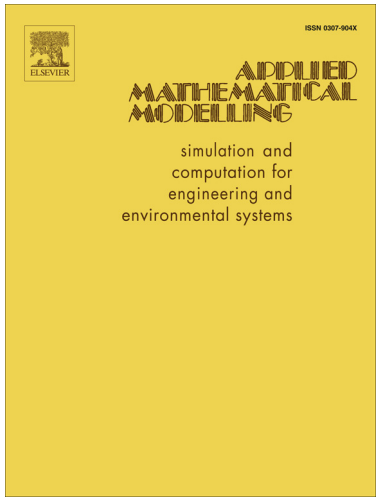

Mixture Estimation with State-Space Components and Markov Model of Switching

Ivan Nagy, Evgenia Suzdaleva

PII: S0307-904X(13)00359-4
DOI: http://dx.doi.org/10.1016/j.apm.2013.05.038
Reference: APM 9528

To appear in: *Appl. Math. Modelling*

Received Date: 14 January 2012
Revised Date: 22 May 2013
Accepted Date: 31 May 2013

Please cite this article as: I. Nagy, E. Suzdaleva, Mixture Estimation with State-Space Components and Markov Model of Switching, *Appl. Math. Modelling* (2013), doi: http://dx.doi.org/10.1016/j.apm.2013.05.038

Mixture Estimation with State-Space Components and Markov Model of Switching

Ivan Nagy[a,b], Evgenia Suzdaleva[b]

[a] *Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 11000 Prague, Czech Republic*
[b] *Department of Adaptive Systems, Institute of Information Theory and Automation of the ASCR, Pod vodárenskou věží 4, 18208 Prague, Czech Republic*

Abstract

The paper proposes a recursive algorithm for estimation of mixtures with state-space components and a dynamic model of switching. Bayesian methodology is adopted. The main features of the presented approach are: (i) recursiveness that enables a real-time performance of the algorithm; (ii) one-pass elaboration of the data sample; (iii) dynamic nature of the model of switching active components; (iv) orientation at explicit solutions with exploitation of numerical procedures only in those parts which cannot be computed analytically; (v) systematic approach to the Bayesian mixture estimation theory.

Keywords: probabilistic dynamic mixtures, probability density function, state-space models, recursive mixture estimation, Bayesian dynamic decision making under uncertainty, Kerridge inaccuracy

1. Introduction

Dynamic systems modifying their behavior by switching several regimes are met in various application areas (industry, medicine, economics, traffic control, etc.). These regimes often differ a lot from each other and then a system must be described by a mixture of several models – *components*. Mixture models are known to be universal approximations for description of such systems [1]. Switching of the active components (which corresponds to switching of the active system regimes) is often described via hidden Markov models (HMM) theory, see e.g., [2].

Within the bounds of individual regimes an unobservable variable can appear – a system state. When the state variable is to be estimated, each regime should be described by a state-space model [3], and the system is modeled by a mixture of state-space components. Estimation of mixtures of state-space models is closely related to tasks of clustering and classification. The developed algorithms can find their application in the form of on-line advising systems which evaluate the active regime and inform an operator or advise a control action.

The following approaches can be found in the area of mixture estimation with state-space components. A series of algorithms is based on Variational Bayes (VB) methods [4, 5, 6, 7] that propose variational approximations to maximize the lower bound of the likelihood. An alternative approach found for switching state-space models is numerical iterative techniques

Email addresses: nagy@utia.cas.cz (Ivan Nagy), suzdalev@utia.cas.cz (Evgenia Suzdaleva)

based on Markov Chain Monte Carlo (MCMC) methods [8, 9, 10, 11]. However, the mentioned algorithms rely on completely numerical solutions and are not fully on-line.

In this paper, we look for a systematic approach to the mixture estimation theory with possibility of further development of algorithms and their applications in practice. We follow a methodology of Bayesian dynamic decision-making under uncertainty adopted in [12] providing the algorithms with high potential for practical applications, see e.g., [13]. Here, we continue a line proposed in [14] and [15] that, using a similar approach, develop the recursive algorithms for estimation of dynamic mixtures of autoregression models and dynamic transition [14] and for state estimation of hybrid systems [15]. The presented paper demonstrates that this unified approach in the frame of the Bayesian methodology can also be extended to the mixture estimation with state-space components and dynamic switching.

The main contributions of the presented algorithm are:

- A dynamic model of switching based on the dependence of the current active component on the previous one. This is a key point for estimation of the active regime of the system.
- A fixed computational complexity during one-pass elaboration of the data sample. It is reached by the fact that the posterior probability density functions (pdfs) are approximately self-reproducing and preserve forms of the prior pdfs. This can be decisive for on-line applications or working with extensive databases.
- A universal real-time easily computable approximation based on the Kerridge inaccuracy [16]. The Kerridge inaccuracy is a part of the Kullback-Leibler divergence [17] which, in the form different from that used in VB methods, is known to be an optimal tool within the adopted Bayesian methodology (see proof in [18]). This universality is significant for further development of the algorithms.
- Explicit solutions with the exception of the Kerridge inaccuracy.

Layout of the paper is as follows. Section 2 formulates a problem. Section 3 provides basic known facts about state-space models, state estimation and Markov model estimation. It also introduces a mixture model. Section 4 constructs the joint pdf of all observed and unknown variables, decomposes it into the models and the prior pdfs and introduces the independence assumptions. Section 5 is devoted to the algorithm derivation. Approximation that is necessary to complete the proposed solution is presented in Section 6. The resulting estimation algorithm is proposed in Section 7. Section 8 provides illustrative experiments and comparison with theoretical counterparts. Conclusion can be found in Section 9. Detailed derivations of the proposed formulas are available in Appendix 10.

2. Problem formulation

Let us consider a system which produces an observable output variable d_t, a state variable x_t which cannot be measured and (optionally) an input variable u_t at discrete time instants $t = \{1, \ldots, T\} \equiv t^*$. Let this system work in several randomly switching regimes. These regimes are supposed to be so different, that it is not possible to describe them all by a single model. A mixture of state-space models is an appropriate tool for description of such a system.

The *main task addressed in this paper* is to estimate recursively the unobserved state x_t and the active regime of the system in dependence on the previous one. Thus, the fully dynamic state-space mixture estimation will be considered.

3. Preliminaries

Let us first recollect a state-space model describing a single component and a standard way for the state estimation.

3.1. State-space model

The state-space model is composed of two parts. The first one, the state model, describes a time evolution of the state. The second part, the observation model, determines how measurements are related to the state. We will assume these models in the form

$$f\left(x_t|x_{t-1}\right) \quad \text{and} \quad f\left(d_t|x_t\right), \tag{1}$$

where $f(\cdot|\cdot)$ is a conditional probability (density) function denoted by pdf throughout this paper, x_t and x_{t-1} are the actual and the past states and d_t is the observation. The input variable u_t is omitted here for brevity reasons and its presence in both the models brings no complication.

3.2. General solution to state estimation

General probabilistic solution of Bayesian filtering [19] is used to estimate the unobserved state x_t. Denoting a set of measurements by $D(t) = (d_1, \ldots, d_t)$, we can express a prior (initial) state pdf at time t as $f\left(x_{t-1}|D(t-1)\right)$. Bayesian filtering takes this prior pdf, predicts for the time instant t and corrects the pdf by incorporation of actually measured data. Then it results in the updated posterior state pdf $f\left(x_t|D(t)\right)$ keeping its initial form. This recursion is easily derived with the help of construction of the joint pdf $f(d_t, x_t|D(t-1))$ of data and the unknown state, application of Bayes rule (see Appendix 10) and an operation of marginalization:

$$f(x_t|D(t)) \;=\; \frac{f(d_t, x_t|D(t-1))}{\int_{x^*} f(d_t, x_t|D(t-1))dx_t} = \frac{\int_{x^*} f(d_t, x_t, x_{t-1}|D(t-1))dx_{t-1}}{\int_{x^*} f(d_t, x_t|D(t-1))dx_t}, \tag{2}$$

then by decomposition via the chain rule (see Appendix 10) and making the independence assumptions [19]:

$$f(x_t|D(t)) \;=\; \frac{\int_{x^*} f(d_t|x_t) f(x_t|x_{t-1}) f(x_{t-1}|D(t-1))dx_{t-1}}{\int_{x^*} f\left(d_t|x_t\right) \int_{x^*} f\left(x_t|x_{t-1}\right) f\left(x_{t-1}|D(t-1)\right) dx_{t-1}dx_t}, \tag{3}$$

where x^* is a set of all possible values of x_t, $\forall t$, and the denominator represents the data prediction:

$$\int_{x^*} f\left(d_t|x_t\right) \int_{x^*} f\left(x_t|x_{t-1}\right) f\left(x_{t-1}|D(t-1)\right) dx_{t-1}dx_t = f\left(d_t|D(t-1)\right). \tag{4}$$

The obtained recursion (3) consists of models (1) and the prior pdf $f\left(x_{t-1}|D(t-1)\right)$, whose form should be preserved. This general approach is used throughout the presented paper. The recursion starts with the prior pdf $f\left(x_0|D(0)\right)$, which expresses the subjective prior knowledge about the initial state.

