Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Mixture Ratio Modelling of Dynamic Systems

Miroslav Kárný, Marko Ruman

The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic, http://www.utia.cas.cz/AS
Email: school@utia.cas.cz, marko.ruman@gmail.com
Correspondence:
Miroslav Kárný, school@utia.cas.cz

Summary

Any knowledge extraction relies (possibly implicitly) on a hypothesis about the modelled-data dependence. The extracted knowledge ultimately serves to a decision making (DM). DM always faces uncertainty and this makes probabilistic modelling adequate. The inspected black-box modelling deals with "universal" approximators of the relevant probabilistic model. Finite mixtures with components in the exponential family are often exploited. Their attractiveness stems from their flexibility, the cluster interpretability of components and the existence of algorithms for processing high-dimensional data streams. They are even used in dynamic cases with mutually dependent data records while regression and auto-regression mixture components serve to the dependence modelling. These dynamic models, however, mostly assume data-independent component weights, i.e. memoryless transitions between dynamic mixture components. Such *mixtures are not universal approximators* of dynamic probabilistic models. Formally, this follows from the fact that the set of finite probabilistic mixtures is not closed with respect to the conditioning, which is the key estimation and predictive operation. The paper overcomes this drawback by using *ratios of finite mixtures* as universally approximating dynamic parametric models. The paper motivates them, elaborates their approximate Bayesian recursive estimation and reveals their application potential.

KEYWORDS:
universal approximation, black-box dynamic model, mixture model, approximate Bayesian estimation, data stream processing, forgetting, Kullback-Leibler divergence

1 | INTRODUCTION

The paper serves to *dynamic* decision making (DM) understood as a targeted choice among available actions (options) influencing dynamically evolving system. DM seen in this way strongly overlaps with machine learning [1], signal processing, [2], hypothesis testing [3], classification [4], knowledge sharing [5], reinforcement learning [6], control [7], adaptive control [?], etc. All these areas benefit from an enrichment of the domain knowledge by that hidden in recorded data. This "data mining" is an extremely broad domain with many results and applications, e.g. [8]. Even surveys specialise, for instance, to Internet of Things [9], cyber-security [10], etc.

Paper Focus: The paper contributes to clustering of data streams [11] with dynamically bonded data records [12]. Importance of such a processing is well seen on specific applications, for instance in geophysics [13] or on a stream clustering of independent data records [14]. Our paper provides universal approximators[1] of dynamic models and their recursive estimation.

[1] We adopt this intuitive notion used in connection with neural networks [15]. Formally, it means that universal approximators form a dense subset of a set of functions they approximate. Their rigorous specification in various sets of approximated functions are, for instance, in [16,17,18,19]. They contain a range of other common references.

DM View on Data Stream Processing: A solution of a DM task leads to a strategy, i.e. a collection of decision rules mapping the evolving knowledge on actions[20]. A good prescriptive DM theory should provide a strategy, which meets user's DM aims in the best possible way. The adopted Bayesian DM[21] is such a theory coping with incomplete knowledge, uncertainty and randomness of the system to which actions relate. Bayesian DM links DM consequences to the acquired knowledge and actions by conditional distributions. Here, conditional probability densities (pd[2]) describe them.

Actions are chosen sequentially. This enables to operate on gradually enriched knowledge, to learn better the system model in a single pass only. Adaptive control[?,23] fully relies on this feature. Bayesian data stream processing updates a statistic, ideally a sufficient one. Thus, it may serve as a knowledge, feature or cluster extractor suitable to all areas mentioned above.

Addressed Modelling Problem: An estimation is possible iff the inspected relations hold during the knowledge accumulation. The most general case relies on invariant mappings describing stochastic state-space model and leads to stochastic filtering[24]. This paper deals with a simpler case[3] and uses parametric, black-box, probabilistic models relating a few adjacent data records. It is a priori unknown, which parameter points to the best model. Bayesian paradigm offers the unambiguous deductive estimation via Bayes' rule[26]. It accumulates the knowledge into the posterior pd and redistributes this belief about the model adequacy[27].

The modelling and thus DM quality depend on the model set over which the belief redistributes. *For a fixed condition*, any smooth pd, describing data-records dependence, can be arbitrarily-well approximated by a finite mixture of conditional pds (components)[18]. Components may have the same functional form, say Gaussian one, but differ in their parameters. Various versions of this ability to approximate "universally" were proved in neural-network context[28]. Zero-memory models with a fixed trivial condition are universally approximable by mixtures with constant component weights[29]. In the inspected dynamic case with a non-trivial condition, the component weights should depend on it, too. The common use of the condition-independent weights of dynamic components goes against the expansion logic and surely violate universal approximation property. Rare exceptions that considered condition-dependent weight[30,31,32] indicate how much is lost when using constant component weights.

The paper offers *ratios* of finite mixtures with components in the exponential family (EF)[33] as black-box,[34], universally approximating, dynamic models[15]. It develops, inevitably approximate, but feasible Bayesian recursive estimation of models from this *extremely rich but yet unconsidered model set*. Note that the thesis[35] and the conference paper[36] consider it but they are direct predecessors of this paper.

Layout: The paper relies on an approximate Bayesian recursive estimation[37]. Its recall in Sec. 2 makes the paper self-containing and prepares notations. Sec. 3 justifies the mixture ratio models and Sec. 4 elaborates their estimation. Sec. 5 illustrates the theory. Sec. 6 outlines the gained application potential.

Common Notation: The bold symbol $\boldsymbol{X}$ denotes a set of Xs. It is a subset of either finite-dimensional real space or it consists of pds. Its exact specification is only given when needed. $|\boldsymbol{X}|$ marks cardinality of $\boldsymbol{X}$. Random variables, their realisations and function arguments are undistinguished. The meaning follows from the context. Mappings, marked by san serif fonts, are taken as different if labels of their arguments differ. Mnemonic symbols prevail: M means a parametric model, J is a joint parametric pd of data vectors, P is posterior pd, O marks *o*bservations, A labels *a*ctions, etc. Decoration $\tilde{}$ marks intermediate objects, $\hat{}$ marks estimates and approximations. $\equiv$ means defining equality.

2 | APPROXIMATE BAYESIAN RECURSIVE ESTIMATION

This section summarises the approximate recursive estimation published in[37]. It serves us for handling of the novel model proposed in Sec. 3.

The inspected parametric model M consists of the conditional pds at discrete-time moments labelled by $t \in \boldsymbol{t} = \{1, 2, \dots, |\boldsymbol{t}|\}$

$$\mathsf{M}(O_t|A_t, D^{t-1}, \Theta), \quad D_t \equiv (O_t, A_t), \quad D^{t-1} \equiv (D_{t-1}, \dots, D_1, D_0). \tag{1}$$

The pd (1) relates the observation $O_t \in \boldsymbol{O}$ to the action $A_t \in \boldsymbol{A}$ and to the past data records D^{t-1} extended by a prior knowledge D_0. A time-invariant $\Theta \in \boldsymbol{\Theta}$ parameterises these pds. The employed decision rules modelled by pds $\mathsf{R}(A_t|D^{t-1}, \Theta)$ are unaware of the adequate Θ. Thus, these decision rules meet natural conditions of control[38]

$$\mathsf{R}(A_t|D^{t-1}, \Theta) = \mathsf{R}(A_t|D^{t-1}). \tag{2}$$

[2]Pd is a common abbreviation covering the usual ones for probability density and mass functions (pdf and pmf). Pds are evaluated with respect to a dominating, typically Lebesque's or counting, product measure[22]. Lebesque's notation $\int \dots \mathrm{d}\bullet$ is mostly used here.

[3]The results also suit to stochastic state-space models, aka hidden Markov models[25]. The addressed universal approximation by mixture ratio, which is closed with respect to conditioning, is applicable to them. The more complex notation and inevitable technical details related to them would, however, mask the basic modelling idea.

The posterior pd $\mathsf{P}_{t-1}(\Theta) \equiv \mathsf{P}(\Theta|A_t, D^{t-1}) \overbrace{=}^{(2)} \mathsf{P}(\Theta|D^{t-1})$ quantifies the knowledge about the parameter $\Theta \in \boldsymbol{\Theta}$ available at time $t-1$. Bayes' rule[26] updates this knowledge by the data-record realisation $D_t \equiv (O_t, A_t)$ to the posterior pd

$$\begin{aligned}\tilde{\mathsf{P}}_t(\Theta) &\equiv \frac{\mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta)}{\mathsf{F}(O_t|A_t, D^{t-1})} \propto \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta)\\ \mathsf{F}(O_t|A_t, D^{t-1}) &\equiv \int_{\boldsymbol{\Theta}} \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta)\mathrm{d}\Theta,\end{aligned} \tag{3}$$

where the value of the forecasting pd F is unexpressed when using the proportionality $\propto$. The decision rule R cancels due to (2).

With a 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[4] pds, has to be evolved. Mostly, the posterior pd $\tilde{\mathsf{P}}_t \notin \mathbf{P}$ even if the pd $\mathsf{P}_{t-1} \in \mathbf{P}$. To stay feasible, the pd $\tilde{\mathsf{P}}_t$ in (3) is to be projected on $\mathbf{P}$. Works[39,40] provide weak, generically met, conditions under which the minimiser $\hat{\mathsf{P}}_t \in \mathbf{P}$ of Kerridge's inaccuracy[41]

$$\mathsf{K}(\tilde{\mathsf{P}}_t||\hat{\mathsf{P}}_t) \equiv -\int_{\boldsymbol{\Theta}} \tilde{\mathsf{P}}_t(\Theta) \ln\left(\hat{\mathsf{P}}_t(\Theta)\right)\mathrm{d}\Theta \tag{4}$$

is the adequate Bayesian projection (minimising an adequate expected utility) of the pd $\tilde{\mathsf{P}}_t$ on $\hat{\mathsf{P}}_t \in \mathbf{P}$.

