Noname manuscript No.
(will be inserted by the editor)

Mixture Ratio Modelling of Dynamic Relations

Miroslav Kárný · Marko Ruman

Received: date / Accepted: date

Abstract Any knowledge extraction from data sets implicitly or explicitly relies on hypothesis about relations between data records. The inspected black-box methodology uses "universally" approximating mapping describing a wide class of relations. The knowledge incompleteness, the ever present uncertainty and the ultimate use in a subsequent decision making (DM) single out probabilistic models as adequate. Among them, the use of finite mixtures with components in the exponential family (EF) dominate. Their dominance stems from their flexibility, cluster interpretability and the availability of a range sophisticated algorithms suitable for high-dimensional data streams. They are even used in dynamic situations with mutually dependent data records. The dependence is modelled by employing regression and auto-regression components. However, with a few specialised exceptions, these dynamic models assume data-independent weights of mixture components. Their use is illogical as it assumes independent (memoryless) transitions between dynamic mixture components. The restricted nature of such mixtures follows from the fact that the *set of finite probabilistic mixtures is not closed with respect to the conditioning*, which is the key learning operation. The paper overcomes this restriction by using *ratios of finite mixtures* as dynamic parametric models. It motivates them, elaborates their approximate Bayesian *stream* learning. The paper reveals their application potential stemming predominantly from their inherent ability to process unbalanced data sets. This is vital when a the portion of interesting data, reflecting for instance failures or attacks, is small.

Keywords Mixture Models · Approximate Bayesian Learning · Stream Learning · Kullback-Leibler Divergence

1 Introduction

The paper topic falls to a broad domain of decision making (DM) understood as a targeted choice among available options. This covers many domains including machine learning, Mitchell (1997), signal processing, Sadghi et al. (2008), filtering,

The Czech Academy of Sciences, Institute of Information Theory and Automation, POB 18, 182 08 Prague 8, Czech Republic, E-mail: school@utia.cas.cz, E-mail: marko.ruman@gmail.com

Jazwinski (1970), hypothesis testing, Guyon et al. (2010), classification and pattern recognition, Bishop (2006), knowledge sharing, Mason et al. (2014), reinforcement learning, Wiering and van Otterlo (2012), control, Guan et al. (2012), etc.

All these areas need to enrich the underlying domain knowledge by that hidden in, often excessive, amount of data records: they need data mining. Unsurprisingly, the amount of relevant theoretical, algorithmic and application results is excessive, see e.g. in Holmes and Liu (2015). This makes us to provide reference sample.

Nowadays, data mining is an extremely broad area with many results and applications. Even surveys specialise to Internet of Things, Tsai et al. (2014), cyber-security, Buczak and Guven (2016), etc.

Paper Focus: The paper contributes to data mining focused on clustering of data streams, Nguyen et al. (2015), with dynamically related data records, Mirsky et al. (2015). The contribution concerns modelling and conditioning. Importance of this data-mining type is seen on specific applications, e.g. in geophysics, Appice et al. (2015), or stream clustering of independent data records, Silva et al. (2014).

DM Perspective on Knowledge Processing: A solution of a DM problem leads to a strategy, a collection of decision rules mapping the evolving available knowledge on actions, Wald (1950). The strategy should meet DM aims in the best way under the faced circumstances. The adopted Bayesian DM, Savage (1954), proved to be an adequate methodology coping with incomplete knowledge, uncertainty and randomness of the environment to which actions relate. Bayesian paradigm links DM consequences to the acquired knowledge and the used actions by conditional distributions, here described by conditional probability densities (pd[1]).

Actions are chosen sequentially (in a single pass). This enables DM to enrich its knowledge, to learn a better environment model. This opportunity is exploited by adaptive controllers, Åström and Wittenmark (1989). The same processing of data streams serves as a statistic (knowledge, feature) extractor. The learning is possible iff the learnt relations practically do not change during the knowledge accumulation. The most general used structure structure of such relations leads to stochastic filtering, Jazwinski (1970). It is mostly applied to hidden Markov models, Elliot et al. (1995).

The Addressed Modelling Problem: The achievable modelling and thus DM quality are predetermined by the inspected model set. One-pass Bayesian learning of general hidden Markov models is possible but hard. The treatment of parametric models we focus on can be conceptually extended to them but it is out of scope of this paper. The considered relations of the observed data are described by parametric probabilistic models. Which parameter corresponds to the best model is a priori unknown. Bayesian learning offers the unambiguous deductive Bayes' rule, Berger (1985), which redistributes the belief (pd) about model adequacy, Berec and Kárný (1997). It accumulates the knowledge into the posterior pd.

As usual in the targeted DM, we rely on black-box models. They approximate joint pd of long sequences of observed data records by a stationary Markov model operating on a fixed-dimensional state. Focus on them is enabled by a common

[1] Pds are evaluated with respect to a dominating, typically Lebesque or counting, product measure, Rao (1987). Lebesque's notation $\int \dots \mathrm{d}\bullet$ is mostly used.

weak dependence of data realised at distant time moments and by the possibility to include time into the model state. Then, a state transition pd fully determines the needed joint pd for any given initial state.

For a fixed condition, any continuous transition pd can be arbitrarily-well approximated by a finite mixture of conditional pds (components). Components differ in their parameters but may have the same functional form, say Gaussian one. Various versions of this "universal approximation property" have been proved in neural-network context, Park and Sandberg (1991), as well as in probabilistic modelling, McLachlan and Peel (2000). Such expansions perfectly work and are widely used across domains for zero-memory models, McNicholas (2017).

By construction, the component weights should depend on the condition. The common use of the condition-independent component weights goes against the expansion logic. Rare ad hoc exceptions, Catania (2016); Frigessi et al. (2002), indicate how much is lost when considering constant component weights.

This paper goes beyond the common practice and offers *ratios* of finite mixtures with components in the exponential family (EF), Barndorff-Nielsen (1978), as black-box, Bohlin (1991), universally approximating models, Haykin (1994). It develops, inevitably approximate, but feasible sequential Bayesian learning of models from this *extremely rich but yet unconsidered model set.*

Layout and Common Notation: The paper relies on an approximate sequential Bayesian learning, Kárný (2014). Its recall in Sec. 2 prepares the used notation. Sec. 3 justifies the mixture ratio models in detail. Sec. 4 elaborates their learning. Sec. 5 numerically illustrates the theory. Sec. 6 outlines the model applicability.

Concerning notation, $\boldsymbol{X}$ is a set of Xs. It is either a non-empty subset of a finite-dimensional real space or a subset of pds. $|\boldsymbol{X}|$ means cardinality of $\boldsymbol{X}$. Random variables, their realisations and function arguments are undistinguished. The context suffices to grasp the correct meaning. San serif fonts mark mappings, which are taken as different if their arguments differ, but mnemonic symbols are preferred. For instance, M denotes a parametric model, J is a joint parametric pd of data vectors, P is posterior pd, O marks observations, A labels actions, etc. Decoration $\tilde{}$ denotes intermediate objects and $\hat{}$ estimates.

2 Approximate Sequential Bayesian Learning

This section summarises the used approximate sequential learning, Kárný (2014), and introduces basic treated objects.

Parametric model M links observations $O_t \in \boldsymbol{O}$ to actions $A_t \in \boldsymbol{A}$ at discrete-time moments labelled by $t \in \boldsymbol{t} = \{1, 2, \ldots, |\boldsymbol{t}|\}$. Data records $D_t = (O_t, A_t)$ are sequentially processed. Parametric model M is pd

$$\mathsf{M}(O_t|A_t, D^{t-1}, \Theta), \quad D^{t-1} = (D_{t-1}, \ldots, D_1, D_0) \tag{1}$$

that relates observation O_t to current action A_t and past data records D^{t-1}, extended by prior knowledge D_0. An unknown $\Theta \in \boldsymbol{\Theta}$ parameterises this pd.

The knowledge about parameter Θ at time $t-1$ is expressed by posterior pd $\mathsf{P}_{t-1}(\Theta) = \mathsf{P}(\Theta|A_t, D^{t-1}) = \mathsf{P}(\Theta|D^{t-1})$. The last equality is implied by natural

conditions of control, Peterka (1981), see also Assumption 1 below. Bayes' rule updates this knowledge by the realisation of data record D_t to

$$\tilde{\mathsf{P}}_t(\Theta) = \frac{\mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta)}{\int_{\boldsymbol{\Theta}} \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta)\mathrm{d}\Theta} \propto \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta), \qquad (2)$$

where $\propto$ is proportionality. With growing t, the analytic form of the posterior pd $\tilde{\mathsf{P}}_t$ generically becomes an excessively complex function of many variables $\Theta \in \boldsymbol{\Theta}$. Then, $\mathsf{P}_t \in \mathbf{P}$, with $\mathbf{P}$ containing computationally feasible[2] pds, has to be evolved.