For a normal prior pdf and linear normal models (1) Bayesian filtering coincides with Kalman filter [20, 19, 3].

3.3. Kalman filter

Normal linear models (1) take a form

$$f\left(x_t|x_{t-1}\right) = \mathcal{N}_{x_t}\left(\underbrace{Ax_{t-1}}_{mean}, \overbrace{R_w}^{variance}\right), \qquad f\left(d_t|x_t\right) = \mathcal{N}_{d_t}\left(Cx_t, R_v\right), \tag{5}$$

where $\mathcal{N}$ denotes normal distribution of a variable; A, C are parameters supposed to be known of appropriate dimensions; R_w and R_v are process and measurement noise covariance matrices respectively assumed to be known and time-invariant. The prior state pdf is also chosen as normal distribution, i.e., for $t = 1$

$$f\left(x_{t-1}|D(t-1)\right) = \mathcal{N}\left(\xi_{t-1|t-1}, R_{t-1|t-1}\right) \tag{6}$$

with mean $\xi_{t-1|t-1}$ and covariance matrix $R_{t-1|t-1}$.

The Kalman filter [3] with models (5) and the prior distribution (6) includes the following equations:

Time updating

$$\begin{aligned} R_{t|t-1} &= R_w + AR_{t-1|t-1}A', & (7)\\ \xi_{t|t-1} &= A\xi_{t-1|t-1}, & (8) \end{aligned}$$

Data updating

$$\begin{aligned} R_y &= R_v + CR_{t|t-1}C', & (9)\\ R_{t|t} &= R_{t|t-1} - R_{t|t-1}C'R_y^{-1}CR_{t|t-1}, & (10)\\ K_G &= R_{t|t}C'R_v^{-1}, & (11)\\ \xi_{t|t} &= \xi_{t|t-1} + K_G\left(d_t - C\xi_{t|t-1}\right), & (12) \end{aligned}$$

where $\xi_{t|t}$ and $R_{t|t}$ determine the resulting normal posterior pdf $f\left(x_t|D(t)\right)$.

The data predictive pdf (4) in this case is denoted by $\mathcal{L}_{d_t}$ and computed as

$$f\left(d_t|D(t-1)\right) = (2\pi)^{-\frac{Y}{2}}|R_y|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}[d_t - C\xi_{t|t}]'R_y^{-1}[d_t - C\xi_{t|t}]\right\} \equiv \mathcal{L}_{d_t}, \tag{13}$$

where Y is a dimension of the vector d_t.

3.4. Model of switching

In this paper, a mixture model consists of a set of n_c state-space components (1)

$$\left\{f\left(x_t|x_{t-1}, c\right), \quad f\left(d_t|x_t, c\right)\right\}_{c=1}^{n_c}, \tag{14}$$

where c labels individual components. In order to describe a switching of the components, this label is introduced as a discrete random process

$$\left\{c_t\right\}_{t=1}^{T}, \;\; c_t = \{1, 2, \ldots, n_c\} \equiv c^*, \tag{15}$$

which is called a *pointer*. Its realization c_t at each time instant t points at the active component that corresponds to the active regime of the system.

A dynamic pointer can be described by a Markov model defined by the conditional pdf

$$f\left(c_t|c_{t-1},\alpha\right)=\alpha_{c_t|c_{t-1}}, \tag{16}$$

where $\alpha_{c_t|c_{t-1}}$ is a transition probability, $c_t|c_{t-1}$ is a multi-index and it holds

$$\alpha_{c_t|c_{t-1}}\in\alpha^*=\left\{\alpha_{c_t|c_{t-1}}\geq 0,\sum_{c\in c^*}\alpha_{c|c_{t-1}}=1,\,\forall c_t,c_{t-1}\in c^*,\,\forall t\in t^*\right\}. \tag{17}$$

3.5. Markov model estimation

In the case of a known active component, according to [12], model (16) can be estimated in the analytical way using the conjugate prior pdf $f\left(\alpha|D(t-1)\right)$ in the Dirichlet form, i.e.,

$$f\left(\alpha|D(t-1)\right)=\mathcal{D}_\alpha\left(\nu_{t-1}\right)=\frac{1}{B(\nu_{t-1})}\prod_{i|j}\alpha_{i|j}^{\nu_{i|j;t-1}-1}, \tag{18}$$

where $i|j$ is a multi-index with $i,j\in c^*$, and $\nu_{i|j;t-1}$ is a prior statistics of estimation for time t and $B\left(\nu_{t-1}\right)$ is a normalization constant, which has the form of multivariate beta function [12]

$$B(\nu)=\prod_{j\in c^*}\frac{\prod_{i\in c^*}\Gamma\left(\nu_{i|j}\right)}{\Gamma\left(\sum_{i\in c^*}\nu_{i|j}\right)}. \tag{19}$$

For given c_t and c_{t-1}, the statistics entries $\nu_{i|j;t}$ evolve in time according to the formula

$$\nu_{i|j;t}=\nu_{i|j;t-1}+\delta\left(c_t|c_{t-1};i|j\right),\quad i,j\in c^*,\quad t=1,2,\cdots, \tag{20}$$

where δ is Kronecker delta such that $\delta\left(c_t|c_{t-1};i|j\right)=1$ for $c_t=i$ and $c_{t-1}=j$ and equals to zero otherwise. Using statistics (20), the point estimates of $\alpha_{i|j}$ can be computed as follows:

$$\hat{\alpha}_{i|j;t}\equiv\int_{\alpha^*}\alpha_{i|j}f\left(\alpha|D(t)\right)d\alpha=\frac{\nu_{i|j;t}}{\sum_i\nu_{i|j;t}}. \tag{21}$$

These preliminaries provide the well known solutions for the known active components. In the case of the unknown pointer value c_t the presented solutions lose self-reproductivity of prior pdfs and become unfeasible due to arising sums in posterior pdfs (it will be explained later).

However, it can be shown that applying Bayes rule, just similar as for recursions (3) and (4), we can derive an algorithm of recursive estimation of the state-space mixture. Thus, the main steps of the algorithm derivation will include: (i) construction of the joint pdf of data and all unknown variables; (ii) its decomposition according to the chain rule; (iii) application of assumptions about the conditional independence; (iv) expression of evolution of the joint pdf in time, using the models and the prior pdfs; (v) marginalization of the joint pdf; (vi) necessary approximations to obtain the posterior pdfs.

This problem solved in subsequent sections is the main contribution of the paper.

4. Joint pdf construction

The system is described by the mixture of the state-space components (14). Throughout this paper, parameters of the state-space components are supposed to be known. The unknown variables to be estimated are: the n-dimensional state x_t, values of the pointer process c_t and the transition table α whose entries $\alpha_{i|j}$ are the stationary probabilities of switching from the jth component to the ith one.

