the data in the component models (3), computing the model pdf values and normalizing them so that their sum equals to one. No information is used about the evolution of the pointer c_t. The simple estimator is obtained by enforcing a uniform prior pdf $f\left(c_t|d\left(t-1\right)\right)$ into the Bayes rule

$$\begin{aligned} f\left(c_t|d\left(t\right)\right) &\propto f\left(d_t,c_t|d\left(t-1\right)\right) = f\left(d_t|c_t,d\left(t-1\right)\right) f\left(c_t|d\left(t-1\right)\right) \\ &\propto f\left(d_t|c_t,d\left(t-1\right)\right), \ \ c_t \in c^*, \ \ t \in t^* \end{aligned} \tag{22}$$

for $f\left(c_t|d\left(t-1\right)\right)$ uniform.

Remark. Comparing the relations (22) and (15), one can see, that the introduced simple estimator is just one part of the dynamic-pointer estimator, namely it is given by the factor $f\left(d_t|c_t,d\left(t-1\right)\right)$. This factor deduces the probability of the active component only from the location of the respective component models. The neglected factor in (15) is $\sum_{c_{t-1}\in c^*} f\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}$. It reflects the probability that a component labeled c_{t-1} was active at previous time instant $t-1$ and the probabilities of its transition to the component c_t. Both probabilities defining this factor have been estimated on the basis of the past data, up to time $t-1$. In the considered simple case with known components, this term often does not play a significant role. For overlapping component models and strong, almost deterministic, switching, this prior pdf, exploiting the model of the pointer evolution, can be decisive. It indicates its importance in general case with recursively estimated components.

Simulated data

A data sample generated for the experiments contains 300 items of realizations from two-dimensional random vector. The generating mixture model has three normal components with fixed expectations μ_c and variances Σ_c shown in Table 1 for $c \in c^* = \{1,2,3\}$.

c	1	2	3
μ_c	$[1,\ 1]'$	$[1,\ 5]'$	$[5,\ 3]'$
Σ_c	$\begin{bmatrix} 0.64 & ,-0.77 \\ -0.77 & ,\ 6.68 \end{bmatrix}$	$\begin{bmatrix} 16 & ,-6.08 \\ -6.08 & ,\ 2.95 \end{bmatrix}$	$\begin{bmatrix} 0.64\,, & 3.2 \\ 3.2\,, & 21.76 \end{bmatrix}$

Tab. 1. Components of mixture model.

The explicit form of the component model, labeled by c, is

$$f\left(d_t|c_t,d\left(t-1\right)\right) = \left(2\pi\sqrt{|\Sigma_c|}\right)^{-1} \exp\left\{-\frac{1}{2}\left(d_t-\mu_c\right)'\Sigma_c^{-1}\left(d_t-\mu_c\right)\right\}, \tag{23}$$

where $d_t = [d_{1;t}, d_{2;t}]'$ and μ_c, Σ_c are from Table 1.

Table 2 provides the parameterized transition probabilities for generating values of the pointer.

$f(c_t \mid c_{t-1}, \alpha)$	$c_t = 1$	$c_t = 2$	$c_t = 3$
$c_{t-1} = 1$	p	$1 - 2p$	p
$c_{t-1} = 2$	p	p	$1 - 2p$
$c_{t-1} = 3$	$1 - 2p$	p	p

Tab. 2. Generating of pointer values.

The parameter p is respectively set to $p = 0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$. One can note that the parameter p influences the degree of "randomness" of the pointer generating. For the value $p = 0$ the pointer model becomes a deterministic one and generates the active components in the fixed order

$$1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \rightarrow 2 \text{ etc.}$$

The greater the value of p, the more uncertain the model outcome. For p about 0.3 and 0.4 in Table 2, one obtains a rather random but correlated behavior of the pointer. For $p = 0.5$, the dependence on the past disappears and the simulation simply jumps among various components with a fixed common probability. By (perhaps not very correct) name "predictions" are denoted the data generated by the estimated models.

Experiments

The layout of the experiments is following. First, we simulate the data sample of the described mixture. Then the estimation is performed for both the models; (i) the static one for the simple estimator and (ii) the dynamic one for the proposed estimation algorithm. Afterwards, new data samples are generated, first with the static model and second with the dynamic one. The generated clusters are compared to the simulation.