Generally, posterior pd $\tilde{\mathsf{P}}_t \notin \mathbf{P}$ even if pd $\mathsf{P}_{t-1} \in \mathbf{P}$. To preserve feasibility, pd $\tilde{\mathsf{P}}_t$ in (2) is to be projected on $\mathbf{P}$. Work of Bernardo (1979) inspecting this approximation task provides weak conditions (met here) under which the minimiser $\hat{\mathsf{P}}_t \in \mathbf{P}$ of Kerridge's inaccuracy, Kerridge (1961),

$$\mathsf{K}(\tilde{\mathsf{P}}_t||\hat{\mathsf{P}}_t) = -\int_{\boldsymbol{\Theta}} \tilde{\mathsf{P}}_t(\Theta) \ln\Big(\hat{\mathsf{P}}_t(\Theta)\Big)\mathrm{d}\Theta \qquad (3)$$

is the proper Bayesian projection of $\tilde{\mathsf{P}}_t$ on $\hat{\mathsf{P}}_t \in \mathbf{P}$. The projection *should not* serve as the prior pd in a further updating: this may cause a divergence of the sequentially projected pds from those projected optimally in the batch mode, Kulhavý (1990). Forgetting, with a data-dependent factor $\lambda_t \in [0, 1]$, is the adequate countermeasure, Kárný (2014). It completes updating of feasible posterior pds

$$\tilde{\mathsf{P}}_t(\Theta) \propto \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta), \quad \hat{\mathsf{P}}_t = \arg\min_{\hat{\mathsf{P}}\in\mathbf{P}} \mathsf{K}(\tilde{\mathsf{P}}_t||\hat{\mathsf{P}})$$

$$\mathsf{P}_t(\Theta) \propto \hat{\mathsf{P}}_t^{\lambda_t}(\Theta)\mathsf{P}_{t-1}^{1-\lambda_t}(\Theta), \quad \lambda_t = \frac{\left[\int_{\boldsymbol{\Theta}} \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta)\mathrm{d}\Theta\right]^2}{\int_{\boldsymbol{\Theta}} \mathsf{M}^2(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta)\mathrm{d}\Theta}. \qquad (4)$$

The above plausible choice of forgetting factor λ_t is in Kárný et al. (2014).

The next assumption summarises conditions under which (4) is applicable. It adds reasons for them and serves for referencing.

Assumption 1 (Conditions for Applicability of (4))

✓ *Parameter $\Theta \in \boldsymbol{\Theta}$ is unknown to the employed decision rules, described by pds $\mathsf{R}(A_t|D^{t-1}, \Theta)$. Thus, they meet natural conditions of control, Peterka (1981),*

$$\mathsf{R}(A_t|D^{t-1}, \Theta) = \mathsf{R}(A_t|D^{t-1}) \;\Leftrightarrow\; \mathsf{P}(\Theta|A_t, D^{t-1}) = \mathsf{P}(\Theta|D^{t-1}). \qquad (5)$$

Only such decision rules cancel in (2).

✓ *Parametric model $\mathsf{M}(O_t|A_t, D^{t-1}, \Theta)$ (1) has a Θ-independent support as justification of forgetting (4) needs this assumption, Kárný (2014).*

✓ *The set of feasible pds $\mathbf{P}$ is log-convex, i.e. forgetting (4) applied to its members gives a pd in $\mathbf{P}$. Without this, an additional projection is needed.*

Remarks 1

✓ *Kerridge's inaccuracy K (3) is a shifted version of Kullback-Leibler divergence $\mathsf{D}(\tilde{\mathsf{P}}||\hat{\mathsf{P}})$, Kullback and Leibler (1951)*

$$\mathsf{D}(\tilde{\mathsf{P}}||\hat{\mathsf{P}}) = \int_{\boldsymbol{\Theta}} \tilde{\mathsf{P}}(\Theta) \ln\left(\frac{\tilde{\mathsf{P}}(\Theta)}{\hat{\mathsf{P}}(\Theta)}\right)\mathrm{d}\Theta = \int_{\boldsymbol{\Theta}} \tilde{\mathsf{P}}(\Theta) \ln\big(\tilde{\mathsf{P}}(\Theta)\big)\,\mathrm{d}\Theta + \mathsf{K}(\tilde{\mathsf{P}}||\hat{\mathsf{P}}).$$

It has the same minimiser but copes better with $\tilde{\mathsf{P}}$ containing Dirac's pds.

✓ *The unavoidable projection of $\tilde{\mathsf{P}}_t$ on $\mathbf{P}$ (4) is demanding but feasible, cf. Sec. 3. The extra effort for evaluating forgetting factor λ_t is small.*

[2] An intuitive understanding of this notion suffices. Sec. 3 provides example of such pds.

3 Ratio of Finite Mixtures

This core section provides the proposed universal parametrisation of pds modelling data observed in the closed loop formed by a dynamic environment and a, possibly randomised and dynamic, decision strategy meeting (5).

Markovian Modelling: Joint parametric pd $\tilde{\mathsf{J}}(D^{|\boldsymbol{t}|}|\Theta, D_0)$ of data sequences $D^{|\boldsymbol{t}|} = (D_t)_{t\in\boldsymbol{t}}$ factorises

$$\begin{aligned}\tilde{\mathsf{J}}(D^{|\boldsymbol{t}|}|\Theta, D_0) &= \prod_{t\in\boldsymbol{t}} \overbrace{\mathsf{M}(O_t|A_t, D^{t-1}, \Theta)}^{\text{parametric model}} \overbrace{\mathsf{R}(A_t|D^{t-1}, \Theta)}^{\text{decision rule}} \\ &\overset{(5)}{=} \prod_{t\in\boldsymbol{t}} \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{R}(A_t|D^{t-1}),\end{aligned}$$

i.e. the parametrisation and learning concern the parametric model.

Assumption 2 (Markov Time-Invariant Parametric Model) *The model is* time-invariant*, parameterised by a constant multivariate parameter $\Theta \in \boldsymbol{\Theta}$. It is Markov model of an order $n < \infty$, which means*

$$\begin{gathered}\mathsf{M}(O_t|A_t, D^{t-1}, \Theta) = \mathsf{M}(O_t|\psi_t, \Theta), \text{ with regression vector} \\ \psi_t = \text{function of } \psi_{t-1} \;\&\; \begin{cases} A_t, D_{t-1}, \ldots, D_{t-n} & \text{if } 1 \le n < \infty \\ A_t \text{ or void} & \text{for } n = 0. \end{cases}. \qquad (6) \\ \text{Data entering model } \mathsf{M} \text{ at time } t \in \boldsymbol{t} \text{ form data vector } \Psi_t = (O_t, \psi_t) \in \boldsymbol{\Psi}.\end{gathered}$$

The prior knowledge D_0 determines the model structure and also ψ_0.

Parametric model (6) is ratio of the joint pd $\mathsf{J}(\Psi|\Theta)$ and its marginal

$$\mathsf{M}(O_t|\psi_t, \Theta) = \frac{\mathsf{J}(O_t, \psi_t|\Theta)}{\int_{\boldsymbol{O}} \mathsf{J}(O_t, \psi_t|\Theta)\mathrm{d}O_t} = \frac{\mathsf{J}(\Psi_t|\Theta)}{\int_{\boldsymbol{O}} \mathsf{J}(O_t, \psi_t|\Theta)\mathrm{d}O_t}. \qquad (7)$$

It is time invariant iff joint pd $\mathsf{J}(\Psi_t|\Theta)$ is a time-invariant function multiplied by an arbitrary positive function of A_t, D^{t-1}, Θ, e.g., a variable decision rule. Thus, Assumption 2 practically makes joint pd $\mathsf{J}(\Psi_t|\Theta)$ time-invariant.

As already discussed in Introduction, the Markov property is inevitable for the targeted feasibility of the sequential learning. If not met naturally, it must be enforced via an approximation. The required time invariance can be relaxed by inclusion of time into ψ_t or by considering a time- and data- dependent unknown parameter Θ. The latter case corresponds with hidden Markov models, Elliot et al. (1995), which are out of scope of this paper.

Exponential Family: The proposed model exploits members of EF. Recall that a parametric Markov model $\tilde{\mathsf{M}}(O_t|\psi_t, \Theta)$ belongs to EF if it is described by the pd

$$\tilde{\mathsf{M}}(O_t|\psi_t, \Theta) = \exp\left\langle \mathsf{B}(\Psi_t), \mathsf{C}(\Theta)\right\rangle, \; \Psi_t \in \boldsymbol{\Psi}, \; \Theta \in \boldsymbol{\Theta}. \qquad (8)$$

There, multivariate real-valued functions $\mathsf{B}(\Psi_t)$, $\mathsf{C}(\Theta)$ have a fixed finite dimension and real-valued mapping $\langle \mathsf{B}(\Psi_t), \mathsf{C}(\Theta)\rangle$ is *linear in* $\mathsf{B}(\Psi_t)$*-values*. Definition (8)

admits usual factors depending respectively on Ψ and Θ. It suffices to include constant entries into B and C. The support indicator is dropped for simplicity.