To express the models and the forms of the prior pdfs we have to construct the joint pdf of all the involved variables $d_t, x_t, x_{t-1}, c_t, c_{t-1}, \alpha$. It takes the form

$$f(d_t, x_t, x_{t-1}, c_t, c_{t-1}, \alpha | D(t-1)) \tag{22}$$

that can be factorized via the chain rule as

$$\begin{aligned} &f\left(d_t|x_t, x_{t-1}, c_t, c_{t-1}, \alpha, D(t-1)\right) f\left(x_t|x_{t-1}, c_t, c_{t-1}, \alpha, D(t-1)\right) f\left(x_{t-1}|c_t, c_{t-1}, \alpha, D(t-1)\right) \\ &\times f\left(c_t|c_{t-1}, \alpha, D(t-1)\right) f\left(c_{t-1}|\alpha, D(t-1)\right) f\left(\alpha|D(t-1)\right) \end{aligned} \tag{23}$$

and then results in the form

$$\underbrace{f\left(d_t|x_t, c_t\right) f\left(x_t|x_{t-1}, c_t\right)}_{state-space\ model} \underbrace{f\left(x_{t-1}|D(t-1)\right)}_{prior\ state\ pdf} \underbrace{f\left(c_t|c_{t-1}, \alpha\right)}_{pointer\ model} \underbrace{f\left(c_{t-1}|D(t-1)\right) f\left(\alpha|D(t-1)\right)}_{prior\ pdfs\ for\ pointer\ and\ for\ \alpha}, \tag{24}$$

where the following independence assumptions are made:

$$f\left(d_t|x_t, x_{t-1}, c_t, c_{t-1}, \alpha, D(t-1)\right) = f\left(d_t|x_t, c_t\right), \tag{25}$$

$$f\left(x_t|x_{t-1}, c_t, c_{t-1}, \alpha, D(t-1)\right) = f\left(x_t|x_{t-1}, c_t\right), \tag{26}$$

$$f\left(x_{t-1}|c_t, c_{t-1}, \alpha, D(t-1)\right) = f\left(x_{t-1}|D(t-1)\right), \tag{27}$$

$$f\left(c_t|c_{t-1}, \alpha, D(t-1)\right) = f\left(c_t|c_{t-1}, \alpha\right), \tag{28}$$

$$f\left(c_{t-1}|\alpha, D(t-1)\right) = f\left(c_{t-1}|D(t-1)\right) \tag{29}$$

that omits independent variables from the conditions of the involved pdfs.

5. Recursive estimation of the state-space mixture

To derive the estimation algorithm the joint pdf of the estimated variables should be constructed.

5.1. Joint pdf of estimated variables

The joint pdf of the estimated variables $f\left(x_t, c_t, \alpha|D(t)\right)$ is derived from (24) by summation over c_{t-1} and integration over x_{t-1} and with the help of (3) and (4):

$$\begin{aligned} f\left(x_t, c_t, \alpha|D(t)\right) \propto\ & \underbrace{f\left(d_t|x_t, c_t\right) \int_{x^*} f\left(x_t|x_{t-1}, c_t\right) f\left(x_{t-1}|D(t-1)\right) dx_{t-1}}_{f(d_t, x_t|c_t, D(t-1)) \propto f(x_t|c_t, D(t))} \\ &\times \sum_{c_{t-1} \in c^*} f\left(c_t|c_{t-1}, \alpha\right) f\left(c_{t-1}|D(t-1)\right) f\left(\alpha|D(t-1)\right). \end{aligned} \tag{30}$$

According to [12] and recalling (16), we can recompute a product of the pointer model $f\left(c_t|c_{t-1},\alpha\right)$ and the Dirichlet prior pdf $f\left(\alpha|D(t-1)\right)$ entering (30) as

$$f\left(c_t|c_{t-1},\alpha\right)f\left(\alpha|D(t-1)\right)=\alpha_{c_t|c_{t-1}}\mathcal{D}_\alpha\left(\nu_{t-1}\right)=\hat{\alpha}_{c_t|c_{t-1};t-1}\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right), \tag{31}$$

where

- ν_{t-1} is a $(n_c\times n_c)$-matrix with non-negative items $\nu_{c_t|c_{t-1};t-1}$,
- $\hat{\alpha}_{c_t|c_{t-1};t-1}$ is a point estimate of $\alpha_{c_t|c_{t-1}}$ (21)
- and $\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)$ is a Dirichlet distribution of the same form as $\mathcal{D}_\alpha\left(\nu_{t-1}\right)$ but with the argument

$$\nu_{i|j;t-1}^{c_t|c_{t-1}}=\nu_{i|j;t-1}+\delta\left(c_t|c_{t-1};i|j\right), \tag{32}$$

  where the superscript stresses the dependence on the component labels c_t and c_{t-1}.

Note that this update is similar to (20) but it is not definitive, because the summation in the joint pdf (30) destroys the original form: (30) now takes the form

$$f\left(x_t,c_t,\alpha|D(t)\right)=f\left(d_t,x_t|c_t,D(t-1)\right)\sum_{c_{t-1}\in c^*}\hat{\alpha}_{c_t|c_{t-1};t-1}f\left(c_{t-1}|D(t-1)\right)\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right). \tag{33}$$

The recursive estimation of the unknown variables x_t, c_t and α can be derived via marginalization of the joint pdf (33), which is shown in subsequent sections.

5.2. Estimation of the pointer variable c_t

In order to derive the posterior pdf $f\left(c_t|D(t)\right)$ for estimation of the pointer value c_t, the joint pdf (33) is marginalized in the following way:

$$\begin{aligned}
&f\left(c_t|D(t)\right)\\
&\propto\int_{x^*}\int_{\alpha^*}\underbrace{f\left(d_t,x_t|c_t,D(t-1)\right)\sum_{c_{t-1}\in c^*}\hat{\alpha}_{c_t|c_{t-1};t-1}f\left(c_{t-1}|D(t-1)\right)\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)}_{(33)}d\alpha dx_t\\
&=\sum_{c_{t-1}\in c^*}\underbrace{f\left(d_t|c_t,D(t-1)\right)\hat{\alpha}_{c_t|c_{t-1};t-1}f\left(c_{t-1}|D(t-1)\right)}_{\textit{denoted by }\bar{w}_{c_t|c_{t-1}}}\propto\sum_{c_{t-1}\in c^*}w_{c_t|c_{t-1}}=w_{c_t},
\end{aligned} \tag{34}$$

where $\propto$ means proportionality,

$$w_{c_t|c_{t-1}}=\frac{\bar{w}_{c_t|c_{t-1}}}{\sum_{c_t\in c^*}\sum_{c_{t-1}\in c^*}\bar{w}_{c_t|c_{t-1}}} \tag{35}$$

and where the integration over x_t gives the marginal pdf (it coincides with the data prediction (4) for the cth component), and the integration over α produces 1. The prior pdf $f\left(c_{t-1}|D(t-1)\right)=w_{c_{t-1}}$ and the posterior one $f\left(c_t|D(t)\right)=w_{c_t}$ are just numeric vectors. Consequently, no special form is necessary to be preserved.

5.3. Estimation of the parameter α

Similarly, the posterior pdf $f(\alpha|D(t))$ for estimation of the parameter α is derived via marginalization of (33), i.e., integration over x_t and summation over c_t. According to the data prediction (4), which is a plain computation of a marginal pdf, it is obtained:

$$\begin{aligned} & f(\alpha|D(t)) \\ & \propto \sum_{c_t\in c^*} \int_{x^*} \underbrace{f(d_t, x_t|c_t, D(t-1)) \sum_{c_{t-1}\in c^*} \hat{\alpha}_{c_t|c_{t-1};t-1} f(c_{t-1}|D(t-1)) \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)}_{(33)} dx_t \\ & = \sum_{c_t\in c^*} \sum_{c_{t-1}\in c^*} \underbrace{f(d_t|c_t, D(t-1)) \hat{\alpha}_{c_t|c_{t-1};t-1} f(c_{t-1}|D(t-1))}_{\bar{w}_{c_t|c_{t-1}}} \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right) \\ & \propto \sum_{c_t\in c^*} \sum_{c_{t-1}\in c^*} w_{c_t|c_{t-1}} \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right) \end{aligned} \tag{36}$$

with $w_{c_t|c_{t-1}}$ obtained in (35). Here it can be seen that relation (36) is a sum of pdfs. Number and complexity of the generated posterior pdfs exponentially grow, and the original Dirichlet form of the prior pdf is destroyed. It means that result (36) is not feasible for a recursive evaluation. Thus an approximation restoring the original form of the prior pdf is necessary. The solution to this problem is presented in Section 6.

5.4. Estimation of the state x_t

The posterior pdf for estimation of the state x_t is evolved according to marginalization of (33) with integration over α and summation over c_t, i.e.,

$$\begin{aligned} & f(x_t|D(t)) \\ & \propto \sum_{c_t\in c^*} \int_{\alpha^*} \underbrace{f(d_t, x_t|c_t, D(t-1)) \sum_{c_{t-1}\in c^*} \hat{\alpha}_{c_t|c_{t-1};t-1} f(c_{t-1}|D(t-1)) \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right)}_{(33)} d\alpha \\ & = \sum_{c_t\in c^*} f(x_t, d_t|c_t, D(t-1)) \sum_{c_{t-1}\in c^*} \hat{\alpha}_{c_t|c_{t-1};t-1} f(c_{t-1}|D(t-1)) \\ & = \sum_{c_t\in c^*} f(x_t|c_t, D(t)) \sum_{c_{t-1}\in c^*} \underbrace{f(d_t|c_t, D(t-1)) \hat{\alpha}_{c_t|c_{t-1};t-1} f(c_{t-1}|D(t-1))}_{\bar{w}_{c_t|c_{t-1}}} \\ & = \sum_{c_t\in c^*} f(x_t|c_t, D(t)) \sum_{c_{t-1}\in c^*} w_{c_t|c_{t-1}} = \sum_{c_t\in c^*} w_{c_t} f(x_t|c_t, D(t)), \end{aligned} \tag{37}$$

where the pdf $f(x_t|c_t, D(t))$ naturally follows from Bayesian filtering (3) performed for the component c_t, and $\sum_{c_{t-1}\in c^*} w_{c_t|c_{t-1}} = w_{c_t}$ is presented in (34). Again, the obtained result produces repetitive products of sums and loses its original form. Such a computation is not recursively feasible and thus an approximation is needed in order to restore the prior form. This solution is discussed in Section 6.

