Information Sciences 326 (2016) 188–201

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Recursive estimation of high-order Markov chains: Approximation by finite mixtures

Miroslav Kárný*

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 23 February 2014
Revised 29 June 2015
Accepted 17 July 2015
Available online 23 July 2015

Keywords:
Markov chain
Approximate parameter estimation
Bayesian recursive estimation
Adaptive systems
Kullback–Leibler divergence
Forgetting

ABSTRACT

A high-order Markov chain is a universal model of stochastic relations between discrete-valued variables. The exact estimation of its transition probabilities suffers from the curse of dimensionality. It requires an excessive amount of informative observations as well as an extreme memory for storing the corresponding sufficient statistic. The paper bypasses this problem by considering a rich subset of Markov-chain models, namely, mixtures of low dimensional Markov chains, possibly with external variables. It uses Bayesian approximate estimation suitable for a subsequent decision making under uncertainty. The proposed recursive (sequential, one-pass) estimator updates a product of Dirichlet probability densities (pds) used as an approximate posterior pd, projects the result back to this class of pds and applies an improved data-dependent stabilised forgetting, which counteracts the dangerous accumulation of approximation errors.

1. Introduction

Adaptive systems select their actions while simultaneously learn dynamics of the environment they interact with [5]. The joint acting and learning is their key feature, which singles them from other controllers, estimators, predictors, classifiers, etc. It is the main source of their strength, applicability [40] and universality [27]. It allows them to work with simple, often input–output, models describing their interactions with their environments locally [23]. Recursively estimated black-box models, relating the observed past to future observations, serve to adaptive systems exploited for prediction [4], decision support [42], feedback control [43], etc.[1] All of them de facto perform dynamic decision making under uncertainty, which has Bayesian paradigm [10,49] as the theoretically most elaborated base. This makes us to focus on it.

Often, the modelled observations as well as explanatory variables are discrete or discretised. Then, a high-order Markov chain provides their universal model. It relates the predicted observation to a finite-dimensional regression vector containing the past observations and explanatory variables. Its recursive Bayesian estimation, which provides a lossless compression of the available knowledge, is formally simple. Basically, it counts joint occurrences of the predicted variable and the corresponding regression vector. The applicability is, however, strongly limited by the curse of dimensionality [8] as the size of the occurrence array blows up with the number of possible occurrence instances. Then, the observations insufficiently populate it.

* Tel.: +420266052274.
E-mail address: school@utia.cas.cz

[1] The vast literature on the topic forces us to provide just subjectively selected samples.

http://dx.doi.org/10.1016/j.ins.2015.07.038
0020-0255/

This curse is mostly counteracted by exploiting conditional independence of modelled variables, cf. Bayesian network [21], compositional models [22], etc. This quite powerful way allows coping even with computationally hard designs of decision strategies [12,19], which exploit specific factored representations. The data-based factorisation into conditionally independent groups itself faces, even more pronounced, curse of dimensionality. This limits universality of these techniques and especially their use in exploratory data analysis and black-box-models-based adaptive systems.

The current paper deals with a rich class of estimation problems in which a detailed independence structure is not supplied by experts. The considered regression problems are delimited by a few possible observation values but a long regression vector is allowed. For them, a finite mixture of Markov chains relating the predicted observation to individual entries of the regression vector is employed. Demonstrations of the modelling strength of such a mixture are in [9,45,48]. The employed solutions, however, fully rely on batch processing, which prevents their permanent use to adaptive systems, data stream processing and modelling of non-stationary phenomena [47].

The recursive learning of a finite mixture of Markov chains is inevitably approximate and consequently endangered by accumulation of approximation errors. In parameter estimation, the "natural" error-damping effect of state-space (partially observable) models [13] does not exist and the accumulation-errors-free algorithms [36,37] operate on too narrow class of non-sufficient statistics. A recent inspection of this problem shown that stabilised forgetting [25] provides a general counter-measure against the discussed accumulation.

The stabilised forgetting and handling of Dirichlet pds summarised [26] provide essential ingredients for the inspected *approximate Bayesian recursive estimation of finite Markov-chain mixtures*.

The estimator design is addressed within the unified Bayesian paradigm [11,29], which respects its use in a subsequent Bayesian decision making. Practically, it makes us to "neglect" important techniques, which are of a heuristic nature or less rigorously connected with the decision making. They include classical quasi-Bayes estimator [52] and its dynamic version [26] (heuristic), variational Bayes [51] (improper order of arguments in minimised Kullback–Leibler divergence), expectation propagation [41] or various point estimators (information on precision is missing), often based on expectation-maximisation algorithm [17] possibly combined with variational approach [50]. In summary, the paper:

- fills the niche in an extreme range of variants of estimating of models exploiting Markov-chains and complements toolkit of estimators dealing with mixed and/or hidden variables;
- adopts parsimonious probabilistic model of finite-mixture type [9,45] relating discrete-valued predicted variable to its multiple delayed values and to external discrete-valued variables;
- designs novel recursive Bayesian estimator of this black-box input–output model, which fits to its intended use for adaptive decision making operating solely on discrete-valued observations;
- designs approximate, recursive estimator via a recent systematic, theoretically justified, way [25]: this distinguishes the estimator from its heuristically justified predecessors like quasi-Bayes estimator [26,52];
- serves as an example of the general theory of approximate recursive estimation [25] and improves its only heuristic step encountered.

Concerning the layout, Section 2 formalises the problem. Core Section 3 solves it. Section 4 describes how to counteract accumulation of errors caused by the approximate estimation and improves the heuristic step adopted in [25]. Section 5 illustrates the theory and its notions on a simple case of exploratory analysis. Other examples there provide comparison with a few published examples addressing similar objectives as the current paper. Section 6 adds concluding remarks.

2. Formalisation of the addressed problem

Scalar[2] discrete-valued *observations* $\Delta_t \in \boldsymbol{\Delta} \equiv \{1,\ldots,|\boldsymbol{\Delta}|\}$, $|\boldsymbol{\Delta}| \ll \infty$ are made at discrete time moments $t \in \boldsymbol{t} \equiv \{1,\ldots,T\}$, $T \leq \infty$, on the modelled stochastic environment. The observation Δ_t is assumed to depend on an $\ell_{\boldsymbol{\psi}}$-dimensional *regression vector* $\psi_t \in \boldsymbol{\psi}$ with discrete-valued entries[3] $\psi_{t;i} \in \boldsymbol{\psi}_i \equiv \{1,\ldots,|\boldsymbol{\psi}_i|\}$, $|\boldsymbol{\psi}_i| \ll \infty$, $i \in \boldsymbol{i} \equiv \{1,\ldots,\ell_{\boldsymbol{\psi}}\}$. The unknown dependence is assumed to be time-invariant. Thus, the observations are described by the *high-order Markov chain*, parameterised by time-invariant finite-dimensional array of transition probabilities Ω,

$$\mathrm{p}(\Delta_t|\Omega,\text{observed past}) = \mathrm{p}(\Delta_t|\Omega,\psi_t) \equiv \prod_{\psi\in\boldsymbol{\psi}}\prod_{\Delta\in\boldsymbol{\Delta}} \Omega_{\Delta|\psi}^{\delta(\Delta\Delta_t)\delta(\psi\psi_t)}. \tag{1}$$

There, p denotes *probability density (pd)* of the variable in its argument conditioned on the argument after the conditioning symbol |. Ω contains time-invariant unknown probabilities belonging to an appropriate probabilistic simplex. The subscript $_{\Delta|\psi}$ of its entries stresses that $\Omega_{\Delta|\psi}$ is the transition probability with properties

$$\Omega_{\Delta|\psi} \geq 0,\ \forall\Delta\in\boldsymbol{\Delta},\ \forall\psi\in\boldsymbol{\psi} \quad \text{and} \quad \sum_{\Delta\in\boldsymbol{\Delta}} \Omega_{\Delta|\psi} = 1\ \ \forall\psi\in\boldsymbol{\psi}.$$

[2] The consideration of a scalar-valued observation represents no restriction as the vector case can always be boiled down to it by employing entry-wise modelling [26]. $\boldsymbol{x}$ denotes a set of x values, $|\boldsymbol{x}|$ its cardinality and $\equiv$ means defining equality. Section 5 illustrates the notions being introduced.

[3] The symbol $\ell_{\boldsymbol{x}}$ is reserved for the length of a vector $x \in \boldsymbol{x}$. The time index $t \in \boldsymbol{t}$ is always the first one and semicolon separates it from other indices. The time index indicates that the values of observed and predicted variables or statistic values are meant.

Kronecker symbol $\delta(ab)$ equals to 1 for $a = b$ and to 0 otherwise.

The considered Bayesian estimation consists of evaluating the *posterior pd*

$$\mathsf{p}_t(\Omega) \equiv \mathsf{p}(\,\Omega \,|\, \text{prior knowledge enriched by the observations up to time } t\,).$$

The Markov chain (1) belongs to exponential family [6]. The *Dirichlet pd*

$$\mathsf{p}_0(\Omega) = \mathcal{D}_\Omega(V_0) \equiv \prod_{\psi\in\boldsymbol{\psi}} \frac{\prod_{\Delta\in\boldsymbol{\Delta}} \Omega_{\Delta|\psi}^{V_{0;\Delta|\psi}-1}}{\mathsf{B}(V_{0;\boldsymbol{\Delta}|\psi})} \tag{2}$$

is its *conjugate prior pd* [26]. There,

$$\begin{aligned} V_{0;\boldsymbol{\Delta}|\psi} &\equiv (V_{0;\Delta|\psi})_{\Delta\in\boldsymbol{\Delta}}, \quad V_0 \equiv V_{0;\boldsymbol{\Delta}|\boldsymbol{\psi}} \equiv (V_{0;\boldsymbol{\Delta}|\psi})_{\psi\in\boldsymbol{\psi}}, \quad V_{0;\boldsymbol{\Delta}|\psi} > 0, \\ \mathsf{B}(\nu_{\boldsymbol{k}}) &\equiv \frac{\prod_{k\in\boldsymbol{k}}\Gamma(\nu_k)}{\Gamma(\sum_{k\in\boldsymbol{k}}\nu_k)}, \quad \boldsymbol{k} = \{1,\dots,\boldsymbol{k}\}, \quad \Gamma(\nu_k) \equiv \int_0^\infty z^{\nu_k-1}\exp(-z)\,\mathrm{d}z. \end{aligned} \tag{3}$$

B is the multivariate beta function [1]. It acts on a $|\boldsymbol{k}|$-vector $\nu_{\boldsymbol{k}}$ with positive entries ν_k, $k \in \boldsymbol{k}$. Here, the repeatedly used notation is adopted

$$\begin{aligned} \mathit{variable}_{\mathsf{set}} &\equiv (\mathit{variable}_{\mathsf{entry}})_{\mathsf{entry}\,\in\,\mathsf{set}} \\ \mathit{variable}_{\mathsf{set}|\mathsf{condition}} &\equiv \left(\mathit{variable}_{\mathsf{entry}|\mathsf{condition}}\right)_{\mathsf{entry}\,\in\,\mathsf{set}} \\ \mathit{variable}_{\mathsf{set}|\mathsf{condition}} &\equiv \left(\left(\mathit{variable}_{\mathsf{entry}|\mathsf{condition}}\right)_{\mathsf{entry}\,\in\,\mathsf{set}}\right)_{\mathsf{condition}\,\in\,\mathsf{condition}} \\ \mathit{variable}_t &\equiv \mathit{variable}_{t;\mathsf{set}|\mathsf{condition}},\ t \in \boldsymbol{t}. \end{aligned}$$

These collections of variables are interpreted as column vectors with an arbitrary fixed ordering of entries. This admits the compact notation

$$\sum_{k\in\boldsymbol{k}} \nu_k = \mathbf{1}'\nu_{\boldsymbol{k}}, \tag{4}$$

$\mathbf{1}'$ is the row vector of ones of the length compatible with $\nu_{\boldsymbol{k}}$, $'$ denotes transposition.

