Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering

Mixture Ratios for Decision Making

Podílové směsové modely pro rozhodování

Master's thesis

Author:	**Bc. Marko Ruman**
Supervisor:	**Ing. Miroslav Kárný, DrSc.**
Academic year:	2017/2018

Acknowledgment:
I would like to thank Ing. Miroslav Kárný, DrSc. for his expert guidance and language assistance.
The research was supported by GAČR, the research project GA16-09848S.

Moreover, I would like to dedicate this thesis to my lovely girlfriend Tereza Bosáková, to my closest family - my mother Božena Rumanová and my two sisters Petra Rumanová and Monika Rumanová and to my long-term friends Viktor Kutný, Maroš Lukačišin and Martin Juríček as they all provided me with a lot of love, support and tons of fun during working on this thesis.

Author's declaration:
I declare that this thesis is entirely my own work and I have listed all the used sources in the bibliography.

Prague, May 4, 2018

Bc. Marko Ruman

Název práce:

Podílové směsové modely pro rozhodování

Autor: Bc. Marko Ruman

Obor: Matematické inženýrství

Druh práce: Diplomová práce

Vedoucí práce: Ing. Miroslav Kárný, DrSc., Oddělení Adaptivních systémů, Ústav Teorie Informace a Automatizace

Abstrakt: Směsi konečného počtu hustot pravděpodobností s komponentami z exponenciální rodiny slouží jako flexibilní modely vícerozměrných systémů. Avšak, s výjimkou pár úzce specializovaných případů, používané dynamické modely předpokládají datově nezávislé váhy jednotlivých směsových komponent. Použití takových vah je nelogické a omezuje praktickou použitelnost daných modelů. Tato práce řeší dané omezení použitím podílu konečných směsí jako parametrického modelu. V práci je použití tohoto modelu motivováno, je vybudováno jeho rekurzivní bayeovské odhadování a je ukázaná jeho vhodnost i pro modelování obecných diskrétně-spojitých systémů. Dále je provedeno porovnání podílových směsí se standardními směsmi s konstantními vahami, a to jak na simulovaných datech, tak na reálných datech vývoje cen futures. Pro případ podílu směsí Markovských řetězců je vybudován návrh rozhodovací strategie.

Klíčová slova: Dynamické systémy, bayesovské učení, směsové modely, podíl směsí

Title:

Mixture Ratios for Decision Making

Author: Bc. Marko Ruman

Abstract: Finite mixtures of probability densities with components from exponential family serve as flexible parametric models of high-dimensional systems. However, with a few specialized exceptions, these dynamic models assume data-independent weights of mixture components. Their use is illogical and restricts the modelling applicability. The work overcomes this restriction by considering ratios of finite mixtures as parametric models. It motivates them and elaborates their approximate Bayesian recursive estimation and shows its ability to model also general discrete-continuous systems. It also compares the mixture ratio model with the standard, fixed-weight mixtures, either on simulation data as well as on real futures price data. For the case of mixture ratio of Markov chains, the design of decision strategy is build.

Key words: Dynamic systems, Bayesian learning, mixture models, mixture ratio

Contents

Introduction

Decision making (DM) is a targeted choice among the available options[1]. This broadly understood DM covers many traditional fields including machine learning, [23], signal processing, [24], estimation and filtering, [17], hypothesis testing, [14], classification and pattern recognition, [11], knowledge sharing, [22], reinforcement learning, [27], control, [13], etc. The amount of relevant results is excessive as seen, e.g., in [16]. To cover all of the mentioned fields, this work uses its own vocabulary to formalize a general DM problem.

Any general solution of DM problems leads to a strategy, a collection of decision rules mapping the knowledge on actions, [26]. The chosen strategy should meet the DM aim in the best way under the encountered circumstances. Bayesian DM, [25], proved to be a powerful methodology of DM facing incomplete knowledge, uncertainty and randomness of the dynamic system to which actions relate. This work stays within it. Bayesian DM relates DM consequences to the acquired knowledge and models the used actions by conditional probabilities, here described by conditional probability densities (pd) or conditional probability functions (pf).

Generally, actions are chosen recursively while enriching the available knowledge. This gives a chance to learn, i.e. to improve gradually the system model. The learning redistributes belief (probability of) into the model adequacy within the set of considered models, [9]. It gradually extracts the knowledge and can be used in data-streams processing or employed as a feature extractor.

The need for learning arises whenever a good model is a priori unknown. A learning is possible iff the learnt relations practically do not change during the knowledge accumulation. This is assumed and a set of parametric models, distinguished by a constant multivariate parameter serving as a "pointer" to its respective members, is considered. Which one points to the best model is a priori unknown. Bayesian learning offers the unambiguous deductive way, Bayes' rule, [10], how to redistribute the belief concerning the model quality. It accumulates knowledge into the posterior pd and provides the predictive pd, which serves as the system model needed for DM. The achievable modeling and thus DM quality are predetermined by the considered set of parametric models.

This work offers *ratios* of finite mixtures, [4], with components from exponential family (EF), [7], as such black-box, [12], universally approximating, [15], models. The recursive learning of ratios of finite mixtures is inevitably approximate and consequently endangered by accumulation of approximation errors. The approximate, but feasible recursive Bayesian learning of models, equipped with a counteractor of approximation errors, from this *extremely flexible but yet unconsidered model set*, is developed, [30].

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

[1] A part of this work was submitted for publication [30].

the occurrence array blows up with the number of possible occurrence instances. Then, the observations insufficiently populate it.

The components of mixture of Markov chains belongs to EF, thus such high-order Markov chain can be well approximated by the developed mixture ratio model, which will counteract the curse of dimensionality of general Markov chain models, see [29] for comparison. This work develops the mentioned approximate Bayesian learning for the mixture ratios of Markov chains with conjugated Dirichlet distribution on its parameters.

Another important and commonly solved task is modelling of continuous variables. Almost any continuous probability density can be approximated by a finite mixture of Gaussian distribution to an arbitrary precision, [4]. Again, Gaussian components belong to EF, hence the approximate Bayesian learning of mixture ratios with such components is build in this work. These two types of EF components can be easily combined and can provide a general parametric models of mixed discrete-continuous systems.

Besides, this work provides comparison of learning of mixture ratio and standard mixture (developed in [18]) of Markov chain components in simulation set-up as well as on real futures trading data, where the suitability of mixture ratio model for decision making is demonstrated. The necessary design of decision strategies for the mixture ratio of Markov chain components is build.

Layout of the work is as follows: Chapter 1 introduces mathematical notation, functions and theorems used in the text. Chapter 2 formalizes a general model of a dynamic system and summarizes its approximate recursive Bayesian learning. Chapter 3 introduces finite mixture ratio as a parametric model. Chapter 4 describes the use of general learning algorithm from Chapter 2 for the finite mixture ratio parametric model. Chapter 5 introduces mixture ratio of Markov chains and builds its learning. Chapter 6 specifies mixture ratio of Gaussian components and derives its learning. Chapter 7 provides simulation and practical examples comparing the mixture ratio model with standard mixture model, where also the design of decision strategies for the mixture ratio of Markov chain components is created. Conclusion summarizes results of this work and outlines open questions for the further examination. Appendix provides derivation of some used formulas.

Chapter 1

Mathematical Preliminaries

This chapter introduces notation as well as theorems and functions used throughout the text. The list of notation symbols used in text:

- $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{R}^+$ stand for sets of *natural*, *integer*, *real* and *positive real* numbers, respectively,
- $\mathbb{R}^n$ denotes n-dimensional vector space, $\mathbb{R}^{nm}$ denotes vector space of matrices with dimensions $n \times m$; $n, m \in \mathbb{N}$,
- $a = (a_1, ..., a_n) \in \mathbb{R}^n$ stands for a n-dimensional column vector,
- for a matrix $A \in \mathbb{R}^{nn}$, $\mathrm{tr}(A)$ denotes its trace, $\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$,
- the bold symbols, for instance $\mathbf{A}$, will stand for the sets of all possible values of its members $A \in \mathbf{A}$,
- $|\mathbf{A}|$ denotes cardinality of $\mathbf{A}$,
- $\chi_{\mathbf{A}}(x)$ is a characteristic function of a set $\mathbf{A}$

$$\chi_{\mathbf{A}}(x) = \begin{cases} 1, & \text{if } x \in \mathbf{A}, \\ 0, & \text{otherwise,} \end{cases}$$

- $[a, b]^n$ stands for a n-ary Cartesian product, where $a, b \in \mathbb{R}$, $a < b$, $n \in \mathbb{N}$ and $[a, b]$ is a real interval,
- $\mathbb{N}^n$ stands for the n-ary Cartesian product of natural numbers, i.e. $\mathbb{N}^n = \bigtimes_{i=1}^{n} \mathbb{N}$,
- $\langle a \,|\, b \rangle$ denotes a scalar product
  - for vectors $a, b \in \mathbb{R}^n$, $\langle a \,|\, b \rangle$ is the dot product, $\langle a \,|\, b \rangle = \sum_{i=1}^{n} a_i b_i$
  - for matrices $a, b \in \mathbb{R}^{nn}$, $\langle a \,|\, b \rangle = \mathrm{tr}(ab)$
- 0^0 is defined as $0^0 = 1$.

Throughout the text, the functions with different arguments are assumed to be different, e.g. $\mathrm{B}(\psi_t)$ and $\mathrm{B}(\psi_{t;c})$.

If it is not necessary to distinguish realization and random variable, for the probability of the observation of a discrete random variable $\tilde{x}$ with a value $\tilde{x} = x$, the symbol $\mathrm{P}(x)$ will be used[1]. The conditional

[1] For discrete valued random variables $\mathrm{P}(x)$ is a probability function (pf), for a continuous random variables $\mathrm{P}(x)$ is used for the probability density function (pd) of a random variable $\tilde{x}$. Alternatively, the symbols $\mathrm{M}(x), \mathrm{S}(x), \mathrm{J}(x)$ will be used for pfs or pds.

probability of the observation of a random variable $\tilde{x}$ with a value $\tilde{x} = x$, which depends on a random variable $\tilde{y}$ with a known realization $\tilde{y} = y$, the symbol $\mathrm{P}(x|y)$ is used, more precisely:

$$\mathrm{P}(x|y) = \frac{\mathrm{P}(x,y)}{\mathrm{P}(y)},$$

where $\mathrm{P}(y) > 0$ denotes the probability, that random variable $\tilde{y}$ will be realized with a value $\tilde{y} = y$, similarly $\mathrm{P}(x,y)$ stands for the probability, that random variables $\tilde{x}$ and $\tilde{y}$ will be realized with a value $\tilde{y} = y$ and $\tilde{x} = x$.

The conditional probability $\mathrm{P}(a,b|c)$ where $\tilde{a},\tilde{b}$ and $\tilde{c}$ are random variables, can be calculated via chain rule:

$$\mathrm{P}(a,b|c) = \mathrm{P}(a|b,c)\mathrm{P}(b|c)\,. \tag{1.1}$$

The marginal probability $\mathrm{P}(a|c)$ can be calculated as:

$$\mathrm{P}(a|c) = \int_{\mathbf{b}} \mathrm{P}(a,b|c)db\,, \tag{1.2}$$

where $\mathbf{b}$ is a set of all values of the random variable $\tilde{b}$.

Let $t \in \mathbf{t}$ be a time index from the set of the time indices $\mathbf{t} \subset \mathbb{N}$. The notation $\mathrm{P}_t(\Theta) = \mathrm{P}(A_t|B_t,\Theta)$ will be used whenever realization of A_t, B_t is inserted into the conditional pf/pd $\mathrm{P}(A_t|B_t,\Theta)$, i.e. when it is treated as likelihood function, Θ is usually a parameter vector.