The projection *should not* serve as the prior pd in a further updating. Its use could cause a divergence of the projected pds from those projected optimally in the batch mode[42]. Forgetting, with a data-dependent factor $\lambda_t \in [0,1]$, is the adequate countermeasure[37]. An application of the minimum *expected* Kullback-Leibler principle[43] shows it. It recommends to select

$$\mathsf{P}_t = \arg\min_{\mathsf{P}\in\mathbf{P}}[\lambda_t \mathsf{D}(\mathsf{P}||\hat{\mathsf{P}}_t) + (1-\lambda_t)\mathsf{D}(\mathsf{P}||\mathsf{P}_{t-1})] \propto \hat{\mathsf{P}}_t^{\lambda_t}\mathsf{P}_{t-1}^{1-\lambda_t}. \tag{5}$$

The Kullback-Leibler divergence[44] $\mathsf{D}(\mathsf{P}||\hat{\mathsf{P}})$

$$\mathsf{D}(\mathsf{P}||\hat{\mathsf{P}}) \equiv \int_{\boldsymbol{\Theta}} \mathsf{P}(\Theta)\ln\left(\frac{\mathsf{P}(\Theta)}{\hat{\mathsf{P}}(\Theta)}\right)\mathrm{d}\Theta = \int_{\boldsymbol{\Theta}} \mathsf{P}(\Theta)\ln\left(\mathsf{P}(\Theta)\right)\mathrm{d}\Theta + \mathsf{K}(\mathsf{P}||\hat{\mathsf{P}}) \tag{6}$$

is a shifted version of Kerridge's inaccuracy (4). The weight $\lambda_t \in [0,1]$ is the belief that $\hat{\mathsf{P}}_t$ is the best prior guess of the unknown P_t and $(1-\lambda_t)$ is the belief into P_{t-1}. The beliefs are "naturally" proportional to forecasting-pds values

$$\hat{\mathsf{F}}(O_t|A_t, D^{t-1}) \equiv \int_{\boldsymbol{\Theta}} \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\hat{\mathsf{P}}_t(\Theta)\mathrm{d}\Theta \text{ and } \mathsf{F}(O_t|A_t, D^{t-1}), \text{ cf. (3).} \tag{7}$$

A priori there are no reasons to expect $\hat{\mathsf{P}}_t$ be better than P_{t-1} as the knowledge innovation brought by D_t may be spoiled by the projection of $\tilde{\mathsf{P}}_t$ on $\hat{\mathsf{P}}_t \in \mathbf{P}$. This motivates the used equal prior belief into $\hat{\mathsf{P}}_t$ and P_{t-1}.

Bayes' rule, the projection via Kerridge inaccuracy, minimum expected Kullback-Leibler principle and Bayes' rule updating beliefs into $\hat{\mathsf{P}}_t$, P_{t-1} provide the updating of a feasible posterior pd P_{t-1} by $D_t = (O_t, A_t)$

$$\begin{aligned}&\tilde{\mathsf{P}}_t(\Theta) \propto \mathsf{M}(O_t|A_t, D^{t-1}, \Theta)\mathsf{P}_{t-1}(\Theta), \quad \hat{\mathsf{P}}_t \equiv \arg\min_{\hat{\mathsf{P}}\in\mathbf{P}} \mathsf{K}(\tilde{\mathsf{P}}_t||\hat{\mathsf{P}})\\ &\mathsf{P}_t(\Theta) \propto \hat{\mathsf{P}}_t^{\lambda_t}(\Theta)\mathsf{P}_{t-1}^{1-\lambda_t}(\Theta), \quad \lambda_t \equiv \left[1 + \mathsf{F}(O_t|A_t, D^{t-1})/\hat{\mathsf{F}}(O_t|A_t, D^{t-1})\right]^{-1}, \text{ cf. (3),(7).}\end{aligned} \tag{8}$$

Remarks

✓ It is worth stressing why the optimisation (4) runs over the second argument while (5) over the first argument. It corresponds with two conceptually different tasks. In the first one, an *approximation of a known* pd $\tilde{\mathsf{P}}_t$ is constructed. The complementary analyses in[39] and[40] recommend both the optimised functional and this order of arguments. The optimisation (5) *completes a partial knowledge about an unknown* pd P_t. Works[45,40,43] axiomatically justify that this completion should be done via minimum (expected) Kullback-Leibler principle that optimises over the first argument.

Note the referred works clarify the relation of the variational Bayes[46] and the expectation propagation[47] techniques.

✓ The employed log-convex $\mathbf{P}$ guarantees that P_t stays in $\mathbf{P}$. Otherwise, an extra projection has to follow the forgetting.

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

✓ Kerridge's inaccuracy K (4) has the same minimiser as Kulback-Leibler divergence but copes with pds $\tilde{\mathsf{P}}$ having Dirac's constituents.

✓ The unavoidable projection of the pd $\tilde{\mathsf{P}}_t$ on the set $\mathbf{P}$ (8) is demanding but feasible, Sec. 3. The extra effort for evaluating λ_t is relatively small.

✓ The next implicit choice of the forgetting factor is possible and meaningful

$$\lambda_t = \left[1 + \mathsf{F}(O_t|A_t, D^{t-1})/\mathsf{F}_{\lambda_t}(O_t|A_t, D^{t-1})\right]^{-1}, \quad \mathsf{F}_{\lambda_t}(O_t|A_t, D^{t-1}) \equiv \int_{\mathbf{\Theta}} \mathsf{M}(O_t|A_t, D^{t-1}, \Theta) \frac{\hat{\mathsf{P}}_t^{\lambda_t}(\Theta)\mathsf{P}_{t-1}^{1-\lambda_t}(\Theta)}{\int_{\mathbf{\Theta}} \hat{\mathsf{P}}_t^{\lambda_t}(\tilde{\Theta})\mathsf{P}_{t-1}^{1-\lambda_t}(\tilde{\Theta})\mathrm{d}\tilde{\Theta}} \mathrm{d}\Theta.$$

It is here unused to avoid the related high computational costs.

3 | RATIO OF FINITE MIXTURES

This core section justifies a universal parametrisation of pds modelling dependent data records. They may be observed in a closed loop formed by a dynamic system and a randomised strategy that meets (2).

Markovian Modelling of a Joint PD: A joint parametric pd $\tilde{\mathsf{J}}(D^{|t|}|\Theta, D_0)$ of data-records sequences $D^{|t|} = (D_t)_{t\in \boldsymbol{t}}$ factorises[38]

$$\tilde{\mathsf{J}}(D^{|t|}|\Theta, D_0) \overset{(2)}{=} \prod_{t\in \boldsymbol{t}} \overbrace{\mathsf{M}(O_t|A_t, D^{t-1}, \Theta)}^{\text{parametric model}} \overbrace{\mathsf{R}(A_t|D^{t-1})}^{\text{decision rule}}.$$

Thus, the parametrisation and estimation only concern the system model.

Assumption 1 (Markov Time-Invariant Parametric Model)**.** The system model is *time-invariant*, parameterised by a constant multivariate parameter $\Theta \in \mathbf{\Theta}$. It is Markov model of an order $n < \infty$. It means

$\mathsf{M}(O_t|A_t, D^{t-1}, \Theta) \equiv \mathsf{M}(O_t|\rho_t, \Theta)$, with time invariant M. The *regression vector*, denoted ρ_t,

ρ_t is a known function of ρ_{t-1} and of $\begin{cases} A_t, D_{t-1}, \dots, D_{t-n} & \text{if } 1 \le n < \infty, \\ A_t \text{ or void} & \text{for } n = 0. \end{cases}$

Data entering the model M at time $t \in \boldsymbol{t}$ form the *data vector* $\Psi_t \equiv (O_t, \rho_t) \in \mathbf{\Psi}$, extending the regression vector ρ_t.

The knowledge D_0 provides the model structure and the regression vector ρ_0. (9)

The Markov parametric model (9) is ratio of the joint pd $\mathsf{J}(\Psi|\Theta)$ and its marginal

$$\mathsf{M}(O_t|\rho_t, \Theta) = \frac{\mathsf{J}(O_t, \rho_t|\Theta)}{\int_{\boldsymbol{O}} \mathsf{J}(O_t, \rho_t|\Theta)\mathrm{d}O_t} = \frac{\mathsf{J}(\Psi_t|\Theta)}{\int_{\boldsymbol{O}} \mathsf{J}(O_t, \rho_t|\Theta)\mathrm{d}O_t}. \tag{10}$$

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

Remark

✓ The Markov property is inevitable for the targeted feasibility of the recursive estimation. If not met naturally, it must be enforced via an approximation. The required time invariance can be relaxed by including time into ρ_t or by considering a time– and data– dependent unknown parameter Θ. The latter case leads to the untreated hidden Markov models[25].

Exponential Family: Our model exploits EF members described by the pd

$$\mathsf{M}(O_t|\rho_t, \Theta) \equiv \exp\langle \mathsf{B}(\Psi_t), \mathsf{C}(\Theta)\rangle, \ \Psi_t \in \mathbf{\Psi}, \ \Theta \in \mathbf{\Theta}, \quad \text{where} \tag{11}$$

vector-valued functions $\mathsf{B}(\Psi_t)$, $\mathsf{C}(\Theta)$ have a finite dimension. The real-valued mapping $\langle \mathsf{B}(\Psi_t), \mathsf{C}(\Theta)\rangle$ is *linear in* $\mathsf{B}(\Psi_t)$*-values*.