The conjugate pd (2) reproduces during the Bayesian updating, i.e.

$$\mathsf{p}_t(\Omega) = \mathcal{D}_\Omega(V_t),\ V_{t;\Delta|\psi} = V_{t-1;\Delta|\psi} + \delta(\Delta\Delta_t)\delta(\psi\psi_t),\ \Delta\in\boldsymbol{\Delta},\ \psi\in\boldsymbol{\psi},\ t\in\boldsymbol{t}.$$

The sufficient statistic $V_t \equiv (V_{t;\,\Delta|\psi})_{\Delta\,\in\,\boldsymbol{\Delta},\,\psi\,\in\,\boldsymbol{\psi}}$ counts occurrences of the following number of possible configurations corresponding to the length of the parameter Ω

$$\ell_{\boldsymbol{\Omega}} = |\boldsymbol{\Delta}| \prod_{i\in\boldsymbol{i}} |\boldsymbol{\psi}_i|. \tag{5}$$

An informative experiment should fill the majority of V-entries by several observations. This needs limit applicability of the model (1) to very small cardinalities $|\boldsymbol{\psi}_i|$ and memory lengths $\ell_{\boldsymbol{\psi}} = |\boldsymbol{i}|$ determining the length $\ell_{\boldsymbol{\Omega}}$ (5) of the sufficient statistics V_t, $t \in \boldsymbol{t}$. Assuming, for instance,

$$|\boldsymbol{\Delta}| = |\boldsymbol{\psi}_{\mathbf{1}}| = \cdots = |\boldsymbol{\psi}_{\boldsymbol{\ell}_{\boldsymbol{\psi}}}| \equiv \omega \geq 2 \ \Rightarrow\ \ell_{\boldsymbol{\Omega}} = \omega^{\ell_{\boldsymbol{\psi}}+1}, \tag{6}$$

the exponential growth is clearly visible. This manifests the encountered curse of dimensionality and motivates the addressed problem:

A Markov chain model and its estimator are searched for, which escape the curse of dimensionality while losing a little of the model universality.

3. Solution

The solution consists of a simplified modelling [45] and of a newly designed approximate recursive Bayesian estimation presented here in a detail.

3.1. Mixture model

The proposed solution models the stochastic environment by the finite time-invariant *probabilistic mixture*

$$\mathsf{p}(\Delta_t|\Theta,\psi_t) \equiv \sum_{i\in\boldsymbol{i}} \alpha_i \Theta_{\Delta_t|\psi_{t;i}}, \quad \Delta_t\in\boldsymbol{\Delta}, \quad \psi_{t;i}\in\boldsymbol{\psi}_i. \tag{7}$$

It is parameterised by a constant array $\Theta \equiv (\alpha, (\Theta_i)_{i\in \boldsymbol{i}})$, where

$$\alpha \in \boldsymbol{\alpha} \equiv \left\{\alpha_i > 0,\ i \in \boldsymbol{i},\ \mathbf{1}'\alpha \overset{(4)}{=} \sum_{i\in\boldsymbol{i}} \alpha_i = 1\right\}$$

are the probabilistic *component weights*, while the *components* $(\Theta_i)_{i\in\boldsymbol{i}}$ are the transition probabilities, $i \in \boldsymbol{i}$, $\psi_i \in \boldsymbol{\psi}_i$,

$$\Theta_i \equiv \Theta_{\boldsymbol{\Delta}|\boldsymbol{\psi}_i},\ \Theta_{\boldsymbol{\Delta}|\psi_i} \in \boldsymbol{\Theta}_{\boldsymbol{\Delta}|\psi_i} = \left\{\Theta_{\Delta|\psi_i} > 0,\ \Delta \in \boldsymbol{\Delta},\ \mathbf{1}'\Theta_{\boldsymbol{\Delta}|\psi_i} = 1\right\}.$$

The amount $\ell_{\boldsymbol{\Theta}}$ of Θ-entries, decisively influencing the extent of observations needed for their estimation, is

$$\ell_{\boldsymbol{\Theta}} = \ell_{\boldsymbol{\psi}} + |\boldsymbol{\Delta}| \sum_{i=1}^{\ell_{\boldsymbol{\psi}}} |\boldsymbol{\psi}_i|, \quad \text{which compares favourably with} \quad \ell_{\boldsymbol{\Omega}} \overset{(5)}{=} |\boldsymbol{\Delta}| \prod_{i=1}^{\ell_{\boldsymbol{\psi}}} |\boldsymbol{\psi}_i|.$$

The number of mixture parameters $\ell_{\boldsymbol{\Theta}}$ grows with the problem cardinalities and dimensions substantially slower than the number $\ell_{\boldsymbol{\Omega}}$ corresponding to the unrestricted parametrisation of high-order Markov chains (1). It is more obvious when considering the special case reflected in (6)

$$|\boldsymbol{\Delta}| = |\boldsymbol{\psi}_{\mathbf{1}}| = \cdots = |\boldsymbol{\psi}_{\boldsymbol{\ell}_{\boldsymbol{\psi}}}| \equiv \omega \geq 2 \ \text{ with } \ \ell_{\boldsymbol{\Omega}} = \omega^{\ell_{\boldsymbol{\psi}}+1} \ \text{ and } \ \ell_{\boldsymbol{\Theta}} = \ell_{\boldsymbol{\psi}}(1+\omega^2).$$

At the same time, the considered mixture-type parametrisation (7) respects the possible dependency of the observation Δ on the whole regression vector ψ. It also exhibits a high modelling efficiency [9,45,48].

Remarks

- Consideration of more complex combinations of the regression vector entries can make the model arbitrarily rich [44]. Such model has the same structure as (7), but its components are conditioned on more entries taken from the regression vector. This extension is possible up to the level at which the dimensionality curse recurs.
  In this case, the complexity increase can be slow down by parameterising individual components as product of independent factors [18].
- A restriction of component supports [54] is a complementary way of decreasing of the problem dimensionality.
- The conditioning of component weights on individual entries of the regression vector is possible with a mild increase of complexity. This extension describes better the dynamics of the modelled environment.

3.2. Prior PD and approximate posterior PDs

Section 2 motivates the adopted choice of the prior pd, describing unknown parameters as the product of pds conjugated to the individual components and their weights. The product factors are Dirichlet pds, cf. (2).

The Bayes rule multiplies the prior pd by the mixture (7) with inserted observations. It leads to the posterior pd with blowing number of summands and makes an approximation inevitable. Thus, the updated pd has to be projected on the same form as the prior pd and the recursive estimation works with *approximate posterior pds*. The accent $\hat{}$ used below stresses this.

At time $t-1 \in \boldsymbol{t}$, the following description of the unknown parameter, $\Theta \equiv (\alpha, (\Theta_i)_{i\in\boldsymbol{i}})$, is dealt with

$$\begin{aligned}
\hat{\mathrm{p}}_{t-1}(\Theta) &\equiv \hat{\mathrm{p}}_{\kappa_{t-1}V_{t-1}}(\Theta) = \mathcal{D}_\alpha(\kappa_{t-1}) \prod_{i\in\boldsymbol{i}} \prod_{\psi_i\in\boldsymbol{\psi}_i} \mathcal{D}_{\Theta_{\boldsymbol{\Delta}|\psi_i}}(V_{t-1;\boldsymbol{\Delta}|\psi_i}) \qquad (8)\\
\kappa_{t-1} &= \kappa_{t-1;\boldsymbol{i}} = (\kappa_{t-1;i})_{i\in\boldsymbol{i}},\ \kappa_{t-1;i} > 0,\ i \in \boldsymbol{i},\\
V_{t-1} &= \left(V_{t-1;\boldsymbol{\Delta}|\psi_i}\right)_{\psi_i\in\boldsymbol{\psi}_i, i\in\boldsymbol{i}},\ V_{t-1;\Delta|\psi_i} > 0,\ \Delta \in \boldsymbol{\Delta},\ \psi_i \in \boldsymbol{\psi}_i,\ i \in \boldsymbol{i}.
\end{aligned}$$

The prior knowledge is described by the positive optional values κ_0, V_0.

3.3. Bayesian prediction

The parameter estimation serves for constructing probabilistic *predictor* of the observation $\Delta \in \boldsymbol{\Delta}$ for an arbitrary regression vector $\psi \in \boldsymbol{\psi}$. For the approximate posterior pds $\hat{\mathrm{p}}_{\kappa V}(\Theta)$ (8), the predictor has the form

$$\hat{\mathrm{p}}_{\kappa V}(\Delta|\psi) = \int_{\boldsymbol{\Theta}} \mathrm{p}(\Delta|\psi,\Theta)\hat{\mathrm{p}}_{\kappa V}(\Theta)\,\mathrm{d}\Theta = \sum_{i\in\boldsymbol{i}} \overbrace{\frac{\kappa_i}{\mathbf{1}'\kappa_{\boldsymbol{i}}}}^{\hat{\alpha}_i} \overbrace{\frac{V_{\Delta|\psi_i}}{\mathbf{1}'V_{\boldsymbol{\Delta}|\psi_i}}}^{\hat{\Theta}_{\Delta|\psi_i}}. \qquad (9)$$

Formula (9) can be verified by using (3) and equality $\Gamma(x+1) = x\Gamma(x)$. It fits the interpretation of an event probability as the number of its realisations normalised by the number of all possible events.