6. Approximation

A valued feature of the suggested estimation algorithms is their real-time performance. The state estimation and its classification into the components run in time as the data are measured. According to the adopted Bayesian approach, the unknown objects (the state x_t, the parameter α and the pointer value c_t) are described through the recursive evolution of their "prior $\rightarrow$ posterior" pdfs. Due to the recursive nature of this evaluation, it is necessary to guarantee their self-reproducibility which means that the prior pdf and the corresponding posterior one are structurally identical. Evolution is allowed only in their numerical characteristics. If this property is not guaranteed, complexity of the posterior pdfs during their evolution grows until they become unfeasible.

The recursive nature of computation appears at all three considered recursions (34), (36) and (37). The pdf of the pointer $f\left(c_t|D(t)\right)$ in (34) is a mere vector and its structure is not disturbed. However, formulas (36) and (37) show that they are not self-reproducing. Thus, the self-reproducing property must be secured only for the state x_t and the parameter α. A remedy applied here is to approximate the arising sums immediately at each step of the estimation and restore the original form of the corresponding prior pdfs.

6.1. Approximation for estimation of α

Let us consider relation (36) from Section 5.3, where the prior pdf $f\left(\alpha|D(t-1)\right)$ should preserve the Dirichlet distribution, but its form is destroyed due to the summation. To restore it, an approximating pdf $\hat{f}\left(\alpha|(D(t)\right)$ is chosen with the Dirichlet distribution

$$\hat{f}\left(\alpha|D(t)\right) = \mathcal{D}_\alpha\left(\nu_t\right), \tag{38}$$

minimizing the Kerridge inaccuracy [16]

$$K\left(f\left(\alpha|D(t)\right) \| \hat{f}\left(\alpha|D(t)\right)\right) = \int_{\alpha^*} f\left(\alpha|D(t)\right) \ln \frac{1}{\hat{f}\left(\alpha|D(t)\right)}\, d\alpha, \tag{39}$$

where $f(\alpha|D(t))$ is computed according to (36):

$$f(\alpha|D(t)) = \sum_{c_t \in c^*} \sum_{c_{t-1} \in c^*} w_{c_t|c_{t-1}} \mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right). \tag{40}$$

The Kerridge inaccuracy [16] is a part of the Kullback-Leibler divergence [17] that in the form

$$\int_{\alpha^*} f\left(\alpha|D(t)\right) \ln \frac{f\left(\alpha|D(t)\right)}{\hat{f}\left(\alpha|D(t)\right)}\, d\alpha$$

(in the considered context) is known to be an optimal tool within the adopted Bayesian approach (see proof in [18]). As during derivation the function in the nominator of the logarithm is reduced and the minimization results both for the Kerridge inaccuracy and the Kullback-Leibler divergence are identical, we can use the first of them.

The approximating pdf (38) minimizing the Kerridge inaccuracy (39) over the statistics $\nu_t = [\nu_{i|j;t}]_{i,j \in c^*}$ is defined by this statistics solving the equation

$$\Xi\left(\nu_{i|j;t}\right) = \sum_{c_t \in c^*} \sum_{c_{t-1} \in c^*} w_{c_t|c_{t-1}} \Xi\left(\nu_{i|j;t-1}^{c_t|c_{t-1}}\right), \tag{41}$$

where $\nu_{i|j;t-1}^{c_t|c_{t-1}}$ are introduced in (32) and $w_{c_t|c_{t-1}}$ in (35), and the following definitions are used:

$$\Xi\left(\nu_{i|j}\right) = \Psi\left(\nu_{i|j}\right) - \Psi\left(\sum_k \nu_{k|j}\right) = \int_{\alpha^*} \ln\left(\alpha_{i|j}\right) \mathcal{D}_\alpha\left(\nu\right)\, d\alpha \tag{42}$$

with the following Ψ function

$$\Psi\left(z\right) = \frac{d}{dz} \ln \Gamma\left(z\right). \tag{43}$$

A numerical solution to equation (41) must be performed at each step of the estimation. However, it causes neither increased complexity nor longer computational time due to the fact, that the minimized function (39) is a convex function of ν_t. Thus, the search for an extreme is straightforward and the extreme found is always the global minimum.

Proof is available in Appendix 10.4.

6.2. Approximation for the state estimation

Here we deal with formula (37) from Section 5.4, where the updated pdf $f(x_t|D(t))$ is a weighted mixture model. In case of its normal components (5), each posterior pdf $f(x_t|c_t, D(t))$ is evolved via the Kalman filter (7)–(12) applied for individual components. The approximation task is to replace this mixture by a single normal pdf $\hat{f}\left(x_t|D(t)\right)$ so that the Kerridge inaccuracy

$$K\left(f\left(x_t|D(t)\right) \parallel \hat{f}\left(x_t|D(t)\right)\right) = \int_{x^*} f\left(x_t|D(t)\right) \ln \frac{1}{\hat{f}\left(x_t|D(t)\right)} dx_t \tag{44}$$

reaches its minimum. With denotations

$$f\left(x_t|c_t, D(t)\right) = \mathcal{N}\left(\xi_{c_t;t|t}, R_{c_t;t|t}\right) \text{ and } \hat{f}\left(x_t|D(t)\right) = \mathcal{N}\left(\hat{\xi}_{t|t}, \hat{R}_{t|t}\right) \tag{45}$$

and w_{c_t} defined in (34) the result of the approximation is

$$\hat{\xi}_{t|t} = \sum_{c_t \in c^*} w_{c_t} \xi_{c_t;t|t}, \tag{46}$$

$$\hat{R}_{t|t} = \sum_{c_t \in c^*} w_{c_t} R_{c_t;t|t} + \sum_{c_t \in c^*} w_{c_t} \left(\hat{\xi}_{t|t} - \xi_{c_t;t|t}\right) \left(\hat{\xi}_{t|t} - \xi_{c_t;t|t}\right)'. \tag{47}$$

The derivation of the result can be found in Appendix 10.3.

6.3. The joint pdf after approximation

The desired posterior joint pdf (33) involving the results of approximation now takes the following form:

$$f\left(x_t, c_t, \alpha|D(t)\right) \approx \hat{f}\left(x_t|D(t)\right) f\left(c_t|D(t)\right) \hat{f}\left(\alpha|D(t)\right), \tag{48}$$

where the normal approximated pdf $\hat{f}\left(x_t|D(t)\right)$ is determined by (46)–(47), the Dirichlet pdf $\hat{f}\left(\alpha|D(t)\right)$ is given by the statistics $\nu_{i|j;t}$ computed as the numerical solution of equation (41), and the pointer pdf $f\left(c_t|D(t)\right)$ is provided in (34). Relation (48) represents a feasible form of the discussed recursive estimation of the mixture of state-space components.

7. 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_c, parameters of the normal state-space model (5) for each component and set initial values of the prior state estimate (6).
- Choose initial values of the statistics $\nu_{i|j;0}$ with $i, j \in c^*$ and compute the initial point estimate $\hat{\alpha}_{c_t|c_{t-1};0}$ according to (21).
- Set initial values of probabilities in the pointer vector $w_{c_0},\ c_0 \in c^*$.