Definition 1 (Statistic)**.** A *statistic* V is here (generally, the range of the statistic can be arbitrary) a finite-dimensional, measurable function of a data sequence
$D^t = \{D_t, D_{t-1}, ..., D_1, D_0\} \in \mathbf{D}^t$:

$$\mathrm{V} : \mathbf{D}^t \mapsto \mathbb{R}^n$$

The notation $\mathrm{V}_t = \mathrm{V}(D^t)$ will be used. □

Definition 2 (Sufficient statistic, [21])**.** A statistic V is *sufficient* for a parameter Θ iff:

$$\mathrm{P}(\Theta|\mathrm{V}_t, D^t) = \mathrm{P}(\Theta|\mathrm{V}_t)$$

□

Definition 3 (Exponential family)**.** Let Θ be a finite-dimensional parameter vector (or matrix), $x = (x_1, ..., x_k)$ a k-dimensional random variable, then the conditional pd $\mathrm{P}(x|\Theta)$ belongs to *Exponential family* of distributions iff has the following form:

$$\mathrm{P}(x|\Theta) = \exp\langle \mathrm{B}(x) \,|\, \mathrm{C}(\Theta)\rangle, \tag{1.3}$$

where B and C are known finite-dimensional vector or matrix functions, $\langle \mathrm{B}(x) \,|\, \mathrm{C}(\Theta)\rangle$ denotes a scalar product. □

Definition 4 (Gamma function)**.** The gamma function Γ is deifned for $x \in \mathbb{R}^+$ as follows

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

□

Remark 1. Gamma function satisfies $\Gamma(x+1) = x\Gamma(x)$, [19]. □

Definition 5 (Digamma function)**.** Digamma function $\psi(x)$ is defined for $x \in \mathbb{R}^+$ as follows:

$$\psi(x) = \frac{\partial}{\partial x}\ln(\Gamma(x)).$$

□

Remark 2. Digamma function satisfies $\psi(x+1) = \frac{1}{x} + \psi(x)$, [19]. □

Definition 6 (Beta function,[19])**.** Beta function $\mathrm{Be}(x,y) : \mathbb{R}^+ \bigtimes \mathbb{R}^+ \mapsto \mathbb{R}^+$ is defined as follows

$$\mathrm{Be}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt,$$

or equivalently

$$\mathrm{Be}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

The multivariate beta function $\mathrm{Be}(v) : \bigtimes_{i=1}^{n} \mathbb{R}^+ \mapsto \mathbb{R}^+$ is defined as follows

$$\mathrm{Be}(v) = \frac{\prod_{i=1}^{n}\Gamma(v_i)}{\Gamma(\sum_{i=1}^{n} v_i)}.$$

□

Definition 7 (Kronecker delta function)**.** The Kronecker delta function δ_{xy} is defined on $\mathbb{Z} \bigtimes \mathbb{Z}$ as

$$\delta_{xy} = \begin{cases} 1 & x = y, \\ 0 & x \neq y. \end{cases} \tag{1.4}$$

For a set $\mathbf{c} = \{1,..,|\mathbf{c}|\}$, the symbol $\delta(c)$ will stand for the following vector:

$$\delta(c) = (\delta_{1c},...,\delta_{|\mathbf{c}|c}).$$

□

Definition 8 (Kullback–Leibler divergence, [1])**.** *Kullback–Leibler divergence* P to Q, KL(P||Q), is a real-valued function mapping of two pfs[2] P and Q to a real number and it is defined as follows:

$$\mathrm{KL}(\mathrm{P}||\mathrm{Q}) = \sum_{a \in \mathbf{a}} \mathrm{P}(a)\ln\frac{\mathrm{P}(a)}{\mathrm{Q}(a)}. \tag{1.5}$$

□

Definition 9 (Kerridge inaccuracy, [20])**.** *Kerridge inaccuracy* K(P||Q) is a real-valued function mapping of two pds P and Q to a real number and it is defined as:

$$\mathrm{K}(\mathrm{P}||\mathrm{Q}) = -\int_{\mathbf{\Theta}} \mathrm{P}(\Theta)\ln(\mathrm{Q}(\Theta))d\Theta. \tag{1.6}$$

□

[2] It can also be defined for pds via its integral form, but this work uses only the version for pfs.

Theorem 1 (Bayes' rule, [10])**.** Let a and b be continuous random variables[3] and $\mathbf{a}$ a set of all possible values of random variable a. Then the conditional pd $\mathrm{P}(a|b)$ can be expressed as follows (for $\int_{\mathbf{a}} \mathrm{P}(b|a)\mathrm{P}(a)da > 0$):

$$\mathrm{P}(a|b) = \frac{\mathrm{P}(b|a)\mathrm{P}(a)}{\int_{\mathbf{a}} \mathrm{P}(b|a)\mathrm{P}(a)da} \propto \mathrm{P}(b|a)\mathrm{P}(a)\,, \tag{1.7}$$

where $\propto$ denotes equality up to normalizing constant. □

[3] In the text, Bayes' rule will be used for continuous random variable a, thus the set $\mathbf{a}$ is a subset of the real numbers $\mathbb{R}$, the random variable $\tilde{b}$ can be either a discrete valued or continuous.

Chapter 2

Dynamic System Model and Its Approximate Learning

A model of a dynamic system in DM theory is formalized in Bayesian way as follows. Observations $O_t \in \mathbf{O}$, stimulated by actions $A_t \in \mathbf{A}$, are made at the discrete-time moments $t \in \mathbf{t} = \{1, 2, ..., |\mathbf{t}|\}$. A data sequence $D^t \in \mathbf{D}$ is formed by data records: $D^t = \{D_t, D_{t-1}, ..., D_1, D_0\}$, where a data record $D_t = (O_t, A_t)$, $t \in \mathbf{t}$, is formed by an observation O_t and an action A_t, D_0 stands for the initial knowledge about the system. considered variables are random so that modelling must be done by pds.

To describe a behaviour of the dynamic system, the parametric pd of the observation O_t is used.

Definition 10 (Parametric system model)**.** The parametric system model is a pd of the observation $O_t \in \mathbf{O}$ conditioned on the data sequence $D^{t-1} \in \mathbf{D}$, the action $A_t \in \mathbf{A}$ and on an unknown parameter $\Theta \in \mathbf{\Theta}$:

$$\mathrm{M}(O_t|A_t, D^{t-1}, \Theta), \tag{2.1}$$

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

The goal of DM theory is to influence the dynamic system by taking appropriate actions to achieve desired observations (generally, states of the modelled system). To describe the next observation of the system, the pd $\mathrm{P}(O_t|D^{t-1})$ is used. Using (1.1) and (1.2) and incorporating the parametric system model (2.1), it can be expressed as follows:

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

where $\pi(A_t|D^{t-1}, \Theta)$ is a pd that describes a randomized decision rule[1], $\mathrm{P}(\Theta|D^{t-1})$ is a posterior pd storing actual knowledge about the parameter $\Theta \in \mathbf{\Theta}$; the notation $\mathrm{P}_{t-1}(\Theta) = \mathrm{P}(\Theta|D^{t-1})$ will be used. Generally, the assumption of *natural conditions of control*, [32], is adopted, i.e. $\pi(A_t|\Theta, D^{t-1}) = \pi(A_t|D^{t-1})$. It means that parameters are unknown to considered decision rules.

[1] The action generator influences the pd ¶$(O_t|D^{t-1})$ via decision rule $\pi(A_t|D^{t-1}, \Theta)$ according to a given goal.

After the interaction with the dynamic system, a new data record $D_t = (O_t, A_t)$ is created and $\mathrm{P}_{t-1}(\Theta)$ is updated via Bayes' rule (1.7) as follows[2]:

$$\tilde{\mathrm{P}}_t(\Theta) = \frac{\mathrm{M}_t(\Theta)\pi(A_t|D^{t-1})\mathrm{P}_{t-1}(\Theta)}{\int_{\boldsymbol{\Theta}} \mathrm{M}_t(\Theta)\pi(A_t|D^{t-1})\mathrm{P}_{t-1}(\Theta)d\Theta} = \frac{\mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)}{\int_{\boldsymbol{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta} \propto \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta), \qquad \forall\Theta \in \boldsymbol{\Theta}. \tag{2.2}$$

To ensure the computational feasibility of updating the pd $\mathrm{P}_{t-1}(\Theta)$ during the whole interaction with a dynamic system, the set of feasible pds $\mathbf{P}$ needs to be considered. Generally, the posterior pd $\tilde{\mathrm{P}}_t(\Theta)$ obtained by (2.2) does not belong to the set of feasible pds $\mathbf{P}$ and with growing $t \in \mathbf{t}$ it can become more and more complex function of Θ. Therefore, a projection of $\tilde{\mathrm{P}}_t(\Theta)$ on pd $\hat{\mathrm{P}}_t(\Theta) \in \mathbf{P}$ from this set of feasible pds has to be made.

Works [1] and [2] suggest, that the adequate projection $\hat{\mathrm{P}}_t(\Theta)$ is the minimizer of Kerridge inaccuracy (see Definition 9):

$$\hat{\mathrm{P}}_t(\Theta) = \underset{\mathrm{P}\in\mathbf{P}}{\mathrm{K}}(\tilde{\mathrm{P}}_t||\mathrm{P}) = \underset{\mathrm{P}\in\mathbf{P}}{\operatorname{argmin}} \int_{\boldsymbol{\Theta}} -\tilde{\mathrm{P}}_t(\Theta)\ln(\mathrm{P}(\Theta))d\Theta \tag{2.3}$$

The updating from $\mathrm{P}_{t-1}(\Theta)$ to $\hat{\mathrm{P}}_t(\Theta)$ can be interpreted as the application of Bayes' rule (1.7) on $\mathrm{P}_{t-1}(\Theta)$ but using an unknown, different model than $\mathrm{M}_t(\Theta)$. Therefore, using $\hat{\mathrm{P}}_t(\Theta)$ as a prior pd for the next learning step may, in general, cause divergence $\hat{\mathrm{P}}_t(\Theta)$ from $\tilde{\mathrm{P}}_t(\Theta)$ obtained via (2.2) without projection (2.3). The solution preventing from the divergence is to include a data-depending forgetting factor $\lambda_t \in [0, 1]$. [3] implies the most plausible choice of the forgetting factor λ_t.

Algorithm 1 (Learning algorithm)**.** One learning step of the learning algorithm is summarized as follows:

1) $\tilde{\mathrm{P}}_t(\Theta) \propto \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)$, 2) $\hat{\mathrm{P}}_t = \underset{\mathrm{P}\in\mathbf{P}}{\operatorname{argmin}}\, \mathrm{K}(\tilde{\mathrm{P}}_t||\mathrm{P})$,

3) $\mathrm{P}_t(\Theta) \propto \hat{\mathrm{P}}_t^{\lambda_t}(\Theta)\mathrm{P}_{t-1}^{1-\lambda_t}(\Theta)$, where $\lambda_t = \dfrac{\left(\int_{\boldsymbol{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta\right)^2}{\int_{\boldsymbol{\Theta}} \mathrm{M}_t^2(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta}$,

where $\mathrm{P}_{t-1}(\Theta)$ is the prior pd, $\mathrm{K}(\tilde{\mathrm{P}}_t||\mathrm{P})$ denotes Kerridge's inaccuracy (2.3) and $\mathrm{P}_t(\Theta)$ is used as a prior pd for the next learning step. □

[2] This particular form of Bayes' rule is valid only under the mentioned natural conditions of control.

