Dynamic Mixture Ratio Model

Marko Ruman
Department of Adaptive Systems
The Czech Academy of Sciences, ÚTIA
18208 Prague 8, Czech Republic
marko.ruman@gmail.com

Miroslav Kárný
Department of Adaptive Systems
The Czech Academy of Sciences, ÚTIA
18208 Prague 8, Czech Republic
school@utia.cas.cz

***Abstract*—Finite mixtures of probability densities with components from exponential family serve as flexible parametric models of high-dimensional systems. However, with a few specialized exceptions, these dynamic models assume data-independent weights of mixture components. Their use is illogical and restricts the modeling applicability. The requirement for closeness with respect to conditioning, the basic learning operation, leads to a novel class of models: the mixture ratios. The paper justified them and shows their ability to model truly dynamic systems.**

***Keywords*—Dynamic systems, Bayesian learning, mixture models, mixture ratio**

I. Introduction

Decision making (DM) chooses actions for reaching a specific aim. Many fields including machine learning, [23], signal processing, [24], estimation and filtering, [17], hypothesis testing, [14], classification and pattern recognition, [11], knowledge sharing, [22], reinforcement learning, [27], control, [13], etc., can all be seen as DM. This makes the amount of relevant results excessive, [16], and volatile vocabularies. This leads to the specific vocabulary used when formalising DM.

A solution of a DM problem leads to a strategy, a collection of decision rules mapping the knowledge on actions, [26]. The chosen strategy should meet the DM aim in the best possible way. The adopted Bayesian paradigm, [25], is a powerful tool whenever DM faces incomplete knowledge and uncertainty concerning the dynamic system to which actions relate. Bayesian DM relates DM consequences to the acquired knowledge and the used actions by conditional probabilities, here given by conditional probability densities (pd).

Actions are generally chosen recursively while enriching the available knowledge. This gives a chance to improve gradually the system model. This learning redistributes probability of the model adequacy within the set of used models, [9].

The need for learning arises if a good model is a priori unknown and it is possible iff the learnt relations practically do not change during the knowledge extraction. This makes us to focus on a set of parametric models. Their constant multivariate parameter serves as a "pointer" to the set members. The best model is a priori unknown. Bayesian learning offers the unambiguous deductive way, Bayes' rule, [10], of redistributing the probability of (belief in) the model quality. It maps the knowledge on the posterior pd and provides the predictive pd serving as the system model used by DM. The achievable modeling and thus DM quality are determined by the employed set of parametric models.

The research was supported by the research project LTC18075 and CA16228.

This work offers *ratios* of finite mixtures, [4], with components from exponential family (EF), [7], as such black-box, [12], universally approximating, [15], models. The recursive learning of ratios of finite mixtures is inevitably approximate and endangered by accumulation of approximation errors. This paper provides a simulation study, which inspects the approximative learning developed in [32] from this perspective. The reader gets the chance to consider this extremely flexible but yet unconsidered model set for solving her/his difficult DM task. The paper focuses on cases, which can formally be covered by a high-order Markov chain, which relates observations to a finite-dimensional regression vector containing the past observations and explanatory variables. Both are discrete or discretised.

Recursive Bayesian estimation the high-order Markov chain [29], a lossless compression of the knowledge, simple counts joint occurrences of the predicted variable and its regression vector. The applicability of this formally universal way is, however, strongly limited by the curse of dimensionality, [8]. The size of the occurrence array blows up with the number of possible data-vectors instances and the observations insufficiently populate it. The advocated model counteracts this curse of dimensionality. The presented simulation results demonstrate improvement caused by the used mixture ratio comparing to the standard mixture model addressing the same dimensionality problem in [18].

The text uses the following notation:

- $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}^+$ stand for sets of *natural*, *real* and *positive real* numbers, respectively,
- the bold symbols, for instance $\mathbf{A}$, stands for the set of its members $A \in \mathbf{A}$,
- $\mathbf{A}^n$ stands for the n-ary Cartesian product of $\mathbf{A}$,
- $|\mathbf{A}|$ denotes cardinality of $\mathbf{A}$,
- $(a_1, ..., a_n) \in \mathbf{A}^n$ stands for a n-dimensional vector,
- $\left\langle a \,|\, b \right\rangle$ denotes a dot product of two vectors $a, b \in \mathbb{R}^n$, $\left\langle a \,|\, b \right\rangle = \sum_{i=1}^{n} a_i b_i$, 0^0 is defined as $0^0 = 1$.

II. Formalisation of a Dynamic System Model and Its Approximate Learning

A model of a dynamic system in the decision-making (DM) theory is formalized in Bayesian way as follows: observations $O_t \in \mathbf{O}$, stimulated by actions $A_t \in \mathbf{A}$, are observed at the discrete-time moments $t \in \mathbf{t} = \{1, 2, ..., |\mathbf{t}|\}$. A data sequence $D^t \in \mathbf{D}$ is formed by data records: $D^t = \{D_t, D_{t-1}, ..., D_1, D_0\}$, where a data record $D_t = \{O_t, A_t\}$, $t \in \mathbf{t}$, is formed by an observation O_t and an action A_t, D_0 means the prior knowledge.

Definition 1 (Parametric system model): The parametric system model is a pd of the observation $O_t \in \mathbf{O}$ conditioned on the data sequence $D^{t-1} \in \mathbf{D}$, the action $A_t \in \mathbf{A}$ and on an unknown parameter $\Theta \in \mathbf{\Theta}$:

$$\mathrm{M}(O_t|A_t, D^{t-1}, \Theta), \tag{1}$$

where $\mathbf{\Theta}$ is a *parametric space*. It is a subset of a real d-dimensional vector space, i.e. $\mathbf{\Theta} \subset \mathbb{R}^d$, $d \in \mathbb{N}$. The notation $\mathrm{M}_t(\Theta) = \mathrm{M}(O_t|A_t, D^{t-1}, \Theta)$ will be used whenever realization of A_t, D^{t-1} is inserted into the parametric system model, i.e. when it is treated as the likelihood function. □

The goal of DM theory is to influence the dynamic system by taking appropriate actions to achieve desired observations (generally, states) of the modeled system. To predict the next observation of the system, the pd $\mathrm{F}(O_t|A_t, ^{t-1})$ is used. Using chain rule, marginalisation and incorporating the parametric system model (1), it can be expressed as follows:

$$\mathrm{F}(O_t|A_t, D^{t-1}) = \int_{\mathbf{\Theta}} \mathrm{M}(O_t|A_t, D^{t-1}, \Theta)\mathrm{P}(\Theta|D^{t-1})d\Theta, \tag{2}$$

where $\mathrm{P}(\Theta|D^{t-1})$ is the posterior pd storing actual knowledge about the parameter $\Theta \in \mathbf{\Theta}$. The short-hand version $\mathrm{P}_{t-1}(\Theta) = \mathrm{P}(\Theta|D^{t-1})$ is used. Formula (2) is valid under adopted *natural conditions of control*, [31], expressing that the parameter is unknown to the randomised action generator, i.e. the pd describing it meets $\mathrm{S}(A_t|D^{t-1}, \Theta) = \mathrm{S}(A_t|D^{t-1})$ $\Leftrightarrow$ $\mathrm{P}(\Theta|A_t, D^{t-1}) = \mathrm{P}(\Theta|D^{t-1})$.