3.4. Bayesian updating

The Bayesian updating of the pd (8) provides the posterior pd, which is a *mixture of Dirichlet pds*[4]

$$\begin{aligned}\tilde{\mathrm{p}}_t(\Theta) &\propto \sum_{j\in\boldsymbol{i}} \frac{\alpha_j \Theta_{\Delta_t|\psi_{t;j}}}{\mathrm{B}(\kappa_{t-1})} \prod_{i\in\boldsymbol{i}} \alpha_i^{\kappa_{t-1;i}-1} \prod_{\psi_i\in\boldsymbol{\psi}_i} \frac{\prod_{\Delta\in\boldsymbol{\Delta}} \Theta_{\Delta|\psi_i}^{V_{t-1;\Delta|\psi_i}-1}}{\mathrm{B}(V_{t-1;\boldsymbol{\Delta}|\psi_i})}\\ &= \sum_{j\in\boldsymbol{i}} \frac{\mathrm{B}(\tilde{\kappa}_{t;\boldsymbol{i}|j})}{\mathrm{B}(\kappa_{t-1})} \frac{\prod_{i\in\boldsymbol{i}} \alpha_i^{\overbrace{\kappa_{t-1;i}+\delta(ij)}^{\tilde{\kappa}_{t;i|j}}-1}}{\mathrm{B}(\tilde{\kappa}_{t;\boldsymbol{i}|j})}\\ &\quad\times \prod_{l\in\boldsymbol{i}} \prod_{\psi_l\in\boldsymbol{\psi}_l} \frac{\mathrm{B}(\tilde{V}_{t;\boldsymbol{\Delta}|\psi_l j})}{\mathrm{B}(V_{t-1;\boldsymbol{\Delta}|\psi_l})} \frac{\prod_{\Delta\in\boldsymbol{\Delta}} \Theta_{\Delta|\psi_l}^{\overbrace{V_{t-1;\Delta|\psi_l}+\delta(lj)\delta(\Delta\Delta_t)\delta(\psi_l\psi_{t;l})}^{\tilde{V}_{t;\Delta|\psi_l j}}-1}}{\mathrm{B}(\tilde{V}_{t;\boldsymbol{\Delta}|\psi_l j})}\\ &= \sum_{j\in\boldsymbol{i}} \gamma_{t;j} \mathcal{D}_\alpha(\tilde{\kappa}_{t;\boldsymbol{i}|j}) \prod_{i\in\boldsymbol{i}} \prod_{\psi_i\in\boldsymbol{\psi}_i} \mathcal{D}_{\Theta_{\boldsymbol{\Delta}|\psi_i}}(\tilde{V}_{t;\boldsymbol{\Delta}|\psi_i j}) \end{aligned} \tag{10}$$

$$\begin{aligned}&\gamma_{t;j} \equiv \frac{\hat{\alpha}_{t-1;j}\hat{\Theta}_{t-1;\Delta_t|\psi_{t;j}}}{\sum_{i\in\boldsymbol{i}} \hat{\alpha}_{t-1;i}\hat{\Theta}_{t-1;\Delta_t|\psi_{t;i}}}, \quad \hat{\alpha}_{t-1;j} \equiv \frac{\kappa_{t-1;j}}{\mathbf{1}'\kappa_{t-1;\boldsymbol{i}}}, \quad \hat{\Theta}_{t-1;\Delta|\psi_j} \equiv \frac{V_{t-1;\Delta|\psi_j}}{\mathbf{1}'V_{t-1;\boldsymbol{\Delta}|\psi_j}}\\ &\tilde{\kappa}_{t;i|j} = \kappa_{t-1;i} + \delta(ij), \quad \tilde{V}_{t;\Delta|\psi_i j} = V_{t-1;\Delta|\psi_i} + \delta(ij)\delta(\Delta\Delta_t)\delta(\psi_i\psi_{t;i})\\ &\quad i,j\in\boldsymbol{i},\ \Delta\in\boldsymbol{\Delta},\ \psi_i\in\boldsymbol{\psi}_i. \end{aligned} \tag{11}$$

In (11), the identity $\Gamma(x+1) = x\Gamma(x)$ is again used.

Note that (9) interprets the normalisation factor of $\gamma_{t;j}$ in (11)

$$\tilde{\Theta}_{t-1;\Delta_t|\psi_t} \equiv \sum_{i\in\boldsymbol{i}} \hat{\alpha}_{t-1;i}\hat{\Theta}_{t-1;\Delta_t|\psi_{t;i}} = \hat{\mathrm{p}}_{\kappa_{t-1}V_{t-1}}(\Delta_t|\psi_t), \quad \hat{\alpha}_t = \hat{\alpha}_{\kappa_t},\ \hat{\Theta}_t = \hat{\Theta}_{V_t}. \tag{12}$$

3.5. Projection of the posterior PD on the product form

The mixture of Dirichlet pds $\tilde{\mathrm{p}}_t(\Theta)$ (10) is to be projected on the product form $\hat{\mathrm{p}}_t(\Theta) \equiv \hat{\mathrm{p}}_{\kappa_t V_t}(\Theta)$ (8). It is done by minimising the *Kullback–Leibler divergence* $\mathrm{D}(\tilde{\mathrm{p}}_t||\hat{\mathrm{p}}_{\kappa V})$ (KLD, [39]) of $\tilde{\mathrm{p}}_t(\Theta)$ to the optimised pd $\hat{\mathrm{p}}_{\kappa V}(\Theta)$

$$\mathrm{D}\left(\tilde{\mathrm{p}}||\hat{\mathrm{p}}_{\kappa V}\right) \equiv \int_{\boldsymbol{\Theta}} \tilde{\mathrm{p}}_t(\Theta) \ln\left(\frac{\tilde{\mathrm{p}}_t(\Theta)}{\hat{\mathrm{p}}(\Theta)_{\kappa V}}\right) \mathrm{d}\Theta.$$

The results [11,29] axiomatically justify this projection way, which was applied to mixtures of dynamic Gaussian components in [3]. The current paper deals with Markov-chain components and adds forgetting as a general counter-measure against the accumulation of errors caused by the repetitive approximation of the posterior pd, see [25] and Section 4.

The minimisation of this KLD version reduces to the minimisation of the Kerridge inaccuracy [34], cf. (8). It leads to the choice

$$\begin{aligned}[\kappa_t \vdots V_t] &= \arg\min_{\kappa>0,V>0} -\int_{\boldsymbol{\Theta}} \tilde{\mathrm{p}}_t(\Theta) \ln\left(\hat{\mathrm{p}}_{\kappa V}(\Theta)\right) \mathrm{d}\Theta\\ &= \arg\min_{\kappa>0,V>0} -\int_{\boldsymbol{\Theta}} \tilde{\mathrm{p}}_t(\Theta) \left[\ln\left(\mathcal{D}_\alpha(\kappa)\right) + \sum_{i\in\boldsymbol{i}} \sum_{\psi_i\in\boldsymbol{\psi}_i} \ln\left(\mathcal{D}_{\Theta_{\boldsymbol{\Delta}|\psi_i}}(V_{\boldsymbol{\Delta}|\psi_i})\right)\right] \mathrm{d}\Theta\\ &= \left[\arg\min_{\kappa>0} \left\{\ln\left(\mathrm{B}(\kappa)\right) - \sum_{i\in\boldsymbol{i}} (\kappa_i - 1) \sum_{j\in\boldsymbol{i}} \gamma_{t;j} \int_{\boldsymbol{\alpha}} \mathcal{D}_\alpha(\tilde{\kappa}_{t;\boldsymbol{i}|j}) \ln(\alpha_i)\, \mathrm{d}\alpha\right\} \vdots\right.\\ &\quad \left.\arg\min_{V>0} \sum_{i\in\boldsymbol{i}} \sum_{\psi_i\in\boldsymbol{\psi}_i} \left\{\ln\left(\mathrm{B}(V_{\boldsymbol{\Delta}|\psi_i})\right)\right.\right. \end{aligned} \tag{13}$$

[4] The accent $\tilde{}$ marks *intermediate quantities*. $\propto$ is equality up to Θ-independent factor. Its use allowed us to insert normalising beta functions so that Dirichlet pds in this posterior pd are readily seen.

$$
\begin{aligned}
&- \sum_{\Delta\in\boldsymbol{\Delta}} (V_{\Delta|\psi_i} - 1) \sum_{j\in\boldsymbol{i}} \gamma_{t;j} \int_{\boldsymbol{\Theta}_{\boldsymbol{\Delta}|\psi_i}} \mathcal{D}_{\Theta_{\boldsymbol{\Delta}|\psi_i}}(\tilde{V}_{t;\boldsymbol{\Delta}|\psi_i j}) \ln\left(\Theta_{\Delta|\psi_i}\right) \mathrm{d}\Theta_{\boldsymbol{\Delta}|\psi_i} \Big\} \Bigg] \\
= &\left[\arg\min_{\kappa>0} \left\{ \ln\left(\mathsf{B}(\kappa)\right) - \sum_{i\in\boldsymbol{i}} (\kappa_i - 1) \sum_{j\in\boldsymbol{i}} \gamma_{t;j} \left[\Psi(\tilde{\kappa}_{t;\boldsymbol{i}|j}) - \Psi(\mathbf{1}'\tilde{\kappa}_{t;\boldsymbol{i}|j}) \right] \right\} \right. \vdots \\
&\left. \arg\min_{V>0} \sum_{i\in\boldsymbol{i}} \sum_{\psi_i\in\boldsymbol{\psi}_i} \left\{ \ln\left(\mathsf{B}(V_{\boldsymbol{\Delta}|\psi_i})\right) - \sum_{\Delta\in\boldsymbol{\Delta}} (V_{\Delta|\psi_i} - 1) \sum_{j\in\boldsymbol{i}} \gamma_{t;j} \left[\Psi\left(\tilde{V}_{t;\Delta|\psi_i j}\right) - \Psi\left(\mathbf{1}'\tilde{V}_{t;\boldsymbol{\Delta}|\psi_i j}\right) \right] \right\} \right].
\end{aligned}
$$

The evaluation of (13) exploits the following type of the definite integral [26]

$$
\int_{\boldsymbol{\alpha}} \mathcal{D}_{\alpha}(\kappa) \ln\left(\alpha_i\right) \mathrm{d}\alpha = \frac{\partial \ln\left(\mathsf{B}(\kappa)\right)}{\partial \kappa_i} = \Psi(\kappa_i) - \Psi\left(\mathbf{1}'\kappa_{\boldsymbol{i}}\right), \quad \Psi(x) \equiv \frac{\partial \ln\left(\Gamma(x)\right)}{\partial x}. \tag{14}
$$

Zeroing of derivatives of the Kerridge inaccuracy with respect to κ_k and $V_{\Delta|\psi_k}$, $k \in \boldsymbol{i}$, $\Delta \in \boldsymbol{\Delta}$, $\psi_k \in \boldsymbol{\psi}_k$, provides the necessary conditions for extreme

$$
\begin{aligned}
0 &= \Psi(\kappa_{t;k}) - \Psi\left(\mathbf{1}'\kappa_{t;\boldsymbol{i}}\right) - \sum_{j\in\boldsymbol{i}} \gamma_{t;j} \left[\Psi(\tilde{\kappa}_{t;k|j}) - \Psi\left(\mathbf{1}'\tilde{\kappa}_{t;\boldsymbol{i}|j}\right) \right] \\
0 &= \Psi(V_{t;\Delta|\psi_k}) - \Psi\left(\mathbf{1}'V_{t;\boldsymbol{\Delta}|\psi_k}\right) - \sum_{j\in\boldsymbol{i}} \gamma_{t;j} \left[\Psi(\tilde{V}_{t;\Delta|\psi_k j}) - \Psi\left(\mathbf{1}'\tilde{V}_{t;\boldsymbol{\Delta}|\psi_k j}\right) \right].
\end{aligned}
$$