Figure 1 provides an insight into the simulation and predictions. It shows the simulated data for the case $p = 0$. The predictions obtained with the described simple estimator are in the middle figure. The right figure shows the predictions with the proposed dynamic-pointer algorithm.

The middle and right picture in Figure 1 demonstrate a good correspondence of the predictions to the original data for both the algorithms.

For each of the considered values $p = 0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$ 300 data was simulated and the number of wrongly classified active components was evaluated. The results of these numerical evaluations are plotted in Figure 2 and summarized in the table attached to it.

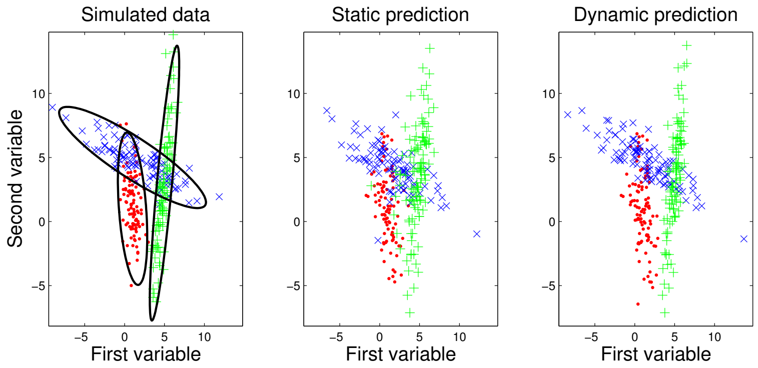


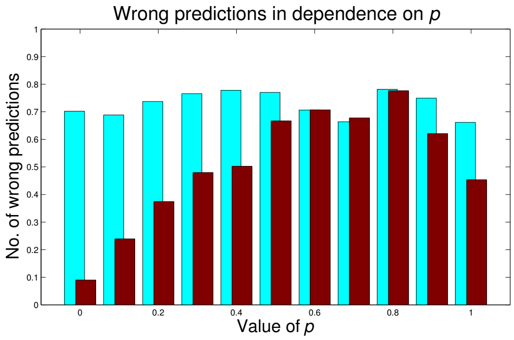
Fig. 1. Left figure shows simulated data. The membership to the components is distinguished by symbols • × + and bound by ellipses, which form the contours of the static Gaussian models used. Data predictions are computed by generating data from all three components and making their average weighted by the estimated probabilities of the component labels. The predictions based on the simple static-pointer estimator are in the middle and those resulting from the proposed dynamic-pointer estimator are on the right.

Discussion

For higher values of p, corresponding to weak dependence of the subsequent pointer values, the simple static-model-based estimator behaves similarly to the proposed mixture estimation based on dynamic model of the pointer evolution. The difference becomes much more pronounced for small values of p, when the transition probabilities are near zero or one. In such truly dynamic cases, the proposed estimator is visibly better than the simple one. Of course, demands on precision and reliability of the estimates strongly depend on the purpose for which the estimation is used. For instance, in safety-sensitive applications any wrong classification is very expensive and high precision is demanded [3, 4]. Then, the gain of the better modelling and adequate processing (without increasing demands on expensive acquisition of informative data), offered by the proposed algorithm, is significant.

The adopted Bayesian framework brings additional benefits. The estimation provides not only point estimates of the transition probabilities α but also information about their uncertainty and consequently about reliability of the classification. According to (33), the variance of α is inversely proportional to the count of time instants for which the particular component has been active during the estimation process. Thus, small values of the estimate of this *activation count* indicate that the point estimate of the corresponding transition probability is unreliable. The discussed feature is illustrated in Figure 3.

Left part of Figure 3 shows the time course of the activation count estimate for



p	0	0.1	0.2	0.3	0.4	0.5
Static	39	31	35	30	33	27
Dynamic	6	24	29	29	33	22

Fig. 2. The counts of incorrectly classified active components provided by the simple estimator (light) and those from the proposed estimator (dark). The number of 300 data items were simulated for each value of the parameter p.

the extremal situation with deterministic, regularly switching pointer, $p = 0$. The information about the transitions is maximally strong and the respective components are visited regularly. Consequently, the achieved precision is high and growing almost monotonically with the number of processed data samples.

The right part of Figure 3 illustrates the opposite case with “very random” pointer, $p = 0.4$. The data are much less informative and consequently the precision is much lower. Also the time evolution of the statistics is far from being monotonic. It shows that a mere processing of a new data item does not always imply an increase in the reliability of estimates.