Under natural conditions of control (5), exponential-family members posses conjugated (self-reproducing) pds, Berger (1985),

$$\mathsf{P}_t(\Theta) = \mathsf{P}(\Theta|D^t) = \mathsf{P}(\Theta|\mathsf{V}_t) = \frac{\exp\langle \mathsf{V}_t, \mathsf{C}(\Theta)\rangle}{\mathsf{N}(\mathsf{V}_t)} \tag{9}$$

$$\mathsf{N}(\mathsf{V}_t) = \int_{\boldsymbol{\Theta}} \exp\langle \mathsf{V}_t, \mathsf{C}(\Theta)\rangle \,\mathrm{d}\Theta \quad \text{with updating } \mathsf{V}_t = \mathsf{V}_{t-1} + \mathsf{B}(\Psi_t).$$

Their use converts Bayes' rule, which is a functional recursion, into algebraic updating of values the fixed-dimensional sufficient statistic V_t. Initial V_0 in (9) describes the conjugated prior pd quantifying prior knowledge D_0. It has to guarantee $\mathsf{N}(\mathsf{V}_0) < \infty$. Such V_0 regularises the learning.

Remarks 2

- ✓ *Under Assumption 1, exponential family exhausts parametric models, smooth in Θ, with a finite-dimensional sufficient statistic $\mathsf{V}_t = \mathsf{V}(D^t)$, Koopman (1936).*
- ✓ *For truly dynamic parametric models with a non-constant regression vector, EF is quite narrow. It essentially contains normal linear-in-regression-coefficient models, for continuous observations, and Markov chain models, for discrete-valued observations and discrete-valued regression vectors. Only for them, the marginal pd of the regression vector, proportional to the normalisation factor in (7), depends on Θ only (not on regression vector). This makes them similar to the static case with $\Psi_t = O_t$, for which EF is much richer.*

Universal Approximation: Joint pds $\mathsf{J}(\Psi) = \mathsf{J}(O,\psi)$ of data vectors $\Psi \in \boldsymbol{\Psi}$, even with non-trivial regression vectors $\psi \in \boldsymbol{\psi}$, can be "universally" approximated by a finite mixture of normal pds, to an arbitrary precision, McLachlan and Peel (2000). Thus, it can be approximated by a finite mixture of pds from EF . This allows us to consider the joint parametric pd in (7) as the finite weighted sum of components, pds $(\mathsf{J}_c(\Psi_t|\Theta_c))_{c\in\boldsymbol{c}}$ on $\boldsymbol{\Psi}$, in EF

$$\mathsf{J}(\Psi|\Theta) = \sum_{c\in\boldsymbol{c}} \alpha_c \overbrace{\exp\langle \mathsf{B}_c(\Psi), \mathsf{C}_c(\Theta_c)\rangle}^{\mathsf{J}_c(\Psi_t|\Theta_c)}, \quad \boldsymbol{c} = \{1,\dots,|\boldsymbol{c}|\},\ \alpha = (\alpha_c)_{c\in\boldsymbol{c}}$$

$$\alpha \in \boldsymbol{\alpha} = \left\{\alpha_c \geq 0, \sum_{c\in\boldsymbol{c}} \alpha_c = 1\right\},\ \Theta = (\alpha_c, \Theta_c)_{c\in\boldsymbol{c}} \in \boldsymbol{\Theta} = (\boldsymbol{\alpha}, (\boldsymbol{\Theta}_c)_{c\in\boldsymbol{c}})\,. \tag{10}$$

Insertion of (10) into (7) gives the parametric *ratio model*, which is needed in learning via Bayes' rule (2),

$$\mathsf{M}(O_t|\psi_t,\Theta) = \sum_{c\in\boldsymbol{c}} \frac{\alpha_c \exp\left\langle \mathsf{B}_c(O_t,\psi_t), \mathsf{C}_c(\Theta_c)\right\rangle}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \underbrace{\int_{\boldsymbol{O}} \exp\langle \mathsf{B}_{\tilde{c}}(O_t,\psi_t), \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\rangle \,\mathrm{d}O_t}_{\mathsf{W}_{\tilde{c}}(\psi_t,\Theta_{\tilde{c}})}} \tag{11}$$

$$= \sum_{c\in\boldsymbol{c}} \underbrace{\frac{\alpha_c \mathsf{W}_c(\psi_t,\Theta_c)}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \mathsf{W}_{\tilde{c}}(\psi_t,\Theta_{\tilde{c}})}}_{\mathsf{w}_c(\psi_t,\Theta)} \underbrace{\frac{\exp\langle \mathsf{B}_c(\Psi_t), \mathsf{C}_c(\Theta_c)\rangle}{\mathsf{W}_c(\psi_t,\Theta_c)}}_{\mathsf{M}_c(O_t|\psi_t,\Theta_c)} = \sum_{c\in\boldsymbol{c}} \mathsf{w}_c(\psi_t,\Theta)\mathsf{M}_c(O_t|\psi_t,\Theta_c).$$

Assumption 2 and *universal* approximation by finite mixtures *for void* ψ imply:

Ratio (11) approximates any *dynamic*, Markov, time-invariant environment model.

Remarks 3

- ✓ *Let us stress that functions* $\mathsf{W}_c(\psi_t, \Theta_c)$ *result from the integration over* $\boldsymbol{O}$ *only, not over* $\boldsymbol{\Psi}$*. This often takes components* $(\mathsf{M}_c(O_t|\psi_t, \Theta_c))_{c\in\boldsymbol{c}}$ *out of EF .*
- ✓ *The second row of (11) shows that the model is a fully dynamic finite mixture. Both components and their weights* $\mathsf{w}_c(\psi_t, \Theta)$ *depend on* ψ_t*. The data-dependence* is not arbitrary *and needs* no *extra parameter.*
- ✓ *Components* $\mathsf{J}_c(\Psi_t|\Theta_c) = \exp\langle \mathsf{B}_c(\Psi_t), \mathsf{C}_c(\Theta_c)\rangle$ *in (10) are pds on* $\boldsymbol{\Psi}$*. These pds may contain parameter-independent factors, i.e. pds*

$$\mathsf{J}_c(\Psi_t|\Theta_c) = \exp\langle \mathsf{B}_c(\Psi_t), \mathsf{C}_c(\Theta_c)\rangle = \exp\langle \mathsf{B}_c(\Psi_{t;c}), \mathsf{C}_c(\Theta_c)\rangle\, \mathsf{G}_c(\underline{\psi}_{t;c}) \tag{12}$$

model $\Psi_{t;c} = (O_t, \psi_{t;c})$*. There* $\psi_{t;c}$ *is a sub-vector of* ψ_t *and* $\mathsf{G}_c(\underline{\psi}_{t;c})$ *is a parameter-free pd on its complement* $\underline{\psi}_{t;c}$ *to* ψ_t*. Their use mimics mixtures of principal component analysers, Tipping and Bishop (1999), and diminishes the dimensionality curse, Kárný (2016).*

4 Learning with Mixture Ratios

This section applies learning (4) to mixture ratio model (11) and arrives at its feasible and justified sequential Bayesian learning.

Choice of Set $\mathbf{P}$ *of Feasible PDs:* The component weights in (10) define pd $\mathsf{M}(c_t = c|\Theta) = \alpha_c$, $c \in \boldsymbol{c}$, of an *unobserved* pointer, $c_t \in \boldsymbol{c}$, to the active component $\mathsf{J}_{c_t}(\Psi_t|\Theta_{c_t})$ (12) "generating" data vector Ψ_t. This pointer model is from EF as

$$\mathsf{M}(c_t|\Theta) = \exp\Big[\sum_{c\in\boldsymbol{c}} \delta(c, c_t), \ln(\alpha_c)\Big] = \exp\langle \delta(c_t), \ln(\alpha)\rangle$$

$$\delta(c, c_t) = \begin{cases} 1 \text{ if } c = c_t \\ 0 \text{ otherwise} \end{cases}, \quad \delta(c_t) = [\delta(1, c_t), \ldots, \delta(|\boldsymbol{c}|, c_t)]$$

$$\ln(\alpha) = [\ln(\alpha_1), \ldots, \ln(\alpha_{|\boldsymbol{c}|})], \quad \langle \delta(c_t), \ln(\alpha)\rangle = \sum_{c\in\boldsymbol{c}} \delta(c, c_t) \ln(\alpha_c).$$