Chapter 3

Finite Mixture Ratio Model

To achieve computational feasibility of the recursive learning summarized in Algorithm 1., the following simplifying assumption is made about the model (2.1):

Assumption 1 (Markov property of the system model)**.** The system model (2.1) is assumed to be time-invariant n-order Markov model ($n \in \mathbb{N}$), i.e.

$$\mathrm{M}(O_t|A_t, D^{t-1}, \Theta) = \mathrm{M}(O_t|\psi_t, \Theta) \text{ , where} \tag{3.1}$$

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

The symbol $\phi_t \in \boldsymbol{\phi}$ will denote a data vector $\phi_t = (O_t, \psi_t)$. □

As stated in Introduction, this work focuses on modelling the dynamic system via mixtures of pds. There are many works on this topic, e.g. [4], however, most of them assume data-independence of the weights of the mixture components. The following approach overcome this limitation by introducing a ratio of finite mixtures.

The model (3.1) can be expressed as follows (using (1.1) and (1.2)):

$$\mathrm{M}(O_t|\psi_t, \Theta) = \frac{\mathrm{J}(O_t, \psi_t|\Theta)}{\int_{\mathbf{O}} \mathrm{J}(O_t, \psi_t|\Theta) dO}. \tag{3.2}$$

Almost any practically met time-invariant joint pd $\mathrm{J}(O_t, \psi_t|\Theta)$ in (3.2) can be approximated by a finite mixture of Gaussian pds, [19], to an arbitraty precision, [4]. A Gaussian pd belongs to the exponential family (EF) (see Definition 3), thus, following EF mixture form of $\mathrm{J}(O_t, \psi_t|\Theta)$ can be considered:

Definition 11 (Joint pd as a mixture)**.** The joint pd $\mathrm{J}(O_t, \psi_t|\Theta)$ is modelled via a mixture of pds from EF (see Definition 3) as follows:

$$\begin{aligned} \mathrm{J}(O_t, \psi_t|\Theta) = \mathrm{J}(\phi_t|\Theta) &= \sum_{c \in \mathbf{c}} \alpha_c \mathrm{J}_c(\phi_t|\omega_c) = \sum_{c \in \mathbf{c}} \alpha_c \exp \langle \mathrm{B}_c(\phi_t) \,|\, \mathrm{C}_c(\omega_c) \rangle \\ &= \sum_{c \in \mathbf{c}} \alpha_c \exp \langle \mathrm{B}_c(\phi_{t;c}) \,|\, \mathrm{C}_c(\omega_c) \rangle \mathrm{P}_c(\phi_{t;c}^{\mathsf{C}}), \end{aligned} \tag{3.3}$$

where

- $\mathrm{J}_c(\phi_t|\omega_c) = \mathrm{J}_c(O_t, \psi_t|\omega_c)$ is the c-th mixture component, it is a member of EF and it is a pd on $\boldsymbol{\phi}$. It is assumed to have the form: $\mathrm{J}_c(\phi_t|\omega_c) = \exp \langle \mathrm{B}_c(\phi_t) \,|\, \mathrm{C}_c(\omega_c) \rangle = \exp \langle \mathrm{B}_c(\phi_{t;c}) \,|\, \mathrm{C}_c(\omega_c) \rangle \mathrm{P}_c(\phi_{t;c}^{\mathsf{C}})$

with $\phi_{t;c}$ being a component-specific subvector of $\phi_{t;c}$ and $\mathrm{P}_c(\phi^{\mathsf{c}}_{t;c})$ is a non-parametrized pd on its complement $\phi^{\mathsf{c}}_{t;c}$ with respect to ϕ_t; $\mathrm{B}_c(\phi_t)$, $\mathrm{C}_c(\omega_c)$ and $\mathrm{B}_c(\phi_{t;c})$ are known, real-valued vector functions, $\mathrm{B}_c(\phi_{t;c}) = (\mathrm{B}_{c1}(\phi_{t;c})\ldots\mathrm{B}_{cm_c}(\phi_{t;c}))$, $\mathrm{C}_c(\omega_c) = (\mathrm{C}_{c1}(\omega_c),\ldots\mathrm{C}_{cm_c}(\omega_c))$

- $\mathbf{c} = \{1,...,|\mathbf{c}|\}$, $|\mathbf{c}|$ denotes the number of mixture components
- the parameter vector Θ has the following form: $\Theta = \left(\alpha_c, (\omega_{cj})_{j=1}^{m_c}\right)_{c\in\mathbf{c}} \in \left\{\boldsymbol{\alpha} \bigtimes_{c\in\mathbf{c}} \boldsymbol{\omega}_c\right\} = \boldsymbol{\Theta}$, where:
  - $\alpha = (\alpha_1,...,\alpha_{|\mathbf{c}|}) \in \boldsymbol{\alpha}$ is the weight vector, $\boldsymbol{\alpha} = \left\{(\alpha_1,...,\alpha_{|\mathbf{c}|}) \,|\, \alpha_c \geq 0, \forall c \in \mathbf{c},\; \sum_{c\in\mathbf{c}} \alpha_c = 1\right\}$
  - $\omega_c = (\omega_{c1},...,\omega_{cm_c}) \in \boldsymbol{\omega}_c$ is the component-specific parameter vector and m_c stands for the number of parameters of the c-th mixture component.

□

The general learning algorithm summarized in the Chapter 2, Algorithm 1, uses the system model $\mathrm{M}(O_t|\psi_t,\Theta)$. By inserting right-hand side of (3.3) into (3.2), the desired parametric system model is obtained:

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

$$= \sum_{c\in\mathbf{c}} \underbrace{\frac{\alpha_c \mathrm{W}_c(\psi_t,\omega_c)}{\sum_{d\in\mathbf{c}}\alpha_d \mathrm{W}_d(\psi_t,\omega_d)}}_{\mathrm{w}_c(\psi_t,\Theta_c)} \underbrace{\frac{\mathrm{J}_c(O_t,\psi_t|\omega_c)}{\mathrm{W}_c(\psi_t,\omega_c)}}_{\mathrm{M}_c(O_t|\psi_t,\Theta_c)} = \sum_{c\in\mathbf{c}} \mathrm{w}_c(\psi_t,\Theta_c)\mathrm{M}_c(O_t|\psi_t,\Theta_c), \tag{3.4}$$

where

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

Equation (3.4) shows, that the parametric model can be interpreted as the mixture with data-dependent weights, while the data-dependence is not arbitrary and *does not introduce new parameters*.

As majority of joint pds is approximated by mixture and model (3.4) is derived from it by deductive rules of probability theory, it "universally" approximate any probabilistic dynamic model operating on finite-dimensional data vector.

Chapter 4

Learning with Mixture Ratio Model

This chapters uses the general approximate learning algorithm proposed in Chapter 2 and summarized in Algorithm 1 for the parametric system model in the form (3.3).

First of all, the set of computationally feasible pds $\mathbf{P}$ has to be specified.

Each component $\exp\langle \mathrm{B}(\phi_{c,t}) \,|\, \mathrm{C}_c(\omega_c)\rangle$ in (3.3) depends only on a part of the parameter vector Θ. For each component, following *conjugated prior pds* to the respective components[1], determined by sufficient statistics $\mathrm{V}_{t;c}$ (see Definition 2), are considered:

$$\mathrm{P}(\omega_c|D^t) = \mathrm{P}(\omega_c|\mathrm{V}_t) = \mathrm{P}_t(\omega_c) = \frac{\exp\langle \mathrm{V}_{t;c} \,|\, \mathrm{C}_c(\omega_c)\rangle}{\mathrm{N}_c(\mathrm{V}_{t;c})}, \qquad \mathrm{V}_{t;c} = (\mathrm{V}_{t;c1}, \dots \mathrm{V}_{t;cm_c}) \tag{4.1}$$

$$\mathrm{N}_c(\mathrm{V}_{t;c}) = \int_{\omega_c} \exp\langle \mathrm{V}_{t;c} \,|\, \mathrm{C}_c(\omega_c)\rangle \, d\omega_c.$$

Contiguously, the component weights in (3.3) can be seen as a pd $\mathrm{M}(c_t = c|\Theta) = \alpha_c$ of an unobserved pointer $c_t \in \mathbf{c}$ to an active component. This pd belongs to EF:

$$\mathrm{M}(c_t|\Theta) = \exp\langle \delta(c_t) \,|\, \ln(\alpha)\rangle$$

$$\delta(c_t) = (\delta_{1c_t}, ..., \delta_{|\mathbf{c}|c_t}) \quad \ln(\alpha) = (\ln(\alpha_1), ..., \ln(\alpha_{|\mathbf{c}|})) \quad \langle \delta(c_t) \,|\, \ln(\alpha)\rangle = \sum_{c\in\mathbf{c}} \delta_{cc_t} \ln(\alpha_c),$$

where δ_{cc_t} is Kronecker delta function (see Definition 7).

The conjugated prior for this pd is Dirichlet distribution, determined by the sufficient vector statistic $\mathrm{v} = \mathrm{v}_t$, in the following form:

$$\mathrm{P}(\alpha|D^t) = \mathrm{P}(\alpha|\mathrm{v}_t) = \mathrm{P}_t(\alpha) = \frac{\exp\langle \mathrm{v}_t - 1 \,|\, \ln(\alpha)\rangle}{\mathrm{Be}(\mathrm{v}_t)}, \qquad \mathrm{v}_t = \left(\mathrm{v}_{t;1}, \dots \mathrm{v}_{t;|\mathbf{c}|}\right), \;\; \mathrm{v}_c > 0, \;\; \forall c \in \mathbf{c}, \tag{4.2}$$

where $\mathrm{Be}(\mathrm{v}_t)$ is the multi-dimensional Beta function (see Definition 6).

The marginal pds $\mathrm{P}_t(\alpha)$ (4.2), $\mathrm{P}_t(\omega_c)$, $c \in \mathbf{c}$ (4.1) do not determine the joint pd $\mathrm{P}_t(\Theta)$ unambiguously, [5]. This allows freedom in choosing the set of feasible joint pds $\mathbf{P}$. Since the information about relations between subparts of the parameter vector Θ is missing, the preferable choice of the set $\mathbf{P}$ is the set of joint pds $\mathrm{P}_t(\Theta)$ that have the product form of their marginals:

$$\mathbf{P} = \left\{ \mathrm{P}_t(\Theta) \;\middle|\; \mathrm{P}_t(\Theta) = \mathrm{P}_t(\alpha) \prod_{c\in\mathbf{c}} \mathrm{P}_t(\omega_c) \right\}. \tag{4.3}$$

[1] The term *conjugated pd* means that the pd preserves its functional form during the updating via Bayes' rule (1.7). The pds $P_t(\omega_c), c \in \mathbf{c}$, here are conjugated to the respective components $\mathrm{J}_c(O_t, \psi_t|\omega_c)$ from Definition 11.

Thus, a pd $\mathrm{P}_t(\Theta) \in \mathbf{P}$ (4.3) has the following form:

$$\mathrm{P}_t(\Theta) = \frac{\exp\langle \mathrm{v}_t - 1 \,|\, \ln(\alpha)\rangle \prod_{c\in\mathbf{c}} \exp\langle \mathrm{V}_{t;c} \,|\, \mathrm{C}_c(\omega_c)\rangle}{\mathrm{Be}(\mathrm{v}_t) \prod_{c\in\mathbf{c}} \mathrm{N}_c(\mathrm{V}_{t;c})} \tag{4.4}$$

Since each member of the set of feasible pds $\mathbf{P}$ (4.3) is determined by the statistics $\mathrm{V}_t = (\mathrm{v}_t, (\mathrm{V}_{t;c})_{c\in\mathbf{c}})$, it remains to convert the general algorithm summarized in Algorithm 1 to an algorithm updating V_{t-1} to V_t.