6. CONCLUSION

An efficient recursive (one-pass) Bayesian estimation of a mixture model with dynamically evolving pointers to the active components has been presented. Its principle is shown on the case with known components.

A lot remains to be done:

- The algorithm has to be extended to the practically significant case with incompletely known components. This extension is straightforward and will be presented soon.

- The algorithm has to be elaborated especially for mixed – discrete and continuous – valued data. This is important especially for data bases applications. The standard control and signal processing domains need it badly, too. A

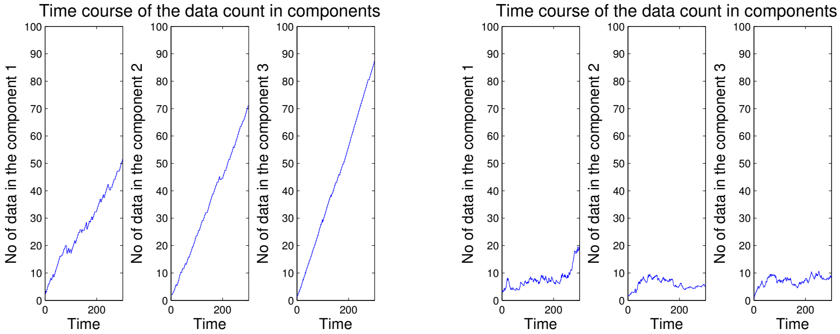


Fig. 3. The individual time courses of the estimates of the activation counts. The left figure corresponds to the deterministic transitions, the right one to the very random transitions. The sharp growth of statistics indicates an increase in the precision of estimation and it occurs for informative data. The slow growth or even stagnation means that data are poorly informative and the gained estimates are unreliable. It allows monitoring the estimation and classification quality.

closer look at the proposed algorithm indicates that it is well possible using the similar approximation technique.

- One-pass treatment suffers loss of quality caused by the use of approximate posterior pdf as the prior pdf for further step. Local iterations with a sort of forgetting promise at least partial remedy of this problem, studied for instance in [14].

In spite of the width of the open problem, there are definite contributions worth stressing:

- The approximation way is universally applicable to fully dynamic models. It opens a way for development of specialized estimation algorithms relying on variety of component and pointer models. Application width of the resulting dynamic clustering, classification and prediction is enormous.

- The adopted Bayesian treatment provides tools for monitoring reliability of the obtained estimates or predictions.

- The Kullback–Leibler divergence is used to measure quality of the constructed recursively feasible approximate estimation. Its prominent role in Bayesian context, analyzed in [1], is often neglected by significant approaches to estimation (for instance, in the so-called functional or mean field approximations [18]).

- Recursive (one-pass) estimation is common in control domain [15]. Its high potential for treatment of extensive data-bases under time constraints is still incompletely exploited. The availability of efficient well justified algorithms can contribute to the positive change of this state.

- The dynamic dependence between adjacent data records is commonly respected in control and signal processing domains [9, 17]. This feature is mostly neglected when handling data bases. At the same time, the commonly accepted assumption on conditional independence of data records is often rather crude approximation. The proposed algorithm can handle possible dependence.

7. APPENDIX

7.1. Proof of the basic equality

Here, we are going to prove the equality

$$\alpha_{c_t|c_{t-1}}\mathcal{D}_\alpha\left(\nu_{t-1}\right) = \hat{\alpha}_{c_t|c_{t-1};t-1}\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right). \tag{24}$$

Taking into account the definition of the Dirichlet distribution

$$\mathcal{D}_\alpha\left(\nu_{t-1}\right) = \frac{1}{B\left(\nu_{t-1}\right)}\prod_{i\in c*}\prod_{j\in c^*}\alpha_{i|j}^{\nu_{i|j;t-1}-1} \equiv \frac{b_\alpha\left(\nu_{t-1}\right)}{B\left(\nu_{t-1}\right)},$$

where

$$b_\alpha\left(\nu\right) = \prod_{i\in c*}\prod_{j\in c^*}\alpha_{i|j}^{\nu_{i|j}-1}, \tag{25}$$

$$\text{and}\quad B\left(\nu\right) = \prod_{j\in c^*}\frac{\prod_{i\in c^*}\Gamma\left(\nu_{i|j}\right)}{\Gamma\left(\sum_{i\in c*}\nu_{i|j}\right)} \tag{26}$$

it is possible to write the right hand side of the proved equality (24)