On-line part (time cycle of the algorithm)

1. Load the current data item d_t.
2. For individual components $c_t = 1, 2, \cdots, n_c$, run the Kalman filter (7)–(12), get the state posterior pdfs with means $\xi_{c_t;t|t}$ and covariance matrices $R_{c_t;t|t}$ and compute the data predictive pdfs $\mathcal{L}_{c_t;d_t}$ according to (13).
3. Update the vector of the pointer w_{c_t} according to (34), i.e.,

$$\bar{w}_{c_t|c_{t-1}} = \mathcal{L}_{c_t;d_t}\hat{\alpha}_{c_t|c_{t-1};t-1}w_{c_{t-1}},$$

$$w_{c_t|c_{t-1}} = \frac{\bar{w}_{c_t|c_{t-1}}}{\sum_{c_t\in c^*}\sum_{c_{t-1}\in c^*}\bar{w}_{c_t|c_{t-1}}},$$

$$w_{c_t} = \sum_{c_{t-1}\in c^*} w_{c_t|c_{t-1}}.$$

4. Use a numerical method to solve equation (41) and obtain the optimal statistics $\nu_{i|j;t}$ for $i, j \in c^*$ defining the approximated pdf (38).
5. Compute the point estimate $\hat{\alpha}_{c_t|c_{t-1};t} = \frac{\nu_{i|j;t}}{\sum_i \nu_{i|j;t}}$ with the obtained statistics $\nu_{i|j;t}$.
6. Using the updated vector of the pointer w_{c_t}, make approximation (46)–(47) and obtain the state approximated posterior pdf with the mean $\hat{\xi}_{t|t}$ and the covariance matrix $\hat{R}_{t|t}$.
7. Compute the point estimate of the active component(s) on the basis of the vector w_{c_t}, for instance, as the most probable component (if necessary).
8. Use statistics of the obtained posterior pdfs, i.e., $\hat{\xi}_{t|t}$, $\hat{R}_{t|t}$, ν_t, $\hat{\alpha}_{c_t|c_{t-1};t}$ and w_{c_t} for the prior ones in the next step of the recursions.

8. Experiments

The proposed algorithm (in this section denoted by MF, i.e., the mixture filter) was tested on simulated data. For comparison, the well known Monte Carlo particle filter (PF) and the Rao-Blackwellised particle filter (RBPF), see e.g., [21, 22], were chosen, where the last was interpreted as an efficient stochastic mixture of Kalman filters. The PF and RBPF software implementation including the efficient state-of-the-art generic resampling routines available at www.cs.ubc.ca/∼nando/software.html was used for the experiments which compared the performance of all the mentioned algorithms. Difference in results of all three filters was not significant. However, the computational time differed a lot. Here we report the results of the series of the experiments performed using the PF built-in data generator.

8.1. Simulation

To generate the 2-dimensional Gaussian state x_t and the scalar observation d_t for 3 components (i.e., $n_c = 3$) the following parameters of the state-space components were used:

$$A_1 = [0.3\ \ -0.3;\ 0.5\ 0.1],\ C_1 = [0.9\ 0.4],\ F_1 = [-5\ 25]',\ R_{w_1} = [4\ 0;\ 0\ 4],\ R_{v_1} = 0.4,$$

$$A_2 = [0.05\ \ -.3;\ 0.4\ \ -0.1],\ C_2 = [2.1\ 2.8],\ F_2 = [25\ 5]',\ R_{w_2} = [12\ 0;\ 0\ 12],\ R_{v_2} = 0.2,$$

$$A_3 = [0.6\ \ -0.1;\ 0.8\ \ -0.5],\ C_3 = [1.9\ \ -0.4],\ F_3 = [5\ \ -25]',\ R_{w_3} = [2\ 0;\ 0\ 2],\ R_{v_3} = 0.5,$$

where subscripts express the number of the corresponding component, F is a constant added to the state evolution model in (5), i.e., the mean value in (5) is taken as $Ax_{t-1} + F$. 1000 data items were generated. The table of the transition probabilities α was set using a random generator. 50 simulations were performed.

The data sample generated during each simulation was used for the state and the pointer estimation via PF, RBPF and MF. 200 particles were used for the estimation both with PF and RBPF. The prior state estimates were chosen with $\xi_{t-1|t-1} = [0;\ 0]$ and $R_{t-1|t-1} = [10\ 0;\ 0\ 10]$. The initial estimates of the transition table were selected randomly.

8.2. Results

Results of the performed 50 experiments were rather similar. One of the typical results of the state estimation obtained using MF, PF and RBPF is shown in Figure 1. For better

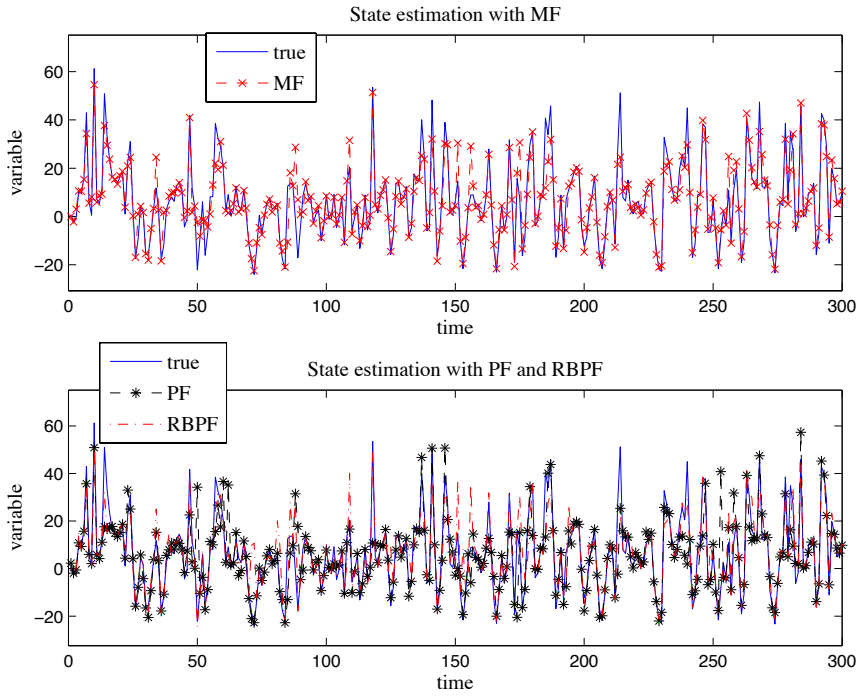

Figure 1: Comparison of the state estimation

Note a difference of the results between 50 and 100 time periods and near 250 time periods.

illustration, a fragment with 300 data items from 1000 is plotted, because larger number of data in the figure worsens visibility.

The results are shown for the first entry of the state vector. For the second state, they are is of a similar quality. The state estimates of all three algorithms in Figure 1 are very close to the simulated values. The difference can be seen, for example, between 50 and 100 time periods and near 250 time periods, where the MF results are closer to the simulated values.

The typical results of the data prediction are presented in Figure 2 for a fragment of 300 data items. PF and RBPF produced identical results, thus Figure 2 (top) presents the MF data predictions, and Figure 2 (bottom) – the RBPF data predictions. The predictions of all the filters are rather similar, however, the difference in favor of MF can be seen around 40, 80 and from 180 to 225 time periods.

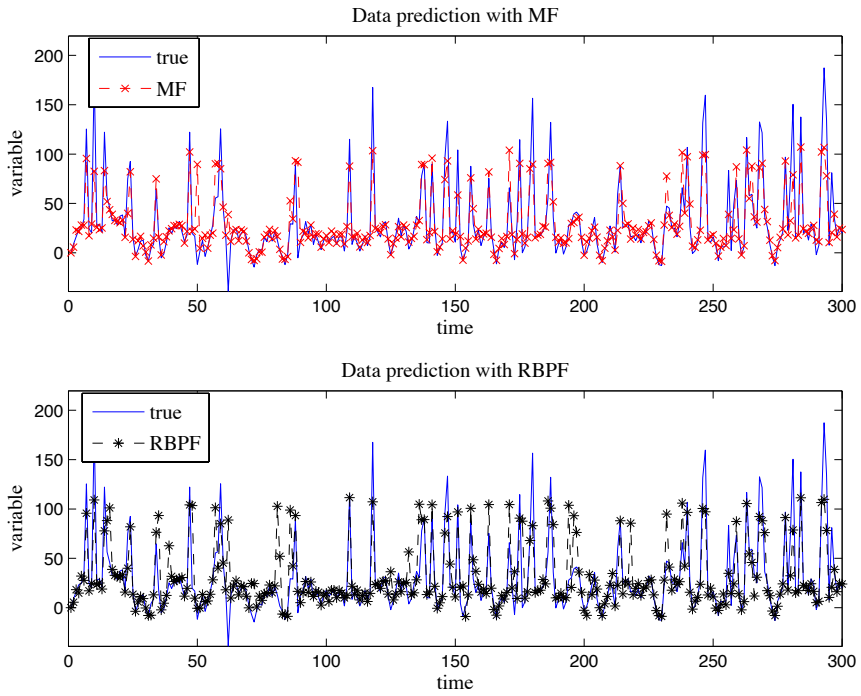

Figure 2: Comparison of the data prediction
Notice the difference of the results around 40, 80 and from 180 to 225 time periods.