Under (2), EF members posses conjugated (self-reproducing) pds[26],

$$\mathsf{P}_t(\Theta) \equiv \mathsf{P}(\Theta|D^t) = \mathsf{P}(\Theta|\mathsf{V}_t) = \frac{\exp\langle \mathsf{V}_t, \mathsf{C}(\Theta)\rangle}{\mathsf{N}(\mathsf{V}_t)} \tag{12}$$

$$\mathsf{N}(\mathsf{V}_t) \equiv \int_{\boldsymbol{\Theta}} \exp\langle \mathsf{V}_t, \mathsf{C}(\Theta)\rangle \, \mathrm{d}\Theta. \text{ The sufficient statistic } \mathsf{V}_t \text{ evolves}$$

$$\mathsf{V}_t = \mathsf{V}_{t-1} + \mathsf{B}(\Psi_t) \text{ for } t \in \boldsymbol{t} \text{ while } \mathsf{V}_0 \text{ determines the prior pd.}$$

Remarks

- ✓ Definition (11) 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.
- ✓ The use of EF converts functional Bayes' rule into the algebraic evolution of the values of the sufficient statistic V_t.
- ✓ V_0 in (12) gives the conjugated prior pd, stores a prior knowledge D_0, regularises the estimation and is to make $\mathsf{N}(\mathsf{V}_0) < \infty$.
- ✓ EF exhausts parametric models with Θ-independent support, smooth in Θ and having a finite-dimensional sufficient statistic $\mathsf{V}_t = \mathsf{V}(D^t)$[48].
- ✓ 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 in (10), depends on Θ only (not on the regression vector ρ). Static models in EF are much richer.

Universal Approximation: Smooth *joint* pds $\mathsf{J}(\Psi) = \mathsf{J}(O, \rho)$ of data vectors $\Psi \in \boldsymbol{\Psi}$ can be universally approximated by a finite mixture of normal pds, to an arbitrary precision[18]. It holds even when the regression vectors $\rho \in \boldsymbol{\rho}$ are non-void. Thus, the joint pds of data vectors can be approximated by a finite mixture of pds from EF. This allows us to consider the joint parametric pd in (10) as the finite weighted sum of EF components, pds $(\mathsf{J}_c(\Psi|\Theta_c))_{c\in\boldsymbol{c}}$ on $\boldsymbol{\Psi}$,

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

$$\alpha \in \boldsymbol{\alpha} \equiv \left\{\alpha_c \geq 0,\ \sum_{c\in\boldsymbol{c}} \alpha_c = 1\right\},\ \Theta \equiv (\alpha, (\Theta_c)_{c\in\boldsymbol{c}}) \in \boldsymbol{\Theta} \equiv \left(\boldsymbol{\alpha}, (\boldsymbol{\Theta}_c)_{c\in\boldsymbol{c}}\right). \tag{13}$$

The insertion of (13) into (10) gives the non-standard parametric *ratio model*

$$\mathsf{M}(O_t|\rho_t, \Theta) = \sum_{c\in\boldsymbol{c}} \frac{\alpha_c \exp\left\langle \mathsf{B}_c(O_t, \rho_t), \mathsf{C}_c(\Theta_c)\right\rangle}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \underbrace{\int_{\boldsymbol{O}} \exp\langle \mathsf{B}_{\tilde{c}}(O_t, \rho_t), \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\rangle \, \mathrm{d}O_t}_{\mathsf{N}_{\tilde{c}}(\rho_t, \Theta_{\tilde{c}})}} \tag{14}$$

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

Assumption 1 and the *universal* approximation by finite mixtures *for void* ρ imply the quite strong property of the model (14):

> The ratio (14) approximates any *dynamic*, Markov, time-invariant model of data.

Remarks

- ✓ The 2nd row of (14) represents the ratio model as a fully dynamic finite mixture. The component weights $\mathsf{w}_c(\rho_t, \Theta)$ depend on the regression vector ρ_t in *non-ambiguous way*, which needs *no extra parameter*.

✓ The components $\mathsf{J}_c(\Psi_t|\Theta_c) = \exp\langle \mathsf{B}_c(\Psi_t), \mathsf{C}_c(\Theta_c)\rangle$ in (13) are joint pds on the set $\boldsymbol{\Psi}$. These pds may contain parameter-independent factors G_c

$$\mathsf{J}_c(\Psi_t|\Theta_c) = \exp\langle \mathsf{B}_c(\Psi_t), \mathsf{C}_c(\Theta_c)\rangle = \exp\left\langle \mathsf{B}_c(\Psi_{t;c}), \mathsf{C}_c(\Theta_c)\right\rangle \mathsf{G}_c(\underline{\rho}_{t;c}). \tag{15}$$

The parametric factor models the data vector $\Psi_{t;c} \equiv (O_t, \rho_{t;c})$, where $\rho_{t;c}$ is a sub-vector of ρ_t. The pd $\mathsf{G}_c(\underline{\rho}_{t;c})$ is a parameter-free pd on the complement $\underline{\rho}_{t;c}$ of $\rho_{t;c}$ to ρ_t. This mimics mixtures of principal component analysers[49] and diminishes the dimensionality curse[50].

4 | ESTIMATION OF MIXTURE RATIOS

This section applies the approximate estimation (8) to the model (14). This yields the *universal and feasible probabilistic clustering of dynamic data streams*.

Choice of the Set $\mathbf{P}$ *of Feasible PDs:* The weights α in (13) define the pd $\mathsf{M}(c_t = c|\Theta) = \alpha_c$, $c \in \boldsymbol{c}$, of a thought *unobserved* pointer, $c_t \in \boldsymbol{c}$, to the active component $\mathsf{J}_{c_t}(\Psi_t|\Theta_{c_t})$ (15) "generating" the data vector Ψ_t. This model is from EF

$$\begin{aligned}
\mathsf{M}(c_t|\Theta) &= \exp\left[\sum_{c\in\boldsymbol{c}} \delta(c, c_t), \ln(\alpha_c)\right] = \exp\langle \delta(c_t), \ln(\alpha)\rangle \\
\delta(c, c_t) &\equiv \begin{cases} 1 \text{ if } c = c_t \\ 0 \text{ otherwise} \end{cases}, \ \delta(c_t) \equiv [\delta(1, c_t), \dots, \delta(|\boldsymbol{c}|, c_t)] \\
\ln(\alpha) &\equiv [\ln(\alpha_1), \dots, \ln(\alpha_{|\boldsymbol{c}|})], \quad \langle \delta(c_t), \ln(\alpha)\rangle = \sum_{c\in\boldsymbol{c}} \delta(c, c_t)\ln(\alpha_c).
\end{aligned}$$

For the *observed* pointer $c_t \in \boldsymbol{c}$ to the active component, Dirichlet's pd[51] determined by a $|\boldsymbol{c}|$-vector statistic v_t, is the self-reproducing pd

$$\begin{aligned}
\mathsf{P}_t(\alpha) &\equiv \mathsf{P}(\alpha|\mathsf{v}_t) \equiv \frac{\exp\langle \mathsf{v}_t - 1, \ln(\alpha)\rangle}{\mathsf{Be}(\mathsf{v}_t)} \ \text{ with } \mathsf{v}_t = \mathsf{v}_{t-1} + \delta(c_t), \ \mathsf{v}_0 > 0 \\
\mathsf{Be}(\mathsf{v}) &\equiv \frac{\prod_{c\in\boldsymbol{c}} \Gamma(\mathsf{v}_c)}{\Gamma\left(\sum_{c\in\boldsymbol{c}} \mathsf{v}_c\right)}, \quad \Gamma(\upsilon) \equiv \int_0^\infty z^{\upsilon-1}\exp(-z)\mathrm{d}z, \ \upsilon > 0.
\end{aligned} \tag{16}$$

For normal and Markov chain components or independent data vectors with components (15), the conjugate pds $\mathsf{P}_t(\Theta_c) = \mathsf{P}(\Theta_c|D^t, c_t, \dots, c_1)$, $c \in \boldsymbol{c}$, are

$$\mathsf{P}_t(\Theta_c) = \frac{\exp\left\langle \mathsf{V}_{t;c}, \mathsf{C}_c(\Theta_c)\right\rangle}{\mathsf{N}_c(\mathsf{V}_{t;c})}, \ \mathsf{N}_c(\mathsf{V}_{t;c}) = \int_{\boldsymbol{\Theta}_c} \exp\left\langle \mathsf{V}_{t;c}, \mathsf{C}_c(\Theta_c)\right\rangle \mathrm{d}\Theta_c. \tag{17}$$

They are "natural" candidates for creating the set $\mathbf{P}$ of feasible pds used in (8). They have to be given by *statistic* $(\mathsf{v}_t, (\mathsf{V}_{t;c})_{c\in\boldsymbol{c}})$ with values dependent on the observed data, not on the unobserved $(c_t, \dots, c_1)$.

The pds $\mathsf{P}_t(\alpha)$ (16), $(\mathsf{P}_t(\Theta_c))_{c\in\boldsymbol{c}}$ (17) do not determine the joint pd $\mathsf{P}_t(\Theta)$, $\Theta \equiv (\alpha, (\Theta_c)_{c\in\boldsymbol{c}})$, unambiguously[52]. Their product coupling is, however, preferable as it models the faced lack of information about mutual relations of the component parameters and their weights. This motivates the choice

$$\begin{aligned}
\mathbf{P} \equiv \Big\{ \mathsf{P}_t(\Theta) : \ \mathsf{P}_t(\Theta) = \mathsf{P}_t(\alpha)\prod_{c\in\boldsymbol{c}} \mathsf{P}_t(\Theta_c) \text{ with factors given by (16), (17),} \\
\text{which are given by the optional statistics values } \left(\mathsf{v}_t, \mathsf{V}_t \equiv (\mathsf{V}_{t;c})_{c\in\boldsymbol{c}}\right) \Big\}.
\end{aligned} \tag{18}$$

Evaluations of Kerridge's Inaccuracy (4): With the chosen $\mathbf{P}$ (18), it remains to map (8) onto the updating of $\mathsf{v}_{t-1}, \mathsf{V}_{t-1} \equiv (\mathsf{V}_{t-1;c})_{c\in\boldsymbol{c}}$ to $\mathsf{v}_t, \mathsf{V}_t \equiv (\mathsf{V}_{t;c})_{c\in\boldsymbol{c}}$.