Updating step of Algorithm 1 The first step in general learning algorithm (Algorithm 1) is to update the feasible pd $\mathrm{P}_{t-1}(\Theta) \in \mathbf{P}$, determined by statistics $\mathrm{V}_{t-1} = (\mathrm{v}_{t-1}, (\mathrm{V}_{t-1;c})_{c\in\mathbf{c}})$, to $\tilde{\mathrm{P}}_t(\Theta)$ via Bayes' rule (1.7) as follows:

$$\begin{aligned}
\tilde{\mathrm{P}}_t(\Theta) &\overbrace{=}^{(1.7)} \frac{\mathrm{M}(O_t|\psi_t,\Theta)\mathrm{P}_{t-1}(\Theta)}{\int_\Theta \mathrm{M}(O_t|\psi_t,\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta} \qquad (4.5)\\
&\overbrace{\propto}^{(3.4)} \sum_{c\in\mathbf{c}} \frac{\alpha_c \mathrm{J}_c(O_t,\psi_t|\omega_c)\mathrm{P}_{t-1}(\Theta)}{\sum_{d\in\mathbf{c}} \alpha_d \mathrm{W}_d(\psi_t,\omega_d)}\\
&\overbrace{\propto}^{(4.4)} \sum_{c\in\mathbf{c}} \frac{\alpha_c \mathrm{P}_c(\phi_{t;c}^{\mathsf{c}})}{\sum_{d\in\mathbf{c}} \alpha_d \mathrm{W}_d(\psi_t,\omega_d)} \exp\langle \mathrm{B}_c(\phi_{t;c}) \,|\, \mathrm{C}_c(\omega_c)\rangle \exp\langle \mathrm{v}_{t-1} - 1 \,|\, \ln(\omega)\rangle \prod_{d\in\mathbf{c}} \exp\langle \mathrm{V}_{t-1;d} \,|\, \mathrm{C}_d(\omega_d)\rangle\\
&= \sum_{c\in\mathbf{c}} \frac{\mathrm{P}_c(\phi_{t;c}^{\mathsf{c}})}{\sum_{d\in\mathbf{c}} \alpha_d \mathrm{W}_d(\psi_t,\omega_d)} \exp\langle \mathrm{v}_{t-1} - 1 + \delta(c) \,|\, \ln(\omega)\rangle \prod_{d\in\mathbf{c}} \exp\langle \mathrm{V}_{t-1;d} + \delta_{cd}\mathrm{B}_c(\phi_{t;c}) \,|\, \mathrm{C}_d(\omega_d)\rangle\\
&= \sum_{c\in\mathbf{c}} \frac{\gamma_{t;c}\mathrm{P}_c(\phi_{t;c}^{\mathsf{c}})\mathrm{Q}_{t;c}(\Theta)}{\sum_{d\in\mathbf{c}} \alpha_d \mathrm{W}_d(\psi_t,\omega_d)},
\end{aligned}$$

where $\mathrm{Q}_{t;c}(\Theta)$, $c \in \mathbf{c}$, are pds of $\Theta \in \boldsymbol{\Theta}$ that equal

$$\begin{aligned}
\mathrm{Q}_{t;c}(\Theta) &= \frac{1}{\gamma_{t;c}} \exp\langle \mathrm{v}_{t-1} - 1 + \delta(c) \,|\, \ln(\alpha)\rangle \prod_{d\in\mathbf{c}} \exp\langle \mathrm{V}_{t-1;d} + \delta_{cd}\mathrm{B}_c(\phi_{t;c}) \,|\, \mathrm{C}_d(\omega_d)\rangle \qquad (4.6)\\
\gamma_{t;c} &= \mathrm{Be}(\mathrm{v}_{t-1} + \delta_c) \prod_{d\in\mathbf{c}} \mathrm{N}_d(\mathrm{V}_{t-1;d} + \delta_{cd}\mathrm{B}_c(\phi_{t;c})) \qquad c \in \mathbf{c}.
\end{aligned}$$

$\tilde{\mathrm{P}}_t(\Theta)$ then can be expressed as follows:

$$\tilde{\mathrm{P}}_t(\Theta) = \frac{\mathrm{H}_t(\Theta) \sum_{c\in\mathbf{c}} \gamma_{t;c}\mathrm{P}_c(\phi_{t;c}^{\mathsf{c}})\mathrm{Q}_{t;c}(\Theta)}{\mathrm{L}_{-1}}, \qquad \text{where} \tag{4.7}$$

$$\mathrm{H}_t(\Theta) = \mathrm{H}(\psi_t,\Theta) = \frac{1}{\sum_{d\in\mathbf{c}} \alpha_d \mathrm{W}_d(\psi_t,\omega_d)} \tag{4.8}$$

$$\mathrm{L}_{-1} = \sum_{c\in\mathbf{c}} \gamma_{t;c}\mathrm{P}_c(\phi_{t;c}^{\mathsf{c}}) \int_{\boldsymbol{\Theta}} \mathrm{H}_t(\Theta)\mathrm{Q}_{t;c}(\Theta)d\Theta \tag{4.9}$$

Projecting step of Algorithm 1 The second step of the learning algorithm (Algorithm 1) is to find a projection of the pd $\tilde{\mathrm{P}}_t$ to a pd $\hat{\mathrm{P}}_t$ from the feasible set $\mathbf{P}$ determined by the statistic $\hat{\mathrm{V}}_t = (\hat{\mathrm{v}}_t, (\hat{\mathrm{V}}_{t;c})_{c\in\mathbf{c}})$, $\hat{\mathrm{v}}_t = (\hat{\mathrm{v}}_{t;1}, ..., \hat{\mathrm{v}}_{t;|\mathbf{c}|})$, $\hat{\mathrm{V}}_{t;c} = (\hat{\mathrm{V}}_{t;c1}, ..., \hat{\mathrm{V}}_{t;cm_c})$, $c \in \mathbf{c}$.

Kerridge's inaccuracy (Definition 9) $\tilde{\mathrm{P}}_t$ to $\mathrm{P} \in \mathbf{P}$ reads:

$$
\begin{aligned}
\mathrm{K}(\tilde{\mathrm{P}}_t||\mathrm{P}) &= -\int_{\boldsymbol{\Theta}} \tilde{\mathrm{P}}_t(\Theta)\ln(\mathrm{P}(\Theta))d\Theta = -\int_{\boldsymbol{\Theta}} \tilde{\mathrm{P}}_t(\Theta)\ln\left(\frac{\exp\langle \mathrm{v}-1\,|\,\ln(\alpha)\rangle \prod_{c\in\mathbf{c}} \exp\langle \mathrm{V}_c\,|\,\mathrm{C}(\Theta_c)\rangle}{\mathrm{Be}(\mathrm{v})\prod_{c\in\mathbf{c}}\mathrm{N}_c(\mathrm{V}_c)}\right)d\Theta \\
&= \ln(\mathrm{Be}(\mathrm{v})) + \sum_{c\in\mathbf{c}} \ln(\mathrm{N}_c(\mathrm{V}_c)) - \left\langle \mathrm{v}-1 \,\middle|\, \int_{\boldsymbol{\Theta}} \ln(\alpha)\tilde{\mathrm{P}}_t(\Theta)d\Theta\right\rangle - \sum_{c\in\mathbf{c}}\left\langle \mathrm{V}_c \,\middle|\, \int_{\boldsymbol{\Theta}} \mathrm{C}_c(\omega_c)\tilde{\mathrm{P}}_t(\Theta)d\Theta\right\rangle \\
&= \ln(\mathrm{Be}(\mathrm{v})) + \sum_{c\in\mathbf{c}} \ln(\mathrm{N}_c(\mathrm{V}_c)) - \sum_{c\in\mathbf{c}}(\mathrm{v}_c - 1)\int_{\boldsymbol{\Theta}} \ln(\alpha_c)\tilde{\mathrm{P}}_t(\Theta)d\Theta \\
&\quad - \sum_{c\in\mathbf{c}}\sum_{j=1}^{m_c} \mathrm{V}_{cj}\int_{\boldsymbol{\Theta}} \mathrm{C}_{cj}(\omega_c)\tilde{\mathrm{P}}_t(\Theta)d\Theta.
\end{aligned}
\tag{4.10}
$$

Evaluation of Kerridge's inaccuracy (4.10) requires evaluation of the following integrals[2] (see (4.7)):

$$\mathrm{L}_{0c} = \sum_{d\in\mathbf{c}} \gamma_{t;d}\mathrm{P}_d(\phi^{\mathsf{c}}_{t;d})\int_{\boldsymbol{\Theta}} \ln(\alpha_c)\mathrm{H}_t(\Theta)\mathrm{Q}_{t;d}(\Theta)d\Theta \tag{4.11}$$

$$\mathrm{L}_{cj} = \sum_{d\in\mathbf{c}} \gamma_{t;d}\mathrm{P}_d(\phi^{\mathsf{c}}_{t;d})\int_{\boldsymbol{\Theta}} \mathrm{C}_{cj}(\omega_c)\mathrm{H}_t(\Theta)\mathrm{Q}_{t;d}(\Theta)d\Theta. \tag{4.12}$$

Inserting them into (4.10) the statistics $\hat{\mathrm{V}}_t$ determining the projection $\hat{\mathrm{P}}_t(\Theta)$ is obtained as follows:

$$\hat{\mathrm{V}}_t = \left(\hat{\mathrm{v}}_t, (\hat{\mathrm{V}}_{t;c})_{c\in\mathbf{c}}\right) = \underset{\left(\mathrm{v},(\mathrm{V}_c)_{c\in\mathbf{c}}\right)}{\operatorname{argmin}}\left[\ln(\mathrm{Be}(\mathrm{v})) + \sum_{c\in\mathbf{c}}\ln(\mathrm{N}_c(\mathrm{V}_c)) - \left\langle \mathrm{v}-1 \,\middle|\, \frac{\mathrm{L}_0}{\mathrm{L}_{-1}}\right\rangle - \sum_{c\in\mathbf{c}}\left\langle \mathrm{V}_c \,\middle|\, \frac{\mathrm{L}_c}{\mathrm{L}_{-1}}\right\rangle\right]$$

$$\hat{\mathrm{v}}_t = \underset{\mathrm{v}}{\operatorname{argmin}}\left(\ln(\mathrm{Be}(\mathrm{v})) - \left\langle \mathrm{v}-1 \,\middle|\, \frac{\mathrm{L}_0}{\mathrm{L}_{-1}}\right\rangle\right) \tag{4.13}$$

$$\hat{\mathrm{V}}_{t;c} = \underset{\mathrm{V}_c}{\operatorname{argmin}}\left(\ln(\mathrm{N}_c(\mathrm{V}_c)) - \left\langle \mathrm{V}_c \,\middle|\, \frac{\mathrm{L}_c}{\mathrm{L}_{-1}}\right\rangle\right) \qquad \forall c \in \mathbf{c} \tag{4.14}$$

Thus, $|\mathbf{c}| + 1$ minimization tasks are solved. The minimisation (4.14) depends on specific form of N_c and will be done in Chapter 5 for Markov chains and in Chapter 6 for Gaussian components.