For the pointer estimation, MF typically produced results better than PF, but worse than RBPF. These results expressing the point estimates of active components are provided in Figure 3 (top), (middle) and (bottom) for MF, PF and RBPF respectively. A fragment for 100 data items is chosen for better illustration.

The average errors of the state estimation (SEE), the data prediction (DPE) and the pointer estimation (PEE) over 50 experiments (denoted by N) with 1000 data items for all three filters are reported in Table 1. SEE and DPE were calculated as follows:

$$\text{SEE} = \frac{1}{NT} \sum_{t=1}^{T=1000} \sum_{n=1}^{N=50} (x_{t,n} - \hat{\xi}_{t|t,n})'(x_{t,n} - \hat{\xi}_{t|t,n}),$$

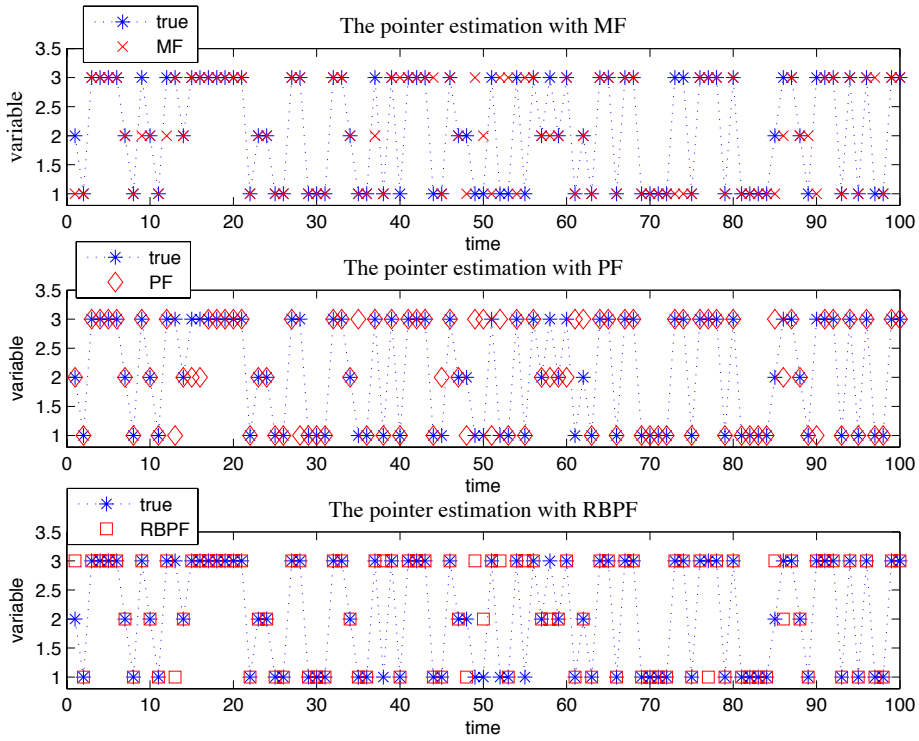

Figure 3: Comparison of the pointer estimation
Notice a difference of the estimates around 12, 35 and 60 time periods.

$$\text{DPE} = \frac{1}{NT}\sum_{t=1}^{T=1000}\sum_{n=1}^{N=50}(d_{t,n}-\hat{d}_{t,n})^2,$$

where a subscript n denotes the number of the experiment, $\hat{d}_t, n$ is the predicted data item. PEE is computed as the count of errors of the point estimates averaged over the number of the experiments N. Table 1 also provides the average computation time (CT) compared with the help of the Matlab functions `tic` and `toc`.

Table 1: The average errors and the computation time

	SEE	DPE	PEE	CT
MF	0.15	0.62	160.4	0.42
PF	0.31	0.88	216	25.27
RBPF	0.27	0.88	129	47.7

Figure 4 demonstrates the state-space components plotted in the form of clusters, where it can be seen that, in general, the difference in the results is not too significant.

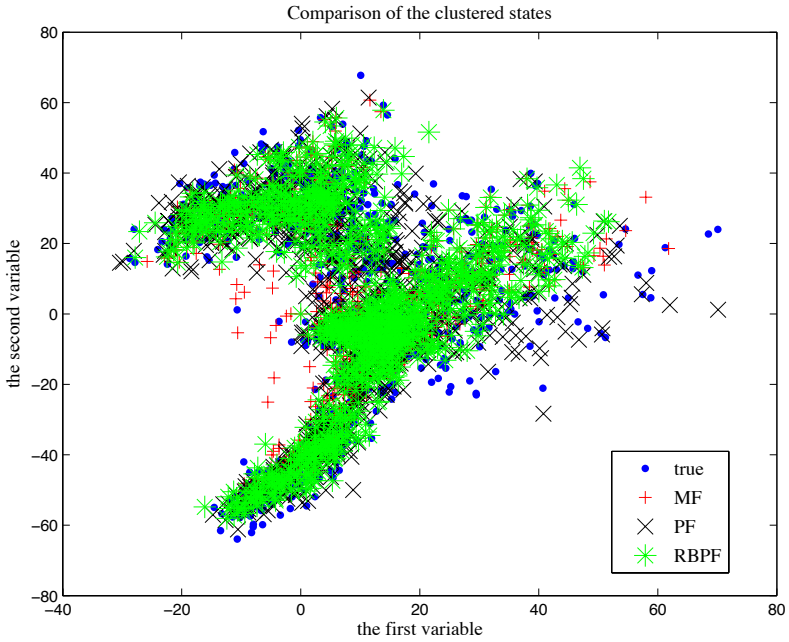

Figure 4: Comparison of the clustered states

8.3. Discussion

Surely it is necessary to consider that these experiments were performed with the simulated data, and the estimation quality of the algorithms can differ for real measurements. Nevertheless, the results presented in Table 1 are promising: MF provides the smallest average errors of the state estimation and the data prediction in comparison with PF and RBPF, and it has the second smallest average error of the pointer estimation after RBPF.

The difference of the computation time reported in Table 1 is substantial: MF is significantly faster. This can be decisive for such application areas, where a real-time filtering for dynamic systems with multiple working regimes is desired (for instance, traffic control, where the queue length at crossroads is often modelled as the unmeasurable state, and its real-time estimation is strongly needed for the traffic light control).

9. Conclusion

The paper proposed the recursive algorithm for estimation of mixtures with state-space components and the dynamic switching model. The so called fully dynamic mixture is considered where the switching of active components is modeled in dependence on the last active component. The algorithm represents an extension of solutions used for the estimation of mixtures of autoregression models and for the state estimation of hybrid systems, inheriting such advantages as one-pass estimation, real-time performance and generality of approach. The presented illustrative experiments show the promising results of the proposed algorithm and the significantly shorter computation time in comparison with two particle filters.

Further extension of this approach can be applied for prediction of future active components that is enabled by the dynamic model of switching. This complicated problem inspired by an

anonymous reviewer will be the subject of further research. The open problems remained here include also (but not limited to) usage of non-linear models of the pointer.

Acknowledgements

The research was supported by the project TAČR TA01030123.

10. Appendix

10.1. Chain rule

The chain rule [12] takes the form

$$f(a,b|c) = f(a|b,c)f(b|c), \tag{49}$$

which decomposes the joint pdf $f(a,b|c)$ into a product of the conditional pdfs for any random variables a, b and c.

10.2. Bayes rule

The Bayes rule [12] claims that

$$f\left(a|b,c\right) \propto f\left(b,a|c\right) \tag{50}$$

or with application of the chain rule

$$f\left(a|b,c\right) \propto f\left(b|a,c\right) f\left(a|c\right) \tag{51}$$

for any random variables a, b and c.