$$\alpha_{c_t|c_{t-1}}\mathcal{D}_\alpha\left(\nu_{t-1}\right) = \alpha_{c_t|c_{t-1}}\frac{1}{B\left(\nu_{t-1}\right)}\prod_{i\in c*}\prod_{j\in c^*}\alpha_{i|j}^{\nu_{i|j;t-1}-1}$$

$$= \frac{1}{B\left(\nu_{t-1}\right)}\prod_{i\in c*}\prod_{j\in c^*}\alpha_{i|j}^{\nu_{i|j;t-1}+\delta_{c_t|c_{t-1}}-1} = \frac{1}{B\left(\nu_{t-1}\right)}\prod_{i\in c*}\prod_{j\in c^*}\alpha_{i|j}^{\nu_{i|j;t-1}^{c_t|c_{t-1}}-1}$$

$$= \frac{b_\alpha\left(\nu_{t-1}^{c_t,|c_{t-1}}\right)}{B\left(\nu_{t-1}\right)} \tag{27}$$

where $\delta_{c_t|c_{t-1}}$ is a zero matrix with the same dimensions as v_{t-1} with the number one on the position $c_t|c_{t-1}$, and

$$\nu_{i|j;t-1}^{c_t|c_{t-1}} = \nu_{i|j;t-1} + \delta_{c_t|c_{t-1}}.$$

After expanding the fraction (27) by dividing and multiplying by the beta function $B\left(\nu_{t-1}^{c_t|c_{t-1}}\right)$ we obtain

$$\alpha_{c_t|c_{t-1}}\mathcal{D}_\alpha\left(\nu_{t-1}\right) = \frac{B\left(\nu_{t-1}^{c_t|c_{t-1}}\right)}{B\left(\nu_{t-1}\right)}\frac{b_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)}{B\left(\nu_{t-1}^{c_t|c_{t-1}}\right)} = \frac{B\left(\nu_{t-1}^{c_t|c_{t-1}}\right)}{B\left(\nu_{t-1}\right)}\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right). \quad (28)$$

Now, what remains is to show, that the ratio of the beta functions in the previous expression is equal to $\hat{\alpha}_{t-1}$ which is the point estimate of α. A straightforward but rather long way how to do it is to use the definition of the beta function. We will prove it in the following way. Apply the integral over α to the first and last term of the previous equation. We obtain

$$\int_{\alpha^*}\alpha_{c_t|c_{t-1}}\mathcal{D}_\alpha\left(\nu_{t-1}\right)\mathrm{d}\alpha = \frac{B\left(\nu_{t-1}^{c_t|c_{t-1}}\right)}{B\left(\nu_{t-1}\right)}\int_{\alpha^*}\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)\mathrm{d}\alpha.$$

The left hand side of the equation is the expectation of α with Dirichlet distribution which is $\hat{\alpha}_{t-1}$ and the integral on the right hand side is equal to one. This completes the proof.

7.2. Derivation of the approximate posterior pdf

The chosen approximation according to (16) is based on a minimization of the Kullback–Leibler divergence (17).

The pdfs entering the minimization are: (*i*) The updated posterior pdf (16) which is to be approximated

$$\tilde{f}\left(\alpha|d\left(t\right)\right) \propto \sum_{c_t\in c^*} f\left(d_t|c_t, d\left(t-1\right)\right)\sum_{c_{t-1}\in c^*} f\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right) \quad (29)$$

and (*ii*) the posterior pdf $f\left(\alpha|d\left(t\right)\right) = \mathcal{D}_\alpha\left(\nu_t\right)$ (17) which is the approximation.

The Kullback–Leibler divergence to be minimized is

$$KL = \int_{\alpha^*}\tilde{f}\left(\alpha|d\left(t\right)\right)\ln\frac{\tilde{f}\left(\alpha|d\left(t\right)\right)}{f\left(\alpha|d\left(t\right)\right)}\mathrm{d}\alpha \propto \int_{\alpha^*}\tilde{f}\left(\alpha|d\left(t\right)\right)\ln\frac{1}{f\left(\alpha|d\left(t\right)\right)}\mathrm{d}\alpha = K.$$

As K is proportional to KL, we will further deal with K which is called Kerridge inaccuracy. The minimization of K is equivalent to the minimization of KL.