4.1 Numerical Approximation of L's

The function $\mathrm{H}_t(\Theta)$ (4.8) depends on the whole parameter vector $\Theta \in \boldsymbol{\Theta}$ and the integrals in (4.9), (4.11) and (4.12) cannot be expressed analytically. Light tails of pds $\mathrm{Q}_{t;d}(\Theta)$ make integrands in (4.9), (4.11) and (4.12) well localized, thus the approximation based on linear Taylor expansion of $\mathrm{H}_t(\Theta)$ around expected values of the parameter vector Θ (according to $\mathrm{Q}_{t;d}(\Theta)$, $d \in \mathbf{c}$) is used. The expected

[2] Please note, that L_{-1} (4.9), L_{0c} (4.11) and L_{cj} (4.12) depend on time t. To simplify the notation, this dependence is not stressed.

values of parameters equals (see (4.6)):

$$\tilde{\alpha}^d_{t;c} = \int_{\boldsymbol{\Theta}} \alpha_c \mathrm{Q}_{t;d}(\Theta) d\Theta = \int_{\alpha} \alpha_c \frac{\exp\langle \mathrm{v}_{t-1} - 1 + \delta(c) \,|\, \ln(\alpha)\rangle}{\mathrm{Be}(\mathrm{v}_{t-1} + \delta_d)} d\alpha = \frac{\mathrm{v}_{t-1;c} + \delta_{cd}}{\sum_{c\in\mathbf{c}} \mathrm{v}_{t-1;c} + 1} \qquad c,d \in \mathbf{c} \tag{4.15}$$

$$\tilde{\omega}^d_{t;cj} = \int_{\boldsymbol{\Theta}} \omega_{cj} \mathrm{Q}_{t;d}(\Theta) d\Theta = \int_{\omega_c} \omega_{cj} \frac{\exp\langle \mathrm{V}_{t-1;d} + \delta_{cd} \mathrm{B}_c(\phi_{t;c}) \,|\, \mathrm{C}_d(\omega_d)\rangle}{\mathrm{N}_d(\mathrm{V}_{t-1;d} + \delta_{cd} \mathrm{B}_c(\phi_{t;c})} d\omega_c \quad c,d \in \mathbf{c},\ j \in \{1,2,...,m_c\} \tag{4.16}$$

$$\tilde{\alpha}^d_t = (\tilde{\alpha}^d_{t;1}, \tilde{\alpha}^d_{t;2}, ..., \tilde{\alpha}^d_{t;|\mathbf{c}|}) \qquad \tilde{\omega}^d_{t;c} = (\tilde{\omega}^d_{t;c1}, \tilde{\omega}^d_{t;c2}, ..., \tilde{\omega}^d_{t;cm_c}) \qquad \tilde{\Theta}^d_t = (\tilde{\alpha}^d_t, (\tilde{\omega}^d_{t;c})_{c\in\mathbf{c}}).$$

Hence, the function $\mathrm{H}(\psi_t, \Theta)$ is approximated as follows (depending on particular pd $\mathrm{Q}_{t;d}(\Theta)$, $d \in \mathbf{c}$, used in the integration):

$$\begin{aligned}\mathrm{H}_t(\Theta) &\approx \mathrm{H}_t(\tilde{\Theta}^d_t) + \nabla \mathrm{H}_t(\tilde{\Theta}^d_t)(\Theta - \tilde{\Theta}^d_t) \\ &= \mathrm{H}_t(\tilde{\Theta}^d_t) + \sum_{c\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}^d_t)}{\partial \alpha_c}(\alpha_c - \tilde{\alpha}^d_{t;c}) + \sum_{c\in\mathbf{c}} \sum_{j=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}^d_t)}{\partial \omega_{cj}}(\omega_{cj} - \tilde{\omega}^d_{t;cj}),\end{aligned} \tag{4.17}$$

The L's are then computed as follows ($\forall c \in \mathbf{c}, j \in \{1,..,m_c\}$):

$$\begin{aligned}\mathrm{L}_{-1} &= \sum_{d\in\mathbf{c}} \gamma_{t;d} \mathrm{P}_d(\phi^{\mathrm{c}}_{t;d}) \int_{\boldsymbol{\Theta}} \Big(\mathrm{H}_t(\tilde{\Theta}^d_t) + \nabla \mathrm{H}_t(\tilde{\Theta}^d_t)(\Theta - \tilde{\Theta}^d_t)\Big) \mathrm{Q}_{t;d}(\Theta) d\Theta \\ &= \sum_{d\in\mathbf{c}} \gamma_{t;d} \mathrm{P}_d(\phi^{\mathrm{c}}_{t;d}) \mathrm{H}_t(\tilde{\Theta}^d_t),\end{aligned} \tag{4.18}$$

$$\begin{aligned}\mathrm{L}_{0c} &= \sum_{d\in\mathbf{c}} \gamma_{t;d} \mathrm{P}_d(\phi^{\mathrm{c}}_{t;d}) \int_{\boldsymbol{\Theta}} \ln(\alpha_c) \Big(\mathrm{H}_t(\tilde{\Theta}^d_t) + \nabla \mathrm{H}_t(\tilde{\Theta}^d_t)(\Theta - \tilde{\Theta}^d_t)\Big) \mathrm{Q}_{t;d}(\Theta) d\Theta \\ &= \sum_{d\in\mathbf{c}} \gamma_{t;d} \mathrm{P}_d(\phi^{\mathrm{c}}_{t;d}) \int_{\boldsymbol{\Theta}} \ln(\alpha_c) \Big(\mathrm{H}_t(\tilde{\Theta}^d_t) + \sum_{e\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}^d_t)}{\partial \alpha_e}(\alpha_e - \tilde{\alpha}^d_{t;e})\Big) \mathrm{Q}_{t;d}(\Theta) d\Theta,\end{aligned} \tag{4.19}$$

$$\begin{aligned}\mathrm{L}_{cj} &= \sum_{d\in\mathbf{c}} \gamma_{t;d} \mathrm{P}_d(\phi^{\mathrm{c}}_{t;d}) \int_{\boldsymbol{\Theta}} \mathrm{C}_{cj}(\omega_c) \Big(\mathrm{H}_t(\tilde{\Theta}^d_t) + \nabla \mathrm{H}_t(\tilde{\Theta}^d_t)(\Theta - \tilde{\Theta}^d_t)\Big) \mathrm{Q}_{t;d}(\Theta) d\Theta \\ &= \sum_{d\in\mathbf{c}} \gamma_{t;d} \mathrm{P}_d(\phi^{\mathrm{c}}_{t;d}) \int_{\boldsymbol{\Theta}} \mathrm{C}_{cj}(\omega_c) \Big(\mathrm{H}_t(\tilde{\Theta}^d_t) + \sum_{i=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}^d_t)}{\partial \omega_{ci}}(\omega_{ci} - \tilde{\omega}^d_{t;ci})\Big) \mathrm{Q}_{t;d}(\Theta) d\Theta.\end{aligned} \tag{4.20}$$

4.2 Forgetting

The final step of the learning algorithm (Algorithm 1) is to compute forgetting factors λ_t, $\lambda_{t;c}$, $c \in \mathbf{c}$, and to obtain the final statistic $\mathrm{V}_t = (\mathrm{v}_t, (\mathrm{V}_{t;c})_{c\in\mathbf{c}})$ determining the posterior pd $\mathrm{P}_t(\Theta)$. It remains to convert the general formula from the third step in Algorithm 1 for the parametric model (3.4).

Work [6] suggests that the forgetting should be applied component-wise. Therefore, the updated $\mathrm{P}_t(\Theta) = \mathrm{P}_t(\alpha) \prod_{c \in \mathbf{c}} \mathrm{P}_t(\omega_c)$ is obtained as follows:

$$\begin{aligned}
\mathrm{P}_t(\alpha) \propto \hat{\mathrm{P}}_t^{\lambda_t}(\alpha)\mathrm{P}_{t-1}^{1-\lambda_t}(\alpha) &= \left(\exp \langle \hat{\mathrm{v}}_t - 1 \,|\, \ln(\alpha)\rangle\right)^{\lambda_t} \left(\exp \langle \mathrm{v}_{t-1} - 1 \,|\, \ln(\alpha)\rangle\right)^{1-\lambda_t} \\
&= \exp \langle \lambda_t \hat{\mathrm{v}}_t + (1-\lambda_t)\mathrm{v}_{t-1} - 1 \,|\, \ln(\alpha)\rangle
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}_t(\omega_c) \propto \hat{\mathrm{P}}_t^{\lambda_{t;c}}(\omega_c)\mathrm{P}_{t-1}^{1-\lambda_{t;c}}(\omega_c) &= \left(\exp \left\langle \hat{\mathrm{V}}_{t;c} \,|\, \mathrm{C}_c(\omega_c)\right\rangle\right)^{\lambda_{t;c}} \left(\exp \langle \mathrm{V}_{t-1;c} \,|\, \mathrm{C}_c(\omega_c)\rangle\right)^{1-\lambda_{t;c}} \\
&= \exp \left\langle \lambda_{t;c}\hat{\mathrm{V}}_{t;c} + (1-\lambda_{t;c})\mathrm{V}_{t-1;c} - 1 \,|\, \mathrm{C}_c(\omega_c)\right\rangle \qquad c \in \mathbf{c}.
\end{aligned}$$

Thus, the final statistic $\mathrm{V}_t = (\mathrm{v}_t, (\mathrm{V}_{t;c})_{c \in \mathbf{c}})$ is computed as follows:

$$\begin{aligned}
\mathrm{v}_t &= \lambda_t \hat{\mathrm{v}}_t + (1-\lambda_t)\mathrm{v}_{t-1} \\
\mathrm{V}_{t;c} &= \lambda_{t;c}\hat{\mathrm{V}}_{t;c} + (1-\lambda_{t;c})\mathrm{V}_{t-1;c} \qquad c \in \mathbf{c}.
\end{aligned}$$

It remains to evaluate the forgetting factors λ_t, $\lambda_{t;c}$, $c \in \mathbf{c}$. They are computed according to the formula in the third step of Algorithm 1 as follows:

$$\lambda_t = \frac{\left(\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta\right)^2}{\int_{\mathbf{\Theta}} \mathrm{M}_t^2(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta} \tag{4.21}$$

$$\lambda_{t;c} = \frac{\left(\int_{\boldsymbol{\omega}_c} \mathrm{M}_{t;c}(\omega_c)\mathrm{P}_{t-1}(\omega_c)d\omega_c\right)^2}{\int_{\boldsymbol{\omega}_c} \mathrm{M}_{t;c}^2(\omega_c)\mathrm{P}_{t-1}(\omega_c)d\omega_c}, \tag{4.22}$$

where $\mathrm{M}_t(\Theta) = \mathrm{M}(O_t|\psi_t, \Theta)$ is the parametric model (3.4) and $\mathrm{M}_{t;c}(\omega_c) = \mathrm{M}_c(O_t|\psi_{t;c}, \omega_c)$ is the c-th component of the parametric model (3.4), which equals:

$$\mathrm{M}_c(O_t|\psi_{t;c}, \omega_c) = \frac{\mathrm{J}_c(O_t, \psi_{t;c}|\omega_c)}{\int_{\mathbf{O}} \mathrm{J}_c(O_t, \psi_{t;c}|\omega_c)dO_t} = \frac{\exp \langle \mathrm{B}_c(O_t, \psi_{t;c}) \,|\, \mathrm{C}_c(\omega_c)\rangle}{\int_{\mathbf{O}} \exp \langle \mathrm{B}_c(O_t, \psi_{t;c}) \,|\, \mathrm{C}_c(\omega_c)\rangle \, dO_t}. \tag{4.23}$$

The denominator in (4.21) can be expressed by using the updated posterior pd $\tilde{\mathrm{P}}_t(\Theta) = \frac{\mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)}{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta}$ obtained via Bayes' rule (4.5). The forgetting factor λ_t then equals:

$$\lambda_t = \frac{\left(\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta\right)^2}{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta \int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\tilde{\mathrm{P}}_t(\Theta)d\Theta} = \frac{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta}{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\tilde{\mathrm{P}}_t(\Theta)d\Theta} \approx \frac{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta}{\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\hat{\mathrm{P}}_t(\Theta)d\Theta}, \tag{4.24}$$

where the last approximation ≈ arises by using $\hat{\mathrm{P}}_t(\Theta)$ instead of $\tilde{\mathrm{P}}_t(\Theta)$ [3].