10.3. Approximation of normal mixture by normal distribution

Let $f(x_t|D(t)) = \sum_{c_t=1}^{n_c} w_{c_t} f\left(x_t|c_t, D(t)\right)$ be a mixture of normal components $f\left(x_t|c_t, D(t)\right) = \mathcal{N}\left(\xi_{c_t;t|t}, R_{c_t;t|t}\right)$, where $\xi_{c_t;t|t}$ are expectations and $R_{c_t;t|t}$ are covariance matrices. The goal is to construct a single normal pdf $\hat{f}\left(x_t|D(t)\right) = \mathcal{N}\left(\hat{\xi}_{t|t}, \hat{R}_{t|t}\right)$, which is as close as possible to the mixture $f(x_t|D(t))$ in the sense of minimization of the Kerridge inaccuracy

$$K\left(f\left(x_t|D(t)\right) \parallel \hat{f}\left(x_t|D(t)\right)\right) = \int_{x^*} f\left(x_t|D(t)\right) \ln \frac{1}{\hat{f}\left(x_t|D(t)\right)} dx_t.$$

Let us firstly compute the Kerridge inaccuracy between the normal distribution of the c_tth component $f\left(x_t|c_t, D(t)\right) = \mathcal{N}\left(\xi_{c_t;t|t}, R_{c_t;t|t}\right)$ and the desired $\hat{f}\left(x_t|D(t)\right) = \mathcal{N}\left(\hat{\xi}_{t|t}, \hat{R}_{t|t}\right)$ of the X-dimensional random vector x_t. Denoted by K_{c_t}, it takes the form

$$K_{c_t} = \int_{x^*} \mathcal{N}\left(\xi_{c_t;t|t}, R_{c_t;t|t}\right) \ln \frac{1}{(2\pi)^{-X/2}|\hat{R}_{t|t}|^{-0.5} \exp\left\{-0.5\left(x_t - \hat{\xi}_{t|t}\right)' \hat{R}_{t|t}^{-1}\left(x_t - \hat{\xi}_{t|t}\right)\right\}} \, dx_t$$

$$= \int_{x^*} \left[-\ln\left((2\pi)^{-X/2}\right) + 0.5 \ln|\hat{R}_{t|t}| + 0.5\left(x_t - \hat{\xi}_{t|t}\right)' \hat{R}_{t|t}^{-1}\left(x_t - \hat{\xi}_{t|t}\right)\right] \mathcal{N}\left(\xi_{c_t;t|t}, R_{c_t;t|t}\right) \, dx_t$$

$$= -\ln\left((2\pi)^{-X/2}\right) + 0.5\ln|\hat{R}_{t|t}|$$

$$+0.5\int_{x^*}\left(\left\{x_t-\xi_{c_t;t|t}\right\}+\left\{\xi_{c_t;t|t}-\hat{\xi}_{t|t}\right\}\right)'\hat{R}_{t|t}^{-1}\left(\left\{x_t-\xi_{c_t;t|t}\right\}+\left\{\xi_{c_t;t|t}-\hat{\xi}_{t|t}\right\}\right)\mathcal{N}\left(\xi_{c_t;t|t},R_{c_t;t|t}\right)\,dx_t$$

$$= -\ln\left((2\pi)^{-X/2}\right) + 0.5\left\{\ln\left(|\hat{R}_{t|t}|\right)+\mathrm{tr}\left\{R_{c_t;t|t}\hat{R}_{t|t}^{-1}\right\}+\left(\xi_{c_t;t|t}-\hat{\xi}_{t|t}\right)'\hat{R}_{t|t}^{-1}\left(\xi_{c_t;t|t}-\hat{\xi}_{t|t}\right)\right\}. \tag{52}$$

One can compute the final Kerridge inaccuracy K for the considered mixture $f(x_t|D(t)) = \sum_{c_t=1}^{n_c} w_{c_t} f\left(x_t|c_t,D(t)\right)$ as follows:

$$K=\sum_{c_t=1}^{n_c} w_{c_t}K_{c_t},\ \sum_{c_t=1}^{n_c} w_{c_t}=1. \tag{53}$$

To minimize K, it is necessary to compute derivatives of the individual inaccuracies K_{c_t} according to $\hat{\xi}_{t|t}$ and $\hat{R}_{t|t}$. One obtains them as follows:

$$\partial K_{c_t}/\partial\hat{\xi}_{t|t} = \hat{R}_{t|t}^{-1}\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right), \tag{54}$$

$$\partial K_{c_t}/\partial\hat{R}_{t|t} = 0.5\hat{R}_{t|t}^{-1}\left[\hat{R}_{t|t}-R_{c_t;t|t}-\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right)\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right)'\right]\hat{R}_{t|t}^{-1}. \tag{55}$$

For a minimum of (53) it holds

$$\sum_{c_t=1}^{n_c} w_{c_t}\hat{R}_{t|t}^{-1}\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right)=0, \tag{56}$$

and

$$0.5\sum_{c_t=1}^{n_c} w_{c_t}\hat{R}_{t|t}^{-1}\left[\hat{R}_{t|t}-R_{c_t;t|t}-\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right)\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right)'\right]\hat{R}_{t|t}^{-1}=0. \tag{57}$$

The desired normal distribution $\mathcal{N}\left(\hat{\xi}_{t|t},\hat{R}_{t|t}\right)$ minimizing the Kerridge inaccuracy is thus given by its characteristics

$$\hat{\xi}_{t|t}=\sum_{c_t=1}^{n_c} w_{c_t}\xi_{c_t;t|t}, \tag{58}$$

$$\hat{R}_{t|t}=\sum_{c_t=1}^{n_c} w_{c_t}R_{c_t;t|t}+\sum_{c_t=1}^{n_c} w_{c_t}\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right)\left(\hat{\xi}_{t|t}-\xi_{c_t;t|t}\right)' \tag{59}$$

to be used for the state approximation in Section 6.2.

10.4. Approximation of a mixture of the Dirichlet pdfs by a single Dirichlet pdf

Let

$$f(\alpha|D(t))=\sum_{c_t\in c^*}\sum_{c_{t-1}\in c^*} w_{c_t|c_{t-1}}\mathcal{D}_\alpha\left(\nu_{t-1}^{c_t|c_{t-1}}\right). \tag{60}$$

be a mixture of the Dirichlet pdfs. The task is to find a single Dirichlet pdf $\hat{f}\left(\alpha|D(t)\right)=\mathcal{D}_{\alpha}\left(\nu_{t}\right)$ of the form

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

which minimizes the Kerridge inaccuracy

$$K_{\mathcal{D}}\left(f\left(\alpha|D(t)\right)\parallel\hat{f}\left(\alpha|D(t)\right)\right)=\int_{\alpha^{*}}f\left(\alpha|D(t)\right)\ln\frac{1}{\hat{f}\left(\alpha|D(t)\right)}\,d\alpha \tag{61}$$

over the statistics $\nu_{t}=[\nu_{i|j;t}]_{i,j\in c^{*}}$.

Substituting $f(\alpha|D(t))$ and $\hat{f}\left(\alpha|D(t)\right)$ in relation (61) denoted by $K_{\mathcal{D}}$, it is obtained

$$K_{\mathcal{D}}=\int_{\alpha^{*}}\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\mathcal{D}_{\alpha}\left(\nu_{t-1}^{c_{t}|c_{t-1}}\right)\ln[\mathcal{D}_{\alpha}\left(\nu_{t}\right)]^{-1}\,d\alpha \tag{62}$$

$$=\int_{\alpha^{*}}\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\frac{1}{\mathcal{B}(\nu_{t-1}^{c_{t}|c_{t-1}})}\prod_{i|j}\alpha_{i|j}^{\nu_{i|j;t-1}^{c_{t}|c_{t-1}}-1}\ln\left[\frac{1}{\mathcal{B}(\nu_{t})}\prod_{k|l}\alpha_{k|l}^{\nu_{k|l;t}-1}\right]^{-1}\,d\alpha$$

$$=\int_{\alpha^{*}}\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\frac{1}{\mathcal{B}(\nu_{t-1}^{c_{t}|c_{t-1}})}\prod_{i|j}\alpha_{i|j}^{\nu_{i|j;t-1}^{c_{t}|c_{t-1}}-1}\left[\ln\mathcal{B}(\nu_{t})-\ln\prod_{k|l}\alpha_{k|l}^{\nu_{k|l;t}-1}\right]\,d\alpha$$

$$=\ln\mathcal{B}(\nu_{t})\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\int_{\alpha^{*}}\frac{1}{\mathcal{B}(\nu_{t-1}^{c_{t}|c_{t-1}})}\prod_{i|j}\alpha_{i|j}^{\nu_{i|j;t-1}^{c_{t}|c_{t-1}}-1}\,d\alpha$$

$$-\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\int_{\alpha^{*}}\frac{1}{\mathcal{B}(\nu_{t-1}^{c_{t}|c_{t-1}})}\prod_{i|j}\alpha_{i|j}^{\nu_{i|j;t-1}^{c_{t}|c_{t-1}}-1}\sum_{k|l}\ln\alpha_{k|l}^{\nu_{k|l;t}-1}\,d\alpha$$