After the interaction with the dynamic system, a new data record $D_t = \{O_t, A_t\}$ is created and $\mathrm{P}_{t-1}(\Theta)$ is updated via Bayes' rule, $\forall \Theta \in \mathbf{\Theta}$, as follows:

$$\begin{aligned}\tilde{\mathrm{P}}_t(\Theta) &= \frac{\mathrm{M}_t(\Theta)\mathrm{S}(A_t|D^{t-1})\mathrm{P}_{t-1}(\Theta)}{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{S}(A_t|D^{t-1})\mathrm{P}_{t-1}(\Theta)d\Theta} \\ &= \frac{\mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)}{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta} \propto \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta).\end{aligned} \tag{3}$$

To ensure the computational feasibility of updating the pd $\mathrm{P}_{t-1}(\Theta)$ during the whole interaction with a dynamic system, the set of feasible pds $\mathbf{P}$ needs to be considered. Generally, the posterior pd $\tilde{\mathrm{P}}_t(\Theta)$ obtained by (3) does not belong to the set of feasible pds $\mathbf{P}$ and with growing $t \in \mathbf{t}$ it can become more and more complex function of Θ. Therefore, a projection of $\tilde{\mathrm{P}}_t(\Theta)$ on pd $\hat{\mathrm{P}}_t(\Theta) \in \mathbf{P}$ from this set of feasible pds has to be made. [1] and [2] suggest, that the projection $\hat{\mathrm{P}}_t(\Theta)$ is the minimizer of Kerridge inaccuracy:

$$\hat{\mathrm{P}}_t(\Theta) = \underset{\mathrm{P}\in\mathbf{P}}{\mathrm{K}}(\tilde{\mathrm{P}}_t||\mathrm{P}) = \underset{\mathrm{P}\in\mathbf{P}}{\operatorname{argmin}} \int_{\mathbf{\Theta}} -\tilde{\mathrm{P}}_t(\Theta)\ln(\mathrm{P}(\Theta))d\Theta \tag{4}$$

The updating from $\mathrm{P}_{t-1}(\Theta)$ to $\hat{\mathrm{P}}_t(\Theta)$ can be interpreted as the application of Bayes' rule on $\mathrm{P}_{t-1}(\Theta)$ but using an unknown, different model than $\mathrm{M}_t(\Theta)$. Therefore, using $\hat{\mathrm{P}}_t(\Theta)$ as a prior pd for the next learning step may, in general, cause divergence $\hat{\mathrm{P}}_t(\Theta)$ from $\tilde{\mathrm{P}}_t(\Theta)$ obtained via (3) without projection (4). The solution preventing from the divergence is to include a data-depending forgetting factor $\lambda_t \in [0, 1]$. [3] implies the most plausible choice of the forgetting factor λ_t.

Remark 1 (Learning algorithm): One learning step of the learning algorithm is summarized as follows:

$$\tilde{\mathrm{P}}_t(\Theta) \propto \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta),\ \hat{\mathrm{P}}_t = \underset{\mathrm{P}\in\mathbf{P}}{\operatorname{argmin}}\, \mathrm{K}(\tilde{\mathrm{P}}_t||\mathrm{P}),$$

$$\mathrm{P}_t(\Theta) \propto \hat{\mathrm{P}}_t^{\lambda_t}(\Theta)\mathrm{P}_{t-1}^{1-\lambda_t}(\Theta),\ \lambda_t = \frac{\left(\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta\right)^2}{\int_{\mathbf{\Theta}} \mathrm{M}_t^2(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta}.$$

$\mathrm{P}_{t-1}(\Theta)$ is the prior pd, $\mathrm{K}(\tilde{\mathrm{P}}_t||P)$ is Kerridge's inaccuracy (4) and $\mathrm{P}_t(\Theta)$ serves as a prior pd for the next learning step.

III. Finite Mixture Ratio Model

To achieve computational feasibility of the recursive learning summarized in Remark 1., the following simplifying assumption is made about the model (1):

Assumption 1 (Markov property of the system model): The system model (1) is assumed to be time-invariant n-order Markov model ($n \in \mathbb{N}$), i.e.

$$\mathrm{M}(O_t|A_t, D^{t-1}, \Theta) = \mathrm{M}(O_t|\psi_t, \Theta)\ \text{, where} \tag{5}$$

$$\psi_t = \begin{cases} A_t, D_{t-1}, ..., D_{t-n} & n \geq 1, \\ A_t & n = 0. \end{cases} \text{ is a } \textit{regression vector}.$$

As stated in Introduction, this work focuses on modelling the dynamic system via mixtures of pds. There are many works on this topic, e.g. [4], however, most of them assume data-independence of the weights of the mixture components. The following approach overcomes this limitation by introducing a ratio of finite mixtures. The model (5) can be expressed as follows (using chain rule and marginalisation):

$$\mathrm{M}(O_t|\psi_t, \Theta) = \frac{\mathrm{J}(O_t, \psi_t|\Theta)}{\int_{\mathbf{O}} \mathrm{J}(O_t, \psi_t|\Theta)dO}. \tag{6}$$

Almost any practically met time-invariant joint pd $\mathrm{J}(O_t, \psi_t|\Theta)$ can be approximated by a finite mixture of Gaussian pds, [19], to an arbitraty precision, [4]. A Gaussian pd belongs to the exponential family (EF), thus, following EF mixture form of $\mathrm{J}(O_t, \psi_t|\Theta)$ can be considered:

Definition 2 (Joint pd as a mixture): The joint pd $\mathrm{J}(O_t, \psi_t|\Theta)$ is modeled via a mixture of pds from EF:

$$\begin{aligned}\mathrm{J}(O_t, \psi_t|\Theta) = \mathrm{J}(\phi_t|\Theta) &= \sum_{c\in\mathbf{c}} \alpha_c \mathrm{J}_c(\phi_t|\Theta_c) \\ &= \sum_{c\in\mathbf{c}} \alpha_c \exp\left\langle \mathrm{B}_c(\phi_{t;c}) \mid \mathrm{C}_c(\Theta_c)\right\rangle \mathrm{G}_c(\phi_{t;c}^{\mathbf{c}}),\ \text{where}\end{aligned} \tag{7}$$

- $\phi_t \in \boldsymbol{\phi}$ is the data vector $\phi_t = (O_t, \psi_t)$.
- $\mathrm{J}_c(\phi_t|\Theta_c) = \mathrm{J}_c(O_t, \psi_t|\Theta_c)$ is the c-th mixture component, it is a member of EF and it is a pd on $\boldsymbol{\phi}$. Its assumed form is: $\mathrm{J}_c(\phi_t|\Theta_c) = \exp\langle \mathrm{B}_c(\phi_t) \,|\, \mathrm{C}_c(\Theta_c)\rangle = \exp\langle \mathrm{B}_c(\phi_{t;c}) \,|\, \mathrm{C}_c(\Theta_c)\rangle \, \mathrm{G}_c(\phi^{\mathsf{c}}_{t;c})$ with $\phi_{t;c}$ being a component-specific subvector of $\phi_{t;c}$ and $\mathrm{G}_c(\phi^{\mathsf{c}}_{t;c})$ is a non-parametrized pd on its complement $\phi^{\mathsf{c}}_{t;c}$ with respect to ϕ_t; $\mathrm{B}_c(\phi_t)$, $\mathrm{C}_c(\Theta_c)$ and $\mathrm{B}_c(\phi_{t;c})$ are known, real-valued vector functions, $\mathrm{C}_c(\omega_c) = (\mathrm{C}_{c1}(\Theta_c), \dots \mathrm{C}_{cm_c}(\Theta_c))$
- $\mathbf{c} = \{1, ..., |\mathbf{c}|\}$, $|\mathbf{c}|$ is the number of components
- the parameter vector Θ has the following form:
  $\Theta = \left(\alpha_c, (\Theta_{cj})_{j=1}^{m_c}\right)_{c\in\mathbf{c}} \in \left\{\boldsymbol{\alpha} \underset{c\in\mathbf{c}}{\times} \boldsymbol{\Theta}_c\right\} = \boldsymbol{\Theta}$, where:
  - ✓ $\alpha = (\alpha_1, ..., \alpha_{|\mathbf{c}|}) \in \boldsymbol{\alpha}$ is the weight vector, $\boldsymbol{\alpha} = \left\{(\alpha_1, ..., \alpha_{|\mathbf{c}|}) \,|\, \alpha_c \geq 0, \ \sum_{c\in\mathbf{c}} \alpha_c = 1\right\}$
  - ✓ $\Theta_c = (\Theta_{c1}, ..., \Theta_{cm_c}) \in \boldsymbol{\Theta}_c$ is the component-specific parameter vector and m_c stands for the number of parameters of the c-th mixture component.