For the *observed* pointer $c_t \in \boldsymbol{c}$ to the active component, Dirichlet's pd, determined by $|\boldsymbol{c}|$-vector statistic v_t, is the corresponding conjugated pd, Berger (1985),

$$\mathsf{P}_t(\alpha) = \mathsf{P}(\alpha|\mathsf{v}_t) = \frac{\exp\langle \mathsf{v}_t - 1, \ln(\alpha)\rangle}{\mathsf{Be}(\mathsf{v}_t)} \quad \text{with update } \mathsf{v}_t = \mathsf{v}_{t-1} + \delta(c_t),\ \mathsf{v}_0 > 0$$

$$\mathsf{Be}(\mathsf{v}) = \frac{\prod_{c\in\boldsymbol{c}} \Gamma(\mathsf{v}_c)}{\Gamma\left(\sum_{c\in\boldsymbol{c}} \mathsf{v}_c\right)}, \qquad \Gamma(v) = \int_0^\infty z^{v-1} \exp(-z)\mathrm{d}z,\ v > 0. \tag{13}$$

For normal and Markov chain components or independent data vectors with components (12), conjugate pds $\mathsf{P}_t(\Theta_c) = \mathsf{P}(\Theta_c|D^t, c_t, \ldots, c_1)$, $c \in \boldsymbol{c}$, are

$$\mathsf{P}_t(\Theta_c) = \frac{\exp\langle \mathsf{V}_{t;c}, \mathsf{C}_c(\Theta_c)\rangle}{\mathsf{N}_c(\mathsf{V}_{t;c})}, \quad \mathsf{N}_c(\mathsf{V}_{t;c}) = \int_{\boldsymbol{\Theta}_c} \exp\langle \mathsf{V}_{t;c}, \mathsf{C}_c(\Theta_c)\rangle\, \mathrm{d}\Theta_c,\ c \in \boldsymbol{c}. \tag{14}$$

They are "natural" candidates for creating set $\mathbf{P}$ of feasible pds. They have to be given by *statistic* $(\mathsf{v}_t, (\mathsf{V}_{t;c})_{c\in\boldsymbol{c}})$ with values given by the observed data, not by unobserved pointers $(c_t, \ldots, c_1)$. Marginal pds $\mathsf{P}_t(\alpha)$ (13), $(\mathsf{P}_t(\Theta_c))_{c\in\boldsymbol{c}}$ (14) of $\mathsf{P}_t(\Theta)$, $\Theta = (\alpha, (\Theta_c)_{c\in\boldsymbol{c}})$, do not determine joint pd $\mathsf{P}_t(\Theta)$ unambiguously, Nelsen (1999). Under the faced lack of information about mutual relations of component parameters $(\alpha, (\Theta_c)_{c\in\boldsymbol{c}})$, the product combination of marginal pds is preferable. This motivates the relatively universal choice

$$\mathbf{P} = \left\{ \mathsf{P}_t(\Theta) = \mathsf{P}_t(\alpha) \prod_{c\in\boldsymbol{c}} \mathsf{P}_t(\Theta_c) \right\}, \quad \text{with} \quad \mathsf{V}_t = (\mathsf{v}_t,\ (\mathsf{V}_{t;c})_{c\in\boldsymbol{c}}), \quad \text{in (13), (14).} \tag{15}$$

Evaluation of Kerridge's Inaccuracy (3): With the chosen $\mathbf{P}$, it remains to convert updating (4) into the updating $\mathsf{v}_{t-1}, \mathsf{V}_{t-1}$ to $\mathsf{v}_t, \mathsf{V}_t$. Kerridge's inaccuracy of pd $\tilde{\mathsf{P}}_t$ (2) to pd $\hat{\mathsf{P}}_t \in \mathbf{P}$ (15), given by $\hat{\mathsf{V}}_t = \left(\hat{\mathsf{v}}_t, (\hat{\mathsf{V}}_{t;c})_{c\in\boldsymbol{c}}\right)$ (13), (14), reads

$$\begin{aligned} \mathsf{K}(\tilde{\mathsf{P}}_t||\hat{\mathsf{P}}_t) &= \ln(\mathsf{Be}(\hat{\mathsf{v}}_t)) + \sum_{c\in\boldsymbol{c}} \ln(\mathsf{N}_c(\hat{\mathsf{V}}_{t;c})) \\ &\quad - \left\langle \hat{\mathsf{v}}_t - 1, \int_{\boldsymbol{\Theta}} \ln(\alpha)\tilde{\mathsf{P}}_t(\Theta)\mathrm{d}\Theta \right\rangle - \sum_{c\in\boldsymbol{c}} \left\langle \hat{\mathsf{V}}_{t;c}, \int_{\boldsymbol{\Theta}} \mathsf{C}_c(\Theta_c)\tilde{\mathsf{P}}_t(\Theta)\mathrm{d}\Theta \right\rangle. \end{aligned} \tag{16}$$

Bayes' rule (2) for (11), (12) and pd $\mathsf{P}_{t-1} \in \mathbf{P}$, (13), (14), and (15), gives $\tilde{\mathsf{P}}_t(\Theta) \propto$

$$\sum_{c\in\boldsymbol{c}} \frac{\alpha_c \exp\langle \mathsf{B}_c(\Psi_t), \mathsf{C}_c(\Theta_c)\rangle \exp\langle \mathsf{v}_{t-1} - 1, \ln(\alpha)\rangle \prod_{\tilde{c}\in\boldsymbol{c}} \exp\left\langle \mathsf{V}_{t-1;\tilde{c}}, \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\right\rangle}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \mathsf{W}_{\tilde{c}}(\psi_t, \Theta_{\tilde{c}})} \tag{17}$$

$$\sum_{c\in\boldsymbol{c}} \frac{\exp\langle \mathsf{v}_{t-1} + \delta(c) - 1, \ln(\alpha)\rangle \prod_{\tilde{c}\in\boldsymbol{c}} \exp\left\langle \mathsf{V}_{t-1;\tilde{c}} + \delta(c,\tilde{c})\mathsf{B}_c(\Psi_{t;c}), \mathsf{C}_{\tilde{c}}(\Theta_c)\right\rangle \mathsf{G}_c(\underline{\psi}_{t;c})}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \mathsf{W}_{\tilde{c}}(\psi_t, \Theta_{\tilde{c}})}.$$

Next steps exploit auxiliary pds arising from summands in (17), cf. (13), (14),

$$\begin{aligned} \tilde{\mathsf{Q}}_{t;c}(\Theta) &= \frac{\exp\langle \mathsf{v}_{t-1} + \delta(c) - 1, \ln(\alpha)\rangle}{\beta_{t;c}} \prod_{\tilde{c}\in\boldsymbol{c}} \exp\left\langle \mathsf{V}_{t-1;\tilde{c}} + \delta(c,\tilde{c})\mathsf{B}_c(\Psi_{t;c}), \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\right\rangle \\ \beta_{t;c} &= \mathsf{Be}(\mathsf{v}_{t-1} + \delta(c)) \prod_{\tilde{c}\in\boldsymbol{c}} \mathsf{N}(\mathsf{V}_{t-1;\tilde{c}} + \delta(c,\tilde{c})\mathsf{B}_c(\Psi_{t;c}))/\mathsf{G}_c(\underline{\psi}_{t;c}). \end{aligned} \tag{18}$$

They independently model α and $(\Theta_c)_{c\in\boldsymbol{c}}$. Dependence is brought into $\tilde{\mathsf{P}}_t$ via

$$\begin{aligned} \mathsf{H}(\psi_t, \Theta) &= \frac{1}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \mathsf{W}_{\tilde{c}}(\psi_t, \Theta)} \quad \text{as} \quad \tilde{\mathsf{P}}_t(\Theta) = \frac{\mathsf{H}(\psi_t, \Theta) \sum_{c\in\boldsymbol{c}} \beta_{t;c}\tilde{\mathsf{Q}}_{t;c}(\Theta)}{\mathsf{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})} \\ \mathsf{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1}) &= \int_{\boldsymbol{\Theta}} \mathsf{H}(\psi_t, \Theta) \sum_{c\in\boldsymbol{c}} \beta_{t;c}\tilde{\mathsf{Q}}_{t;c}(\Theta)\mathrm{d}\Theta. \end{aligned} \tag{19}$$

The determination of Kerridge's inaccuracy (16) needs to evaluate the integrals

$$\begin{aligned} \mathsf{I}_{0c}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1}) &= \int_{\boldsymbol{\Theta}} \ln(\alpha_c)\mathsf{H}(\psi_t, \Theta) \sum_{\tilde{c}\in\boldsymbol{c}} \beta_{t;\tilde{c}}\tilde{\mathsf{Q}}_{t;\tilde{c}}(\Theta)\mathrm{d}\Theta, \quad \mathsf{I}_0 = (\mathsf{I}_{0c})_{c\in\boldsymbol{c}}, \\ \mathsf{I}_c(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1}) &= \int_{\boldsymbol{\Theta}} \mathsf{C}_c(\Theta)\mathsf{H}(\psi_t, \Theta) \sum_{\tilde{c}\in\boldsymbol{c}} \beta_{t;\tilde{c}}\tilde{\mathsf{Q}}_{t;\tilde{c}}(\Theta)\mathrm{d}\Theta, \ c \in \boldsymbol{c}. \end{aligned} \tag{20}$$