[3] The use of $\hat{\mathrm{P}}_t(\Theta)$ instead of $\tilde{\mathrm{P}}_t(\Theta)$ is motivated by the wish to make evaluations computationally cheap.

The integrals in (4.24) are computed as follows:

$$\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)d\Theta = \frac{1}{\mathrm{Be}(\mathrm{v}_{t-1})\prod_{d\in\mathbf{c}}\mathrm{N}_d(\mathrm{V}_{t-1;d})}\int_{\mathbf{\Theta}}\Bigg(\mathrm{H}_t(\Theta)\sum_{c\in\mathbf{c}}\omega_c\mathrm{P}_c(\phi^{\mathsf{c}}_{t;c})\exp\left\langle \mathrm{B}_c(\phi_{t;c})\,|\,\mathrm{C}_c(\omega_c)\right\rangle$$

$$\times\exp\left\langle \mathrm{v}_{t-1}-1\,|\,\ln(\omega)\right\rangle\prod_{d\in\mathbf{c}}\exp\left\langle \mathrm{V}_{t-1;d}\,|\,\mathrm{C}_d(\omega_d)\right\rangle d\Theta\Bigg)$$

$$\overbrace{=}^{(4.6)}\frac{1}{\mathrm{Be}(\mathrm{v}_{t-1})\prod_{d\in\mathbf{c}}\mathrm{N}_d(\mathrm{V}_{t-1;d})}\sum_{c\in\mathbf{c}}\mathrm{P}_c(\phi^{\mathsf{c}}_{t;c})\gamma_{t;c}\int_{\mathbf{\Theta}}\mathrm{H}_t(\Theta)\mathrm{Q}_{t;c}(\Theta)d\Theta$$

$$\overbrace{=}^{4.9}\frac{1}{\mathrm{Be}(\mathrm{v}_{t-1})\prod_{d\in\mathbf{c}}\mathrm{N}_d(\mathrm{V}_{t-1;d})}\mathrm{L}_{-1}$$

$$\int_{\mathbf{\Theta}} \mathrm{M}_t(\Theta)\hat{\mathrm{P}}_t(\Theta)d\Theta = \frac{1}{\mathrm{Be}(\hat{\mathrm{v}}_t)\prod_{d\in\mathbf{c}}\mathrm{N}_d(\hat{\mathrm{V}}_{t;d})}\hat{\mathrm{L}}_{-1},$$

where $\hat{\mathrm{L}}_{-1}$ is computed by the same formula as (5.5) with newly computed statistic $\hat{\mathrm{V}}_t=\left(\hat{\mathrm{v}}_t,(\hat{\mathrm{V}}_{t;c})_{c\in\mathbf{c}}\right)$ (4.13), (4.14).

Practically, L_{-1} and $\hat{\mathrm{L}}_{-1}$ are evaluated in the same way as (4.18). The forgetting factor (4.24) is computed as follows:

$$\lambda_t = \frac{\mathrm{Be}(\hat{\mathrm{v}}_t)\prod_{d\in\mathbf{c}}\mathrm{N}_d(\hat{\mathrm{V}}_{t;d})}{\mathrm{Be}(\mathrm{v}_{t-1})\prod_{d\in\mathbf{c}}\mathrm{N}_d(\mathrm{V}_{t-1;d})}\frac{\mathrm{L}_{-1}}{\hat{\mathrm{L}}_{-1}} \approx \frac{\mathrm{Be}(\hat{\mathrm{v}}_t)\prod_{d\in\mathbf{c}}\mathrm{N}_d(\hat{\mathrm{V}}_{t;d})}{\mathrm{Be}(\mathrm{v}_{t-1})\prod_{d\in\mathbf{c}}\mathrm{N}_d(\mathrm{V}_{t-1;d})}\frac{\sum_{d\in\mathbf{c}}\gamma_{t;d}\mathrm{P}_d(\phi^{\mathsf{c}}_{t;d})\mathrm{H}_t(\tilde{\Theta}^d_t)}{\sum_{d\in\mathbf{c}}\hat{\gamma}_{t;d}\mathrm{P}_d(\phi^{\mathsf{c}}_{t;d})\mathrm{H}_t(\hat{\Theta}^d_t)}, \tag{4.25}$$

where

- $\hat{\gamma}_{t;c}=\mathrm{Be}(\hat{\mathrm{v}}_t+\delta_c)\prod_{d\in\mathbf{c}}\mathrm{N}_d(\hat{\mathrm{V}}_{t;d}+\delta_{cd}\mathrm{B}_c(\phi_{t;c}))$,
- $\hat{\Theta}^d_t=\left(\hat{\alpha}^d_t,(\hat{\omega}^d_{t;c})_{c\in\mathbf{c}}\right)$ is expected value of parameter vector Θ computed in the same way as (4.15) and (4.16) with the statistic $\hat{\mathrm{V}}_t=\left(\hat{\mathrm{v}}_t,(\hat{\mathrm{V}}_{t;c})_{c\in\mathbf{c}}\right)$.

Denoting $\mathrm{H}_c(\psi_t,\omega_c)=\frac{1}{\int_{\mathbf{O}}\mathrm{J}_c(O_t,\psi_{t;c}|\omega_c)dO_t}$ and using an analogical derivation as in (4.24) and (4.25), the remaining forgetting factors $\lambda_{t;c}, c\in\mathbf{c}$ are obtained as follows:

$$\lambda_{t;c}=\frac{\int_{\boldsymbol{\omega_c}}\mathrm{M}_{t;c}(\omega_c)\mathrm{P}_{t-1}(\omega_c)d\omega_c}{\int_{\boldsymbol{\omega_c}}\mathrm{M}_{t;c}(\omega_c)\hat{\mathrm{P}}_t(\omega_c)d\omega_c}=\frac{\frac{1}{\mathrm{N}_c(\mathrm{V}_{t-1;c})}\int_{\boldsymbol{\omega_c}}\mathrm{H}_c(\psi_t,\omega_c)\exp\left\langle \mathrm{B}_c(\phi_t)\,|\,\mathrm{C}_c(\omega_c)\right\rangle\,\exp\left\langle \mathrm{V}_{t-1;c}\,|\,\mathrm{C}_c(\omega_c)\right\rangle d\omega_c}{\frac{1}{\mathrm{N}_c(\hat{\mathrm{V}}_{t;c})}\int_{\boldsymbol{\omega_c}}\mathrm{H}_c(\psi_t,\omega_c)\exp\left\langle \mathrm{B}_c(\phi_t)\,|\,\mathrm{C}_c(\omega_c)\right\rangle\,\exp\left\langle \hat{\mathrm{V}}_{t;c}\,|\,\mathrm{C}_c(\omega_c)\right\rangle d\omega_c}$$

$$=\frac{\frac{1}{\mathrm{N}_c(\mathrm{V}_{t-1;c})}\int_{\boldsymbol{\omega_c}}\mathrm{H}_c(\psi_t,\omega_c)\exp\left\langle \mathrm{V}_{t-1;c}+\mathrm{B}_c(\phi_t)\,|\,\mathrm{C}_c(\omega_c)\right\rangle d\omega_c}{\frac{1}{\mathrm{N}_c(\hat{\mathrm{V}}_{t;c})}\int_{\boldsymbol{\omega_c}}\mathrm{H}_c(\psi_t,\omega_c)\exp\left\langle \hat{\mathrm{V}}_{t;c}+\mathrm{B}_c(\phi_t)\,|\,\mathrm{C}_c(\omega_c)\right\rangle d\omega_c}$$

$$\approx\frac{\mathrm{N}_c(\mathrm{V}_{t-1;c}+\mathrm{B}_c(\phi_t))}{\mathrm{N}_c(\mathrm{V}_{t-1;c})}\frac{\mathrm{N}_c(\hat{\mathrm{V}}_{t;c})}{\mathrm{N}_c(\hat{\mathrm{V}}_{t;c}+\mathrm{B}_c(\phi_t))}\frac{\mathrm{H}_c(\psi_t,\tilde{\omega}^c_c)}{\mathrm{H}_c(\psi_t,\hat{\omega}^c_c)}. \tag{4.26}$$

Algorithm 2 (Learning algorithm for mixture ratio model)**.** One complete step of the learning algorithm applied to the mixture ratio model is summarized as follows:

1. Compute the expected values of the parameter vector $\tilde{\Theta}_t^d$ via (4.15) and (4.16).

2. Evaluate L_{-1}, L_{0c}, L_{ci}, $c \in \mathbf{c}, i \in \{1, ..., m_c\}$ (4.9, 4.11, 4.12) via (4.18-4.20).

3. Perform minimization tasks (4.13, 4.14), see Chapter 5 and Chapter 5.

4. Apply forgetting with factors (4.25) and (4.26).

□

Remark 3. Please note, that the dynamic forgetting factors (4.25) and (4.25) derived in this section will eventually not be used, instead the following algorithms will use fixed ones $\lambda_t = \lambda_{t;c} = 1$, $c \in \mathbf{c}$. The reason is that the proposed forgetting was tested in the research project in simulation learning of Markov chain and it did not provide sufficiently good results. The proper choice of the forgetting factor is subject to an ongoing research which is out of the scope of this work. □

Chapter 5

Mixture of Markov Chain Components

One of the most important mixtures with the components from the exponential family are Markov chain mixtures, [18]. A Markov chain operates with discrete valued observations $O \in \mathbf{O} \subset \mathbb{N}$ as well as discrete valued regression vectors $\psi_t \in \boldsymbol{\psi} \subset \mathbb{N}^n$, the joint pds $\mathrm{J}_c(\psi_t|\omega_c)$, $\mathrm{J}(\psi_t|\Theta)$ as well as the parametric system model $\mathrm{M}(O_t|\psi_t, \Theta)$ are pfs (see Definition 11, (3.4)).

This chapter specialises the general mixture ratio model (Chapter 3) and its learning (Chapter 4) for the mixture ratio of Markov chain components with conjugated Dirichlet pd $\mathrm{P}_t(\omega_c)$, $c \in \mathbf{c}$. The specialisation allows to work with specific function forms and to perform minimisation of Kerridge's inaccuracy used in projections of $\tilde{\mathrm{P}}$ on $\hat{\mathrm{P}}$.