□

The general learning algorithm proposed in Section II uses the system model $\mathrm{M}(O_t|\psi_t, \Theta)$. By inserting right-hand side of (7) to (6), the desired parametric system model is obtained:

$$\mathrm{M}(O_t|\psi_t,\Theta) = \sum_{c\in\mathbf{c}} \frac{\alpha_c \mathrm{J}_c(O_t,\psi_t|\Theta_c)}{\sum_{d\in\mathbf{c}} \alpha_d \underbrace{\int_{\mathbf{O}} \mathrm{J}_d(O_t,\psi_t|\Theta_d) dO_t}_{\mathrm{W}_d(\psi_t,\Theta_d)}} \tag{8}$$

$$= \sum_{c\in\mathbf{c}} \underbrace{\frac{\alpha_c \mathrm{W}_c(\psi_t,\Theta_c)}{\sum_{d\in\mathbf{c}} \alpha_d \mathrm{W}_d(\psi_t,\Theta_d)}}_{\mathrm{w}_c(\psi_t,\Theta_c)} \underbrace{\frac{\mathrm{J}_c(O_t,\psi_t|\Theta_c)}{\mathrm{W}_c(\psi_t,\Theta_c)}}_{\mathrm{M}_c(O_t|\psi_t,\Theta_c)}$$

$$= \sum_{c\in\mathbf{c}} \mathrm{w}_c(\psi_t,\Theta_c)\mathrm{M}_c(O_t|\psi_t,\Theta_c), \quad \text{where}$$

- $\mathrm{W}_c(\psi_t,\Theta_c) = \int_{\mathbf{O}} \mathrm{J}_c(O_t,\psi_t|\Theta_c) dO_t$
- $\mathrm{w}_c(\psi_t,\Theta_c) = \frac{\alpha_c \mathrm{W}_c(\psi_t,\Theta_c)}{\sum_{d\in\mathbf{c}} \alpha_d \mathrm{W}_d(\psi_t,\Theta_d)}$
- $\mathrm{M}_c(O_t|\psi_t,\Theta_c) = \frac{\mathrm{J}_c(O_t,\psi_t|\Theta_c)}{\mathrm{W}_c(\psi_t,\Theta_c)}$

Equation (8) shows, that the model can be interpreted as the mixture with data-dependent weights, while the data-dependence is not arbitrary and *does not introduce new parameters*.

Learning algorithm described by Remark 1 has been elaborated for (8). All details can be found in [32].

IV. Mixture of Markov Chain Components

One of the most important mixtures with the components from the exponential family are Markov chain mixtures, [18]. A Markov chain operates with discrete valued observations $O \in \mathbf{O} \subset \mathbb{N}$ as well as discrete valued regression vectors $\psi_t \in \boldsymbol{\psi} \subset \mathbb{N}^n$, the joint pds $\mathrm{J}_c(\psi_t|\Theta_c)$, $\mathrm{J}(\psi_t|\Theta)$ as well as the parametric system model $\mathrm{M}(O_t|\psi_t, \Theta)$ are probability functions.

This part specifies the general mixture ratio model (Section III) and its learning for the mixture ratio of Markov chain components with conjugated Dirichlet pd $\mathrm{P}_t(\Theta_c)$, $c \in \mathbf{c}$.

Definition 3 (Joint probability of Markov chain mixture model): The joint probability of a mixture of Markov chain models is defined as follows:

$$\mathrm{J}(O_t,\psi_t|\Theta) = \mathrm{J}(\phi_t|\Theta) = \sum_{c\in\mathbf{c}} \alpha_c \mathrm{J}_c(O_t,\psi_{t;c}|\Theta_c) \tag{9}$$

$$= \sum_{c\in\mathbf{c}} \alpha_c \frac{1}{|\boldsymbol{\psi}^{\mathsf{c}}_c|} \prod_{j=1}^{m_c} \left(\frac{\Theta_{cj}}{\tilde{K}_{cj}}\right)^{\Delta_{cj}} \quad \text{where}$$

- $\mathbf{c}$ is the set of component indexes $c \in \mathbf{c}$, the symbol $|\mathbf{c}|$ stands for its cardinality, $\mathbf{c} = \{1, 2, .., |\mathbf{c}|\}$,
- let $\boldsymbol{\phi}$ be a set of all possible values of the data-vector $\phi \in \boldsymbol{\phi}$, then $\boldsymbol{\phi}_c$ is the subset of $\boldsymbol{\phi}$ modeled by the c-th component in the mixture model (9); the set $\boldsymbol{\phi}_c$ can be rewritten as $\boldsymbol{\phi}_c = \mathbf{O} \times \boldsymbol{\psi}_c$, where $\mathbf{O}$ and $\boldsymbol{\psi}_c$ are the sets of all possible values of the observation O and the component-specific regression vector ψ_c respectively, the $\boldsymbol{\psi}^{\mathsf{c}}_c$ denotes the complement of the set $\boldsymbol{\psi}_c$ with respect to the set $\boldsymbol{\psi}$,
- α_c is the weight of the c-th component; the set of all possible values of the weight-vectors $\alpha = \left(\alpha_1, \alpha_2, ..., \alpha_{|\mathbf{c}|}\right) \in \boldsymbol{\alpha}$ is defined as follows - $\boldsymbol{\alpha} = \left\{\alpha \subset [0,1]^{|\mathbf{c}|} \,|\, \sum_{c\in\mathbf{c}} \alpha_c = 1\right\}$,
- $\mathrm{J}_c(O_t,\psi_{t;c}|\Theta_c) = \mathrm{G}_c(\psi^{\mathsf{c}}_c) \prod_{j=1}^{m_c} \left(\frac{\Theta_{cj}}{\tilde{K}_{cj}}\right)^{\Delta_{cj}}$ $= \frac{1}{|\boldsymbol{\psi}^{\mathsf{c}}_c|} \prod_{j=1}^{m_c} \left(\frac{\Theta_{cj}}{\tilde{K}_{cj}}\right)^{\Delta_{cj}}$ is the c-th component in the mixture model (9), $m_c \in \mathbb{N}$ denotes the number of parameters of the c-th component; the non-parametric probability $\mathrm{G}_c(\psi^{\mathsf{c}}_c)$ (7) is chosen as uniform on ψ^{c}_c, i.e. $\mathrm{G}_c(\psi^{\mathsf{c}}_c) = \frac{1}{|\boldsymbol{\psi}^{\mathsf{c}}_c|}$,
- Θ denotes the parameter vector, $\boldsymbol{\Theta}$ is the set of all possible values of Θ and they have the form:

$$\Theta = (\alpha, (\Theta_c)_{c\in\mathbf{c}}) \qquad \boldsymbol{\Theta} = \boldsymbol{\alpha} \underset{c\in\mathbf{c}}{\times} \boldsymbol{\Theta}_c, \tag{10}$$

  where $\Theta_c = (\Theta_{c1}, \Theta_{c2}, ..., \Theta_{cm_c})$ is the parameter vector specific for the c-th component of the mixture and $\boldsymbol{\Theta}_c$ the set of all its possible values, $\boldsymbol{\Theta}_c = \left\{\Theta_c \in [0,1]^{m_c} \,|\, \sum_{j=1}^{m_c} \Theta_{cj} = 1\right\}$,
- $\Delta_{cj} = \Delta_{cj}(O_t, \psi_{t;c})$ is the indicator function[1] for the parameter Θ_{cj}, generally it has the form of a combination (sums, multiplications and compositions) of the Kronecker delta functions on subparts of the vector $(O_t, \psi_{t;c})$; the indicator outputs 0 or 1 for each vector $(O_t, \psi_{t;c})$; the indicator vector $\Delta_c(O_t, \psi_{t;c}) = (\Delta_{c1}, \Delta_{c2}, ..., \Delta_{cm_c})$ outputs 1 on at most one position, 0 on all remaining positions.

[1]The use of Δ_{cj} instead of B_{cj} (as in Definition 2) stresses that observations O and regression vectors ψ are discrete valued.

- $\tilde{K}_{cj}$ is the normalizing constant belonging to the parameter Θ_{cj}, generally, $\tilde{K}_{cj}$ is the number of values of the vector $(O_t, \psi_{t;c})$ for which $\Delta_{cj}(O_t, \psi_{t;c})$ outputs 1, i.e.

$$\tilde{K}_{cj} = \sum_{(O_t,\psi_{t;c})\in\mathbf{O}\times\boldsymbol{\psi}_c} \Delta_{cj}(O_t, \psi_{t;c}). \tag{11}$$

The notation $K_{cj} = |\boldsymbol{\psi}_c^{\mathsf{C}}|\tilde{K}_{cj}$ will be used. □

Remark 2: The corresponding parametric model of the observation $O_t \in \mathbf{O}$ (cf. Definition 1) has the following mixture ratio form:

$$\mathrm{M}(O_t|\psi_t,\Theta) = \mathrm{M}_t(\Theta) = \mathrm{H}_t(\Theta)\sum_{c\in\mathbf{c}}\alpha_c\prod_{j=1}^{m_c}\left(\frac{\Theta_{cj}}{K_{cj}}\right)^{\Delta_{cj}}, \tag{12}$$

$$\text{where}\quad \mathrm{H}_t(\Theta) = \frac{1}{\sum_{O_t\in\mathbf{O}}\sum_{c\in\mathbf{c}}\alpha_c\prod_{j=1}^{m_c}\left(\frac{\Theta_{cj}}{K_{cj}}\right)^{\Delta_{cj}}}$$

V. Monte Carlo Comparison

The simulation comparison of the Markov chain *mixture ratio model* discussed in Section IV and the *standard* Markov chain *mixture model*, where the parametric conditional pd $\mathrm{M}_C(O_t|\psi_t,\bar{\Theta})$ (compared to the joint pd $\mathrm{J}(O_t,\psi_t|\Theta)$) is modeled, see [18], was made. The comparison was made as follows:

- A sequence of observations $(O_t)_{t=1}^n$ was generated by the pd (12) with known parameters Θ_R (i.e. $\mathrm{M}_R(O_t|\psi_t) = \mathrm{M}(O_t|\psi_t,\Theta_R)$; $\mathrm{M}_R(O_t|\psi_t)$ is referred as the *real model*), actions were generated from the pd $S(A_t|\overline{\psi}_{t-1})$, where $\overline{\psi}_{t-1} = O_{t-1}, A_{t-1}, ..., O_{t-n-1}, A_{t-n-1}$ (see (5)),
- both, the mixture ratio model and the standard mixture model were learnt recursively,
- the quality of both models during the simulation was compared using KullbackLeibler divergence of pairs[2] of joint pds $\mathrm{P}(O_1,...,O_t,A_1,...,A_t|\psi_0)$, $t\in\mathbf{t}$.

Let $\mathrm{M}_J(O_t|\psi_t,\underline{D}^{t-1})$ and $\mathrm{M}_C(O_t|\psi_t,\underline{D}^{t-1})$ be learnt predictors for the mixture ratio model and standard mixture model, respectively. The predictors depends on particular data realizations $\underline{D}^{t-1} = (\underline{O}_\tau,\underline{A}_\tau)_{\tau=1}^{t-1}$ and are computed as follows:

$$\mathrm{M}_N(O_t|\psi_t,\underline{D}^{t-1}) = \int_{\boldsymbol{\Theta}}\mathrm{M}_N(O_t|\psi_t,\Theta,\underline{D}^{t-1})\mathrm{P}(\Theta|\underline{D}^{t-1})d\Theta,$$
$$N\in\{J,C\}.$$

To simplify the notation, the dependence on particular data realization $\underline{D}^{t-1}$ will not be stressed, i.e. $\mathrm{M}_J(O_t|\psi_t,\underline{D}^{t-1}) = \mathrm{M}_J(O_t|\psi_t)$, $\mathrm{M}_C(O_t|\psi_t,\underline{D}^{t-1}) = \mathrm{M}_C(O_t|\psi_t)$.

Furthermore, let $\mathrm{P}_R(O_1,...,O_t,A_1,...,A_t|\psi_0)$, $\mathrm{P}_J(O_1,...,O_t,A_1,...,A_t|\psi_0)$, $\mathrm{P}_C(O_1,...,O_t,A_1,...,A_t|\psi_0)$, $t\in\mathbf{t}$, be joint pds for the real model, the mixture ratio model and the standard mixture model respectively. They equal:

$$\mathrm{P}_N(O_1,...,O_t,A_1,...,A_t|\psi_0) = \prod_{\tau=1}^{t}\mathrm{M}_N(O_\tau|\psi_\tau)\mathrm{S}(A_\tau|\overline{\psi}_{\tau-1}) \tag{13}$$
$$N\in\{R,J,C\}.$$

KullbackLeibler divergences of the joint pds then equal[3]:

$$\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_N) = \sum_{\substack{O_1,...O_t\in\mathbf{O}\\A_1,...,A_t\in\boldsymbol{A}}}\prod_{\tau=1}^{t}\mathrm{M}_R(O_\tau|\psi_\tau)\mathrm{S}(A_\tau|\overline{\psi}_{\tau-1}) \tag{14}$$
$$\times\ln\left(\prod_{\tau=1}^{t}\frac{\mathrm{M}_R(O_\tau|\psi_\tau)\mathrm{S}(A_\tau|\overline{\psi}_{\tau-1})}{\mathrm{M}_N(O_\tau|\psi_\tau)\mathrm{S}(A_\tau|\overline{\psi}_{\tau-1})}\right),$$
$$N\in\{J,C\}, t\in\mathbf{t}.$$