The definition of statistics $\tilde{\kappa}$, $\tilde{V}$ (11), the recursive formula

$$
\ln\left(\Gamma(x+1)\right) = \ln\left(x\Gamma(x)\right) \;\Rightarrow\; \Psi(x+1) = \Psi(x) + \frac{1}{x},
$$

the identity $\mathbf{1}'\gamma_{t;\boldsymbol{i}} = 1$ and the definition $\bar{\Theta}_{t-1;\boldsymbol{\Delta}_t|\psi_t} \equiv \sum_{i\in\boldsymbol{i}} \hat{\alpha}_{t-1;i} \hat{\Theta}_{t-1;\Delta_t|\psi_{t;i}}$ (12) give the final form of the necessary conditions determining the statistics of the approximate posterior pd $\kappa_{t;k}$, $V_{t;\Delta|\psi_k}$ for $k \in \boldsymbol{i}$, $\Delta \in \boldsymbol{\Delta}$, $\psi_k \in \boldsymbol{\psi}_k$

$$
\begin{aligned}
&\Psi(\kappa_{t;k}) - \Psi\left(\mathbf{1}'\kappa_{t;\boldsymbol{i}}\right) = \Psi(\kappa_{t-1;k}) - \Psi\left(\mathbf{1}'\kappa_{t-1;\boldsymbol{i}}\right) + \frac{\hat{\Theta}_{t-1;\Delta_t|\psi_{t;k}} - \bar{\Theta}_{t-1;\boldsymbol{\Delta}_t|\psi_t}}{\bar{\Theta}_{t-1;\boldsymbol{\Delta}_t|\psi_t} \mathbf{1}'\kappa_{t-1;\boldsymbol{i}}} \equiv R_k \\
&\Psi(V_{t;\Delta|\psi_k}) - \Psi\left(\mathbf{1}'V_{t;\boldsymbol{\Delta}|\psi_k}\right) = \Psi(V_{t-1;\Delta|\psi_k}) - \Psi\left(\mathbf{1}'V_{t-1;\boldsymbol{\Delta}|\psi_k}\right) \\
&\quad + \delta(\psi_k\psi_{t;k}) \hat{\alpha}_{t-1;k} \frac{\delta(\Delta\Delta_t) - \hat{\Theta}_{t-1;\Delta_t|\psi_{t;k}}}{\bar{\Theta}_{t-1;\boldsymbol{\Delta}_t|\psi_t} \mathbf{1}'V_{t-1;\boldsymbol{\Delta}|\psi_{t;k}}} \equiv R_{\Delta|\psi_k}.
\end{aligned} \tag{15}
$$

3.6. Numerical search for the best approximation

The discussed necessary conditions (15) are also sufficient as the non-linear part of the minimised function, $\ln\left(\mathsf{B}(\nu)\right)$, $\nu \in \{\kappa, V\}$, is convex function [2]. Thus, a simple numerical algorithm using the first-order Taylor expansion suffices for solving (15). For the positive vectors $\nu \in \left\{\kappa, \{V_{\boldsymbol{\Delta}|\psi_k}\}_{\psi_k\in\boldsymbol{\psi}_k, k\in\boldsymbol{k}}\right\}$, with the corresponding $\ell_{\boldsymbol{\nu}} \in \{\ell_{\boldsymbol{\psi}}, \{|\boldsymbol{\Delta}| \times |\boldsymbol{\psi}_k|\}_{k\in\boldsymbol{k}}\}$, the solved equations become

$$
\Psi(\nu_k) - \Psi\left(\mathbf{1}'\nu_{\boldsymbol{k}}\right) = \rho_k, \quad k \in \boldsymbol{k} \equiv \{1, \ldots, |\boldsymbol{k}|\}. \tag{16}
$$

There, the right-hand sides $\rho_k = R_k$, or $\rho_k = R_{\boldsymbol{\Delta}|\psi_k}$ of (15) occur.

Denoting $\zeta(x) \equiv \frac{\mathrm{d}\Psi(x)}{\mathrm{d}x}$ and ν_τ the approximate solution obtained at τth discrete time of the iterative solution $\tau \in \boldsymbol{\tau} \equiv \boldsymbol{t}$, the linear Taylor expansion at $\nu_{\tau-1}$ maps (16) on the system of $|\boldsymbol{k}|$ linear equations for $\nu_{\tau;\boldsymbol{k}}$

$$
\begin{aligned}
&\overbrace{\left\{ \mathrm{diag}\Big[\underbrace{\zeta\left(\nu_{\tau-1;1}\right), \ldots, \zeta\left(\nu_{\tau-1;|\boldsymbol{k}|}\right)}_{[1/g_1,\ldots,1/g_{|\boldsymbol{k}|}]} \Big] - \zeta\left(\mathbf{1}'\nu_{\tau;\boldsymbol{k}}\right) \mathbf{1}\mathbf{1}' \right\}}^{\mathsf{A}} \left(\nu_{\tau;\boldsymbol{k}} - \nu_{\tau-1;\boldsymbol{k}}\right) \\
&\quad = \rho_{\boldsymbol{k}} - \Psi(\nu_{\tau-1;\boldsymbol{k}}) + \Psi\left(\mathbf{1}'\nu_{\tau-1;\boldsymbol{k}}\right).
\end{aligned} \tag{17}
$$

It can be directly verified that inversion A^{-1} of the matrix A reads

$$
\mathsf{A}^{-1} = \mathsf{D} + \frac{\zeta\left(\mathbf{1}'\nu_{\tau-1;\boldsymbol{k}}\right) g_{\boldsymbol{k}} g_{\boldsymbol{k}}'}{1 - \zeta\left(\mathbf{1}'\nu_{\tau-1;\boldsymbol{k}}\right) \mathbf{1}' g_{\boldsymbol{k}}}, \quad \mathsf{D} = \mathrm{diag}[g_{\boldsymbol{k}}].
$$

This makes the iterations (17) computationally cheap. No problems with the convergence from the "natural" starting point of iterations $\nu_{\tau=0;\boldsymbol{k}} = \nu_{t-1;\boldsymbol{k}}$ are expected. The positivity of intermediate ν_τ is achieved by clipping its values at a small positive value. The sufficiency of this simple measure follows from a relatively steepest shape of $\Psi(\nu_{\boldsymbol{k}}) - \Psi(\mathbf{1}'\nu_{\boldsymbol{k}})$ for $\nu_k > 0$ and $\nu_k \approx 0$.

4. Counteracting accumulation of errors

The proposed estimation intertwines data updating and projection to the class of approximate posterior pds (8). Commonly, the resulting approximate posterior pd serves as a prior pd for the next estimation step. Under such treatment, the approximation errors propagate and may cause divergence: the approximate posterior pd gradually deviates too much from the exact pd, which would be obtained without the projection. This is a common danger of recursive estimation. The papers [36,37] completely characterise a narrow class of approximate recursive estimators, which do not suffer this problem. The proposed estimator is out of this class and needs an active counteracting of the error accumulation.

Recently, a systematic counteracting has been proposed [25]. It coincides with the *stabilised forgetting* [26,38] determined by a data-dependent forgetting factor. The projection $\hat{\mathsf{p}}_t = \hat{\mathsf{p}}_{\kappa_t V_t}$, described in Section 3.5, is only taken as an *intermediate approximate pd* $\tilde{\hat{\mathsf{p}}}_t \equiv \hat{\mathsf{p}}_{\kappa_t V_t}$ and it is forgotten before it is used in the subsequent data updating. The forgotten pd $\hat{\mathsf{p}}_{\lambda_t}$ is "transferred" to further updating step as its prior pd

$$\hat{\mathsf{p}}_t \equiv \hat{\mathsf{p}}_{\lambda_t} \propto \tilde{\hat{\mathsf{p}}}_t^{\lambda_t} \hat{\mathsf{p}}_{t-1}^{1-\lambda_t}.$$

The existence of $\lambda_t \in [0, 1]$ indeed counteracting the errors was proved in [25] and λ_t solving

$$0 = \int_{\Theta} (\tilde{\mathsf{p}}_t(\Theta) - \hat{\mathsf{p}}_{\lambda_t}(\Theta)) \ln\left(\frac{\tilde{\hat{\mathsf{p}}}_t(\Theta)}{\hat{\mathsf{p}}_{t-1}(\Theta)}\right) \mathrm{d}\Theta, \tag{18}$$

where $\tilde{\mathsf{p}}_t(\Theta)$ is the outcome of the Bayesian update (10), is the recommended choice of the forgetting factor. Below, an alternative and computationally simpler choice of λ_t is justified and used.

The pd $\tilde{\hat{\mathsf{p}}}_t$ is the best projection of the *approximate* posterior pd $\tilde{\mathsf{p}}_t$, which differs from the *optimal projection* $\hat{\mathsf{p}}$ of the *exact* (fully reflecting unreduced information) posterior pd p_t. The intermediate approximate pd $\tilde{\hat{\mathsf{p}}}_t$ is expected to be a better approximation of the best projection $\hat{\mathsf{p}}$ than its prior guess $\hat{\mathsf{p}}_{t-1}$, i.e. $\mathsf{D}(\hat{\mathsf{p}}||\tilde{\hat{\mathsf{p}}}_t) \le \mathsf{D}(\hat{\mathsf{p}}||\hat{\mathsf{p}}_{t-1})$. The minimum KLD principle recommends to select $\hat{\mathsf{p}}_t$ as the optimal projection of the exact pd p_t (given by the available knowledge) as follows

$$\hat{\mathsf{p}}_t = \arg\min_{\hat{\mathsf{p}}\in\hat{\mathbf{p}}} \mathsf{D}(\hat{\mathsf{p}}||\underline{\mathsf{p}}), \tag{19}$$

$$\hat{\mathbf{p}} = \{\hat{\mathsf{p}}(\Theta) = \hat{\mathsf{p}}_{\kappa V}(\Theta) \text{ of form } (8), \kappa, V > 0, \text{ such } \mathsf{D}(\hat{\mathsf{p}}||\tilde{\hat{\mathsf{p}}}_t) \le \mathsf{D}(\hat{\mathsf{p}}||\hat{\mathsf{p}}_{t-1})\},$$

where the pd $\underline{\mathsf{p}}$ is a prior guess of the optimal projection $\hat{\mathsf{p}}$ selected without considering the condition on 2nd line of (19). This prior guess of $\hat{\mathsf{p}}$ is here chosen as an intuitively plausible, weighted geometric mean of $\hat{\mathsf{p}}_{t-1}$ and $\tilde{\hat{\mathsf{p}}}_t$

$$\underline{\mathsf{p}} \propto \tilde{\hat{\mathsf{p}}}_t^{\underline{\lambda}} \hat{\mathsf{p}}_{t-1}^{1-\underline{\lambda}}, \quad \underline{\lambda} \in [0, 1], \tag{20}$$

where $\underline{\lambda}$ should reflect quality of the combined factors $\tilde{\hat{\mathsf{p}}}_t, \hat{\mathsf{p}}_{t-1}$, typically, quality of predictors constructed from them. Without sufficient reasons $\underline{\lambda} = 0.5$ is to be chosen. Note that the geometric mean (20) has the form (8).