Kerridge's inaccuracy of the pd $\tilde{\mathsf{P}}_t$ (3) to a pd $\hat{\mathsf{P}}_t \in \mathbf{P}$ (18), given by the statistic $\hat{\mathsf{V}}_t \equiv \left(\hat{\mathsf{v}}_t, (\hat{\mathsf{V}}_{t;c})_{c\in\boldsymbol{c}}\right)$, reads

$$\begin{aligned}
\mathsf{K}(\tilde{\mathsf{P}}_t||\hat{\mathsf{P}}_t) &= \ln(\mathsf{Be}(\hat{\mathsf{v}}_t)) + \sum_{c\in\boldsymbol{c}} \ln(\mathsf{N}_c(\hat{\mathsf{V}}_{t;c})) \\
&- \left\langle \hat{\mathsf{v}}_t - 1, \int_{\boldsymbol{\Theta}} \ln(\alpha)\tilde{\mathsf{P}}_t(\Theta)\mathrm{d}\Theta \right\rangle - \sum_{c\in\boldsymbol{c}} \left\langle \hat{\mathsf{V}}_{t;c}, \int_{\boldsymbol{\Theta}} \mathsf{C}_c(\Theta_c)\tilde{\mathsf{P}}_t(\Theta)\mathrm{d}\Theta \right\rangle.
\end{aligned} \tag{19}$$

Bayes' rule (3) for the mixture ratio (14) with EF components (15) and a prior pd $\mathsf{P}_{t-1} \in \mathbf{P}$ (18), gives

$$\begin{aligned}\tilde{\mathsf{P}}_t(\Theta) \propto & \sum_{c\in\boldsymbol{c}} \frac{\alpha_c \exp\langle \mathsf{B}_c(\Psi_t), \mathsf{C}_c(\Theta_c)\rangle \exp\langle \mathsf{v}_{t-1}-1, \ln(\alpha)\rangle \prod_{\tilde{c}\in\boldsymbol{c}} \exp\left\langle \mathsf{V}_{t-1;\tilde{c}}, \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\right\rangle}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \mathsf{N}_{\tilde{c}}(\rho_t, \Theta_{\tilde{c}})} \\ = & \sum_{c\in\boldsymbol{c}} \frac{\exp\langle \mathsf{v}_{t-1} + \delta(c) - 1, \ln(\alpha)\rangle}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \mathsf{N}_{\tilde{c}}(\rho_t, \Theta_{\tilde{c}})} \\ & \times \prod_{\tilde{c}\in\boldsymbol{c}} \exp\left\langle \mathsf{V}_{t-1;\tilde{c}} + \delta(c,\tilde{c})\mathsf{B}_c(\Psi_{t;c}), \mathsf{C}_{\tilde{c}}(\Theta_c)\right\rangle \mathsf{G}_c(\underline{\rho}_{t;c}). \end{aligned} \tag{20}$$

Further evaluations exploit auxiliary pds arising from summands in (20), cf. (16), (17), via an appropriate normalisation

$$\begin{aligned}\tilde{\mathsf{Q}}_{t;c}(\Theta) &\equiv \frac{\exp\langle \mathsf{v}_{t-1} + \delta(c) - 1, \ln(\alpha)\rangle}{\beta_{t;c}} \prod_{\tilde{c}\in\boldsymbol{c}} \exp\left\langle \mathsf{V}_{t-1;\tilde{c}} + \delta(c,\tilde{c})\mathsf{B}_c(\Psi_{t;c}), \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}})\right\rangle \\ \beta_{t;c} &\equiv \mathsf{Be}(\mathsf{v}_{t-1} + \delta(c)) \prod_{\tilde{c}\in\boldsymbol{c}} \mathsf{N}(\mathsf{V}_{t-1;\tilde{c}} + \delta(c,\tilde{c})\mathsf{B}_c(\Psi_{t;c}))/\mathsf{G}_c(\underline{\rho}_{t;c}). \end{aligned} \tag{21}$$

They independently model α and $(\Theta_c)_{c\in\boldsymbol{c}}$. The dependency enters $\tilde{\mathsf{P}}_t$ via

$$\begin{aligned}\mathsf{H}(\rho_t, \Theta) &\equiv \frac{1}{\sum_{\tilde{c}\in\boldsymbol{c}} \alpha_{\tilde{c}} \mathsf{N}_{\tilde{c}}(\rho_t, \Theta)} \quad \text{because} \quad \tilde{\mathsf{P}}_t(\Theta) = \frac{\mathsf{H}(\rho_t, \Theta) \sum_{c\in\boldsymbol{c}} \beta_{t;c} \tilde{\mathsf{Q}}_{t;c}(\Theta)}{\mathtt{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})} \\ \text{with} \quad \mathtt{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1}) &\equiv \int_{\boldsymbol{\Theta}} \mathsf{H}(\rho_t, \Theta) \sum_{c\in\boldsymbol{c}} \beta_{t;c} \tilde{\mathsf{Q}}_{t;c}(\Theta) \mathrm{d}\Theta. \end{aligned} \tag{22}$$

The determination of Kerridge's inaccuracy (19) needs to evaluate the integrals

$$\begin{aligned}\mathtt{I}_{0c}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1}) &\equiv \int_{\boldsymbol{\Theta}} \ln(\alpha_c) \mathsf{H}(\rho_t, \Theta) \sum_{\tilde{c}\in\boldsymbol{c}} \beta_{t;\tilde{c}} \tilde{\mathsf{Q}}_{t;\tilde{c}}(\Theta) \mathrm{d}\Theta, \quad \mathtt{I}_0 \equiv (\mathtt{I}_{0c})_{c\in\boldsymbol{c}}, \\ \mathtt{I}_c(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1}) &\equiv \int_{\boldsymbol{\Theta}} \mathsf{C}_c(\Theta) \mathsf{H}(\rho_t, \Theta) \sum_{\tilde{c}\in\boldsymbol{c}} \beta_{t;\tilde{c}} \tilde{\mathsf{Q}}_{t;\tilde{c}}(\Theta) \mathrm{d}\Theta, \ c \in \boldsymbol{c}. \end{aligned} \tag{23}$$

Their insertion into Kerridge's inaccuracy (19) gives the best-projection statistic

$$\begin{aligned}\left[\hat{\mathsf{v}}_t, (\hat{\mathsf{V}}_{t;c})_{c\in\boldsymbol{c}}\right] \in & \operatorname{Arg} \min_{\hat{\mathsf{v}}\in\hat{\mathbf{v}}, (\hat{\mathsf{V}}_c\in\hat{\mathbf{V}}_c)_{c\in\boldsymbol{c}}} \left(\ln(\mathsf{Be}(\hat{\mathsf{v}})) + \sum_{c\in\boldsymbol{c}} \ln(\mathsf{N}_c(\hat{\mathsf{V}}_c)) \right. \\ & \left. - \left\langle \hat{\mathsf{v}} - 1, \frac{\mathtt{I}_0(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})}{\mathtt{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})} \right\rangle - \sum_{c\in\boldsymbol{c}} \left\langle \hat{\mathsf{V}}_c, \frac{\mathtt{I}_c(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})}{\mathtt{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})} \right\rangle \right) \\ = & \left[\operatorname{Arg} \min_{\hat{\mathsf{v}}\in\hat{\mathbf{v}}} \left(\ln(\mathsf{Be}(\hat{\mathsf{v}})) - \left\langle \hat{\mathsf{v}} - 1, \frac{\mathtt{I}_0(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})}{\mathtt{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})} \right\rangle \right), \right. \\ & \left. \left(\operatorname{Arg} \min_{\hat{\mathsf{V}}_c\in\hat{\mathbf{V}}_c} \left(\ln(\mathsf{N}_c(\hat{\mathsf{V}}_c)) - \left\langle \hat{\mathsf{V}}_c, \frac{\mathtt{I}_c(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})}{\mathtt{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})} \right\rangle \right) \right)_{c\in\boldsymbol{c}} \right]. \end{aligned} \tag{24}$$

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

Numerical Evaluations of $\mathtt{I}$*'s:* The function $\mathsf{H}(\rho_t, \Theta)$ (22) depends on all involved parameters. It combines influences of respective components. A brute force evaluation of $\mathtt{I}_{-1}$, $\mathtt{I}_0$ and $\mathtt{I}_c$, $c \in \boldsymbol{c}$, say by Monte Carlo (MC), is mostly too demanding in a typical stream processing. At the same time, the averaged functions in (22), (23) are smooth due to their construction. Moreover, their growth over their domains is well suppressed by generally light tails of the pds $\tilde{\mathsf{Q}}_{t;c}$ (21). Thus, we conjecture (and experiments support this conjecture) that the simplest approximation based on the first order Taylor expansion of $\mathsf{H}(\rho, \Theta)$ around expected values

$$\int_{\boldsymbol{\Theta}} \alpha_{\tilde{c}} \tilde{\mathsf{Q}}_{t;c}(\Theta) \mathrm{d}\Theta, \qquad \int_{\boldsymbol{\Theta}} \Theta_{\tilde{c}} \tilde{\mathsf{Q}}_{t;c}(\Theta) \mathrm{d}\Theta, \qquad \tilde{c}, c \in \boldsymbol{c}, \tag{25}$$

suffices. In addition to (25), the approximation requires to evaluate $\forall \tilde{c}, c \in \boldsymbol{c}$

$$\int_{\boldsymbol{\Theta}} \alpha \ln(\alpha_{\tilde{c}}) \tilde{\mathsf{Q}}_{t;c}(\Theta) \mathrm{d}\Theta, \quad \int_{\boldsymbol{\Theta}} \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}}) \tilde{\mathsf{Q}}_{t;c}(\Theta) \mathrm{d}\Theta, \quad \int_{\boldsymbol{\Theta}} \Theta_{\tilde{c}} \mathsf{C}_{\tilde{c}}(\Theta_{\tilde{c}}) \tilde{\mathsf{Q}}_{t;c}(\Theta) \mathrm{d}\Theta. \tag{26}$$