(14) can be rewritten as follows, $N\in\{J,C\}$:

$$\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_L) = \sum_{\tau=1}^{t}\sum_{\substack{O_1,...O_t\in\mathbf{O}\\A_1,...,A_t\in\boldsymbol{A}}}\prod_{\tau=1}^{t}\mathrm{M}_R(O_\tau|\psi_\tau)\mathrm{S}(A_\tau|\overline{\psi}_{\tau-1})$$
$$\times\ln\left(\frac{\mathrm{M}_R(O_\tau|\psi_\tau)}{\mathrm{M}_N(O_\tau|\psi_\tau)}\right)$$
$$=\sum_{\tau=1}^{t}\sum_{\substack{O_\tau\in\mathbf{O}\\\psi_\tau\in\boldsymbol{\psi}}}\mathrm{M}_R(O_\tau|\psi_\tau)\mathrm{P}(\psi_\tau)\times\ln\left(\frac{\mathrm{M}_R(O_\tau|\psi_\tau)}{\mathrm{M}_N(O_\tau|\psi_\tau)}\right). \tag{15}$$

The joint pd $\mathrm{P}(\psi_t)$ is computed recursively, starting with the known $\mathrm{P}(\psi_0)$, as follows:

$$\mathrm{P}(\psi_t) = \sum_{\substack{O_{t-n-1}\in\mathbf{O}\\A_{t-n-1}\in\mathbf{A}}}\mathrm{P}(\psi_t,O_{t-n-1},A_{t-n-1})$$
$$=\sum_{\substack{O_{t-n-1}\in\mathbf{O}\\A_{t-n-1}\in\mathbf{A}}}\mathrm{P}(A_t,O_{t-1},\psi_{t-1})$$
$$=\sum_{\substack{O_{t-n-1}\in\mathbf{O}\\A_{t-n-1}\in\mathbf{A}}}\mathrm{S}(A_t|\overline{\psi}_{t-1})\mathrm{P}(O_{t-1}|\psi_{t-1})\mathrm{P}(\psi_{t-1})$$

with $\psi_t = A_t, O_{t-1}, A_{t-1}, ..., O_{t-n}, A_{t-n}$,
$\overline{\psi}_{t-1} = O_{t-1}, A_{t-1}, ..., O_{t-n}, A_{t-n}$,
$\psi_{t-1} = A_{t-1}, O_{t-2}, A_{t-2}, ..., O_{t-n-1}, A_{t-n-1}$.

As stated above, the quality of both models is compared by comparing the values of $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_J)$ and $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_C)$ computed via (15), where the smaller value of the Kullback-Leibler divergence indicates better predicting and modeling quality of the respective model.

The learnt predictors $\mathrm{M}_J(O_t|\psi_t)$ and $\mathrm{M}_C(O_t|\psi_t)$ depend on the particular data realizations. Thus, to make a reliable comparison of the mentioned variables, it is necessary to make a Monte Carlo (MC) study of them.

[2]In particular, KullbackLeibler divergence was computed for two pairs of the "real" and "learnt" pds (one learnt with mixture ratio model and the other with standard mixture model), the details follow.

[3]The time index $t\in\mathbf{t}$ in $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_N)$ denotes Kullback-Leibler divergence after the t-th observation.

A. Compared Models

In simulation, the system with 5 possible observations $O_t \in \mathbf{O} = \{1, 2, ..., |\mathbf{O}|\}$, where $|\mathbf{O}| = 5$ was modeled. Both mixture ratio model (denoted as $\mathrm{M}_J(O_t|\psi_t, \Theta)$) and standard mixture model (denoted as $\mathrm{M}_C(O_t|\psi_t, \bar{\Theta})$) were 2-component and used the same parametric space $\mathbf{\Theta}$, regression vector ψ_t and respective components operated on the same subsets of data-vector $\boldsymbol{\phi}$ ($\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, see Definition 9), which are specified as follows:

- $\mathbf{\Theta} = \boldsymbol{\alpha} \times \mathbf{\Theta}_1 \times \mathbf{\Theta}_2$,
- $\boldsymbol{\alpha} = \mathbf{\Theta}_1 = \mathbf{\Theta}_2 = \{(a, b) \in [0, 1]^2 \mid a + b = 1\}$,
- $\psi_t = (O_{t-1}, O_{t-2}) \in \boldsymbol{\psi} = \mathbf{O} \times \mathbf{O}$,
- $\phi_t = (O_t, \psi_t) \in \boldsymbol{\phi} = \mathbf{O} \times \mathbf{O} \times \mathbf{O}$,
- $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \mathbf{O} \times \mathbf{O}$,
- $\phi_{t;1} = (O_t, O_{t-1}) \qquad \phi_{t;2} = (O_t, O_{t-2})$.

a) Mixture ratio model: The joint pd of the mixture ratio model equals (cf. (9)):

$$\begin{aligned}\mathrm{J}(\phi_t|\Theta) &= \mathrm{J}(O_t, \psi_t|\Theta) \\ &= \alpha \mathrm{J}(O_t, O_{t-1}|\Theta) + (1-\alpha)\mathrm{J}(O_t, O_{t-2}|\Theta) \\ &= \alpha \frac{1}{25} \prod_{i=1}^{25} \Theta_{1i}^{\Delta_{1i}} + (1-\alpha)\frac{1}{25}\prod_{i=1}^{25}\Theta_{2i}^{\Delta_{2i}}, \quad \text{where}\end{aligned}$$

- $\Theta = \left(\alpha, (\Theta_{1i})_{i=1}^{25}, (\Theta_{2i})_{i=1}^{25}\right) \in \mathbf{\Theta}, \quad \sum_{i=1}^{25}\Theta_{1i} = \sum_{i=1}^{25}\Theta_{2i} = 1$,
- Δ_{1i} and Δ_{2i} are indicators for all combinations (O_t, O_{t-1}) and (O_t, O_{t-2}), respectively.

The corresponding parametric system model then reads (cf. (8), (12)):

$$\begin{aligned}\mathrm{M}_J(O_t|\psi_t, \Theta) &= \frac{\alpha \prod_{i=1}^{25}\Theta_{1i}^{\Delta_{1i}} + (1-\alpha)\prod_{i=1}^{25}\Theta_{2i}^{\Delta_{2i}}}{\underbrace{\alpha \sum_{O_t \in \mathbf{O}}\prod_{i=1}^{25}\Theta_{1i}^{\Delta_{1i}}}_{\mathrm{W}_1(\psi_t,\Theta_1)} + \underbrace{(1-\alpha)\sum_{O_t\in\mathbf{O}}\prod_{i=1}^{25}\Theta_{2i}^{\Delta_{2i}}}_{\mathrm{W}_2(\psi_t,\Theta_2)}} \\ &= \mathrm{w}_1(\psi_t, \Theta)\frac{\prod_{i=1}^{25}\Theta_{1i}^{\Delta_{1i}}}{\mathrm{W}_1(\psi_t, \Theta_1)} \\ &\quad + \mathrm{w}_2(\psi_t, \Theta)\frac{\prod_{i=1}^{25}\Theta_{2i}^{\Delta_{2i}}}{\mathrm{W}_2(\psi_t, \Theta_2)}, \end{aligned} \tag{16}$$