After substitution of $\tilde{f}\left(\alpha|d\left(t\right)\right)$ and $f\left(\alpha|d\left(t\right)\right)$, we obtain

$$K = \sum_{c_t\in c^*} f\left(d_t|c_t, \mathrm{d}\left(t-1\right)\right)\sum_{c_{t-1}\in c^*} f\left(c_{t-1}|\mathrm{d}\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}\int_{\alpha^*}\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)$$
$$\times\ln\frac{B\left(\nu_t\right)}{\prod_i\prod_j\alpha_{i|j}^{\nu_{i|j;t}-1}}\,d\alpha. \quad (30)$$

In the expression (30) one has to evaluate integrals

$$\int_{\alpha^*} \ln\left(\alpha_{i|j}\right) \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right) \mathrm{d}\alpha = \Xi\left(\nu_{i|j;t-1}^{c_t|c_{t-1}}\right),$$

where the function Ξ is defined in (21). The derivation can be found e. g. in [10].

Now, the expression K to be minimized with respect to ν_t takes the form

$$K = \ln\left[B\left(\nu_t\right)\right]\sum_{c_t\in c^*} f\left(d_t|c_t, d\left(t-1\right)\right)\sum_{c_{t-1}\in c^*}\tilde{f}\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}$$

$$-\sum_{c_t\in c^*} f\left(d_t|c_t, d\left(t-1\right)\right)\sum_{c_{t-1}\in c^*}\tilde{f}\left(c_{t-1}|d\left(t-1\right)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}\sum_{i,j\in c^*}\left(\nu_{i|j;t}-1\right)\Xi\left(\nu_{i|j;t-1}^{c_t|c_{t-1}}\right).$$

The zero point of the derivative is searched for. For the derivative the following relation holds

$$\frac{\partial K}{\partial \nu_{i|j;t}} = \Xi\left(\nu_{i|j;t}\right) G_{t-1} - H_{i|j;t-1}, \quad i,j \in c^*,$$

where G_{t-1} and $H_{i|j;t-1}$ are given in (18). In the differentiation, the formula

$$\frac{\partial}{\partial \nu_{i|j}} \ln B\left(\nu\right) = \Xi\left(\nu_{i|j}\right), \quad i,j \in c^*$$

has been used. The proof can be found in [10].

The necessary condition for an extreme is a zero gradient. Here, fulfilling the conditions $\frac{\partial K}{\partial \nu_{i|j;t}} = 0$ for all $i,j \in c^*$ is required. However, if the function to be minimized, is convex and it is defined on a convex domain, the extreme is global minimum, see [20]. It can be proved that this is the case dealt with here – see Section 7.3 of this Appendix. This guarantees a quick and smooth search for the minimum of the Kerridge inaccuracy K_2 (30). The search can be done with the help of standard numerical subroutines. For instance, in Matlab the subroutine `fsolve` has been successfully used. About 3 or 4 iterations were needed for obtaining satisfactory results.

7.3. Convexity of the Kullback–Leibler divergence

The Kullback–Leibler divergence (17) is used for approximation of the part of posterior pdf, which loses its form during the estimation and must be restored to its original Dirichlet form. For this special case, with the approximate pdf being conjugated to the exponential family of pdfs [10], the Kullback–Leibler divergence is a convex function of the optimized parameter. This fact is exploited when looking for a minimum of this function. The search is not only quick and straightforward, but it leads to the global minimum. The uniqueness of the minimum is not guaranteed, however, any global minimizer suits the problem addressed in this paper.

In the following paragraphs, the claimed convexity is demonstrated.

7.4. Exponential family of pdfs

The pdf $f\left(d_t|d\left(t-1\right),\alpha\right)$ belongs to the exponential family if it can be written in the form

$$f\left(d_t|d\left(t-1\right),\alpha\right)\propto\exp\left\{\left\langle C\left(\alpha\right),B(d(t))\right\rangle\right\}, \tag{31}$$

where $C\left(\alpha\right)$ is a function only of the parameter α (not the data $d\left(t\right)=\left[d_t,\,d\left(t-1\right)\right]$) and $B(d(t))$ is just the function of $d(t)$ and not a function of α. Here, $\langle\cdot\rangle$ denotes the scalar product (in our case with column vectors C, B $\langle C,B\rangle=C'B$, where $'$ means the transposition).