$$\underbrace{=}_{\text{integral of the Dirichlet pdf is equal to 1}}$$

$$=\ln\mathcal{B}(\nu_{t})\underbrace{\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}}_{=1}-\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\int_{\alpha^{*}}\frac{1}{\mathcal{B}(\nu_{t-1}^{c_{t}|c_{t-1}})}\prod_{i|j}\alpha_{i|j}^{\nu_{i|j;t-1}^{c_{t}|c_{t-1}}-1}\sum_{k|l}\ln\alpha_{k|l}^{\nu_{k|l;t}-1}\,d\alpha$$

$$=\ln\mathcal{B}(\nu_{t})-\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\sum_{k|l}\int_{\alpha^{*}}\frac{1}{\mathcal{B}(\nu_{t-1}^{c_{t}|c_{t-1}})}\alpha_{k|l}^{\nu_{k|l;t-1}^{c_{t}|c_{t-1}}-1}\ln\alpha_{k|l}^{\nu_{k|l;t}-1}\,d\alpha$$

$$=\ln\mathcal{B}(\nu_{t})-\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\sum_{k|l}(\nu_{k|l;t}-1)\int_{\alpha^{*}}\mathcal{D}_{\alpha}\left(\nu_{t-1}^{c_{t}|c_{t-1}}\right)\ln\alpha_{k|l}\,d\alpha$$

$$\underbrace{=}_{\text{all what does not belong to the } l\text{th row is integrated, giving the marginal Dirichlet pdf}}$$

$$=\ln\mathcal{B}(\nu_{t})-\sum_{c_{t}\in c^{*}}\sum_{c_{t-1}\in c^{*}}w_{c_{t}|c_{t-1}}\sum_{k|l}(\nu_{k|l;t}-1)\int_{\alpha^{*}}\mathcal{D}_{\alpha}\left(\nu_{\cdot|l;t-1}^{c_{t}|c_{t-1}}\right)\ln\alpha_{k|l}\,d\alpha.$$

Using the relation

$$\int_{\alpha^*} \ln \alpha_{i|j} \mathcal{D}_\alpha \left(\nu_{\cdot|j}\right)\, d\alpha = \Xi\left(\nu_{i|j}\right), \tag{63}$$

the Kerridge inaccuracy $K_{\mathcal{D}}$ to be minimized with respect to $\nu_{i|j;t}$ now takes the following form

$$K_{\mathcal{D}} = \ln \mathcal{B}(\nu_t) - \sum_{c_t \in c^*} \sum_{c_{t-1} \in c^*} w_{c_t|c_{t-1}} \sum_{k|l} (\nu_{k|l;t} - 1) \Xi\left(\nu_{k|l;t-1}^{c_t|c_{t-1}}\right). \tag{64}$$

To find the extreme, it is necessary to compute the derivative

$$\begin{aligned}
\frac{\partial}{\partial \nu_{i|j;t}} K_{\mathcal{D}} &= \frac{\partial}{\partial \nu_{i|j;t}} \left\{ \ln \mathcal{B}(\nu_t) - \sum_{c_t \in c^*} \sum_{c_{t-1} \in c^*} w_{c_t|c_{t-1}} \sum_{k|l} (\nu_{k|l;t} - 1) \Xi\left(\nu_{k|l;t-1}^{c_t|c_{t-1}}\right) \right\} \\
&= \frac{\partial}{\partial \nu_{i|j;t}} \ln \mathcal{B}(\nu_t) - \sum_{c_t \in c^*} \sum_{c_{t-1} \in c^*} w_{c_t|c_{t-1}} \frac{\partial}{\partial \nu_{i|j;t}} \sum_{k|l} (\nu_{k|l;t} - 1) \Xi\left(\nu_{k|l;t-1}^{c_t|c_{t-1}}\right) \\
&= \sum_{c_t \in c^*} \sum_{c_{t-1} \in c^*} w_{c_t|c_{t-1}} \Xi\left(\nu_{i|j;t}\right) - \sum_{c_t \in c^*} \sum_{c_{t-1} \in c^*} w_{c_t|c_{t-1}} \Xi\left(\nu_{i|j;t-1}^{c_t|c_{t-1}}\right), \quad i, j \in c^*.
\end{aligned} \tag{65}$$

A zero gradient is a necessary condition for an extreme. Thus, the condition $\frac{\partial K_{\mathcal{D}}}{\partial \nu_{i|j;t}} = 0$ for all $i, j \in c^*$ must be satisfied. According to [23], if a minimized function is convex and it is defined on a convex domain, the extreme is global minimum. This is the case considered here (the detailed explanation can be found in [14]). The mentioned property guarantees a quick and smooth search for the minimum of the Kerridge inaccuracy $K_{\mathcal{D}}$ in (62). 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. The solution can be computed for each row separately.

Remark: By the denotation $\Xi\left(\nu_{i|j}\right)$ we mean (according to definition) $\Psi\left(\nu_{i|j}\right) - \Psi\left(\sum_k \nu_{k|j}\right)$. It means, the argument $\nu_{i|j}$ is decisive, but the whole row $\nu_{\cdot|j}$ is involved. It is necessary to keep it in mind.

References

[1] S. Haykin, Neural Networks: A Comprehensive Foundation, Macmillan, New York, 1994.

[2] M. J. Beal, Z. Ghahramani, C. E. Rasmussen, The infinite hidden markov model, in: Advances in Neural Information Processing Systems, Vol. 14, 2002.

[3] M. Grewal, A. Andrews, Kalman Filtering: Theory and Practice Using MATLAB. 2nd edition, Wiley, 2001.

[4] C. A. McGrory, D. M. Titterington, Variational Bayesian analysis for hidden Markov models, Australian & New Zealand Journal of Statistics 51 (2009) 227244. doi:10.1111/j.1467-842X.2009.00543.x.

[5] V. Šmídl, A. Quinn, The Variational Bayes Method in Signal Processing, Springer, 2005.

[6] S. Chiappa, D. Barber, Dirichlet Mixtures of Bayesian Linear Gaussian State-Space Models: a Variational Approach. Technical Report 161, Max Planck Institute for Biological Cybernetics, 2007.

[7] Z. Ghahramani, G. E. Hinton, Variational learning for switching state-space models, Neural Computation 12 (4) (2000) 831–864.

[8] A. Doucet, C. Andrieu, Iterative algorithms for state estimation of jump markov linear systems, IEEE Transactions on Signal Processing 49 (6) (2001) 1216–1227.

[9] R. Chen, J. S. Liu, Mixture kalman filters, J. R. Statist. Soc. B 62 (2000) 493–508.

[10] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models, 2nd Edition, Springer New York, 2006.

[11] S. Früwirth-Schnatter, Fully bayesian analysis of switching gaussian state space models, Annals of the Institute of Statistical Mathematics 53 (1) (2001) 31–49.

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

[13] FET, IST12088 – ProDaCTool: Decision Support for Complex Industrial Processes Based on Probabilistic Data Clustering, http://www.prodactool.rdg.ac.uk/.

[14] I. Nagy, E. Suzdaleva, M. Kárný, T. Mlynářová, Bayesian estimation of dynamic finite mixtures, Int. Journal of Adaptive Control and Signal Processing 25 (9) (2011) 765–787. doi:10.1002/acs.1239.

[15] E. Suzdaleva, I. Nagy, Recursive state estimation for hybrid systems, Applied Mathematical Modelling. 36 (4) (2012) 1347–1358. doi:10.1016/j.apm.2011.08.042.

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

[17] S. Kullback, R. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–87.

[18] J. M. Bernardo, Expected information as expected utility, The Annals of Statistics 7 (3) (1979) 686–690.

[19] V. Peterka, Bayesian system identification, in: P. Eykhoff (Ed.), Trends and Progress in System Identification, Pergamon Press, Oxford, 1981, pp. 239–304.

[20] M. West, J. Harrison, Bayesian forecasting and dynamic models, 2nd Edition, Springer, 1997.

[21] N. de Freitas, Rao-blackwellised particle filtering for fault diagnosis, in: Aerospace Conference Proceedings, 2002. IEEE, Vol. 4, pp. 4–1767–4–1772. doi:10.1109/AERO.2002.1036890.

[22] A. Doucet, N. de Freitas, N. G. (eds), Sequential Monte Carlo Methods in Practice, Springer-Verlag, 2001.

[23] P. Algoet, C. T., A sandwich proof of the Shannon-McMillan-Breiman theorem, The Annals of Probability 16 (1988) 899–909.