Proposition 1 (Optimal forgetting)**.** *The task (19) with the choice (20) has a unique solution of the form (8) given by*

$$\kappa_{\lambda_t} = \lambda_t \kappa_t + (1 - \lambda_t)\kappa_{t-1}, \quad V_{\lambda_t} = \lambda_t V_t + (1 - \lambda_t) V_{t-1} \tag{21}$$

where $\lambda_t \in [\,\underline{\lambda}, 1]$ is either the unique solution of the following equation

$$\mathsf{e}(\lambda) = 0, \ \mathsf{e}(\lambda) \equiv \int_{\Theta} \hat{\mathsf{p}}_{\kappa_\lambda V_\lambda}(\Theta) \ln\left(\frac{\hat{\mathsf{p}}_{t-1}(\Theta)}{\tilde{\hat{\mathsf{p}}}_t(\Theta)}\right) \mathrm{d}\Theta \tag{22}$$

or has the value $\underline{\lambda}$. The function $\mathsf{e}(\lambda)$ is decreasing on the interval $[0, 1]$ and changes its sign there.

Proof. The minimised function (19) is convex in the optimised $\kappa, V > 0$ (open set) [2]. The corresponding Kuhn–Tucker function with a non-negative multiplier μ can be rewritten for $\lambda = \underline{\lambda} + \mu$ into the form

$$\mathsf{D}(\hat{\mathsf{p}}||\underline{\mathsf{p}}) + \mu(\mathsf{D}(\hat{\mathsf{p}}||\tilde{\hat{\mathsf{p}}}_t) - \mathsf{D}(\hat{\mathsf{p}}||\hat{\mathsf{p}}_{t-1})) = \mathsf{D}(\hat{\mathsf{p}}||\hat{\mathsf{p}}_{\kappa_\lambda V_\lambda}) - \ln(\mathsf{c}_\lambda) + \ln(\mathsf{c}_{\underline{\lambda}})$$

$$\mathsf{c}_\lambda \equiv \int_{\Theta} \tilde{\hat{\mathsf{p}}}_t^{\lambda}(\Theta) \hat{\mathsf{p}}_t^{1-\lambda}(\Theta)\, \mathrm{d}\Theta.$$

The normalisation factor c_λ is finite for $0 \le \lambda \le 1$ and the Kuhn–Tucker function is minimised by $\hat{\mathsf{p}} = \hat{\mathsf{p}}_{\kappa_\lambda V_\lambda}$, (8) and (21). It remains to inspect how to determine $\lambda_t \in [\,\underline{\lambda}, 1]$ meeting the constraint in (19), i.e. to make the next λ-dependent difference zero or negative

$$\mathsf{e}(\lambda) \equiv \mathsf{D}(\hat{\mathsf{p}}_{\kappa_\lambda V_\lambda}||\tilde{\hat{\mathsf{p}}}_t) - \mathsf{D}(\hat{\mathsf{p}}_{\kappa_\lambda V_\lambda}||\hat{\mathsf{p}}_{t-1}) = \begin{cases} -\mathsf{D}(\tilde{\hat{\mathsf{p}}}_t||\hat{\mathsf{p}}_{t-1}) < 0 & \text{for } \lambda = 1 \\ \mathsf{D}(\hat{\mathsf{p}}_{t-1}||\tilde{\hat{\mathsf{p}}}_t) > 0 & \text{for } \lambda \to 0 \end{cases}.$$

A direct evaluation reveals that

$$\frac{\mathrm{d}\mathsf{e}(\lambda)}{\mathrm{d}\lambda} = -\int_{\Theta} \hat{\mathsf{p}}_{\kappa_\lambda V_\lambda} \ln^2\left(\frac{\tilde{\hat{\mathsf{p}}}_t(\Theta)}{\hat{\mathsf{p}}_{t-1}(\Theta)}\right) \mathrm{d}\Theta + \left[\int_{\Theta} \hat{\mathsf{p}}_{\kappa_\lambda V_\lambda}(\Theta) \ln\left(\frac{\tilde{\hat{\mathsf{p}}}_t(\Theta)}{\hat{\mathsf{p}}_{t-1}(\Theta)}\right) \mathrm{d}\Theta\right]^2 < 0,$$

where the negativity follows from the Jensen inequality. Thus, the decreasing continuous function $\mathrm{e}(\lambda)$ crosses zero for some $\lambda_t \in (0, 1]$. If $\lambda_t \in [\underline{\lambda}, 1]$ then it determines the minimiser of the Kuhn–Tucker function meeting the given constraint. If the solution is smaller than $\underline{\lambda}$, then the constraint is not active, i.e. $\mu = 0 \Leftrightarrow \lambda_t = \underline{\lambda}$. In this case $\mathrm{e}(\underline{\lambda}) < 0$, and the non-active constraint is respected. □

Remarks.

- The elementary proof has avoided checking standardly used properties of the function in the constraint (like quasi-convexity). Without a priori restricting the form of the optimised $\hat{\mathrm{p}}$, it is easy to see that the convex functional under convex constraints is minimised and the use of the Kuhn–Tucker functional is the correct one. Then, it remains to "discover" that the minimiser has the required form (8).
- The resulting "forgotten" pd $\hat{\mathrm{p}}_{\kappa_{\lambda_t} V_{\lambda_t}}$ coincides with that proposed in [25] but the equation for λ_t (22) differs substantially from the original heuristic proposal (18) both in its form and non-exploitation of the pd $\tilde{\mathrm{p}}_t$ (10) resulting from the Bayesian update (before the projection).
- Monotonicity of $\mathrm{e}(\lambda)$ on the bounded interval makes a trivial search of its root sufficient while keeping computational overheads low.
- The used formulation generalises by specifying the set $\hat{\mathbf{p}}$ in (19) by several inequalities between KLDs of independent factors of the pd (8). This version is strongly related to *partial* (vector) *forgetting* [16]. The available experience indicates that this generalisation, which is formally superior to the described one, is substantially better practically.

The use of the form (8) of $\hat{\mathrm{p}}_{\kappa_{\lambda_t} V_{\lambda_t}}$, $\hat{\mathrm{p}}_{t-1}$, $\tilde{\tilde{\mathrm{p}}}_t$, of the formula (14) and the fact that $\mathrm{V}_{t;\Delta|\psi_i} = \mathrm{V}_{t-1;\Delta|\psi_i}$ for $\psi_i \neq \psi_{t;i}$, $i \in \boldsymbol{i}$, leads to the following specific form of the function $\mathrm{e}(\lambda)$ whose zeroing argument λ_t is searched for.

$$\begin{aligned}\mathrm{e}(\lambda) = &\ln\left(\frac{\mathrm{B}(\kappa_t)}{\mathrm{B}(\kappa_{t-1})}\right)\\ &+\sum_{i\in\boldsymbol{i}}(\kappa_{t-1;i}-\kappa_{t;i})\left(\Psi(\kappa_{i\lambda})-\Psi\left(\mathbf{1}'\kappa_{\boldsymbol{i}\lambda}\right)\right)+\sum_{i\in\boldsymbol{i}}\ln\left(\frac{\mathrm{B}(V_{t;\boldsymbol{\Delta}|\psi_{t;i}})}{\mathrm{B}(V_{t-1;\boldsymbol{\Delta}|\psi_{t;i}})}\right)\\ &+\sum_{\Delta\in\boldsymbol{\Delta},i\in\boldsymbol{i}}(V_{t-1;\Delta|\psi_{t;i}}-V_{t;\Delta|\psi_{t;i}})\left(\Psi(V_{\Delta|\psi_{t;i}\lambda})-\Psi(\mathbf{1}'V_{\boldsymbol{\Delta}|\psi_{t;i}\lambda})\right)\end{aligned}$$

with $\kappa_{1\lambda}$, $V_{\Delta|\psi\lambda}$ defined in (21).

The used search is based on linear Taylor expansion and needs

$$\begin{aligned}\frac{\mathrm{de}(\lambda)}{\mathrm{d}\lambda} = &-\sum_{i\in\boldsymbol{i}}(\kappa_{t-1;i}-\kappa_{t;i})^2\zeta(\kappa_{i\lambda})+\sum_{i\in\boldsymbol{i}}(\mathbf{1}'\kappa_{t-1;i}-\mathbf{1}'\kappa_{t;i})^2\zeta\left(\mathbf{1}'\kappa_{\boldsymbol{i}\lambda}\right)\\ &-\sum_{\Delta\in\boldsymbol{\Delta},i\in\boldsymbol{i}}(V_{t-1;\Delta|\psi_{t;i}}-V_{t;\Delta|\psi_{t;i}})^2\zeta(V_{\Delta|\psi_{t;i}\lambda})\\ &+\sum_{i\in\boldsymbol{i}}(\mathbf{1}'V_{t-1;\boldsymbol{\Delta}|\psi_{t;i}}-\mathbf{1}'V_{t;\boldsymbol{\Delta}|\psi_{t;i}})^2\zeta\left(\mathbf{1}'V_{\boldsymbol{\Delta}|\psi_{t;i}\lambda}\right),\quad \text{where } \zeta(x)\equiv\frac{\mathrm{d}\Psi(x)}{\mathrm{d}x}.\end{aligned}$$

The intermediate guesses λ_τ, $\tau \in \boldsymbol{\tau}$, of λ_t then evolve

$$\lambda_\tau = \max\left(\underline{\lambda}, \min\left(1, \lambda_{t-1} - \frac{\mathrm{e}(\lambda_{\tau-1})}{\frac{\mathrm{de}(\lambda_{\tau-1})}{\mathrm{d}\lambda}}\right)\right),\quad \lambda_{\tau=0} = 1.$$

5. Examples

This section provides simple examples illustrating the use and properties of the proposed algorithm. The examples also illustrate computational complexity of the proposed estimator. It is fully determined by the number of estimated parameters. Fig. 1 exemplifies that the generic lack of data needed for learning primarily limits a use of fully parameterised high-order Markov chains. The adopted parsimonious parametrisation (7) with $\ell_{\boldsymbol{\Theta}}$ unknowns helps us to overcome this limitation. At the same time, $\ell_{\boldsymbol{\Theta}}$ coincides with cardinality of the collected sub-sufficient statistic $\kappa_{t;\boldsymbol{i}}$, $V_{t;\boldsymbol{\Delta}|\boldsymbol{i}}$ describing the used Bayesian estimate (8). The optimal updating of this statistic requires a solution of non-linear equations expressing the necessary and sufficient conditions on optimal projection (15). These equations have nice analytical properties so that even elementary iterative solution, requiring no matrix inversion, suffices. The number of operations needed per update step is proportional to ${\ell_{\boldsymbol{\Theta}}}^2$. The numerical evaluations of data-dependent forgetting factor (22) also deals with simple necessary and sufficient conditions and their solutions are less demanding even when the most complex – entry-wise applied – partial forgetting is employed.