This family possesses the so-called conjugate prior pdf of the form

$$f(\alpha|V)=\frac{\exp\left\langle C\left(\alpha\right),V\right\rangle}{\int_{\alpha^*}\exp\left\{\left\langle C\left(\alpha\right),V\right\rangle\right\}\mathrm{d}\alpha} \tag{32}$$

determined by a statistic V.

7.4.1. Dirichlet pdf as a member of the exponential family

The Dirichlet distribution $\mathcal{D}_\alpha\left(\nu\right)$ is a member of the exponential family pdf, i. e., it can be given the form

$$f\left(\alpha|\nu\right)=\frac{1}{B\left(\nu\right)}\prod_{i\in c*}\prod_{j\in c^*}\alpha_{i|j}^{\nu_{i|j}-1}=\frac{\exp\left\{\left\langle C\left(\alpha\right),V\right\rangle\right\}}{\int_{\alpha^*}\exp\left\{\left\langle C\left(\alpha\right),V\right\rangle\right\}\mathrm{d}\alpha}=f(\alpha|V),$$

where

$$C\left(\alpha\right)=\left[\ln\alpha_{1|1},\,\ln\alpha_{1|2},\,\ldots,\ln\alpha_{1|n},\,\ln\alpha_{2|1},\,\ldots,\,\ln\alpha_{n|n}\right]',$$

$$V=\left[\nu_{1|1}-1,\,\nu_{1|2}-1,\,\ldots,\,\nu_{1|n}-1,\,\nu_{2|1}-1,\,\ldots,\,\nu_{n|n}-1\right]',$$

and $\int_{\alpha^*}\exp\left\{\left\langle C\left(\alpha\right),V\right\rangle\right\}\mathrm{d}\alpha=B\left(\nu\right).$

7.4.2. Kullback–Leibler divergence for an exponential family pdf

The proof is performed for the Kerridge inaccuracy

$$K\left(\tilde{f}(\cdot)\parallel f(\cdot)\right)=\int\tilde{f}(\cdot)\frac{1}{f(\cdot)}\mathrm{d}\cdot$$

which is proportional to KL divergence and fully expresses the properties of the exponential family.

Consider $f_\alpha\left(V\right)$ as a conjugate pdf to the exponential family

$$f_\alpha\left(V\right)=\frac{\exp\left\{\left\langle C\left(\alpha\right),V\right\rangle\right\}}{\int_{\alpha^*}\exp\left\{\left\langle C\left(\tau\right),V\right\rangle\right\}\mathrm{d}\tau},$$

where $C\left(\alpha\right)$ is the parametric function, V is the data function. Let a pdf $\tilde{f}_\alpha$ be arbitrary pdf of the parameter α. Then the Kerridge inaccuracy

$$K=\int_{\alpha^*}\tilde{f}_\alpha\ln\frac{1}{f_\alpha\left(V\right)}\mathrm{d}\alpha$$

is a convex function in the vector V.

The convexity is shown by inserting the considered conjugate form into the Kerridge inaccuracy and demonstrating that the matrix of second derivatives is positive semi-definite. It holds

$$\begin{aligned} K &= \int_{\alpha^*} \tilde{f}_\alpha \ln \frac{1}{\frac{\exp\{\langle C(\alpha),V\rangle\}}{\int_{\alpha^*}\exp\{\langle C(\tau),V\rangle\}\,\mathsf{d}\tau}} \mathrm{d}\alpha \\ &= -\int_{\alpha^*} \tilde{f}_\alpha \left[\ln\left(\exp\left\{\langle C\left(\alpha\right),V\rangle\right\}\right) - \ln\int_{\alpha^*} \exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau\right]\mathrm{d}\alpha \\ &= -\int_{\alpha^*} \tilde{f}_\alpha \langle C\left(\alpha\right),V\rangle \mathrm{d}\alpha + \ln\int_{\alpha^*}\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau \\ &= -\left\langle \int_{\alpha^*} C\left(\alpha\right),V\right\rangle \mathrm{d}\alpha + \ln\int_{\alpha^*}\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau. \end{aligned}$$

Form of both the needed derivatives is presented for the scalar V. The evaluation of the vector V is the same. Just squares x^2 have to be replaced by dyadic product xx'.

$$\frac{\partial K}{\partial V} = -\int_{\alpha^*} C\left(\alpha\right)\,\mathsf{d}\alpha + \frac{1}{\int_{\alpha^*}\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau}\int_{\alpha^*} C\left(\tau\right)\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau.$$