Remarks

- ✓ The normalising constants Be, N (16), (17) of factors forming $\tilde{\mathsf{Q}}_{t;c}$ (21) can be analytically expressed for Normal-inverse-Wishart and Dirichlet's pds, which are conjugated to normal and Markov chain components[35].
- ✓ $\tilde{c}$th entries of expectations (25), (26) 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 ratio model, on the key novelty brought.
- ✓ This estimation applied to standard mixtures reduces to the projection-based algorithm[37], which is still the state-of-the-art recursive Bayesian estimator.

Forgetting Handling: It consists of the choice of the forgetting factor and its use. The formula for λ_t in (8) depends on the ratio of the forecasting pd divided by the forecasting pd constructed from the already updated posterior pd (both in the data realisation). Independent studies[53] imply that the forgetting should be applied component-wise. The overall forecasting quality then influences only the forgetting of component weights. It gives the forgetting factors λ_t, $(\lambda_{t;c})_{c\in\boldsymbol{c}}$, (8),

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

There $\mathtt{I}_{-1}(\Psi_t, \mathsf{v}_{t-1}, \mathsf{V}_{t-1})$ is approximately evaluated using an expansion around (25). The denominator $\mathtt{I}_{-1}(\Psi_t, \hat{\mathsf{v}}_t, \hat{\mathsf{V}}_t)$ is computed according to the same formulae with $\hat{\mathsf{v}}_t, \hat{\mathsf{V}}_t$ (24) replacing $\mathsf{v}_{t-1}, \mathsf{V}_{t-1}$. $\tilde{\Theta}_{t-1;c}$, $\tilde{\Theta}_{t;c}$ are the expected values of the component parameter Θ_c computed as the cth part of (25) with the component statistics $\mathsf{V}_{t-1;c}$ and $\hat{\mathsf{V}}_{t;c}$, respectively (14). The forgetting completes the updating (8) for the proposed mixture ratio model (14), (15)

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

5 | ILLUSTRATIVE EXAMPLES

This section deals both with the modelling and the stream estimation. The used toy examples serve to the aimed illustration of the theory. A systematic simulation and real-life studies are out of scope of this paper and will be published independently. Some experiments, including real-data application to commodity futures[54] are already available in[35,36].

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*.

On Mixture-Ratio Modelling Strength: The simulated system E generates real scalar observations O_t with no actions and the regression vector $\rho_t = O_{t-1}$. E is the mixture of the joint normal pds $\mathcal{N}_\Psi(\mu_c, \omega_c^{0.5})$, $\Psi_t \equiv (O_t, O_{t-1})$, with expectations μ_c and square roots $\omega_c^{0.5}$ of precision matrices, $c \in \boldsymbol{c} \equiv \{1, 2\}$,

$$\mathsf{E}(\Psi_t) = \frac{1}{2}\mathcal{N}_{\Psi_t}\Bigg(\underbrace{\begin{bmatrix}1\\1\end{bmatrix}}_{\mu_1}, \underbrace{\begin{bmatrix}1 & 3\\ 0 & 0.5\end{bmatrix}}_{\omega_1^{0.5}}\Bigg) + \frac{1}{2}\mathcal{N}_{\Psi_t}\Bigg(\underbrace{\begin{bmatrix}-1\\-1\end{bmatrix}}_{\mu_2}, \underbrace{\begin{bmatrix}1 & -2\\ 0 & 0.5\end{bmatrix}}_{\omega_2^{0.5}}\Bigg). \tag{27}$$

The typical simulation results were sampled from $\mathsf{E}(O_t|O_{t-1})$ while differing only in initial values $O_0 \in \{-0.8, 0, 3\}$, are in Fig. 1. For comparison, the same mixture was simulated as the conditional one, i.e. with fixed component weights.

Discussion: The simulation results demonstrate the dynamic dependence of the components weight on the data realisation. This is the key feature of the truly dynamic model. It allows to model non-linear dynamic effects and unbalanced activations of respective components. The standard mixture model lacks both mentioned features of the mixture ratio model.

On Performance of the Standard and Ratio Mixture Models:

Simulation Conditions: Scalar discrete observations $O_t \in \boldsymbol{O} = \{1, \dots, |\boldsymbol{O}|\}$, $|\boldsymbol{O}| = 5$, and 2^{nd} order regression with $\rho_t = (O_{t-1}, O_{t-2})$ are used. This defines data vectors (9) $\Psi_t = (O_t, O_{t-1}, O_{t-2})$. The *learnt* (estimated) *models* L and the *simulated system models* E are either the mixture ratio model M or the standard mixture S, both with $|\boldsymbol{c}| = 2$,

$$\mathsf{E}, \mathsf{L} \in \{\mathsf{M}, \mathsf{S}\} = \{\text{mixture ratio, standard mixture}\}. \tag{28}$$

This gives four combinations of the simulated and learnt models.

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

$$\mathsf{J}(\Psi_t|\Theta_{\mathsf{M}}) = \alpha_{\mathsf{M}}\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M1}})\mathsf{G}_1(O_{t-2}) + (1-\alpha_{\mathsf{M}})\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M2}})\mathsf{G}_2(O_{t-1}).$$

The joint pds $\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M1}})$ and $\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M2}})$ are parameterised by their values Θ_{M1}, Θ_{M2} assigned to possible pairs (O_t, O_{t-1}), (O_t, O_{t-2}). The model parameter is $\Theta_{\mathsf{M}} = (\alpha_{\mathsf{M}}, \Theta_{\mathsf{M1}}, \Theta_{\mathsf{M2}})$. The parameter-free factors $\mathsf{G}_1(O_{t-2})$, $\mathsf{G}_2(O_{t-2})$ are uniform on $\boldsymbol{O}$. They cancel in the mixture ratio M, which gets the form

$$\mathsf{M}(O_t|\rho_t, \Theta_{\mathsf{M}}) = \frac{\alpha_{\mathsf{M}}\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M1}}) + (1-\alpha_{\mathsf{M}})\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M2}})}{\sum_{O_t \in \boldsymbol{O}} \alpha_{\mathsf{M}}\mathsf{J}_1(O_t, O_{t-1}|\Theta_{\mathsf{M1}}) + (1-\alpha_{\mathsf{M}})\mathsf{J}_2(O_t, O_{t-2}|\Theta_{\mathsf{M2}})}.$$

The standard model $\mathsf{S}(O_t|\rho_t, \Theta_{\mathsf{S}}) = \alpha_{\mathsf{S}}\mathsf{S}_1(O_t|O_{t-1}, \Theta_{\mathsf{S1}}) + (1-\alpha_{\mathsf{S}})\mathsf{S}_2(O_t|O_{t-2}.\Theta_{\mathsf{S2}})$ is parameterised by $\Theta_{\mathsf{S}} = (\alpha_{\mathsf{S}}, \Theta_{\mathsf{S1}}, \Theta_{\mathsf{S2}})$. Its components $\mathsf{S}_1(O_t|O_{t-1}, \Theta_{\mathsf{S1}})$ and $\mathsf{S}_2(O_t|O_{t-2}, \Theta_{\mathsf{S2}})$ are parameterised by probabilities $\Theta_{\mathsf{S1}}, \Theta_{\mathsf{S2}}$ assigned to O_t when conditioned on O_{t-1} and O_{t-2}, respectively.

MC Study over E*:* It consists of 200 runs for $t \in \boldsymbol{t}$, $|\boldsymbol{t}| = 500$, with randomly generated parameters Θ_{E} for both considered simulated model structures $\mathsf{E} \in \{\mathsf{M}, \mathsf{S}\}$ (28). In both cases, the mixture ratio M model and the standard mixture S model are learnt on the same observation sequences.

MC Study with a Fixed $\mathsf{E} = \underline{\mathsf{M}}$*:* The fixed system being the mixture ratio model $\underline{\mathsf{M}}$ given by the fixed parameter $\underline{\Theta}_{\mathsf{M}}$ having truly dynamic weights $\mathsf{w}_c(\rho_t)$ (14) is simulated. Its 200 runs differ in the noise realisations.