- $\mathrm{w}_1(\psi_t, \Theta) = \frac{\alpha\,\mathrm{W}_1(\psi_t,\Theta_1)}{\alpha \mathrm{W}_1(\psi_t,\Theta_1)+(1-\alpha)\mathrm{W}_2(\psi_t,\Theta_2)}$,
- $\mathrm{w}_2(\psi_t, \Theta) = \frac{(1-\alpha)\,\mathrm{W}_2(\psi_t,\Theta_2)}{\alpha \mathrm{W}_1(\psi_t,\Theta_1)+(1-\alpha)\mathrm{W}_2(\psi_t,\Theta_2)}$.

b) Standard mixture model: The standard mixture model had the following form:

$$\mathrm{M}_C(O_t|\psi_t, \bar{\Theta}) = \bar{\alpha}\prod_{i=1}^{25}\bar{\Theta}_{1i}^{\Delta_{1i}} + (1-\bar{\alpha})\prod_{i=1}^{25}\bar{\Theta}_{2i}^{\Delta_{2i}}, \tag{17}$$

- $\bar{\Theta} = \left(\bar{\alpha}, \left(\bar{\Theta}_{1i}\right)_{i=1}^{25}, \left(\bar{\Theta}_{2i}\right)_{i=1}^{25}\right) \in \bar{\mathbf{\Theta}}$,
- $\sum_{j=1}^{5}\Theta_{1(5i+j)} = \sum_{j=1}^{5}\Theta_{2(5i+j)} = 1$, $i \in \{0, 1, 2, 3, 4\}$,
- Δ_{1i} and Δ_{2i} are indicators for all combinations (O_t, O_{t-1}) and (O_t, O_{t-2}), respectively.

B. Results of Comparison

a) Compared quality of models: The quality of models was compared by a MC study. Three sets of 200 stochastic simulations (with identical initial conditions) of the dynamic system were made. In each simulation, both, the mixture ratio and the standard mixture models were learnt recursively on 500 observations. Those were generated by the real model $\mathrm{M}_R(O_t|\psi_t)$, which was different for each set of simulations:

1) The real model was the mixture ratio model with parameters $\underline{\Theta}_R \in \mathbf{\Theta}$ generated randomly for each simulation, i.e. $\mathrm{M}_R(O_t|\psi_t) = \mathrm{M}_J(O_t|\psi_t, \underline{\Theta}_R)$ (16).
2) The real model was the standard mixture model with parameters $\underline{\Theta}_S \in \bar{\mathbf{\Theta}}$ generated randomly for each simulation, i.e. $\mathrm{M}_R(O_t|\psi_t) = \mathrm{M}_C(O_t|\psi_t, \underline{\Theta}_S)$ (17).
3) The real model was the mixture ratio model with parameters $\underline{\Theta}_F \in \mathbf{\Theta}$, which were fixed for all simulations giving true dynamic weights $w_c(\psi_t)$, i.e. $\mathrm{M}_R(O_t|\psi_t) = \mathrm{M}_J(O_t|\psi_t, \underline{\Theta}_F)$ (16).

The initial statistics $\mathrm{V}_0 = (\mathrm{v}_0, (\mathrm{V}_{0;c})_{c\in\mathbf{c}})$ describing prior information about the parameters were set to

$$\mathrm{v}_0 = 1 \qquad \mathrm{V}_{0;c} = 1 \qquad c \in \mathbf{c} \tag{18}$$

for both models, which express no prior information about the parameters. The initial regression vector ψ_0 was set to

$$\psi_0 = (O_0, O_{-1}, O_{-2}) = (|\mathbf{O}|, |\mathbf{O}|, |\mathbf{O}|). \tag{19}$$

The values of Kullback-Leibler divergence for both models (at the end of the simulation), $\mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_J)$ and $\mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_C)$, were studied (see (15)). They were compared by their differences

$$\Delta = \mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_J) - \mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_C). \tag{20}$$

$\mathrm{P}_R(O_1, ..., O_t, A_1, ..., A_t|\psi_0)$, $\mathrm{P}_J(O_1, ..., O_t, A_1, ..., A_t|\psi_0)$ and $\mathrm{P}_C(O_1, ..., O_t, A_1, ..., A_t|\psi_0)$ denote the joint pds for the real model, the mixture ratio model and the standard mixture model respectively (cf. (13)).

The results are summarized in Table I and in Figure 1. For the illustration, Figure 2 displays time-evolution of Kullback-Leibler divergence for both models, $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_J)$ and $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_C)$ (15), in one of the simulations.

To illustrate the truly dynamic property of the mixture ratio model, the Figure 3 shows the time-evolution of the weight of the first component $\mathrm{w}_1(\psi_t, \Theta_R)$, (16), of the real model during one simulation.

b) Compared effects of forgetting: During each of the simulations, the effects of forgetting factor were also compared. The learning with the mixture ratio model was done with dynamic forgetting factor, see in Section II, as well as with fixed forgetting factors $\lambda_t = 1, \lambda_{t;1} = 1, \lambda_{t;2} = 1$ (which implies no forgetting). Let $\overline{\mathrm{P}}_J(O_1, ..., O_t, A_1, ..., A_t)$ denote the joint pd for the model with fixed forgetting factor.

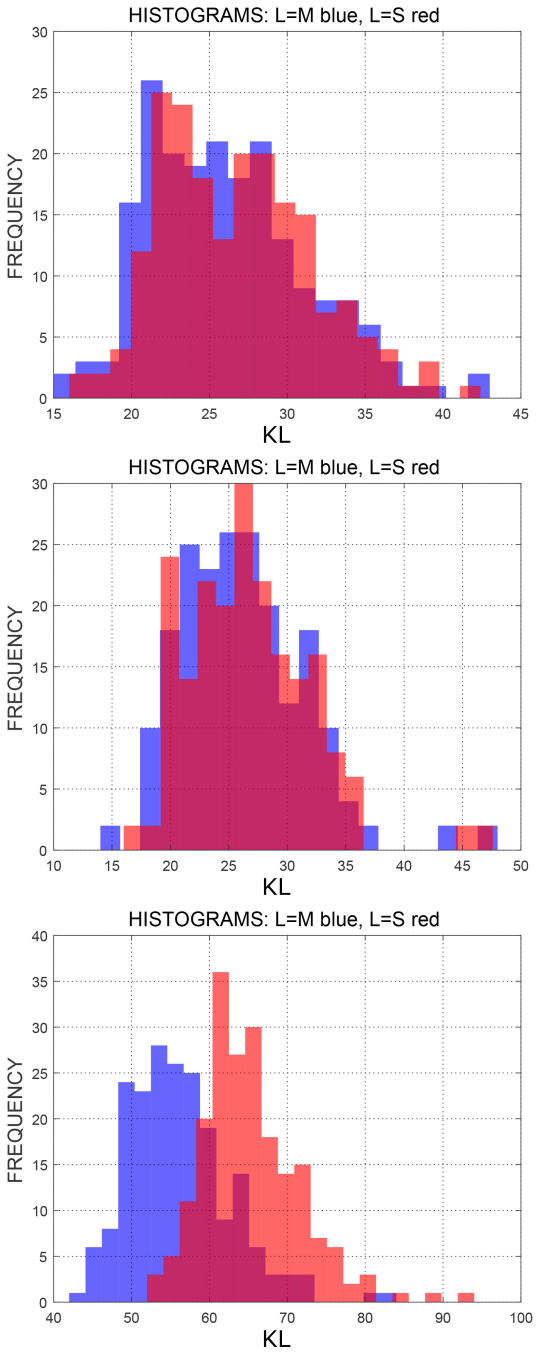

Fig. 1. Histograms of $\mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_L)$ (15), $\mathsf{N} \in \{\mathsf{J}, \mathsf{C}\} = \{\text{blue}, \text{red}\}$. The rows corresponds to the sets of three simulations introduced in the beginning of this section: the 1-st column stands for the mixture ratio model with random parameters $\underline{\Theta}_R$ for each simulation; the 2-nd column stands for the standard mixture model with random parameters $\underline{\Theta}_S$ for each simulation; the 3-rd stands for the mixture ratio model with fixed parameters $\underline{\Theta}_F$ having truly dynamic weights $\mathrm{w}_c(\psi_t)$.