Definition 12 (Joint probability of Markov chain mixture model)**.** The joint probability of a mixture of Markov chain models is defined as follows:

$$\mathrm{J}(O_t, \psi_t|\Theta) = \mathrm{J}(\phi_t|\Theta) = \sum_{c \in \mathbf{c}} \alpha_c \mathrm{J}_c(O_t, \psi_{t;c}|\omega_c)\mathrm{P}_c(\psi_c^{\complement}) = \sum_{c \in \mathbf{c}} \alpha_c \chi_{\boldsymbol{\phi}_c}(\phi_t) \, \frac{1}{|\boldsymbol{\psi}_c^{\complement}|} \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{\tilde{K}_{cj}}\right)^{\Delta_{cj}(O_t, \psi_{t;c})} \tag{5.1}$$

where

- $\mathbf{c}$ is the set of component indices $c \in \mathbf{c}$, the symbol $|\mathbf{c}| < \infty$ stands for its cardinality, $\mathbf{c} = \{1, 2, .., |\mathbf{c}|\}$,
- $\chi_{\boldsymbol{\phi}_c}(\phi_t)$ is the characteristic function of the set $\boldsymbol{\phi}_c$; let $\boldsymbol{\phi}$ be a set of all possible values of the data-vector $\phi \in \boldsymbol{\phi}$, then $\boldsymbol{\phi}_c$ is the subset of $\boldsymbol{\phi}$ modelled by c-th component in the mixture model (5.1); the set $\boldsymbol{\phi}_c$ can be rewritten as $\boldsymbol{\phi}_c = \mathbf{O} \times \boldsymbol{\psi}_c$, where $\mathbf{O}$ and $\boldsymbol{\psi}_c$ are the sets of all possible values of the observation O and the component-specific regression vector ψ_c respectively, the $\boldsymbol{\psi}_c^{\complement}$ denotes the complement of the set $\boldsymbol{\psi}_c$ with respect to the set $\boldsymbol{\psi}$,
- $\mathrm{J}_c(O_t, \psi_{t;c}|\omega_c) = \chi_{\boldsymbol{\phi}_c}(\phi_t) \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{\tilde{K}_{cj}}\right)^{\Delta_{cj}(O_t, \psi_{t;c})}$ is the c-th component in the mixture model (5.1), $m_c \in \mathbb{N}$ denotes the number of parameters of the c-th component,
- the non-parametric probability $\mathrm{P}_c(\psi_c^{\complement})$ (3.3) is chosen as uniform on $\psi_c^{\complement}$, i.e. $\mathrm{P}_c(\psi_c^{\complement}) = \frac{1}{|\psi_c^{\complement}|}$,
- α_c is the weight of c-th component; the set of all possible values of the weight-vectors $\alpha = \left(\alpha_1, \alpha_2, ..., \alpha_{|\mathbf{c}|}\right) \in \boldsymbol{\alpha}$ is defined as follows - $\boldsymbol{\alpha} = \left\{(\alpha_1, ..., \alpha_{|\mathbf{c}|}) \,|\, \alpha_c \geq 0, \forall c \in \mathbf{c}, \ \sum_{c \in \mathbf{c}} \alpha_c = 1\right\}$,
- Θ denotes the parameter vector, $\mathbf{\Theta}$ is the set of all possible values of Θ and they have the form:

$$\Theta = (\alpha, (\omega_c)_{c \in \mathbf{c}}) \qquad \mathbf{\Theta} = \boldsymbol{\alpha} \bigtimes_{c \in \mathbf{c}} \omega_c, \tag{5.2}$$

where $\omega_c = (\omega_{c1}, \omega_{c2}, ..., \omega_{cm_c})$ is the parameter vector specific for the c-th component of the mixture and $\boldsymbol{\omega}_c$ the set of all its possible values, $\boldsymbol{\omega}_c = \left\{\omega_c \in [0,1]^{m_c} \,|\, \sum_{j=1}^{m_c} \omega_{cj} = 1\right\}$,

- $\Delta_{cj}(O_t, \psi_{t;c})$ is the indicator function[1] for the parameter ω_{cj}, generally it has the form of a combination (sums, multiplications and compositions) of the Kronecker delta functions on subparts of the vector $(O_t, \psi_{t;c})$; the indicator outputs 0 or 1 for each vector $(O_t, \psi_{t;c})$; the indicator vector $\Delta_c(O_t, \psi_{t;c}) = (\Delta_{c1}, \Delta_{c2}, ..., \Delta_{cm_c})$ outputs 1 on at most one position, 0 on all remaining positions.

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

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

□

Remark 4. Throughout the subsequent text, following simplified notation will be used:

The values of functions $\chi_{\boldsymbol{\phi}_c}(\phi_t)$ and $\Delta_{cj}(O_t, \psi_{t;c})$ depend on ϕ_t and $(O_t, \psi_{t;c})$ respectively, but their arguments will not be highlighted, i.e. $\chi_{\boldsymbol{\phi}_c}(\phi_t) = \chi_{\boldsymbol{\phi}_c}$, $\Delta_{cj}(O_t, \psi_{t;c}) = \Delta_{cj}$, $\quad \forall c \in \mathbf{c}, j \in \{1, ..., m_c\}$. Consequently, let $\forall c \in \mathbf{c}, j \in \{1, .., m_c\}$, $K_{cj} = |\boldsymbol{\psi}_c^{\mathsf{C}}| \tilde{K}_{cj}$, then the joint probability of the parametric model (5.1) can be rewritten as follows:

$$\mathrm{J}(O_t, \psi_t | \Theta) = \sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \alpha_c \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{K_{cj}} \right)^{\Delta_{cj}}.$$

□

Remark 5. The corresponding parametric model of the observation $O_t \in \mathbf{O}$ (cf. Definition 10) has the following mixture ratio form:

$$\mathrm{M}(O_t | \psi_t, \Theta) = \mathrm{M}_t(\Theta) = \frac{\sum\limits_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \alpha_c \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{K_{cj}} \right)^{\Delta_{cj}}}{\sum\limits_{O_t \in \mathbf{O}} \sum\limits_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \alpha_c \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{K_{cj}} \right)^{\Delta_{cj}}} = \mathrm{H}_t(\Theta) \sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \alpha_c \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{K_{cj}} \right)^{\Delta_{cj}}, \tag{5.3}$$

$$\mathrm{H}_t(\Theta) = \mathrm{H}(\psi_t, \Theta) = \frac{1}{\sum\limits_{O_t \in \mathbf{O}} \sum\limits_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \alpha_c \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{K_{cj}} \right)^{\Delta_{cj}}}$$

□

Remark 6. One practical possibility how to use mixture ratios for modelling high-order Markov chains is to model its joint probability $\mathrm{J}(O_t, \psi_t)$ (5.1) as a mixture, where each component models dependence between output O_t and one variable in regression vector ψ_t, i.e.

$$\mathrm{J}(O_t, \psi_t) = \sum_{c=1}^{|\psi_t|} \alpha_c \mathrm{J}(O_t, \psi_{t;c}) \frac{1}{|\boldsymbol{\psi}_c^{\mathsf{C}}|} \qquad \psi_t = \left(\psi_{t;1}, ..., \psi_{t;|\psi_t|} \right) \in \boldsymbol{\psi}_t$$

□

[1] The use of Δ_{cj} instead of B_{cj} (as in Definition 11) stresses that observations O and regression vectors ψ are discrete valued.

Learning with Ratio of Markov Mixtures

The set of feasible posterior pds (see Algorithm 1 and (4.3)) is chosen as follows:

$$\mathbf{P} = \left\{ \mathrm{P}_t(\alpha) \prod_{c \in \mathbf{c}} \mathrm{P}_t(\omega_c) \right\},$$

where the component-conjugated pds $\mathrm{P}_t(\alpha)$ and $\mathrm{P}_t(\omega_c)$ are determined by finite-dimensional sufficient statistics v_t and $\mathrm{V}_{t;c}$ respectively, $\forall c \in \mathbf{c}$, and they are defined as follows (see (5.2) and Definition 6):

$$\mathrm{P}_t(\alpha) = \mathrm{P}(\alpha|\mathrm{v}_t) = \frac{\prod_{c \in \mathbf{c}} \alpha_c^{\mathrm{v}_{t;c}-1}}{\mathrm{Be}(\mathrm{v}_t)} \qquad \mathrm{P}_t(\omega_c) = \mathrm{P}(\omega_c|\mathrm{V}_{t;c}) = \frac{\prod_{j=1}^{m_c} \omega_{cj}^{\mathrm{V}_{t;cj}-1}}{\mathrm{Be}(\mathrm{V}_{t;c})} \quad \forall c \in \mathbf{c},$$

where $\alpha \in \boldsymbol{\alpha}, \quad \omega_c \in \boldsymbol{\omega}_c, \ \ c \in \mathbf{c}, \quad \mathrm{V}_t = (\mathrm{v}_t, \mathrm{V}_{t;1}, ..., \mathrm{V}_{t;|\mathbf{c}|}),\ \mathrm{v}_t > 0, \mathrm{V}_{t;c} > 0, c \in \mathbf{c}$.

Updating step of Algorithm 1 The posterior pd $\tilde{\mathrm{P}}_t(\Theta)$ (prior to its projection to **P**) is calculated via Bayes' rule (1.7) as follows (cf. (4.5)):

$$\begin{aligned}
\tilde{\mathrm{P}}_t(\Theta) &= \frac{\mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)}{\int_\Theta \mathrm{M}_t(\Theta)\mathrm{P}_{t-1}(\Theta)} \\
&\propto \mathrm{H}(\psi_t, \Theta) \sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \alpha_c \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{K_{cj}} \right)^{\delta_{cj}} \prod_{d \in \mathbf{c}} \alpha_d^{\mathrm{v}_{t-1;d}-1} \prod_{j=1}^{m_d} \omega_{dj}^{\mathrm{V}_{t-1;dj}-1} \\
&= \mathrm{H}(\psi_t, \Theta) \sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} \prod_{d \in \mathbf{c}} \alpha_d^{\mathrm{v}_{t-1;d}-1+\delta_{cd}} \prod_{j=1}^{m_d} \omega_{dj}^{\mathrm{V}_{t-1;dj}-1+\delta_{cd}\Delta_{dj}} \\
&= \mathrm{H}(\psi_t, \Theta) \sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} \gamma_{t;c} \mathrm{Q}_{t;c}(\Theta),
\end{aligned}$$

where, $\forall c \in \mathbf{c}$, $\mathrm{Q}_{t;c}(\Theta)$ is a pd of the parameter $\Theta \in \boldsymbol{\Theta}$,

$$\mathrm{Q}_{t;c}(\Theta) = \frac{\prod_{d \in \mathbf{c}} \alpha_d^{\mathrm{v}_{t-1;d}-1+\delta_{cd}} \prod_{j=1}^{m_d} \omega_{dj}^{\mathrm{V}_{t-1;dj}-1+\delta_{cd}\Delta_{dj}}}{\gamma_{t;c}}$$

$$\gamma_{t;c} = \mathrm{Be}(\mathrm{v}_{t-1} + \delta(c)) \prod_{d \in \mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d} + \delta_{cd}\Delta_c) \tag{5.4}$$

Thus, $\tilde{\mathrm{P}}_t(\Theta)$ equals:

$$\tilde{\mathrm{P}}_t(\Theta) = \frac{\mathrm{H}(\psi_t, \Theta) \sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} \gamma_{t;c} \mathrm{Q}_{t;c}(\Theta)}{L_{-1}},$$

where

$$L_{-1} = \sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} \int_{\boldsymbol{\Theta}} \mathrm{H}(\psi_t, \Theta) \mathrm{Q}_{t;c}(\Theta) d\Theta \tag{5.5}$$