Performance Evaluation: At the time $t \in \boldsymbol{t}$, the simulation and estimation dealt with the underlying model $\mathsf{E}, \mathsf{L} \in \{\mathsf{M}, \mathsf{S}\}$ and assigned the following probabilities to observations

$$\begin{aligned} \mathsf{E}_t(O_t) &= \begin{cases} \underline{\mathsf{M}}_t(O_t) = \mathsf{M}(O_t|\rho_t, \underline{\Theta}_{\mathsf{M}}) & \text{if } \mathsf{E} = \mathsf{M} \text{ with a fixed } \underline{\Theta}_{\mathsf{M}} \\ \underline{\mathsf{S}}_t(O_t) = \mathsf{S}(O_t|\rho_t, \underline{\Theta}_{\mathsf{S}}) & \text{if } \mathsf{E} = \mathsf{S} \text{ with a fixed } \underline{\Theta}_{\mathsf{S}} \end{cases} \\ \mathsf{L}_t(O_t) &= \mathsf{L}_t(O_t|O_{t-1}, \dots, O_0) = \int_{\boldsymbol{\Theta}_{\mathsf{L}}} \mathsf{L}(O_t|\rho_t, \Theta_{\mathsf{L}})\mathsf{P}(\Theta_{\mathsf{L}}|O_{t-1}, \dots, O_0)\mathrm{d}\Theta_{\mathsf{L}}. \end{aligned} \tag{29}$$

The corresponding Kullback-Leibler divergences were sequentially evaluated

$$\begin{aligned} \mathsf{D}_t(\mathsf{E}||\mathsf{L}) &= \sum_{\tau=1}^{t} \mathsf{D}(\mathsf{E}_\tau||\mathsf{L}_\tau), \quad \text{with} \quad \mathsf{D}(\mathsf{E}_\tau||\mathsf{L}_\tau) = \sum_{O_\tau \in \boldsymbol{O}} \mathsf{E}_\tau(O_\tau) \ln\left(\frac{\mathsf{E}_\tau(O_\tau)}{\mathsf{L}_\tau(O_t)}\right) \\ \mathsf{D}(\mathsf{E}||\mathsf{L}) &\equiv \mathsf{D}_{|\boldsymbol{t}|}(\mathsf{E}||\mathsf{L}), \quad \text{and their differences} \quad \Delta = \mathsf{D}(\mathsf{E}||\mathsf{M}) - \mathsf{D}(\mathsf{E}||\mathsf{S}). \end{aligned} \tag{30}$$

They quantified estimation quality within the comparable observation space.

Results: The simulation results are in Fig. 2. The left hand column contains histogram values $\mathsf{D}(\mathsf{E}||\mathsf{L}) = \mathsf{D}_{|t|}(\mathsf{E}||\mathsf{L})$, $\mathsf{L} \in \{\mathsf{M}, \mathsf{S}\}$, at final time $|\boldsymbol{t}| = 500$. The right hand column shows the corresponding sample cumulative distributions. The better outcomes have them shifted to the left. Fig. 3 shows the course of $\mathsf{D}_t(\mathsf{E}||\mathsf{L})$, $t \in \boldsymbol{t}$, $\mathsf{L} \in \{\mathsf{M}, \mathsf{S}\}$ for: (a) $\mathsf{E} = \underline{\mathsf{M}}$, given by a fixed $\underline{\Theta}_{\mathsf{M}}$; (b) $\mathsf{E} = \underline{\mathsf{S}}$, given by a fixed $\underline{\Theta}_{\mathsf{S}}$; (c) $\mathsf{E} = \underline{\mathsf{M}}$ with $\underline{\Theta}_{\mathsf{M}}$ giving the dynamic weights $\mathsf{w}_c(\rho_t)$ (14). Tab. 1 shows sample statistics of increments Δ (30). Negative values of sample means and medians reflect the dominance of the mixture ratio model.

Discussion: Tab. 1 and Figs. 2, 3 confirm the expectation that the mixture ratio M outperforms the standard mixture if the mixture ratio M is simulated. The improvement can be significant. The learnt mixture ratio copies the performance of the standard mixture S when S is simulated. It is even slightly better as it copes better with errors caused by the approximate estimation (8).

TABLE 1 Sample statistics of differences of Kullback-Leibler divergences $\Delta = \mathsf{D}(\mathsf{E}||\mathsf{M}) - \mathsf{D}(\mathsf{E}||\mathsf{S})$ (30), where E is the predictor corresponding to the simulated system, to the learnt predictor with M, which corresponds to the mixture ratio, and S corresponding to the standard mixture. The 1st column shows statistics for the simulated system $\mathsf{E} = \mathsf{M}$ with Θ_{M} varied in 200 MC runs; the 2nd column shows results for the simulated system $\mathsf{E} = \mathsf{S}$ with Θ_{S} varied in 200 MC runs; the 3rd column reflects the case with the simulated $\mathsf{E} = \underline{\mathsf{M}}$ with a fixed $\underline{\Theta}_{\mathsf{M}}$ causing the truly dynamic weights $\mathsf{w}_c(\rho_t)$ (14) (MC runs differ in noise).

Simulated Case	MC over $\mathsf{E} = \mathsf{M}$	MC over $\mathsf{E} = \mathsf{S}$	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 dynamically related data. The next list explicates our gains.

- ✓ The mixture ratios universally approximate non-linear *dynamic* stochastic relations.
- ✓ The mixture ratios handle cases in which rare visits of active components are significant, cf. Fig. 1. This is vital in fault detection[55], detection of non-standard fraud behaviours needed[56], or in cyber-security applications[10]; everywhere, where outlier detection is faced[57]. These cases are hard dynamic versions of estimation with unbalanced data[58].
- ✓ The mixture ratios suit to modelling of mixed (discrete and continuous valued) data. The components are the joint pds factorable into pds of continuous valued data, possibly conditioned on discrete ones, and pds relating discrete data.
- ✓ The mixture ratios may serve as a relatively universal dynamic feature extractor. Indeed, the approximately sufficient statistic collected during estimation, see Sec. 4, are the relevant features. It suffices to use Bayesian structure estimation on mixture ratios with low-dimensional components.
- ✓ The estimation of the mixture ratios provides the stationary joint pd of the data vector (13), which can directly serve for the design of adaptive decision strategies for an infinite decision horizon[59]. The estimated 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[60].

In future, it is desirable:

- ✓ to elaborate numerically robust (factorised) procedures for normal models, sparse Markov-chains and mixed cases in the way mimic to[50];[35] has made definite steps in this respect;
- ✓ to refine the data-dependent choice of forgetting factors;
- ✓ to tailor Bayesian structure estimation, ready for mixtures both with respect to suitable regression vectors ρ_c and the number of mixture components[51] to the mixture-ratio model;
- ✓ to perform real-life tests confirming that the theoretically higher approximating strength of the discussed mixture ratios, see Sec. 3, is mostly undiminished by the approximate estimation; the experiments with commodity futures trading, made in[35], as well as other (yet unpublished) real-data processing, are more than promising.

The theoretical insight, experience from simulations, preliminary processing of real data, width and importance of the outlined contributions to estimation dynamic data relations make this effort worthwhile.

Data Availability The paper deals just with simulations. Authors will provide the experimental code upon a request.

Conflict of interest The authors declare that this is their original work and represents no potential conflict of interests.
Acknowledgement This research was supported by MŠMT ČR LTC18075 and by EU-COST Action CA1622.

References

1. Mitchell T. *Machine Learning*. McGraw Hill . 1997.

2. Sadghi P, Kennedy R, Rapajic P, Shams R. Finite-State Markov Modeling of Fading Channels. *IEEE Signal Processing Magazine* 2008; 57. doi: 10.1109/MSP.2008.926683

3. Guyon I, Saffari A, Dror G, Cawley G. Model Selection: Beyond the Bayesian/Frequentist Divide. *Journal of Machine Learning Research* 2010; 11: 61–87.

4. Bishop C. *Pattern Recognition and Machine Learning*. Springer . 2006.

5. Mason W, Vaughan J, Wallach H. Special Issue: Computational social science and social computing. *Machine Learning* 2014; 96: 257–469.

6. Wiering M, Otterlo vM., eds.*Reinforcement Learning: State-of-the-Art*. Springer . 2012.

7. Guan P, Raginsky M, Willett R. Online Markov Decision Processes with Kullback Leibler Control Cost. *IEEE Trans. on AC* 2014; 59(6): 1423–1438.

8. Holmes G, Liu T. , eds., *Proc. of 7th Asian Conf. on Machine Learning (ACML2015), JMLR Workshop and Conf. Proceedings*. 45; 2015.

9. Tsai C, Lai C, Chiang M, Yang L. Data Mining for Internet of Things: A Survey. *IEEE Communications Surveys & Tutorials* 2014; 16(1): 77-95.

10. Buczak A, Guven E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys Tutorials* 2016; 18(2): 1153-1176.

11. Nguyen H, Woon Y, Ng W. A survey on data stream clustering and classification. *Knowledge and Information Systems* 2015; 45: 535–569.

12. Mirsky Y, Shapira B, Rokach L, Elovici Y. 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*. 9078. LNCS Springer, Cham. 2015 (pp. 230–237).

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

14. Silva J, Faria E, Barros R, Hruschka E, Carvalho A, Gama J. Data Stream Clustering: A Survey. *ACM Computing Surveys* 2014; 46. doi: 10.1145/2522968.2522981

15. Haykin S. *Neural Networks: A Comprehensive Foundation*. N.Y.: Macmillan . 1994.

16. Kolmogorov A. On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk* 1957; 114: 953-956.

17. Ferguson T. Bayesian density estimation by mixtures of normal distributions. In: Rizvi M, Rustagi J, Siegmund D. , eds. *Recent Advances in Statistics*Ac. Press. 1983 (pp. 287âĂŞ-302).