TABLE I
SAMPLE STATISTICS OF Δ (20).

Simulated Case	1)	2)	3)
Mean	-0.7001	-0.6029	-9.1191
Median	-0.7818	-0.5957	-9.0598
Minimum	-3.2138	-4.0519	-16.1047
Maximum	4.9534	1.9341	-4.5774
Standard Deviation	1.0583	0.9377	2.0317

The columns belongs to the simulations carried with different real models discussed in the beginning of this section: the 1-st column stands for the mixture ratio model with random parameters $\underline{\Theta}_R$ for each simulation; the 2-nd column stands for the standard mixture model with random parameters $\underline{\Theta}_S$ for each simulation; the 3-rd stands for the mixture ratio model with fixed parameters $\underline{\Theta}_F$ having truly dynamic weights $\mathrm{w}_c(\psi_t)$.

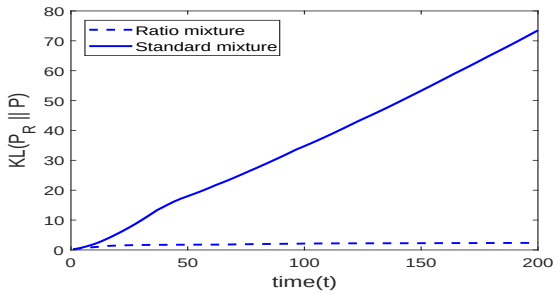

Fig. 2. The figure shows the time-evolution of the Kullback-Leibler divergence of the ratio mixture model (16) $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_J)$ (dashed line) and the standard mixture model (17) $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_C)$ (solid line) (see (15)) in one of the simulations of the MC study for the third type of simulation - the mixture ratio model with fixed parameters $\underline{\Theta}_F$ having truly dynamic weights $\mathrm{w}_c(\psi_t)$.

The Kullback-Leibler divergences $\mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_J)$ and $\mathrm{KL}^{200}(\mathrm{P}_R||\overline{\mathrm{P}}_J)$, as well as the increments [4]

$$\overline{\mathrm{KL}}(\mathrm{P}_R||\mathrm{P}_J) = \mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_J) - \mathrm{KL}^{199}(\mathrm{P}_R||\mathrm{P}_J), \quad (21)$$

$$\overline{\mathrm{KL}}(\mathrm{P}_R||\overline{\mathrm{P}}_J) = \mathrm{KL}^{200}(\mathrm{P}_R||\overline{\mathrm{P}}_J) - \mathrm{KL}^{199}(\mathrm{P}_R||\overline{\mathrm{P}}_J), \quad (22)$$

were studied and the results are summarized in Table II.

[4]The increments $\overline{\mathrm{KL}}(\mathrm{P}_R||\mathrm{P}_J)) = \mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_J) - \mathrm{KL}^{199}(\mathrm{P}_R||\mathrm{P}_J)$ and $\overline{\mathrm{KL}}(\mathrm{P}_R||\overline{\mathrm{P}}_J)) = \mathrm{KL}^{200}(\mathrm{P}_R||\overline{\mathrm{P}}_J) - \mathrm{KL}^{199}(\mathrm{P}_R||\overline{\mathrm{P}}_J)$ describe the quality of the learnt predictors for the last observation.

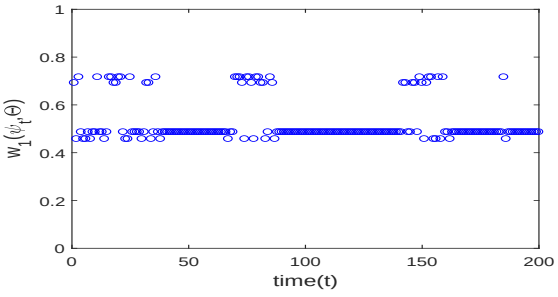

Fig. 3. The figure shows the time-evolution of the first dynamic component weight $\mathrm{w}_1(\psi_t, \Theta_R)$ (16) of the real model in one of the simulations of the MC study.

TABLE II
MC STUDY

	$\mathrm{KL}^{200}(\mathrm{P}_R\|\|\mathrm{P}_J)$ $-\mathrm{KL}^{200}(\mathrm{P}_R\|\|\overline{\mathrm{P}}_J)$	$\overline{\mathrm{KL}}(\mathrm{P}_R\|\|\mathrm{P}_J))$ $-\overline{\mathrm{KL}}(\mathrm{P}_R\|\|\overline{\mathrm{P}}_J))$
mean	0.5733	0.000773
median	0.4800	0.000714
minimum	-0.8155	-0.0035
maximum	3.6706	0.0077
standard deviation	0.7026	0.0024

The table shows outcomes of the MC study comparing the quality of the mixture ratio (16) model with dynamic forgetting and the mixture ratio model with no forgetting. The comparison is made by the values of the differences of the respective Kullback-Leibler divergences $\mathrm{KL}^{200}(\mathrm{P}_R||\mathrm{P}_J)-\mathrm{KL}^{200}(\mathrm{P}_R||\overline{\mathrm{P}}_J)$, as well as the the differences of the Kullback-Leibler divergences for the last observation (see (15), (21) and (22)). The mean, median, minimum, maximum and standard deviation were computed from the outcomes of 100 simulations.

The Figure 4 shows the time-evolution of the Kullback-Leibler divergence $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_J)$ and $\mathrm{KL}^t(\mathrm{P}_R||\overline{\mathrm{P}}_J)$ as well as the time-evolution of the dynamic forgetting factors $\lambda_t, \lambda_{t;1}$ and $\lambda_{t;2}$ in one of the simulations.

c) Discussion: The carried MC study shows several results. If the modelled system is truly dynamic with dynamic weights of the mixture components, the standard mixture model (described in detail in [18]) is not sufficient to model such a system. The results in Table I shows, that even the "minimum" row (corresponding to the best simulation in the MC study for the standard model) suggests significantly higher quality of the mixture ratio model over the standard mixture model. Figure 2 illustrates this result.

Figure 3 demonstrates, that even for relatively simple two-component model, the dynamic weights can vary significantly during a simulation. This property of the ratio mixture model is potentially suitable for modeling non-linear stochastic systems.

The MC comparison of the dynamic forgetting factors proposed in Section II with fixed forgetting factors $\lambda_t = \lambda_{t;c} = 1, c \in \mathbf{c}$ (i.e. without forgetting), summarized in Table II, indicates decreased quality of the model with dynamic forgetting. Thus, the use of proposed dynamic forgetting factors remains as an open problem.