Their insertion into Kerridge's inaccuracy (16) gives the best-projection statistic

$$\begin{aligned}\hat{\mathsf{V}}_t &= \left[\hat{\mathsf{v}}_t, (\hat{\mathsf{V}}_{t;c})_{c\in\boldsymbol{c}}\right] \in \mathrm{Arg}\min_{\hat{\mathsf{v}}\in\hat{\mathbf{v}},(\hat{\mathsf{V}}\in\hat{\mathbf{V}}_c)_{c\in\boldsymbol{c}}}\left(\ln(\mathsf{Be}(\hat{\mathsf{v}})) + \sum_{c\in\boldsymbol{c}}\ln(\mathsf{N}_c(\hat{\mathsf{V}}_c))\right.\\ &\left. - \left\langle \hat{\mathsf{v}}-1, \frac{\mathsf{I}_0(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}{\mathsf{I}_{-1}(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}\right\rangle - \sum_{c\in\boldsymbol{c}}\left\langle \hat{\mathsf{V}}_c, \frac{\mathsf{I}_c(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}{\mathsf{I}_{-1}(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}\right\rangle\right)\\ &= \left[\mathrm{Arg}\min_{\hat{\mathsf{v}}\in\hat{\mathbf{v}}}\left(\ln(\mathsf{Be}(\hat{\mathsf{v}})) - \left\langle \hat{\mathsf{v}}-1, \frac{\mathsf{I}_0(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}{\mathsf{I}_{-1}(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}\right\rangle\right),\right.\\ &\left.\left(\mathrm{Arg}\min_{\hat{\mathsf{V}}\in\hat{\mathbf{V}}_c}\left(\ln(\mathsf{N}_c(\hat{\mathsf{V}}_c)) - \left\langle \hat{\mathsf{V}}_c, \frac{\mathsf{I}_c(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}{\mathsf{I}_{-1}(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}\right\rangle\right)\right)_{c\in\boldsymbol{c}}\right]. \end{aligned} \tag{21}$$

Thus, $|\boldsymbol{c}|+1$ independent minimisations are solved. They are numerically simple. Parts of them have analytical solutions for normal and Markov chain models.

Numerical Evaluation of I*'s:* Function $\mathsf{H}(\psi_t,\Theta)$ (19) depends on all involved parameters and combines influences of respective components. A brute force evaluation of I_{-1}, I_0 and I_c, $c\in\boldsymbol{c}$, say by Monte Carlo, is mostly too demanding in a typical on line learning. At the same time, the averaged functions in (19), (20) are by their construction smooth. Moreover, their growth over their domains is well suppressed by generally light tails of involved pds $\tilde{\mathsf{Q}}_{t;c}$ (18). Thus, we conjecture that the simplest approximation based on the first order Taylor expansion of $\mathsf{H}(\psi,\Theta)$ around expected values

$$\int_{\boldsymbol{\Theta}}\alpha_{\tilde{c}}\tilde{\mathsf{Q}}_{t;c}(\Theta)\mathrm{d}\Theta, \qquad \int_{\boldsymbol{\Theta}}\Theta_{\tilde{c}}\tilde{\mathsf{Q}}_{t;c}(\Theta)\mathrm{d}\Theta, \qquad \tilde{c},c\in\boldsymbol{c}, \tag{22}$$

suffices. In addition to (22), such an approximation requires to evaluate

$$\int_{\boldsymbol{\Theta}}\alpha\ln(\alpha_{\tilde{c}})\tilde{\mathsf{Q}}_{t;c}(\Theta)\mathrm{d}\Theta,\ \int_{\boldsymbol{\Theta}}\mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\tilde{\mathsf{Q}}_{t;c}(\Theta)\mathrm{d}\Theta,\ \int_{\boldsymbol{\Theta}}\Theta_{\tilde{c}}\mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\tilde{\mathsf{Q}}_{t;c}(\Theta)\mathrm{d}\Theta,\ \ \tilde{c},c\in\boldsymbol{c}. \tag{23}$$

Remarks 4

- ✓ *Normalising constants* Be, N *(13), (14) of factors forming* $\tilde{\mathsf{Q}}_{t;c}$ *(18) can be analytically expressed for Normal-inverse-Wishart and Dirichlet's pds, which are conjugated to normal and Markov chain components, Ruman (2018).*
- ✓ $\tilde{c}$*-th entry of expectations (22), (23) at time* t *coincide with those gained at time* $t-1$ *if* $\tilde{c}\neq c$*. This significantly reduces the computational load.*
- ✓ *The used approximation could be refined but often suffices. Its use allowed us to focus on the advocated model, on the novelty brought.*

Forgetting: It consists of the choice of the forgetting factor and its use. The formula for λ_t in (4) arisen as ratio of the predictive pd divided by the predictive pd constructed from the already updated posterior pd (both in data realisation). The studies of partial forgetting, Dedecius et al. (2009), imply that forgetting should be applied component-wise. The overall prediction quality then influences only

forgetting of component weights. The value of I_{-1} (19) is proportional to the value of the predictive pd. This gives forgetting factors λ_t, $(\lambda_{t;c})_{c\in\boldsymbol{c}}$, (4),

$$\lambda_t = \frac{\mathsf{Be}(\hat{\mathsf{v}}_t)\prod_{\tilde{c}\in\boldsymbol{c}}\mathsf{N}(\hat{\mathsf{V}}_{t;\tilde{c}})}{\mathsf{Be}(\mathsf{v}_{t-1})\prod_{\tilde{c}\in\boldsymbol{c}}\mathsf{N}(\mathsf{V}_{t-1;\tilde{c}})}\frac{\mathsf{I}_{-1}(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})}{\mathsf{I}_{-1}(\Psi_t,\hat{\mathsf{v}}_t,\hat{\mathsf{V}}_t)}$$
$$\lambda_{t;c} = \frac{\mathsf{N}(\mathsf{V}_{t-1;c}+\mathsf{B}_c(\Psi_{t;c}))\mathsf{N}(\hat{\mathsf{V}}_{t;c})}{\mathsf{N}(\mathsf{V}_{t-1;c})\mathsf{N}(\hat{\mathsf{V}}_{t;c}+\mathsf{B}_c(\Psi_{t;c}))}\frac{\mathsf{W}_c(\psi_t,\tilde{\Theta}_{t;c})}{\mathsf{W}_c(\psi_t,\tilde{\Theta}_{t-1;c})},\quad c\in\boldsymbol{c}.$$

There $\mathsf{I}_{-1}(\Psi_t,\mathsf{v}_{t-1},\mathsf{V}_{t-1})$ is approximately evaluated using expansion around (22). Denominator $\mathsf{I}_{-1}(\Psi_t,\hat{\mathsf{v}}_t,\hat{\mathsf{V}}_t)$ is computed according to the same formulae with $\hat{\mathsf{v}}_t,\hat{\mathsf{V}}_t$ (21) replacing $\mathsf{v}_{t-1},\mathsf{V}_{t-1}$. Consequently, $\tilde{\Theta}_{t-1;c}$, $\tilde{\Theta}_{t;c}$ are the expected values of component parameter Θ_c computed as the c-th part of (22) with component statistics $\mathsf{V}_{t-1;c}$ and $\hat{\mathsf{V}}_{t;c}$, respectively (11). The forgetting use completes updating (4) for proposed mixture ratio model (11), (12)

$$\mathsf{v}_t = \lambda_t\hat{\mathsf{v}}_t + (1-\lambda_t)\mathsf{v}_{t-1},\qquad \mathsf{V}_{t;c} = \lambda_{t;c}\hat{\mathsf{V}}_{t;c} + (1-\lambda_{t;c})\mathsf{V}_{t-1;c},\ t\in\boldsymbol{t},\ c\in\boldsymbol{c}.$$

5 Illustrative Examples

This section illustrates the modelling theory and the corresponding learning. A systematic simulation and real-life studies are out of scope of this paper and will be published independently.

The first example illustrates the modelling potential of mixture ratios. The second one shows that the standard mixture model is not good when the mixture ratio model is adequate while the mixture ratio is competitive even when the standard mixture model is adequate. Additional experiments, including real-data application to commodity futures, Carter (1999), are in Ruman (2018).