The matrix of the second derivatives $\frac{\partial^2 K}{\partial V^2}$

$$\frac{\int_{\alpha^*}\left[C\left(\tau\right)\right]^2\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau\int_{\alpha^*}\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau - \left[\int_{\alpha^*}C\left(\tau\right)\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau\right]^2}{\left[\int_{\alpha^*}\exp\left\{\langle C\left(\tau\right),V\rangle\right\}\mathrm{d}\tau\right]^2}$$

$$= \int_{\alpha^*}\left[C\left(\tau\right)\right]^2\frac{\exp\left\{\langle C\left(\tau\right),V\rangle\right\}}{\int_{\alpha^*}\exp\left\{\langle C\left(\vartheta\right),V\rangle\right\}\mathrm{d}\vartheta}\mathrm{d}\tau - \left[\int_{\alpha^*}C\left(\tau\right)\frac{\exp\left\{\langle C\left(\tau\right),V\rangle\right\}}{\int_{\alpha^*}\exp\left\{\langle C\left(\vartheta\right),V\rangle\right\}\mathrm{d}\vartheta}\,\mathrm{d}\tau\right]^2$$

$$= E\left[C^2\right] - \left(E\left[C\right]\right)^2 = \mathrm{cov}\left(C\right),$$

where the expectation E is taken over the conjugate pdf.

As the covariance $\mathrm{cov}(C)$ is semi-definite, the Kerridge inaccuracy as well as the Kullback–Leibler divergence are convex functions.

7.5. Estimation of variance of α

The standard estimation of discrete systems consists in recomputing the statistics by adding one to such its item that corresponds to the actual state of the system. Thus, the items of the statistics cannot descend. The statistics of the estimation method, presented during its evolution, can not only grow or stagnate, it can even fall. Thus, the estimation gives not only the point estimates of the parameter α but also its variance which determines the strength of the belief in the estimates. The result reads: the bigger is the sum of all statistics items belonging to a specific component, the smaller is the variance of component parameters and the more confidence they have.

In the following paragraph, the relation between the variance of α-estimates and the count of data belonging to individual components is demonstrated.

In [10] it is proved, that the Dirichlet pdf $\mathcal{D}_\alpha(\nu_t)$ has the following expectation and variance

$$E\left[\alpha_{i|j}|\nu_t, d(t)\right] = \frac{\nu_{i|j;t}}{\sum_i \nu_{i|j;t}},$$

$$\operatorname{var}(\alpha|\nu_t, d(t)) = E\left[\alpha_{i|j}|\nu_t, d(t)\right]^2 \frac{\nu_{i|j;t}^{-1} - \left(\sum_i \nu_{i|j;t}\right)^{-1}}{1 + \left(\sum_i \nu_{i|j;t}\right)^{-1}}.$$

Now, for a fixed j dealing with α_i, we have

$$\operatorname{var}(\alpha|\nu_t, d(t)) = \left(\frac{\nu_{i;t}}{\sum_i \nu_{i;t}}\right)^2 \frac{\nu_{i;t}^{-1} - \left(\sum_i \nu_{i;t}\right)^{-1}}{1 + \left(\sum_i \nu_{i;t}\right)^{-1}}$$

$$= \frac{\nu_{i;t}^2 \left[\nu_{i;t}^{-1} - \left(\sum_i \nu_{i;t}\right)^{-1}\right]}{\left(\sum_i \nu_{i;t}\right)^2 \left[1 + \left(\sum_i \nu_{i;t}\right)^{-1}\right]} = \frac{\nu_{i;t}\left(1 - \hat{\alpha}_{i;t}\right)}{\sum_i \nu_{i;t}\left(\sum_i \nu_{i;t} + 1\right)} = \frac{\hat{\alpha}_{i;t}\left(1 - \hat{\alpha}_{i;t}\right)}{\left(\sum_i \nu_{i;t} + 1\right)}. \tag{33}$$

From this result, it can be seen, that the variance of a particular component is inversely proportional to the sum $\sum_i \nu_{i;t}$ representing the number of how many times the component was active during the estimation.

ACKNOWLEDGEMENTS

This research has been partially supported by grant MŠMT 1M0572, general agreement ÚTIA and Škoda Auto No. ENS/2009/UTIA, projects GAČR 102/08/0567 and TAČR TA01030123.

(Received February 5, 2010)

REFERENCES

[1] J. M. Bernardo: Expected information as expected utility. Ann. Statist. *7* (1979), 3, 686–690.