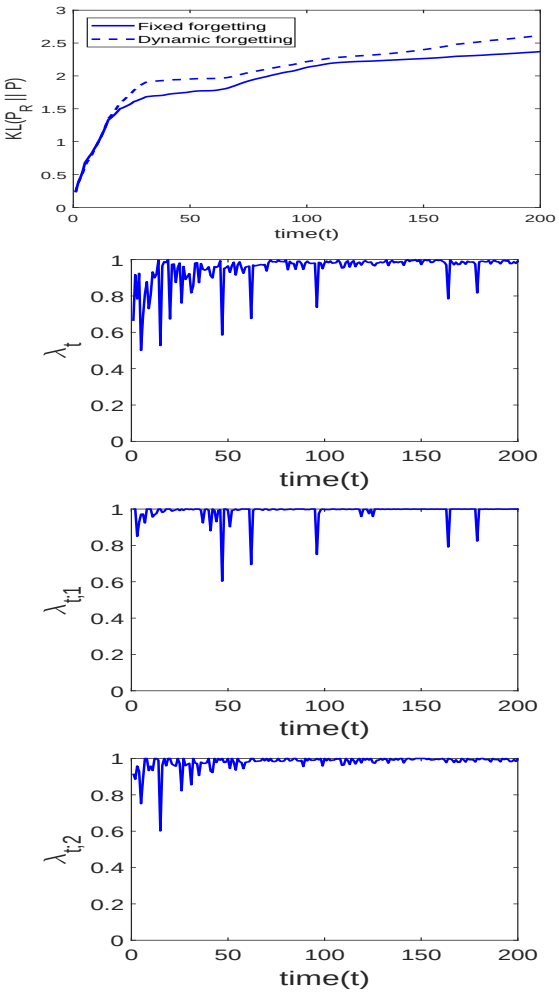

Fig. 4. The first row shows the time-evolution of the Kullback-Leibler divergence of the ratio mixture model with dynamic forgetting, $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_J)$ (dashed line), the ratio mixture model with no forgetting $\mathrm{KL}^t(\mathrm{P}_R||\overline{\mathrm{P}}_J)$ (solid line) (16), the second row shows the time-evolution of the dynamic forgetting factors λ_t, $\lambda_{t;1}$, $\lambda_{t;2}$, respectively. Both illustrates one of the simulations of the MC study.

VI. CONCLUSION

In this work, a very flexible, but yet uncosindered model set of mixture ratios with components from exponential family, was build. The approximate Bayesian learning, presented in [3] and summarized in Section II, was tested for the mixture ratios models.

The mixture ratio models with Markov chain components, as one of the important members of EF, were closely examined. The system model for the Markov chain with the Dirichlet conjugated prior pds with the learning algorithm from [32] was considered.

In Section V, the MC study comparing the mixture ratio model with the standard mixture model, built in [18], was made. The results of the study show that:

- the numerical approximations included in learning, suffice for an effective learning of the simulated system model (at least for low-dimensional systems),
- the standard mixture model is dominated by the proposed mixture ratio model when applied on truly dynamic system with dynamic component weights,
- the time-evolution of the component weights of the ratio mixture model can be significant even for low-dimensional models, which indicates the potential suitability for modeling of non-linear stochastic systems,
- the proposed dynamic forgetting factors, as a counter measure to the accumulation of approximation errors, did not provide better model and needs to be improved in the future.

The Section V also provides a general way how to compare quality of prediction of probabilistic models which is not

limited only to the mixture models.

Many challenging tasks remains to be studied, in particular: i) studying other important members of EF, such as Gaussian mixtures, ii) examining the suitability of the proposed numerical approximations for high-dimensional models, iii) testing dynamic forgetting factors for simulations with more observations and possibly modifying it, iv) testing the proposed ratio mixture model with real data in some real-world scenario, including active interaction with the system and optimization of the strategy of a decision maker.

REFERENCES

[1] J.M. Bernardo. *Expected information as expected utility*. The Annals of Statistics, 7(3): 686-690, 1979.

[2] M. Kárný and T.V. Guy. *On support of imperfect Bayesian participants*. In T.V. Guy, M. Kárný, and D.H. Wolpert, editors, Decision Making with Imperfect Decision Makers, volume 28. Springer, Berlin, 2012. Intelligent Systems Reference Library.

[3] M. Kárný. *Approximate Bayesian recursive estimation*. Information Sciences, 289:100-111, 2014.

[4] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, New York, 2000.

[5] R.B. Nelsen. *An Introduction to Copulas*. Springer, New York, 1999.

[6] K. Dedecius, I. Nagy, M. Kárný, and L. Pavelková. *Parameter estimation with partial forgetting method*. In Proc. of the 15th IFAC Symposium on Identification and System Parameter Estimation - SYSID, 2009.

[7] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, NY, 1978.

[8] R.E. Bellman, *Adaptive Control Processes*, Princeton University Press, NJ, 1961.

[9] L. Berec and M. Kárný. Identification of reality in Bayesian context. In K. Warwick and M. Kárný, editors, *Computer-Intensive Methods in Control and Signal Processing*, pages 181–193. Birkhäuser, 1997.

[10] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, NY, 1985.

[11] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] T. Bohlin. *Interactive System Identification: Prospects and Pitfalls*. Springer, NY, 1991.

[13] P. Guan, M. Raginsky, and R. Willett. Online Markov decision processes with Kullback-Leibler control cost. In *American Control Conference*, pages 1388–1393. IEEE, 2012.

[14] I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: Beyond the Bayesian/frequentist divide. *Journal of Machine Learning Research*, 11:61–87, 2010.

[15] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan, NY, 1994.

[16] G. Holmes and T.Y. Liu, editors. *Proceedings of 7th Asian Conference on Machine Learning (ACML2015), JMLR Workshop and Conference Proceedings*, volume 45, 2015.

[17] A.M. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, NY, 1970.

[18] M. Kárný. Recursive estimation of high-order Markov chains: Approximation by finite mixtures. *Infor. Sciences*, 326:188 – 201, 2016.

[19] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, 2006.

[20] D.F. Kerridge. Inaccuracy and inference. *J. of the Royal Statistical Society*, B 23:284–294, 1961.

[21] R. Koopman. On distributions admitting a sufficient statistic. *Trans. of Am. Math. Society*, 39:399, 1936.

[22] W. Mason, J.W. Vaughan, and H. Wallach. Special issue: Computational social science and social computing. *Machine Learning*, 96:257–469, 2014.

[23] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[24] P. Sadghi, R.A. Kennedy, P.B. Rapajic, and R. Shams. Finite-state Markov modeling of fading channels. *IEEE Signal Processing Magazine*, 57, 2008.

[25] L.J. Savage. *Foundations of Statistics*. Wiley, NY, 1954.

[26] A. Wald. *Statistical Decision Functions*. John Wiley, New York, London, 1950.

[27] M. Wiering and M. van Otterlo, editors. *Reinforcement Learning: State-of-the-Art*. Springer-Verlag, 2012.

[28] I. Mez. *Some infinite sums arising from the Weierstrass Product Theorem*. Applied Mathematics and Computation. 219: 98389846, 2013.

[29] A.A. Markov. *Extension of the limit theorems of probability theory to a sum of variables connected in a chain*. reprinted in Appendix B of: R. Howard. Dynamic Probabilistic Systems, volume 1: Markov Chains. John Wiley and Sons, 1971.

[30] M. Ruman, F. Hůla, M. Kárný, and T.V. Guy. *Deliberation-aware responder in multi- proposer ultimatum game*. In Artificial Neural Networks and Machine Learning - Proceedings ICANN 2016, pages 230-237. Barcelona, 2016.

[31] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[32] M. Kárný, M. Ruman: Mixture Ratio Modelling of Dynamic Environments, Neural Networks, 2019, submitted.