5.1 Mixture-Ratio Modelling Strength

Simulated environment model E, generating real scalar observations O_t with no actions and regression vector $\psi_t = O_{t-1}$, is the mixture of joint normal pds $\mathcal{N}_\Psi(\mu_c,\omega_c^{0.5})$ with expectations μ_c and square roots $\omega_c^{0.5}$ of precision matrices, $c\in\boldsymbol{c}=\{1,2\}$,

$$\mathsf{E}(O_t,O_{t-1}) = \tag{24}$$
$$\frac{1}{2}\mathcal{N}_{\underbrace{O_t,O_{t-1}}_{\Psi_t}}\left(\underbrace{\begin{bmatrix}1\\1\end{bmatrix}}_{\mu_1},\underbrace{\begin{bmatrix}1 & 3\\0 & 0.5\end{bmatrix}}_{\omega_1^{0.5}}\right)+\frac{1}{2}\mathcal{N}_{\underbrace{O_t,O_{t-1}}_{\Psi_t}}\left(\underbrace{\begin{bmatrix}-1\\-1\end{bmatrix}}_{\mu_2},\underbrace{\begin{bmatrix}1 & -2\\0 & 0.5\end{bmatrix}}_{\omega_2^{0.5}}\right).$$

The typical simulation results, differing only in initial values $O_0\in\{-0.8,0,3\}$, are shown in Fig. 1. It demonstrates the dynamic dependence of the components weight on the data realisation. This is the key feature of the truly dynamic model, which obviously allows to model non-linear dynamic effects and unbalanced activations of respective components. For comparison, the same mixture was simulated as the conditional one, i.e. with the constant component weights. It lacks both mentioned features of the mixture ratio model.

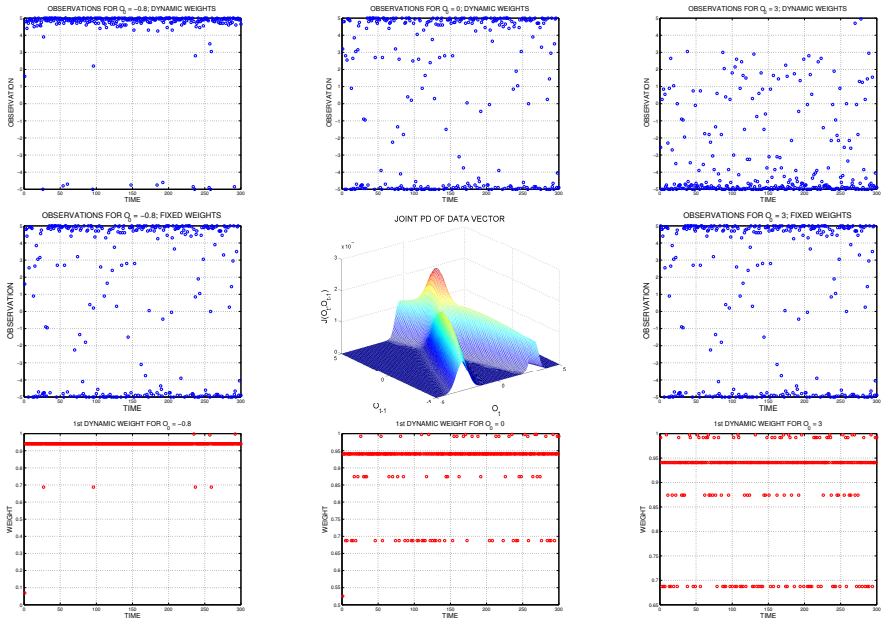

Fig. 1 Simulation of environment (24). Columns differ in the initial value of regression vector $\psi_1 = O_0 \in \{-0.8,\ 0,\ 3\}$. Pseudo-random realisations are the same in all cases. The 1st row shows the realised observations. The 2nd row shows realised observations when the same components as in (24) are used as conditional pds (constant component weights). The practically identical case for $O_0 = 0$ is replaced by the simulated joint pd. The 3rd row shows dynamic weights $(\mathsf{w}_{t;1}(\psi_t))_{t\in\boldsymbol{t}}$ (11) corresponding to the 1st-row.

5.2 Learning with Standard and Ratio Mixture Models

Simulation Conditions: Scalar observations $O_t \in \boldsymbol{O} = \{1, \ldots, |\boldsymbol{O}|\}$, $|\boldsymbol{O}| = 5$, and 2nd order regression vectors $\psi_t = (O_{t-1}, O_{t-2})$ are considered. This defines data vectors (6) $\boldsymbol{\Psi}_t = (O_t, O_{t-1}, O_{t-2})$. Learnt models L and simulated environment E are either mixture ratio model M or standard mixture S with two components $|\boldsymbol{c}| = 2$

$$\mathsf{E}, \mathsf{L} \in \boldsymbol{\mathsf{L}} = \{\mathsf{M}, \mathsf{S}\} = \{\text{mixture ratio}, \text{standard mixture}\}. \tag{25}$$

The structure of joint $\mathsf{J}(\boldsymbol{\Psi}_t|\Theta)$ pd defining the mixture ratio model is

$$\mathsf{J}(\boldsymbol{\Psi}_t|\Theta_{\mathsf{M}}) = \alpha_{\mathsf{M}}\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M}1})\mathsf{G}_1(O_{t-2}) + (1-\alpha_{\mathsf{M}})\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M}2})\mathsf{G}_2(O_{t-1}).$$

Joint pds $\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M}1})$ and $\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M}2})$ are parameterised by their values $\Theta_{\mathsf{M}1}$, $\Theta_{\mathsf{M}2}$ of possible pairs (O_t, O_{t-1}), (O_t, O_{t-2}), and $\Theta_{\mathsf{M}} = (\alpha_{\mathsf{M}}, \Theta_{\mathsf{M}1}, \Theta_{\mathsf{M}2})$. $\mathsf{G}_1(O_{t-2})$ and $\mathsf{G}_2(O_{t-2})$ are uniform pds on $\boldsymbol{O}$. This gives mixture ratio M

$$\mathsf{M}(O_t|\psi_t, \Theta_{\mathsf{M}}) = \frac{\alpha_{\mathsf{M}}\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M}1}) + (1-\alpha_{\mathsf{M}})\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M}2})}{\sum_{O_t\in\boldsymbol{O}} \alpha_{\mathsf{M}}\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M}1}) + (1-\alpha_{\mathsf{M}})\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M}2})}.$$

Used standard mixture model S is parameterised by $\Theta_{\mathsf{S}} = (\alpha_{\mathsf{S}}, \Theta_{\mathsf{S}1}, \Theta_{\mathsf{S}2})$

$$\mathsf{S}(O_t|\psi_t, \Theta_{\mathsf{S}}) = \alpha_{\mathsf{S}}\mathsf{S}_1(O_t|O_{t-1}, \Theta_{\mathsf{S}1}) + (1-\alpha_{\mathsf{S}})\mathsf{S}_2(O_t|O_{t-2}, \Theta_{\mathsf{S}2}),$$

where $\mathsf{S}_1(O_t|O_{t-1}, \Theta_{\mathsf{S}1})$ and $\mathsf{S}_2(O_t|O_{t-2}, \Theta_{\mathsf{S}2})$ are components parameterised by probabilities $(\Theta_{\mathsf{S}1}, \Theta_{\mathsf{S}2})$ of O_t conditioned on O_{t-1} and O_{t-2}, respectively.

Monte Carlo Study: It consists of 200 runs for time steps $t \in \boldsymbol{t}$, $|\boldsymbol{t}| = 500$, with randomly generated parameters Θ_{E} for both model structures $\mathsf{E} \in \mathbf{L} = \{\mathsf{M}, \mathsf{S}\}$ (25). In both cases, the mixture ratio and standard mixture models are sequentially learnt on the identical observations. An individual study is also considered with a fixed data-generating mixture ratio model given by parameter $\underline{\Theta}_{\mathsf{M}}$ with fixed $\underline{\Theta}_{\mathsf{M}}$ causing truly dynamic weights $\mathsf{w}_c(\psi_t)$ (11). These 200 runs differ in realisations.