18. McLachlan G, Peel D. *Finite Mixture Models*. Wiley Series in Probab. & Stat.N.Y.: Wiley . 2000.

19. Yarotsky D. Universal approximations of invariant maps by neural networks. arXiv:1804.10306v1 [cs.NE] 26 Apr (2018).

20. Wald A. *Statistical Decision Functions*. N.Y., London: J. Wiley . 1950.

21. Savage L. *Foundations of Statistics*. Wiley . 1954.

22. Rao M. *Measure Theory and Integration*. J. Wiley . 1987.

23. Mosca E. *Optimal, Predictive, and Adaptive Control*. Prentice Hall . 1994.

24. Jazwinski A. *Stochastic Processes and Filtering Theory*. Ac. Press . 1970.

25. Elliot R, Assoun L, Moore J. *Hidden Markov Models*. N.Y.: Springer-Verlag . 1995.

26. Berger J. *Statistical Decision Theory and Bayesian Analysis*. Springer . 1985.

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

28. Park J, Sandberg I. Universal Approximation using Radial-Basis-Function Networks. *Neural Comput.* 1991; 3: 246–257.

29. McNicholas P. *Mixture model-based classification*. Boca Raton, London, N.Y.: CRC Press . 2017.

30. Catania L. Dynamic Adaptive Mixture Models. 2016. arXiv:1603.01308v1.

31. Nagy I, Suzdaleva E, Kárný M, Mlynářová T. Bayesian estimation of dynamic finite mixtures. *Int. J. Adapt Control Signal Process.* 2011; 25(9): 765-787.

32. Frigessi A, Haug O, Rue H. A Dynamic Mixture Model for Unsupervised Tail Estimation without Threshold Selection. *Extremes* 2002; 5(3): 219 – 235.

33. Barndorff-Nielsen O. *Information and Exponential Families in Statistical Theory*. N.Y.: Wiley . 1978.

34. Bohlin T. *Interactive System Identification: Prospects and Pitfalls*. Springer . 1991.

35. Ruman M. Mixture Ratios for Decision Making. Master's thesis. FJFI, Czech Technical University. Prague: 2018.

36. Ruman M, Kárný M. Dynamic Mixture Ratio Model. In: *Proc. ICCAIRO*, IEEE; 2019: 92 – 99

37. Kárný M. Approximate Bayesian Recursive Estimation. *Inf. Sci.* 2014; 289: 100–111.

38. Peterka V. Bayesian system identification. In: Eykhoff P., ed. *Trends and Progress in System Identification*Perg. Press. 1981.

39. Bernardo J. Expected Information as Expected Utility. *The An. of Stat.* 1979; 7: 686–690.

40. Kárný M, Guy T. On Support of Imperfect Bayesian Participants. In: Guy T, et al., eds. *Decision Making with Imperfect Decision Makers*. 28. Springer, Int. Syst. Ref. Lib. 2012 (pp. 29–56).

41. Kerridge D. Inaccuracy and inference. *J. of the Royal Stat. Soc.* 1961; B 23: 284–294.

42. Kulhavý R. A Bayes-closed approximation of recursive nonlinear estimation. *Int. J. Adapt Control Signal Process.* 1990; 4: 271–285.

43. Kárný M. Minimum Expected Relative Entropy Principle. In: *Proc. of the 18th ECC*, IFAC; 2020; Sankt Petersburg: 35–40.

44. Kullback S, Leibler R. On information and sufficiency. *Ann Math Stat* 1951; 22: 79–87.

45. Shore J, Johnson R. Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. *IEEE Tran. on Inf. Th.* 1980; 26(1): 26–37.

46. Šmídl V, Quinn A. *The Variational Bayes Method in Signal Processing*. Springer . 2005.

47. Minka T. *A family of algorithms for approximate Bayesian inference*. PhD thesis. MIT, 2001.

48. Koopman R. On distributions admitting a sufficient statistic. *Trans. of Am. Math. Society* 1936; 39: 399.

49. Tipping M, Bishop C. Probabilistic principal component analysis. *J. of the Royal Society Series B* 1999; 61: 611–622.

50. Kárný M. Recursive estimation of high-order Markov chains: Approx. by finite mixtures. *Inf. Sci.* 2016; 326: 188–201.

51. Kárný M, Böhm J, Guy T, et al. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer . 2006.

52. Nelsen R. *An Introduction to Copulas*. N.Y.: Springer . 1999.

53. Dedecius K, Nagy I, Kárný M, Pavelková L. Parameter Estimation With Partial Forgetting Method. In: *Proc. of the 15th IFAC SYSID*, 2009.

54. Carter C. Commodity future markets: A survey. *The Australian J. Agricultural and Resource Economics* 1999; 43: 209–247.

55. Polycarpou M, Helmicki A. Automated fault detection and accommodation: A learning systems approach. *IEEE Trans. on Systems, Man, and Cybernetics* 1995; 25(11): 1447-1458.

56. Kou Y, Lu C, Sirwongwattana S, Huang Y. Survey of fraud detection techniques. In: *IEEE Intern. Conf. on Networking, Sensing and Control*, 2; 2004: 749-754

57. Hodge V, Austin J. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review* 2004; 22(2): 85–126.

58. Bekkar M, Alitouche T. Imbalanced Data Learning Approaches Review. *Int. J. of Data Mining & Knowledge Management Process* 2013; 3(4): 15–33.

59. Kushner H. *Introduction to Stochastic Control*. Holt, Rinehart and Winston . 1971.

60. Kárný M, Kroupa T. Axiomatisation of fully probabilistic design. *Inf. Sci.* 2012; 186(1): 105-113.

AUTHORS' BIOGRAPHIES

Miroslav Kárný. Ing. (MSc) Czech Technical University (CTU) Prague, 1973; CSc. (PhD) 1978, DrSc. (DSc) 1990, both the Institute of Information Theory and Automation, the Czechoslovak Academy of Sciences (UTIA CAS) employing him since 1973 in the Department of Adaptive Systems (AS). *Research:* conceptual, theoretical and algorithmic aspects of AS based on Bayesian dynamic decision making (DM) and its original fully probabilistic extension. *Teaching:* the advanced course on DM, CTU since 1991; supervision of 13 defended PhD students (+6 co-supervision) and numerous Bc, MSc theses and research projects. *Publications:* 1 monograph, 6 edited books, ≈ 400 works (≈ 10 chapters, ≈ 100 articles); for the list after 1989 see http://www.utia.cz/people/karny.

Marko Ruman. Ing. (MSc) mathematical engineering at Faculty of Nuclear Sciences and Physical Engineering (FNSPE CTU) Prague, 2018; PhD Student 2018 – at FNSPE CTU, supervisor Dipl. Eng T.V. Guy, PhD. Since 2018 research assistant at AS UTIA CAS. *Research:* adaptive decision making, Markov decision processes, deep reinforcement learning, transferring knowledge between reinforcement learning tasks. *Theoretical skills:* advanced statistical and numerical methods, advanced functional analysis, dynamic decision making, Monte Carlo methods, classical physics, relational data bases. *Programming skills:* since 2006 active design of web-sites using HTML, CSS, PHP, MySQL and JavaSript, see e.g. www.ladywander.com; also an active user of Matlab, C++, Python. *Languages:* fluent English, Slovak, basic Italian, French and German. *Publications:* co-author of 4 publications, see https://www.utia.cas.cz/node/3090.

How to cite this article: Kárný M., Ruman M., (2020), Mixture Ratio Modelling of Dynamic Systems, *Int J Adapt Control Signal Processing*, *2021; @@:@@–@@*.