The interested reader can also consult the report [20], http://library.utia.cas.cz/separaty/2015/AS/hrabak-0440385.pdf, which documents a complex test dealing with prediction of pedestrian decisions during the egress situation and contains scripts of the resulting estimator.

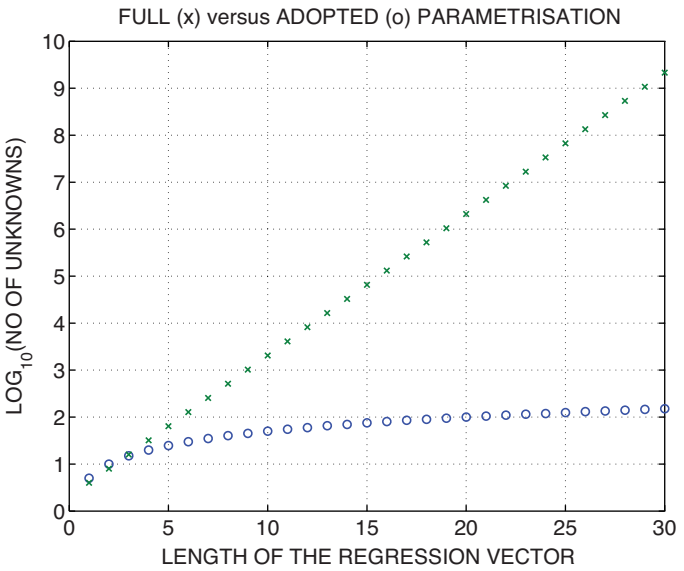

Fig. 1. The decadic logarithms of the number of unknown parameters for full parametrisation (1) and the adopted one (7) as the function of the regression-vector length $\ell_{\boldsymbol{\psi}}$. The curves correspond to the least demanding binary case, $\omega = 2$, see (6).

5.1. Illustration of data exploration

This example illustrates: (i) the general notions; (ii) the usefulness of the proposed estimator even in seemingly simple situations.

The analysed observations should reveal influence of non-economical factors on outcomes of Ultimatum Game [46]. It is an extensively studied scenario used in behavioural economics [7]. Here, its simplest version is used.

Computer acting as the proposer offers to the human responder the monetary offer $\in$ **offer** $\equiv \{1, \ldots, 9\} = \boldsymbol{\psi}_1$, $|\boldsymbol{\psi}_1| = 9$. The responder selects her decision $\in$ **decision** $\equiv$ {reject,accept} $= \boldsymbol{\Delta}$, $|\boldsymbol{\Delta}| = 2$. The choice "reject" means that both players get no reward, otherwise the overall amount 9 is split as proposed. The game repeats. A purely rational responder maximising her reward should accept any offer but almost no human does this.

The behavioural study [35] inspected the influence of the offer fairness on the observed decision Δ_t. The collected data provided the possibility to inspect whether the offer value, responder's age $\in$ **age** $\equiv$ {young,middle,senior} $= \boldsymbol{\psi}_2$, $|\boldsymbol{\psi}_2| = 3$, education $\in$ **education** $\equiv$ {basic,high,university} $= \boldsymbol{\psi}_3$, $|\boldsymbol{\psi}_3| = 3$, and mood $\in$ **mood** $= \{1, \ldots, 5\} \equiv$ {the worst, $\ldots$, the best} $= \boldsymbol{\psi}_4$, $|\boldsymbol{\psi}_4| = 5$, influence the decision.

The inspection falls into the supported problem class. Indeed, when considering the static regression vector $\psi_t = (\text{offer}_t, \text{age}_t, \text{education}_t, \text{mood}_t)$ at tth game ($\ell_{\boldsymbol{\psi}} = 4$) the number of unknown parameters is $\ell_{\boldsymbol{\Omega}} = 810$, while the mixture model (7) has $\ell_{\boldsymbol{\Theta}} = 42$. The former number is not excessive in absolute sense but the available number of observations $T = 100$, corresponding to 10 responders each playing 10 game rounds, makes the latter, less abundant, parametrisation inevitable.

The recorded marginal occurrences were

$$V_{T;\textbf{decision}} = [49, 51],\ V_{T;\textbf{offer}} = [14, 16, 10, 15, 10, 9, 10, 5, 11],$$
$$V_{T;\textbf{age}} = [20, 60, 20],\ V_{T;\textbf{education}} = [20, 0, 80] \text{ and } V_{T;\textbf{mood}} = [20, 0, 60, 20, 0].$$

The prior statistics κ_0, V_0 leading to the equal point parameter estimates with the smallest variance were chosen ($\kappa_{0;i} \approx {\ell_{\boldsymbol{\psi}}}^{-0.5}$, $V_{0;\Delta|\psi_i} \approx |\boldsymbol{\Delta}|^{-0.5}$). For the lower bound $\underline{\lambda} = 0.75$ on the forgetting factor (19) and the partial forgetting, the obtained component weights' estimate (11) for $t - 1 = T$ was

$$\hat{\alpha}_{T;\{\text{offer,age,education,mood}\}} = [0.86,\ 0.05,\ 0.04,\ 0.05].$$

It indicates that only the offer value influences the decision. Bayesian hypotheses testing confirmed it. The point estimate (11) for $t - 1 = T$ of the probabilities of the "reject" for increasing offers within the first significant component was

$$\hat{\Theta}_{T;\text{reject}|\textbf{offer}} = \begin{bmatrix} 0.94 & 0.95 & 0.76 & 0.62 & 0.25 & 0.37 & 0.21 & 0.43 & 0.10 \end{bmatrix}.$$

It confirms the expected tendency to reject low offers. The numerical coincidence with intuitively expected results indicates that the proposed estimation provides useful tool for an exploratory data analysis within the inspected problem class.

General observations implied by additional experiments are as follows.

- The partial forgetting was uniformly better than the globally determined data-dependent forgetting as well as a fixed one.

Table 1
Prediction quality (23) for the best structure of delayed values of the predicted pewee song phrases. Processing of $T = 1327$ data items required 79 s (a standard PC, non-optimised scripts).

Predictor	$\mathbf{Q_0}$	$\mathbf{Q_1}$	$\mathbf{Q_2}$
Markov mixture with delays 1 2 4 6	0.88	0.08	0.03
Trivial predictor	0.05	0.53	0.42
Batch-learnt 1st order Markov chain	0.73	0.26	0.00

- The stabilised forgetting makes the learning robust to over-parametrisation (for instance, in the discussed example, 74 predictions of the modelled decisions $(\Delta_t)_{t=1}^{T=100}$ were correct even with non-significant components included while the properly parameterised one-component model predicted correctly 75 times).
- The results, similarly as in other experiments, were found weakly dependent on the optional lower bound $\underline{\lambda}$ of the forgetting factor.

5.2. Pewee song phrases modelling by high-order Markov chain

The sequences generated by wood pewee – a New England song bird – was analysed e.g. in [9]. The correlation of wood pewee song phrases is a major feature of the data. The main wood pewee phrases are coded into three groups $\Delta \in \mathbf{\Delta} \equiv \{1, 2, 3\}$. The processed sequence is in Appendix A. This is one of a few real-life examples within the considered category to which raw data is available [45].

The intended use of the proposed estimator makes its predictive abilities the decisive characteristics. However, numerical characterisation of prediction quality is mostly missing in references offering raw data. This led us to selection of the following quality criteria (serving in the next example, too)

$$Q_k = \frac{1}{T}\sum_{t\in\boldsymbol{t}} \delta(\mathrm{abs}(\hat{\Delta}_t - \Delta_t)k) \tag{23}$$
$$= \text{the portion of absolute prediction errors with value } k \in \{0, \ldots, |\mathbf{\Delta}| - 1\}.$$

The following point predictions $\hat{\Delta}_t$ of Δ_t were used:

- The maximiser of the proposed predictor (9) given by statistic values κ_{t-1}, V_{t-1}.
- The trivial predictor (tomorrow's weather will be the same as today's) selecting $\hat{\Delta}_t = \Delta_{t-1}$ served as "base-line" for judging quality of the proposed predictor.
- The *batch estimation* of the 1st order Markov chain was performed and maximisers of the corresponding predictor processing the same data were used as point predictions. This provides the hard, processing independent, limit of the 1st order Markov chain serving as predictor.

Only the value Q_0 is relevant in the purely discrete world. Other values are meaningful, when close discrete values reflect closeness of underlying discretised continuous-valued variable.

In the discussed case, high-order mixture models embedded into the model with delayed values $(\Delta_{t-i})_{i=1}^{10}$ of Δ_t were recursively learnt on this sequence and used for one-step-ahead prediction. The processing results achieved for the best regression-vector structure are in Table 1.

General observations are as follows.

- The achieved high prediction quality $Q_0 = 0.88$ fits statements in [9].
- The use of higher order model significantly improved prediction quality.
- The parsimonious mixture parametrisation with the maximum considered delay 10 dealt with $\ell_{\mathbf{\Theta}} = 10 + 3 \times 10 \times 3 = 100$ parameters in contrast with infeasible parametrisation (1) with $\ell_{\mathbf{\Omega}} = 3^{11} = 177,147$ parameters.
- The manual selection of the best structure according to values of the point estimates of component weights suffices as it strongly correlates with prediction quality.
- The results are weakly dependent on the optional lower bound $\underline{\lambda}$ of the forgetting factor.
- Similar unreported experiments with DNA sequences of a mouse have confirmed observations of [9] about their dependence structure.

5.3. Prediction of sale demands

The paper [15] deals with a discretised data about sales demands with observations $\Delta \in \mathbf{\Delta} \in \{1, \ldots, 6\} \equiv$ { no, very slow-moving, slow-moving, standard, fast-moving, very fast-moving }. It provides predicted discretised data, copied in Appendix B, and proposes a heuristically justified estimation of high-order multivariate Markov chain. The estimation should serve for sale predictions and for recognising correlations between the inspected time series.

The use of the estimator and predictor proposed here implies that four time series (denoted B, C, D, E as in [15]) are optimally described by 1st order Markov chain. Only the series A is optimally described by 2nd order model with the 1st delayed value omitted. A strong correlation is found between the series pair A, B, see the first two values of $\mathbf{Q_0}$ in Table 2.

Table 2
Prediction quality (23) of sales' series A for the best higher order Markov chain and for the best regression on values taken from series B. Processing of $T = 269$ data items required 8 s (a standard PC, non-optimised scripts).

Predictor	Q_0	Q_1	Q_2	Q_3	Q_4	Q_5
Markov chain with delay 2	0.46	0.09	0.03	0.02	0.31	0.09
Static dependence on delayed B	0.47	0.12	0.03	0.02	0.33	0.04
Trivial predictor	0.40	0.10	0.03	0.01	0.35	0.10
Batch-learnt 1st order Markov chain	0.46	0.07	0.02	0.01	0.38	0.05

General observations are as follows.