Evaluation: At time $t \in \boldsymbol{t}$, the simulation and the sequential learning dealt with the next *predictors*, with an abused notation pointing to the underlying parametric models,

$$\mathsf{E}_t(O_t) = \begin{cases} \underline{\mathsf{M}}_t(O_t) = \mathsf{M}(O_t|\psi_t, \underline{\Theta}_{\mathsf{M}}) \text{ if } \mathsf{E} = \mathsf{M} \text{ with fixed parameter } \underline{\Theta}_{\mathsf{M}} \\ \underline{\mathsf{S}}_t(O_t) = \mathsf{S}(O_t|\psi_t, \underline{\Theta}_{\mathsf{S}}) \quad \text{if } \mathsf{E} = \mathsf{S} \quad \text{with fixed parameter } \underline{\Theta}_{\mathsf{S}} \end{cases} \tag{26}$$

$$\mathsf{L}_t(O_t) = \mathsf{L}_t(O_t|O_{t-1}, \ldots, O_0) = \int_{\boldsymbol{\Theta}_{\mathsf{L}}} \mathsf{L}(O_t|\psi_t, \Theta_{\mathsf{L}})\mathsf{P}(\Theta_{\mathsf{L}}|O_{t-1}, \ldots, O_0)\mathrm{d}\Theta_{\mathsf{L}}, \ \mathsf{L} \in \mathbf{L}.$$

The sequentially evaluated Kullback-Leibler divergences of environment E and learnt predictors (26)

$$\mathsf{D}_t(\mathsf{E}||\mathsf{L}) = \sum_{\tau=1}^{t} \mathsf{D}(\mathsf{E}_\tau||\mathsf{L}_\tau), \quad \text{with} \quad \mathsf{D}(\mathsf{E}_\tau||\mathsf{L}_\tau) = \sum_{O_\tau \in \boldsymbol{O}} \mathsf{E}_\tau(O_\tau) \ln\left(\frac{\mathsf{E}_\tau(O_\tau)}{\mathsf{L}_\tau(O_t)}\right)$$
$$\mathsf{D}(\mathsf{E}||\mathsf{L}) = \mathsf{D}_{|\boldsymbol{t}|}(\mathsf{E}||\mathsf{L}), \quad \text{and their differences} \quad \Delta = \mathsf{D}(\mathsf{E}||\mathsf{M}) - \mathsf{D}(\mathsf{E}||\mathsf{S}) \tag{27}$$

quantify learning quality in the comparable observation space.

Results: Results of Monte Carlo simulations are in Fig. 2. The left hand column contains histogram values $\mathsf{D}(\mathsf{E}||\mathsf{L}) = \mathsf{D}_{|\boldsymbol{t}|}(\mathsf{E}||\mathsf{L})$, $\mathsf{L} \in \mathbf{L}$, at final time $|\boldsymbol{t}| = 500$. The right hand column shows the corresponding sample cumulative distributions. The better predictors have them shifted to the left. Fig. 3 shows course of $\mathsf{D}_t(\mathsf{E}||\mathsf{L})$, $t \in \boldsymbol{t}$, $\mathsf{L} \in \mathbf{L} = \{\mathsf{M}, \mathsf{S}\}$ for: (a) $\mathsf{E} = \underline{\mathsf{M}}$, given by fixed $\underline{\Theta}_{\mathsf{M}}$; (b) $\mathsf{E} = \underline{\mathsf{S}}$, given by fixed $\underline{\Theta}_{\mathsf{S}}$; (c) $\mathsf{E} = \underline{\mathsf{M}}$ with $\underline{\Theta}_{\mathsf{M}}$ giving dynamic weights $\mathsf{w}_c(\psi_t)$ (11).

Table 1 shows sample statistics of increments Δ (27). Negative values mean the dominance of the mixture ratio model.

Discussion: Fig. s 2, 3 and Table 1 confirm the expectation that the mixture ratio outperforms the standard mixture if the mixture ratio is simulated. It leads to essentially same results when the standard mixture serves as the simulated environment. It is even slightly better in the latter case as it (probably) copes better with errors caused by approximate learning (4).

Table 1 Sample statistics of Δ (27): the 1st column $\mathsf{E} = \underline{\mathsf{M}}$ with $\underline{\Theta}_{\mathsf{M}}$ varied in 200 Monte Carlo runs; the 2nd column $\mathsf{E} = \underline{\mathsf{S}}$ with $\underline{\Theta}_{\mathsf{S}}$ varied in 200 Monte Carlo runs; the 3rd column $\mathsf{E} = \underline{\mathsf{M}}$ with fixed $\underline{\Theta}_{\mathsf{M}}$ causing truly dynamic weights $\mathsf{w}_c(\psi_t)$ (11).

Simulated Case	Monte Carlo $\mathsf{E} = \underline{\mathsf{M}}$	Monte Carlo $\mathsf{E} = \underline{\mathsf{S}}$	Fixed $\mathsf{E} = \underline{\mathsf{M}}$
Mean	-0.7001	-0.5729	-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

6 Potential of the Mixture Ratio Model

The paper brings an important message that the ratios of finite mixtures model well dynamic relations. The main promising features are as follows.

- ✓ The mixture ratios universally describe non-linear dynamic stochastic relations.
- ✓ The mixture ratios may serve as a relatively universal dynamic feature extractor. Indeed, the approximately sufficient statistic collected during learning, see Sec. 4, are the relevant features. Thus, it suffices to apply Bayesian structure estimation to mixture ratios with low-dimensional components.
- ✓ The mixture ratios handle cases in which rare visits of active components are significant, cf. Fig. 1. This is vital in fault detection, Polycarpou and Helmicki (1995), detection of non-standard fraud behaviours needed, Kou et al. (2004), or in cyber-security applications Buczak and Guven (2016), generally everywhere where outlier detection is faced, Hodge and Austin (2004). These cases are hard dynamic versions of learning with unbalanced data, Bekkar and Alitouche (2013).
- ✓ The mixture ratios learns the stationary joint pd of the data vector (10), which can directly serve for the design of adaptive decision strategies for an infinite decision horizon, Kushner (1971). Indeed, the sequentially updated joint pd of data vector can be appropriately factorised and the current model of the stationary decision rule replaced by the rule, which makes the joint pd close to a desired stationary joint pd. This fits well to fully probabilistic design of decision strategies, Kárný and Kroupa (2012).
- ✓ The mixture ratios suit to modelling of mixed (discrete and continuous valued) data as the components are joint pds, which factorise into product of pds describing continuous valued data, possibly conditioned on discrete ones, and pds relating discrete values.

The paper indicates tractability all these cases, but it is necessary:

- ✓ to elaborate numerically robust (factorised) procedures for normal models, sparse Markov-chains and mixed cases in the way mimic to Kárný (2016); Ruman (2018) has made definite steps in this respect;
- ✓ to refine the data-dependent choice of forgetting factors;
- ✓ to tailor Bayesian structure estimation, ready for mixtures, Kárný et al. (2006), to the mixture-ratio model;
- ✓ to perform real-life tests confirming that the theoretically higher approximating strength of the discussed mixture ratios is mostly undiminished by the

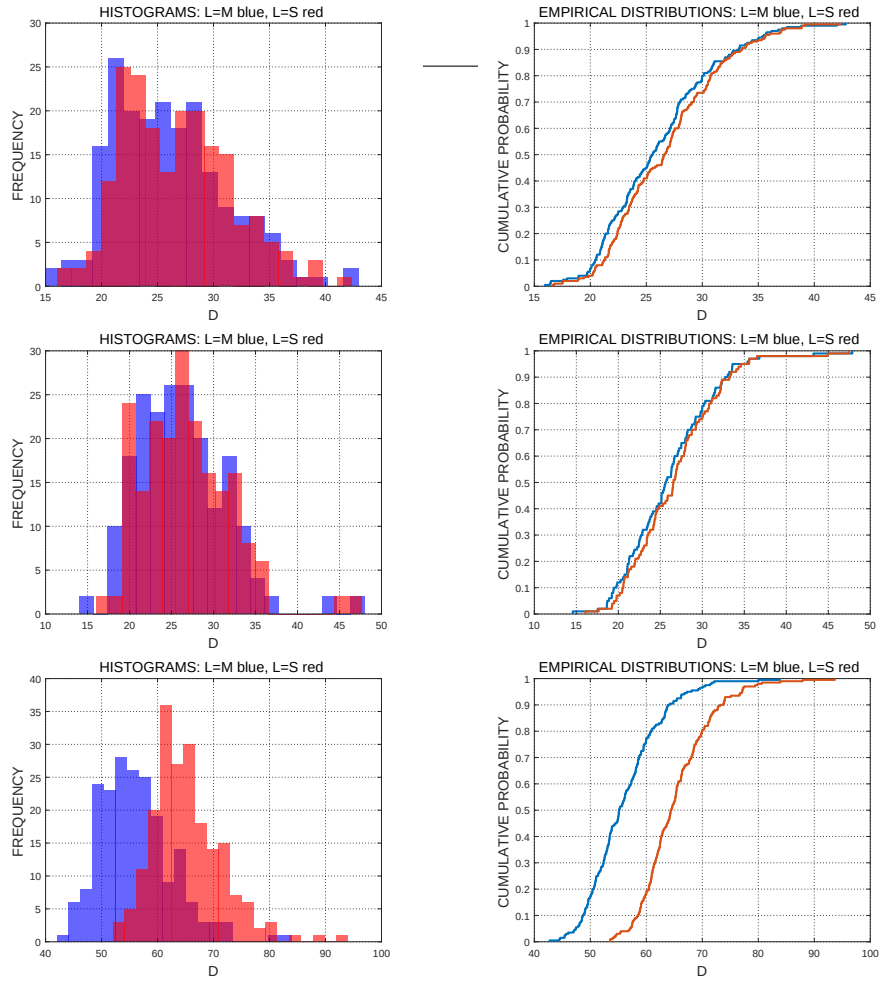

Fig. 2 Histograms of $\mathsf{D}(\mathsf{E}||\mathsf{L})$ ((27), left) and their empirical distribution function of (right), $\mathsf{L} \in \{\mathsf{M}, \mathsf{S}\} = \{\text{blue}, \text{red}\}$. The 1st row $\mathsf{E} = \underline{\mathsf{M}}$ with $\underline{\Theta}_{\mathsf{M}}$ varied in 200 Monte Carlo runs; the 2nd row $\mathsf{E} = \underline{\mathsf{S}}$ with $\underline{\Theta}_{\mathsf{S}}$ varied in 200 Monte Carlo runs; the 3rd row for $\mathsf{E} = \underline{\mathsf{M}}$ with fixed $\underline{\Theta}_{\mathsf{M}}$ causing truly dynamic weights $\mathsf{w}_c(\psi_t)$ (11).

approximate learning; the experiments with commodity futures trading made in Ruman (2018), are more than promising.