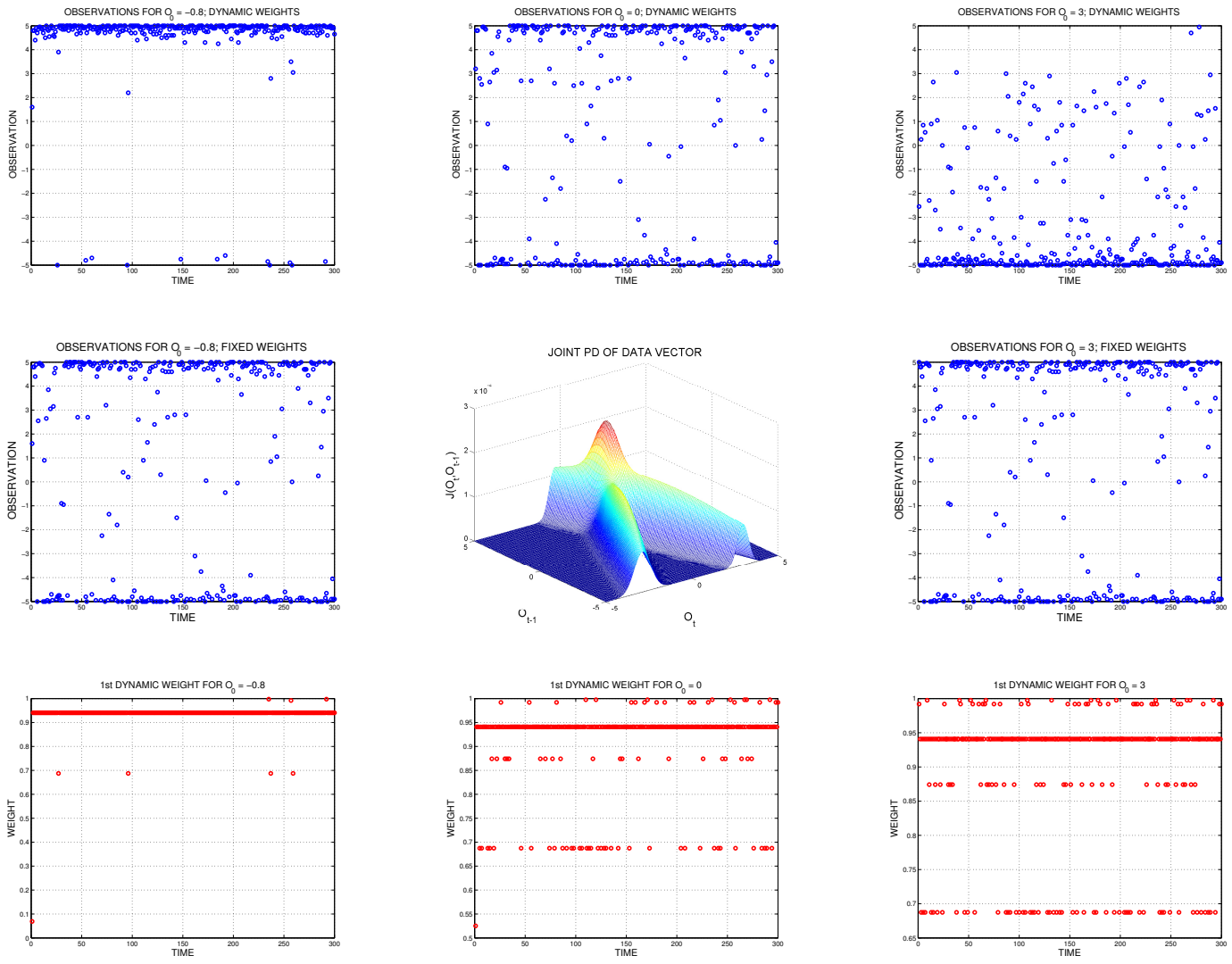

FIGURE 1 Simulation of the system E (27). Its joint pd is shown in the middle. Trajectories in respective columns differ in the initial value of the regression vector $\rho_1 = O_0 \in \{-0.8,\ 0,\ 3\}$, respectively, while pseudo-random noise realisation is the same in all cases. The 1st row of figures shows the realised observations O_t, $t \in \boldsymbol{t} \equiv \{1, \dots, 300\}$. It demonstrates the strong dependence of the observed trajectories on the initial condition $\rho_1 = O_0$. The 2nd row shows realised observations when the same components as in (27) are used as the conditional pds, i.e. when the standard mixture with constant component weights is simulated. The respective trajectories are almost identical. It has allowed us to replace the case corresponding $\rho_1 = O_0 = 0$ by the simulated joint pd (27). The 3rd row shows evolution of dynamic weights $(\mathsf{w}_{t;1}(\rho_t))_{t\in\boldsymbol{t}}$ (14) corresponding to the realised observations shown in the 1st row of figures. Again, the strong dependence on the initial $\rho_1 = O_0$ is clearly seen.

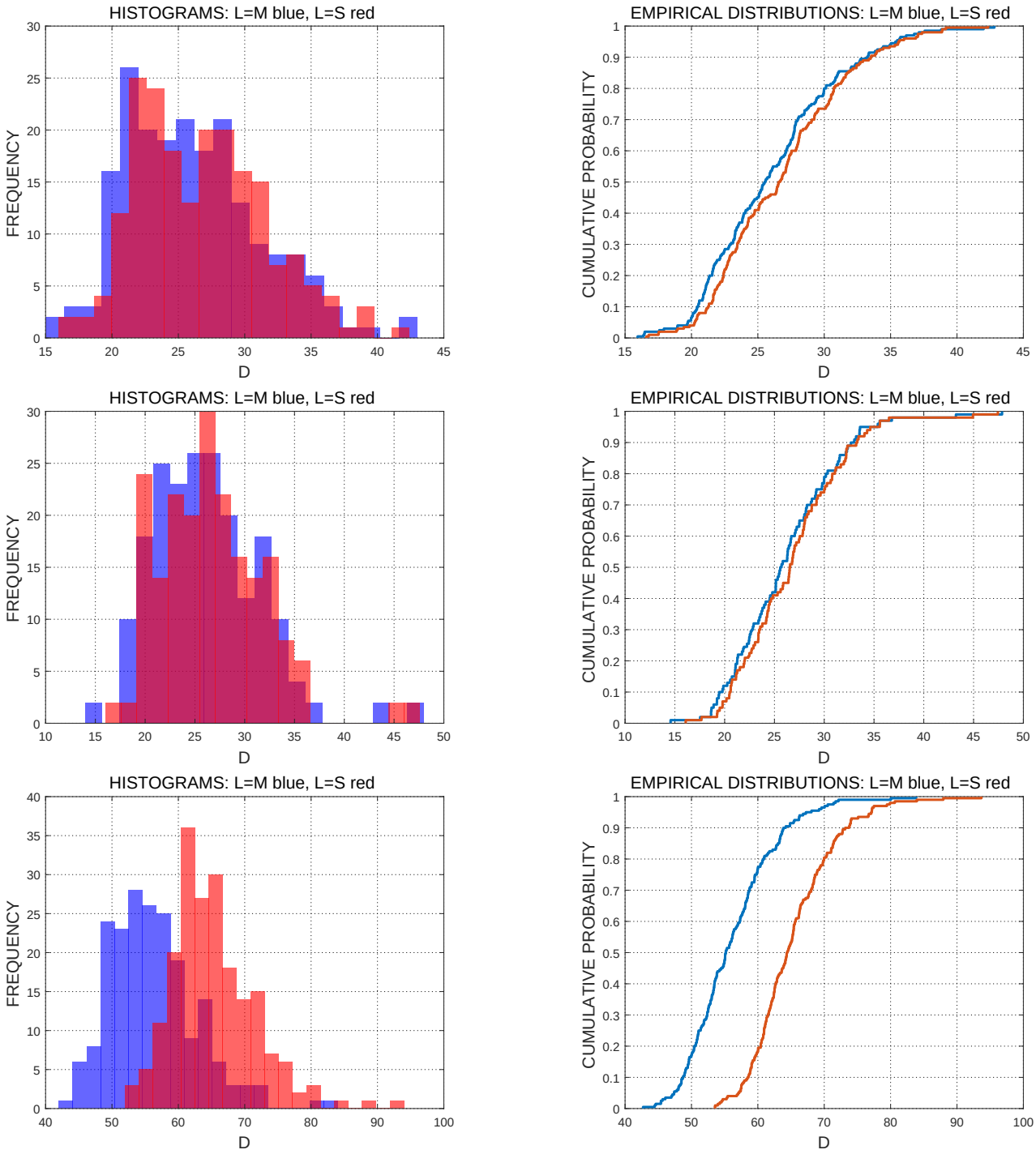

FIGURE 2 Evaluation of Kullback-Leibler divergences $\mathsf{D}(\mathsf{E}||\mathsf{L})$ of the observation predictors corresponding to the simulated systems E to the learnt predictors, see (30), in the Monte Carlo study. The left column shows histograms, the right one shows the empirical distribution functions on which the favourable shift to smaller divergence values is better seen. Colors distinguish the inspected learnt predictor $\mathsf{L} \in \{\mathsf{M},\mathsf{S}\} = \{\text{mixture ratio}, \text{standard mixture}\} = \{\text{blue}, \text{red}\}$. The 1st row deals with the simulated system $\mathsf{E} = \mathsf{M} =$ mixture ratio, with its parameter Θ_{M} varied in 200 MC runs; the 2nd row shows the results with the simulated system $\mathsf{E} = \mathsf{S} =$ standard mixture, with its parameter Θ_{S} varied in 200 MC runs; the 3rd row shows results for $\mathsf{E} = \underline{\mathsf{M}} =$ mixture ratio, with a fixed parameter $\underline{\Theta}_{\mathsf{M}}$ causing truly dynamic evaluation of the weights $\mathsf{w}_c(\rho_t)$ (14), there the MC study runs over 200 different noise realisations.

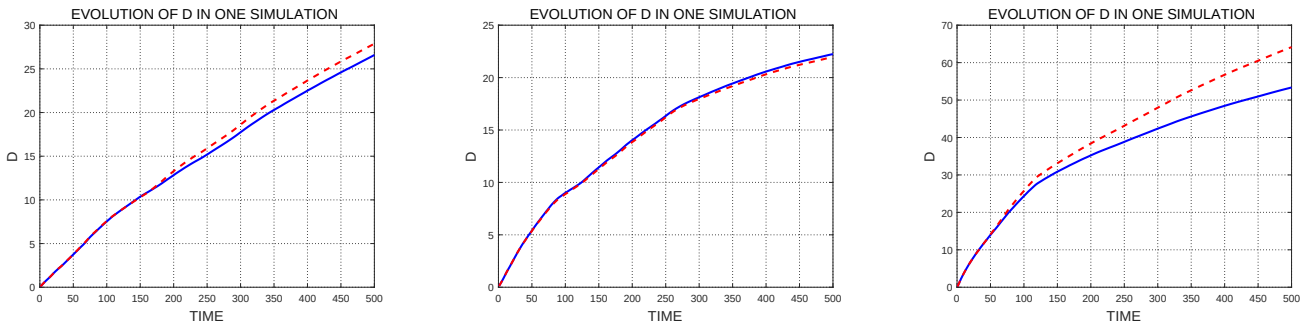

FIGURE 3 Time evolutions of the Kullback-Leibler divergences $\mathsf{D}_t(\mathsf{E}||\mathsf{L})$ (30) od predictors corresponding to the simulated system E and learnt predictors $\mathsf{L} \in \{\mathsf{M},\mathsf{S}\} = \{\text{full blue, dashed red}\} = (\text{mixture ratio, standard mixture})$. The respective figures from left to right correspond to: $\mathsf{E} = \underline{\mathsf{M}} =$ simulated mixture ratio with a randomly chosen $\underline{\Theta}_{\mathsf{M}}$; $\mathsf{E} = \underline{\mathsf{S}} = mboxsimulated standard mixture$ with a randomly chosen $\underline{\Theta}_{\mathsf{S}}$; $\mathsf{E} = \underline{\mathsf{M}} =$ simulated mixture ratio with with the fixed $\underline{\Theta}_{\mathsf{M}}$ causing a truly dynamic weighting $\mathsf{w}_c(\rho_t)$ of respective components (14).