- Comparison of our prediction results with those in [15] is impossible as this work suffers from methodological flaw: prediction quality is evaluated on data used for learning. The batch-learnt 1st order Markov chain, see Section 5.2, provides a firm numerical check of this statement. This makes doubtful conclusions of [15] concerning the model structure and estimator properties. The claimed suitability of the 8th order Markov chain can be valid for fitting (interpolation) but is unconfirmed for prediction (extrapolation). This conclusion concerns also quality of higher order models.
- Concerning correlation aspects, the results of [15] are more relevant. For instance, they found a strong correlation between series A, B, which is in harmony with the physical nature of products they describe. Our corresponding result, demonstrated on predictive power, confirms this, see Table 2. The same coincidence of qualitative correlation results was found for other series, too.
- Again, parsimonious representation has allowed us to inspect joint dependence of individual sales on values of all other series with order up to 2. This case included $\ell_{\boldsymbol{\Theta}} = 2 \times 5 + 2 \times 5 \times 6 \times 6 = 370$ parameters in contrast with unfeasible amount $\ell_{\boldsymbol{Xi}} = 6^{11} = 362,797,056$ of the full parametrisation.
- This discretised case makes inspection of quality via $Q_k, k \geq 0$ meaningful. Closeness of underlying continuous variables for close discretised values can be and was reflected in prior pd of unknown parameters—transitions probabilities between close values of Δ are higher than between distant one. This prior pd has not influenced the final results but helped us meaningfully treat the cases with the number of unknown parameters larger than the number of data.

6. Concluding remarks

An efficient algorithm of recursive Bayesian estimation of high-order Markov chains with possible external explanatory variables in regression was proposed and discussed in the text. It remains to add:

- The recursive learning of the considered black-box model has a high application potential. Diagnostics, fault detection and isolation problems are the areas of an immediate use as they, typically, deal with $|\boldsymbol{\Delta}|, |\boldsymbol{\psi}_i|$ having values 2 or 3 and the number of regressors $\ell_{\boldsymbol{\psi}}$ up to several hundreds.
- Promising tests of the proposed estimator dealing with modelling of a crowd of pedestrians are documented in [20] and the resulting predictor is being built into a system trading with futures. Stream data processing, e.g. [14], is conjectured to be another area with an immediate application potential.
- General applicability is limited by the lack of universal structure estimation algorithm. Bayesian testing of compound hypotheses [10] provides theoretical basis but an algorithm coping with an extreme size of the hypotheses space is to be elaborated. At least, heuristic techniques available for normal regression models [31] and their mixtures [26] can be mimicked.
- Extension of the proposed learning to partially observable Markov chains [13] would extend applicability substantially but it is expected to be much harder.
- The corresponding fully probabilistic design of decision strategies [26,28], generalising the classical Bayesian decision making [30], is conjectured to be convertible into estimation of high-order Markov chain, similarly as in [24]. If this is true then a universal, widely-applicable, adaptive decision strategy can be created.

The paper solves practically important learning problem serving both to advanced adaptive systems and big-data handling, which often requires one-pass (recursive) processing. Subjectively, the *major contribution is*, however, *methodological*. It makes a definite step within a long-term effort to create widely applicable recursive-learning theory, which (i) serves to subsequent dynamic decision making; (ii) respects uncertainty and limited evaluation resources; (iii) is applicable to a wide class of parametric models; (iv) avoids as much as possible heuristic steps; and (v) diminishes an overload of users induced by offering them unnecessary tuning knobs.

The comparison with author's former results related to Markov chains, which are an integral part of general development discussed in the Introduction, well illustrates this progress. The oldest attempts [53] relied on particular parsimonious block-diagonal parametrisation of transition probability and suited only for discretised variables. A more general methodology [32] allowed combination of low-order Markov models but required heuristically motivated 2nd layer model too weakly connected with the original learning task. Markov chain mixtures were also employed, however, their learning relied on heuristic mean-tracking clustering [54] and quasi-Bayes learning [26,33]. These intermediate solutions moreover did not address systematically the problem of accumulation of approximation errors. The recent general theory of approximate recursive learning [25] almost

reached the targeted systematic methodology except a heuristic choice of the data-dependent forgetting factor. The current paper can be seen as an application of this methodology to mixture of low order Markov chain, which moreover decreases the heuristic extent of the forgetting-factor choice. This improvement is applicable generally and thus the dreamt research aim has been (almost) reached.

Acknowledgements

The research reflected here has been supported by GAČR 13-13502S. The paper was improved due to the feedback from the review process.

Appendix A. Wood pewee song phrases time series

The processed data are ordered row-wise.

2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 3 1
2 1 2 1 1 2 1 1 2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1
2 1 3 1 2 1 3 1 2 1 2 1 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 1 3 1 2 1 3 1
2 1 3 1 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 1
2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1
2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 2 1 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1
3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1
2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1
2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1
3 1 2 1 3 1 2 1 1 3 1 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1
2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1
3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 3 1
2 1 3 1 2 3 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 1 2 1 3 1 2 1 3 1
2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3
1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 3 1 2 1 1 2 1 3 1 2 1 2 1 3 1
2 1 3 1 2 1 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 1 3 1 2 1 3 1
2 1 3 1 2 1 3 1 2 1 1 2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 1 2 1 3 1 1 2 1 1 2 1
1 3 2 1 1 2 1 1 3 1 2 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1 3 1 2 1
2 1 3 1 2 1 3 1 2 1 1 2 1 3 1 2 1 3 1 2 1 2 1 3 1 2 1 3 1 2
1 3 1 2 1 1 2 1 3 1 1 2 1 3 1 2 1 3 1 2 1 1 3 1 2 1 1 2 1 1
2 1 1 2 1 1 2 1 1 2 1 1 1 2 1 3 1 2 1 1 2 1 1 2 1 3 1 2 1 3
1 2 1 1 2 1 1 2 1 3 1 2 1 3 1 2 1 2 1 1 3 1 2 1 2 1 2 1 1 2
1 3 1 2 1 2 1 2 1 1 2 3 1 1 2 1 3 1 2 1 1 2 1 3 1 2 1 1 2 1
3 1 2 1 2 1 1 1 2 1 1 2 1 3 1 2 1 2 1 3 1 2 1 2 1 3 1 2 1 1
2 1 1 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2
1 1 1 2 1 1 2

Appendix B. Sales-demands time series

The processed data are ordered row-wise.

Product A

```
6 6 6 6 2 6 2 6 2 2 6 2 6 6 2 6 2 4 4 4 5 6 6 1 2 2 6 6 6 2
6 2 6 6 2 6 2 2 6 2 1 2 2 6 6 6 2 1 2 6 2 6 6 2 2 6 2 2 2 6
2 6 2 2 2 2 2 6 2 2 6 6 6 6 1 2 2 6 2 2 2 2 6 2 2 2 2 3 3 2
3 2 6 6 6 6 2 6 2 6 6 2 6 2 6 6 2 6 6 2 2 3 4 3 3 1 3 1 2 1
6 1 6 6 1 6 6 2 6 2 6 2 2 2 6 6 1 6 2 6 1 2 1 6 2 6 2 2 2 2
6 6 1 6 6 2 2 6 2 2 2 3 4 4 4 6 4 6 1 6 6 1 6 6 6 6 1 6 2 2
2 6 6 6 6 2 6 6 2 2 6 2 6 2 2 2 6 2 2 2 6 6 6 6 3 2 2 6 2 2
2 2 2 2 6 2 6 2 2 2 6 2 2 6 6 2 6 6 6 2 2 2 3 3 3 4 1 6 6 1
6 6 1 6 1 6 6 6 6 1 6 6 6 2 1 2 2 2 2 2 2 2 3 6 6 6 6 6 2 6
```

Product B

```
1 6 6 1 6 1 1 1 1 1 1 6 6 6 1 2 1 6 6 1 1 1 6 6 2 1 6 6 1 1
1 6 1 2 1 6 2 2 2 2 2 6 1 6 6 1 2 1 6 6 6 1 1 1 6 6 1 1 1 1
6 1 1 2 1 6 1 6 1 1 6 2 6 2 6 6 6 3 6 6 1 6 6 2 2 2 3 2 2 6
6 6 1 1 6 2 6 6 2 6 2 6 6 1 3 6 6 1 1 1 2 2 3 2 2 6 2 2 2 1
6 1 6 1 1 6 2 1 1 1 2 2 1 6 1 1 1 1 2 6 1 1 1 1 6 1 6 1 2 1
6 1 6 6 1 6 1 2 2 2 2 3 3 2 2 2 6 6 6 6 2 1 1 6 1 1 1 6 1 6
1 6 1 6 1 1 6 6 2 1 1 6 6 1 1 2 6 2 6 6 6 1 2 6 1 6 1 1 1 1
6 1 6 1 1 6 6 1 6 6 1 6 1 6 6 1 1 6 6 2 2 2 2 2 2 2 2 2 6 6
6 6 1 6 6 6 1 6 6 1 6 6 1 1 6 1 3 3 3 5 1 6 6 6 6 6 6 6 6
```

Product C

```
6 6 6 6 6 6 6 2 6 6 6 6 6 6 6 2 6 6 6 6 2 6 6 6 2 2 6 6 6 6
6 6 6 1 6 2 6 6 6 6 6 6 6 6 6 2 6 6 1 2 6 1 6 6 1 6 2 6 6 6
6 6 6 2 6 6 6 2 6 6 1 6 6 6 6 6 6 6 3 3 6 3 2 1 2 2 1 6 6 1
6 1 6 6 6 6 6 6 1 6 6 6 1 6 6 6 6 6 6 6 6 6 6 6 2 6 6 6 6 6
6 6 6 2 2 6 6 2 6 1 2 6 6 6 2 6 6 2 6 6 2 6 1 6 2 6 2 1 2 6
6 2 2 6 2 6 2 2 6 2 6 6 6 2 2 2 6 6 2 6 6 2 2 6 1 2 1 2 6 6
2 2 6 6 1 2 2 1 6 2 6 2 2 1 1 5 6 3 6 1 6 6 1 2 2 6 1 6 2 6
6 1 6 2 6 2 6 6 6 1 6 1 6 6 2 2 2 1 2 3 6 1 6 1 6 1 6 1 6 6
6 1 1 6 6 6 6 6 1 6 6 6 1 6 1 1 6 6 6 6 6 6 6 6 1 6 6 1 6
```

Product D

```
6 2 2 2 2 3 3 4 4 4 5 4 3 3 6 2 6 6 6 3 4 4 3 3 3 3 3 2 6 6
3 4 4 4 4 3 4 2 6 2 2 6 2 2 6 6 3 4 5 4 4 6 3 6 6 6 2 6 2 6
6 2 2 6 4 4 5 4 3 4 3 4 4 6 2 6 6 2 2 6 2 6 6 2 6 6 2 6 6 2
6 2 6 3 5 5 5 4 4 4 3 6 2 6 6 2 6 2 6 2 2 6 2 6 6 2 6 4 4 4
4 4 4 6 3 6 6 2 6 2 6 2 6 2 6 6 2 2 2 2 2 2 2 2 2 3 3 3 5 5
4 5 3 3 3 6 2 6 6 2 2 6 2 2 2 2 6 2 3 2 2 3 6 3 2 2 3 4 4 4
4 5 5 4 4 6 6 2 6 2 6 2 2 2 2 2 2 2 5 5 4 4 5 5 2 6 2 6 6 2
6 2 6 2 2 3 3 4 4 5 4 4 4 3 4 3 6 2 6 2 2 2 2 2 2 2 2 2 2 2
3 4 4 4 4 5 4 4 4 3 2 2 2 6 2 2 2 6 2 6 2 6 2 2 2 2 2 3 2
```