The theoretical insight, experience from simulations and preliminary processing of real data, width and importance of possible contributions to machine learning and DM applications make this effort worthwhile.

Acknowledgement

This research is supported by GAČR GA16-09848S.

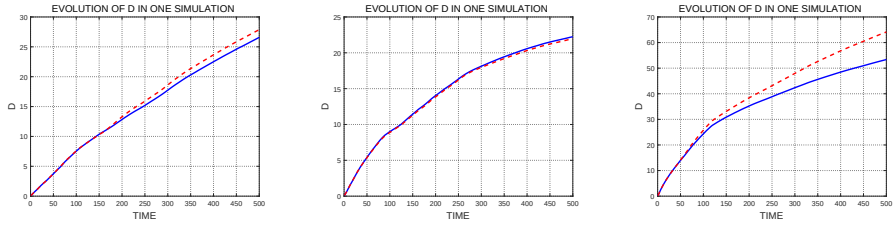

Fig. 3 Time evolutions of $\mathsf{D}_t(\mathsf{E}||\mathsf{L})$ (27), $\mathsf{L} \in \{\mathsf{M}, \mathsf{S}\} = \{\text{blue}, \text{red}\}$. Left to right: $\mathsf{E} = \underline{\mathsf{M}}$ with randomly chosen $\underline{\Theta}_{\mathsf{M}}$; $\mathsf{E} = \underline{\mathsf{S}}$ with randomly chosen $\underline{\Theta}_{\mathsf{S}}$; $\mathsf{E} = \underline{\mathsf{M}}$ with with fixed $\underline{\Theta}_{\mathsf{M}}$ causing truly dynamic weights $\mathsf{w}_c(\psi_t)$ (11).

References

Appice A, A, Ciampi, Malerba D (2015) Summarizing numeric spatial data streams by trend cluster discovery. Data Mining and Knowledge Discovery 29(1):84–136

Åström K, Wittenmark B (1989) Adaptive Control. Addison-Wesley

Barndorff-Nielsen O (1978) Information and Exponential Families in Statistical Theory. Wiley, NY

Bekkar M, Alitouche T (2013) Imbalanced data learning approaches review. Int J of Data Mining & Knowledge Management Process 3(4):15–33

Berec L, Kárný M (1997) Identification of reality in Bayesian context. In: Warwick K, Kárný M (eds) Computer-Intensive Methods in Control and Signal Processing, Birkhäuser, pp 181–193

Berger J (1985) Statistical Decision Theory and Bayesian Analysis. Springer, NY

Bernardo J (1979) Expected information as expected utility. The An of Stat 7(3):686–690

Bishop C (2006) Pattern Recognition and Machine Learning. Springer

Bohlin T (1991) Interactive System Identification: Prospects and Pitfalls. Springer

Buczak A, Guven E (2016) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys Tutorials 18(2):1153–1176

Carter C (1999) Commodity future markets: A survey. The Australian Journal of Agricultural and Resource Economics 43:209–247

Catania L (2016) Dynamic adaptive mixture models. ArXiv:1603.01308v1

Dedecius K, Nagy I, Kárný M, Pavelková L (2009) Parameter estimation with partial forgetting method. In: Proc. of the 15th IFAC SYSID

Elliot R, Assoun L, Moore J (1995) Hidden Markov Models. Springer-Verlag, NY

Frigessi A, Haug O, Rue H (2002) A dynamic mixture model for unsupervised tail estimation without threshold selection. Extremes 5(3):219 – 235

Guan P, Raginsky M, Willett R (2012) Online Markov decision processes with Kullback-Leibler control cost. In: Am. Control Conference, IEEE, pp 1388–1393

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

Haykin S (1994) Neural Networks: A Comprehensive Foundation. Macmillan, NY

Hodge V, Austin J (2004) A survey of outlier detection methodologies. Artificial Intelligence Review 22(2):85–126

Holmes G, Liu T (eds) (2015) Proceedings of 7th Asian Conference on Machine Learning (ACML2015), JMLR Workshop and Conference Proceedings, vol 45
Jazwinski A (1970) Stochastic Processes and Filtering Theory. Ac. Press, NY
Kárný M (2014) Approximate Bayesian recursive estimation. Infor Sciences 289:100–111
Kárný M (2016) Recursive estimation of high-order Markov chains: Approximation by finite mixtures. Infor Sciences 326:188–201
Kárný M, Kroupa T (2012) Axiomatisation of fully probabilistic design. Information Sciences 186(1):105–113
Kárný M, Böhm J, Guy TV, Jirsa L, Nagy I, Nedoma P, Tesař L (2006) Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer
Kárný M, Macek K, Guy T (2014) Lazy fully probabilistic design of decision strategies. In: 11th Int. Symp. on Neural Net., no. 8866 in LNCS, pp 140–149
Kerridge D (1961) Inaccuracy and inference. J of the Royal Stat Society B 23:284–294
Koopman R (1936) On distributions admitting a sufficient statistic. Trans of Am Math Society 39:399
Kou Y, Lu C, Sirwongwattana S, Huang Y (2004) Survey of fraud detection techniques. In: IEEE International Conference on Networking, Sensing and Control, 2004, vol 2, pp 749–754 Vol.2
Kulhavý R (1990) A Bayes-closed approximation of recursive nonlinear estimation. Int J Adaptive Control and Signal Proc 4:271–285
Kullback S, Leibler R (1951) On information and sufficiency. Annals of Mathematical Statistics 22:79–87
Kushner H (1971) Introduction to Stochastic Control. Holt, Rinehart and Winston
Mason W, Vaughan J, Wallach H (2014) Special issue: Computational social science and social computing. Machine Learning 96:257–469
McLachlan G, Peel D (2000) Finite Mixture Models. Wiley Series in Probab. & Stat., Wiley, NY
McNicholas P (2017) Mixture model-based classification. CRC Press, Boca Raton, London, New York
Mirsky Y, Shapira B, Rokach L, Elovici Y (2015) pcstream: A stream clustering algorithm for dynamically detecting and managing temporal contexts. In: Cao T, et al (eds) Advances in Knowledge Discovery and Data Mining, vol 9078, LNCS Springer, Cham, pp 230–237
Mitchell T (1997) Machine Learning. McGraw Hill
Nelsen R (1999) An Introduction to Copulas. Springer, NY
Nguyen H, Woon Y, Ng W (2015) A survey on data stream clustering and classification. Knowledge and Information Systems 45:535–569, DOI https://doi.org/10.1007/s1011
Park J, Sandberg I (1991) Universal approximation using radial-basis-function networks. Neural Computations 3:246–257
Peterka V (1981) Bayesian system identification. In: Eykhoff P (ed) Trends and Progress in System Identification, Pergamon Press, Oxford, pp 239–304
Polycarpou M, Helmicki A (1995) Automated fault detection and accommodation: a learning systems approach. IEEE Transactions on Systems, Man, and Cybernetics 25(11):1447–1458
Rao M (1987) Measure Theory and Integration. John Wiley, NY

Ruman F (2018) Mixture ratios for decision making. Master's thesis, FJFI, Czech Technical University, Prague

Sadghi P, Kennedy R, Rapajic P, Shams R (2008) Finite-state Markov modeling of fading channels. IEEE Signal Processing Magazine 57, DOI 10.1109/MSP.2008.926683

Savage L (1954) Foundations of Statistics. Wiley, NY

Silva J, Faria E, Barros R, Hruschka E, Carvalho A, Gama J (2014) Data stream clustering: A survey. ACM Computing Surveys 46, DOI 10.1145/2522968.2522981

Tipping M, Bishop C (1999) Probabilistic principal component analysis. J of the Royal Society Series B – Stat Methodology 61:611–622

Tsai C, Lai C, Chiang M, Yang L (2014) Data mining for internet of things: A survey. IEEE Communications Surveys & Tutorials 16(1):77–95

Wald A (1950) Statistical Decision Functions. John Wiley, NY, London

Wiering M, van Otterlo M (eds) (2012) Reinforcement Learning: State-of-the-Art. Springer