[2] J. Böhm and M. Kárný: Transformation of user's knowledge into initial values for identification. In: Preprints DYCOMANS Workshop Industrial Control and Management Methods: Theory and Practice (M. Součková and J. Böhm, eds.), ÚTIA AV ČR, Prague 1995, pp. 17–24.

[3] W. Chen and P. Jovanis: Method for identifying factors contributing to driver-injury severity in traffic crashes. Highway And Traffic Safety: Crash Data, Analysis Tools, And Statistical Methods *1717* (2000), 1–9.

[4] G. Chinnaswamy, E. Chirwa, S. Nammi, S. Nowpada, T. Chen, and M. Mao: Benchmarking and accident characteristics of flat-fronted commercial vehicles with respect to pedestrian safety. Internat. J. Crashworthiness *12* (2007), 279–291.

[5] A. P. Dempster, N. Laird, and D. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B (Methodological) *39* (1977), 1, 1–38.

[6] J. D. Hamilton and R. Susmel: Autoregressive conditional heteroskedasticity and changes in regime. J. Econometrics *64* (1994), 307–333.

[7] S. Haykin: Neural Networks: A Comprehensive Foundation. MacMillan, New York 1994.

[8] Jianyong Wang, Yuzhou Zhang, Lizhu Zhou, G. Karypis and Charu C. Aggarwal: Contour: An efficient algorithm for discovering discriminating subsequences. In: Data Mining and Knowledge Discovery, Springer *18* (2009), 1, pp. 1–29.

[9] M. Kárný: Tools for computer-aided design of adaptive controllers. IEE Control Theory Appl. *150* (2003), 6, 643.

[10] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London 2005.

[11] M. Kárný, J. Kadlec, and E. L. Sutanto: Quasi-Bayes estimation applied to normal mixture. in In: Preprints 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing (J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, eds.), ÚTIA AV ČR, Prague 1998, pp. 77–82.

[12] M. Kárný, I. Nagy, and J. Novovičová: Mixed-data multi-modelling for fault detection and isolation. Internat. J. Adaptive Control Signal Process. *16* (2002), 1, 61–83.

[13] D. F. Kerridge: Inaccuracy and Inference. J. Royal Statist. Soc. Ser. B (Methodological) *23* (1961), 1, 184–194.

[14] R. Kulhavý: A Bayes-closed approximation of recursive non-linear estimation. Internat. J. Adaptive Control Signal Process. *4* (1990), 271–285.

[15] L. Ljung: System Identification: Theory for the User. Prentice-Hall, London 1987.

[16] R. Murray-Smith and T. Johansen: Multiple Model Approaches to Modelling and Control. Taylor &Francis, London 1997.

[17] A. Oppenheim and A. Wilsky: Signals and Systems. Englewood Clifts, Jersey 1983.

[18] M. Opper and D. Saad: Advanced Mean Field Methods: Theory and Practice. The MIT Press, Cambridge 2001.

[19] H. B. Qu and B. G. Hu: Variational learning for Generalized Associative Functional Networks in modeling dynamic process of plant growth. Ecological Informatics *4* (2009), 3, 163–176.

[20] W. A. Roberts: Convex Functions. Academic Press, New York 1973.

[21] J. Sander, M. Ester, H.-P. Kriegel and X. Xu: Density-based clustering in spatial databases: The algorithm gdbscan and its applications. In: Data Mining and Knowledge Discovery, Springer, *2* (1998), 2, pp. 169–194.

[22] D. Titterington, A. Smith, and U. Makov: Statistical Analysis of Finite Mixtures. John Wiley, New York 1985.

[23] Xiaowei Xu, J. Jäger and H.-P. Kriegel: A fast parallel clustering algorithm for large spatial databases. In: Data Mining and Knowledge Discovery, Springer, *3* (1999), 3, pp. 263–290.

[24] T. Zhang, R. Ramakrishnan and M. Livny: Birch: A new data clustering algorithm and its applications. In: Data Mining and Knowledge Discovery, Springer, *1* (1997), 2, pp. 141–182.

Ivan Nagy, Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 110 00 Praha 1. Czech Republic.
e-mail: nagy@utia.cas.cz

Evgenia Suzdaleva, Department of Adaptive Systems, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: suzdalev@utia.cas.cz

Miroslav Kárný, Department of Adaptive Systems, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: school@utia.cas.cz