Product E

```
6 2 2 2 2 3 3 4 4 4 5 4 3 3 6 2 6 6 2 3 4 4 3 4 4 3 3 2 2 6
3 4 4 4 4 3 4 2 3 2 2 6 3 3 6 6 3 4 5 4 5 3 3 2 6 6 2 6 2 6
6 2 2 6 4 4 4 4 4 4 5 4 4 6 2 6 6 2 2 6 2 6 6 2 6 6 2 6 6 2
6 2 6 3 4 4 4 4 4 4 4 6 2 6 6 2 6 2 6 6 6 6 2 6 2 2 6 4 4 4
4 4 4 6 3 3 6 2 2 2 6 2 6 2 2 2 2 2 2 2 2 2 2 2 3 6 4 5 5 5
5 5 2 4 6 6 2 6 6 2 2 6 2 2 2 2 6 2 3 2 2 3 6 3 2 2 3 4 4 4
4 5 5 4 3 3 6 2 6 2 2 2 6 3 2 2 2 2 5 5 4 4 4 4 3 6 2 6 6 2
6 2 6 2 2 3 3 4 4 5 4 4 4 4 4 3 6 2 6 2 2 2 6 2 2 2 2 2 2 2
3 4 4 4 4 5 4 4 4 3 2 2 2 6 6 6 2 6 2 6 2 6 2 2 2 2 2 2 2 2
```

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, Dover Publications, New York, 1972.
[2] H. Alzer, Inequalities for the beta function of n variables, ANZIAM J. 44 (2003) 609–623.
[3] J. Andrýsek, Estimation of Dynamic Probabilistic Mixtures (Ph.D. thesis), FJFI, ČVUT, POB 18, 18208 Prague 8, Czech Republic, 2005.
[4] A. Asahara, K. Maruyama, A. Sato, K. Seto, Pedestrian-movement prediction based on mixed Markov-chain model, in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'11), ACM, New York, NY, USA, 2011, pp. 25–33.
[5] K.J. Astrom, B. Wittenmark, Adaptive Control, Addison-Wesley, Massachusetts, 1989.
[6] O. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, Wiley, New York, 1978.

[7] J.N. Bearden, Ultimatum bargaining experiments: the state of the art, Social Science Research Network, 2001 Technical report. http://dx.doi.org/10.2139/ssrn.626183
[8] R.E. Bellman, Adaptive Control Processes, Princeton University Press, NJ, 1961.
[9] A. Berchtold, A. Raftery, The mixture transition distribution model for high-order Markov chains and non-Gaussian time series, Stat. Sci. 17 (3) (2002) 328–356.
[10] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer, New York, 1985.
[11] J.M. Bernardo, Expected information as expected utility, Ann. Stat. 7 (3) (1979) 686–690.
[12] C. Boutilier, R. Dearden, M. Goldszmidt, Stochastic dynamic programming with factored representations, Artif. Intell. 121 (2000) 49–107.
[13] X. Boyen, D. Koller, Approximate learning of dynamic models, in: M. Kearns, S. Solla, D. Cohn (Eds.), Advances in Neural Information Processing Systems, 11, 1998, pp. 396–403.
[14] L. Chen, Q. Mei, Mining frequent items in data stream using time fading model, Inform. Sci. 257 (2014) 54–69.
[15] W.K. Ching, M.K. Ng, E.S. Fung, Higher-order multivariate Markov chains and their applications, Linear Algebra Appl. 428 (2008) 492–507.
[16] K. Dedecius, I. Nagy, M. Kárný, Parameter tracking with partial forgetting method, Int. J. Adaptive Control Signal Process. 26 (1) (2012) 1–12.
[17] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B (Methodol.) 39 (1) (1977) 1–38.
[18] J. Grim, J. Vejvalková, An iterative inference mechanism for the probabilistic expert system PES, Int. J. Gen. Syst. 27 (1999) 373–396.
[19] C. Guestrin, D. Koller, R. Parr, Max-norm projections for factored MDPs, in: International Joint Conference on Artificial Intelligence, IJCAI-01, Seattle, Washington, 2001, pp. 673–680.
[20] P. Hrabák, O. Ticháček, Prediction of pedestrian decisions during the egress situation: application of recursive estimation of high-order Markov chains using the approximation by finite mixtures, ÚTIA AV ČR, PO Box 18, 182 08 Prague 8, Czech Republic, 2015 Technical report 2346.
[21] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer-Verlag, New York, 2001.
[22] R. Jiroušek, J. Vejnarová, Compositional models and conditional independence in evidence theory, Int. J. Approximate Reasoning 52 (2011) 316–334.
[23] M. Kárný, Adaptive systems: local approximators? in: Workshop on Adaptive Systems in Control and Signal Processing, IFAC, Glasgow, 1998, pp. 129–134.
[24] M. Kárný, On approximate fully probabilistic design of decision making strategies, in: T.V. Guy, M. Kárný (Eds.), Proceedings of the 3rd International Workshop on Scalable Decision Making, ECML/PKDD 2013, UTIA AV ČR, Prague, 2013. ISBN 978-80-903834-8-7.
[25] M. Kárný, Approximate Bayesian recursive estimation, Inform. Sci. 289 (2014) 100–111.
[26] M. Kárný, J. Böhm, T.V. Guy, L. Jirsa, I. Nagy, P. Nedoma, L. Tesař, Optimized Bayesian Dynamic Advising: Theory and Algorithms, Springer, 2006.
[27] M. Kárný, J. Böhm, T.V. Guy, P. Nedoma, Mixture-based adaptive probabilistic control, Int. J. Adaptive Control Signal Process. 17 (2) (2003) 119–132.
[28] M. Kárný, T.V. Guy, Fully probabilistic control design, Systems Control Lett. 55 (4) (2006) 259–265.
[29] M. Kárný, T.V. Guy, On support of imperfect Bayesian participants, in: T.V. Guy, M. Kárný, D.H. Wolpert (Eds.), Decision Making with Imperfect Decision Makers, Intelligent Systems Reference Library, vol. 28, Springer, Berlin, 2012.
[30] M. Kárný, T. Kroupa, Axiomatisation of fully probabilistic design, Inform. Sci. 186 (1) (2012) 105–113.
[31] M. Kárný, R. Kulhavý, Structure determination of regression-type models for adaptive prediction and control, in: J.C. Spall (Ed.), Bayesian Analysis of Time Series and Dynamic Models, Marcel Dekker, New York, 1988. (Chapter 12).
[32] M. Kárný, I. Nagy, A. Halousková, Second layer least squares for recursive non-linear estimation, in: M. Blanke, T. Söderström (Eds.), Preprints of the 10th IFAC Symposium on System Identification (SYSID'94), vol. 2, Danish Automation Society, Kopenhagen, 1994, pp. 655–660.
[33] M. Kárný, M. Valečková, H. Gao, Mixtures of adaptive controllers based on Markov chains: a future of intelligent control? in: International Conference on Control'98, IEE, London, 1998, pp. 721–726.
[34] D.F. Kerridge, Inaccuracy and inference, J. R. Stat. Soc. B 23 (1961) 284–294.
[35] Z. Knejflová, G. Avanesyan, T.V. Guy, M. Kárný, What lies beneath players' non-rationality in ultimatum game?, in: T.V. Guy, M. Kárný (Eds.) Proceedings of the 3rd International Workshop on Scalable Decision Making, ECML/PKDD 2013, UTIA AV ČR, Prague, 2013.
[36] R. Kulhavý, Recursive Bayesian estimation under memory limitations, Kybernetika 26 (1990) 1–20.
[37] R. Kulhavý, Implementation of Bayesian parameter estimation in adaptive control and signal processing, Statistician 42 (1993) 471–482.
[38] R. Kulhavý, M.B. Zarrop, On a general concept of forgetting, Int. J. Control 58 (4) (1993) 905–924.
[39] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–87.
[40] J.M. Martín-Sánchez, J.M. Lemos, J. Rodellar, Survey of industrial optimized adaptive control, Int. J. Adaptive Control Signal Process. 26 (10) (2013) 881–918.
[41] T.P. Minka, Expectation propagation for approximate Bayesian inference, in: Proceedings of the Seventeenth conference on Uncertainty in Artificial intelligence, Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.
[42] E.W.T. Ngai, L. Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: a literature review and classification, Expert Syst. Appl. 36 (2009) 2592–2602.
[43] A.S. Poznyak, K. Najim, E. Gomez-Ramirez, Self-Learning Control of Finite Markov Chains, Automation and Control Engineering, Taylor & Francis, 2000.
[44] H.J. Rabitz, O.F. Alis, General foundations of high-dimensional model representations, J. Math. Chem. 25 (1999) 197–233.
[45] A.E. Raftery, S. Tavaré, Estimation and modelling repeated patterns in high-order Markov chains with the mixture transition distribution (MTD) model, J. R. Stat. Soc., Ser. C: Appl. Stat. 43 (1994) 179–200.
[46] A. Rubinstein, Perfect equilibrium in a bargaining model, Econometrica 50 (1) (1982) 97–109.
[47] P. Sadghi, R.A. Kennedy, P.B. Rapajic, R. Shams, Finite-state Markov modeling of fading channels, IEEE Signal Process. Mag. 57 (2008).
[48] L.K. Saul, M.I. Jordan, Mixed memory Markov models: decomposing complex stochastic processes as mixtures of simpler ones, Mach. Learn. 37 (1999) 75–86.
[49] L.J. Savage, Foundations of Statistics, Wiley, NY, 1954.
[50] T. Singliar, M. Hauskrecht, Variational learning for the noisy-or component analysis, in: Proceedings of the SIAM International Data Mining Conference, 2005, pp. 370–379.
[51] V. Šmídl, A. Quinn, The Variational Bayes Method in Signal Processing, Springer, 2005.
[52] A.F.M. Smith, U. Makov, A quasi-Bayes sequential procedures for mixtures, J. R. Stat. Soc. 40 (1) (1978) 106–112.
[53] M. Valečková, M. Kárný, Estimation of Markov chains with reduced parametrisation, in: L. Berec, J. Rojíček, M. Kárný, K. Warwick (Eds.), Preprints of the 2nd European IEEE Workshop on Computer-Intensive Methods in Control and Signal Processing (CMP'96), ÚTIA AV ČR, Prague, 1996, pp. 135–140.
[54] M. Valečková, M. Kárný, E.L. Sutanto, Bayesian M-T clustering for reduced parameterisation of Markov chains used for nonlinear adaptive elements, Automatica 37 (6) (2001) 1071–1078.