Projecting step of Algorithm 1 Subsequently, the projection of $\tilde{\mathrm{P}}_t(\Theta)$ to $\hat{\mathrm{P}}_t(\Theta) \in \mathbf{P}$ is chosen as the minimizer of Kerridge's inaccuracy, which reads (cf. (4.10)):

$$\begin{aligned}
\mathrm{K}(\tilde{\mathrm{P}}_t \| \mathrm{P}) &= -\int_{\boldsymbol{\Theta}} \tilde{\mathrm{P}}_t(\Theta)\ln(\mathrm{P}(\Theta))d\Theta = -\int_{\boldsymbol{\Theta}} \tilde{\mathrm{P}}_t(\Theta)\ln\left(\prod_{c\in\mathbf{c}} \frac{\alpha_c^{\mathrm{v}_c-1}\prod_{j=1}^{m_c}\omega_{cj}^{\mathrm{V}_{cj}-1}}{\mathrm{Be}(\mathrm{v})\mathrm{Be}(\mathrm{V}_c)}\right)d\Theta \\
&= \ln(\mathrm{Be}(\mathrm{v})) + \sum_{c\in\mathbf{c}}\ln(\mathrm{Be}(\mathrm{V}_c)) - \sum_{c\in\mathbf{c}}(\mathrm{v}_c - 1)\int_{\boldsymbol{\Theta}} \tilde{\mathrm{P}}_t(\Theta)\ln(\alpha_c)d\Theta \\
&\quad - \sum_{c\in\mathbf{c}}\sum_{j=1}^{m_c}(\mathrm{V}_{cj} - 1)\int_{\boldsymbol{\Theta}} \tilde{\mathrm{P}}_t(\Theta)\ln(\omega_{cj})d\Theta \\
&= \ln(\mathrm{Be}(\mathrm{v})) + \sum_{c\in\mathbf{c}}\ln(\mathrm{Be}(\mathrm{V}_c)) - \sum_{c\in\mathbf{c}}(\mathrm{v}_c - 1)\frac{L_{0c}}{L_{-1}} - \sum_{c\in\mathbf{c}}\sum_{j=1}^{m_c}(\mathrm{V}_{cj} - 1)\frac{L_{cj}}{L_{-1}},
\end{aligned}$$

where, $\forall c \in \mathbf{c}$,

$$L_{0c} = \sum_{c\in\mathbf{c}}\chi_{\phi_c}\gamma_{t;c}\prod_{j=1}^{m_c}\left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}}\int_{\boldsymbol{\Theta}}\ln(\alpha_c)\mathrm{H}(\psi_t,\Theta)\mathrm{Q}_{t;c}(\Theta)d\Theta \tag{5.6}$$

$$L_{cj} = \sum_{c\in\mathbf{c}}\chi_{\phi_c}\gamma_{t;c}\prod_{j=1}^{m_c}\left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}}\int_{\boldsymbol{\Theta}}\ln(\omega_{cj})\mathrm{H}(\psi_t,\Theta)\mathrm{Q}_{t;c}(\Theta)d\Theta, \quad \forall j \in \{1,2...,m_c\}. \tag{5.7}$$

The projection $\hat{\mathrm{P}}_t(\Theta) \in \mathbf{P}$ is fully determined by the vector statistic $\mathrm{V}_t = (\mathrm{v}_t, (\mathrm{V}_{t;c})_{c\in\mathbf{c}})$. Thus, searching for the desired projection $\hat{\mathrm{P}}_t(\Theta)$ is converted into following convex minimization tasks:

$$\hat{\mathrm{v}}_t = \operatorname*{argmin}_{\mathrm{v}\in(0,+\infty)^{|\mathbf{c}|}}\left(\ln(\mathrm{Be}(\mathrm{v})) - \sum_{c\in\mathbf{c}}(\mathrm{v}_c - 1)\frac{L_{0c}}{L_{-1}}\right) \tag{5.8}$$

$$\hat{\mathrm{V}}_{t;c} = \operatorname*{argmin}_{\mathrm{V}_c\in(0,+\infty)^{m_c}}\left(\ln(\mathrm{Be}(\mathrm{V}_c)) - \sum_{j=1}^{m_c}(\mathrm{V}_{cj} - 1)\frac{L_{cj}}{L_{-1}}\right), \quad \forall c \in \mathbf{c}. \tag{5.9}$$

Numerical Approximation of *L*'s

The projection is possible when knowing L's (5.5), (5.6) and (5.7). Their calculation for the considered Markov-chain case is done here. The numerical approximation of them is based on linear Taylor expansions of the function $\mathrm{H}_t(\Theta) = \mathrm{H}(\psi_t, \Theta)$ (5.3) around expected values $\tilde{\Theta}_t^d = (\tilde{\alpha}_t^d, (\tilde{\omega}_{t;c}^d)_{c\in\mathbf{c}})$ according to pds $\mathrm{Q}_{t;d}(\Theta)$, $\forall d \in \mathbf{c}$, i.e. (cf. Section 4.1):

$$\tilde{\alpha}_{t;c}^d = \int_{\boldsymbol{\Theta}}\alpha_c\mathrm{Q}_{t;d}(\Theta)d\Theta = \int_{\boldsymbol{\alpha}}\alpha_c\frac{\prod_{e\in\mathbf{c}}\alpha_e^{\mathrm{v}_{t-1;e}-1+\delta_{de}}}{\mathrm{Be}(\mathrm{v}_{t-1}+\delta_d)}d\alpha = \frac{\mathrm{v}_{t-1;c}+\delta_{cd}}{\sum_{c\in\mathbf{c}}\mathrm{v}_{t-1;c}+1} \quad c \in \mathbf{c}$$

$$\tilde{\omega}_{t;cj}^d = \int_{\boldsymbol{\Theta}}\omega_{cj}\mathrm{Q}_{t;d}(\Theta)d\Theta = \int_{\omega_c}\omega_{cj}\frac{\prod_{i=1}^{m_c}\omega_{ci}^{\mathrm{V}_{t-1;ci}-1+\delta_{dc}\Delta_{ci}}}{\mathrm{Be}(\mathrm{V}_{t-1;c}+\delta_{dc}\Delta_c)}d\omega_c = \frac{\mathrm{V}_{t-1;cj}+\delta_{dc}\Delta_{cj}}{\sum_{i=1}^{m_c}(\mathrm{V}_{t-1;i}+\delta_{dc}\Delta_{ci})}$$

$$c \in \mathbf{c}, j \in \{1,2,...,m_c\}$$

$$\tilde{\alpha}_t^d = (\tilde{\alpha}_{t;1}^d, \tilde{\alpha}_{t;2}^d, ..., \tilde{\alpha}_{t;|\mathbf{c}|}^d) \quad \tilde{\omega}_{t;c}^d = (\tilde{\omega}_{t;c1}^d, \tilde{\omega}_{t;c2}^d, ..., \tilde{\omega}_{t;cm_c}^d) \quad \tilde{\Theta}_t^d = (\tilde{\alpha}_t^d, (\tilde{\omega}_{t;c}^d)_{c\in\mathbf{c}}). \tag{5.10}$$

Hence, the function $\mathrm{H}_t(\Theta)$ is approximated as follows (depending on particular pd $\mathrm{Q}_{t;d}(\Theta)$, $d \in \mathbf{c}$, used in integration):

$$\mathrm{H}_t(\Theta) \approx \mathrm{H}_t(\tilde{\Theta}_t^d) + \sum_{c\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c}(\alpha_c - \tilde{\alpha}_{t;c}^d) + \sum_{c\in\mathbf{c}} \sum_{j=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{cj}}(\omega_{cj} - \tilde{\omega}_{t;cj}^d),$$

$$\text{where} \quad \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c} = -\mathrm{H}_t^2(\tilde{\Theta}_t^d) \sum_{O_t\in\mathbf{O}} \chi_{\boldsymbol{\phi}_c} \prod_{j=1}^{m_c} \left(\frac{\tilde{\omega}_{cj}^d}{K_{cj}}\right)^{\Delta_{cj}}$$

$$\frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{cj}} = -\mathrm{H}_t^2(\tilde{\Theta}_t^d)\tilde{\alpha}_{t;c}^d \sum_{O_t\in\mathbf{O}} \chi_{\boldsymbol{\phi}_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} 0^{1-\Delta_{cj}}.$$

The L's are then computed as follows (see Appendix for detailed derivation):

$$L_{-1} = \sum_{c\in\mathbf{c}} \chi_{\boldsymbol{\phi}_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{H}_t(\tilde{\Theta}_t^c), \tag{5.11}$$

$$\begin{aligned} L_{0d} = {} & \sum_{c\in\mathbf{c}} \chi_{\boldsymbol{\phi}_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \Bigg[\mathrm{H}_t(\tilde{\Theta}_t^c)\Big(\psi(v_{t-1;d} + \delta_{dc}) - \psi\Big(\sum_{e\in\mathbf{c}} v_{t-1;e} + 1\Big)\Big) \\ & + \frac{1}{\sum_{e\in\mathbf{c}} v_{t-1;e} + 1} \sum_{\substack{e\in\mathbf{c} \\ e\neq d}} \tilde{\alpha}_{t;e}^c \mathrm{H}_t^2(\tilde{\Theta}_t^c) \sum_{O_t\in\mathbf{O}} \Bigg(\chi_{\phi_e} \prod_{j=1}^{m_e} \left(\frac{\tilde{\omega}_{t;ej}^c}{K_{ej}}\right)^{\Delta_{ej}} - \chi_{\phi_d} \prod_{j=1}^{m_d} \left(\frac{\tilde{\omega}_{t;dj}^c}{K_{dj}}\right)^{\Delta_{dj}} \Bigg)\Bigg], \quad d \in \mathbf{c}, \end{aligned} \tag{5.12}$$

$$\begin{aligned} L_{di} = {} & \sum_{c\in\mathbf{c}} \chi_{\boldsymbol{\phi}_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \Bigg\{\mathrm{H}_t(\tilde{\Theta}_t^c)\Big(\psi(V_{t-1;di} + \delta_{dc}\Delta_{di}) - \psi\Big(\sum_{j=1}^{m_d}(V_{t-1;dj} + \Delta_{dj})\Big)\Big) \\ & + \frac{1}{\sum_{j=1}^{m_d}(V_{t-1;dj} + \Delta_{dj})} \sum_{\substack{j=1 \\ j\neq i}}^{m_i} \Bigg[\tilde{\omega}_{t;dj}^c \tilde{\alpha}_{t;j}^c \mathrm{H}_t^2(\tilde{\Theta}_t^c) \sum_{O_t\in\mathbf{O}} \Bigg(\left(\frac{1}{K_{dj}}\right)^{\Delta_{dj}} 0^{1-\Delta_{dj}} - \left(\frac{1}{K_{di}}\right)^{\Delta_{di}} 0^{1-\Delta_{di}}\Bigg)\Bigg]\Bigg\}, \\ & d \in \mathbf{c}, i \in \{1, ..., m_d\}. \end{aligned} \tag{5.13}$$

Remark 7. Since the values of $\gamma_{t;c}$ can be near zero, it is not wise to use them directly for numerical calculations. For practical calculations, the minimization tasks (5.8) and (5.9) requires values of $\frac{L_{0c}}{L_{-1}}$ and $\frac{L_{cj}}{L_{-1}}$, $c \in \mathbf{c}, j \in (1, 2, .., m_c)$.

The solution for this problem is to use attributes of *beta* function ($\gamma_{t;c}$ is the product of *beta* functions, see (5.4)) and expressing the ratios via $\tilde{L}_{-1}$, $\tilde{L}_{0c}$ and $\tilde{L}_{cj}$ while keeping $\frac{L_{0c}}{L_{-1}} = \frac{\tilde{L}_{0c}}{\tilde{L}_{-1}}$ and $\frac{L_{cj}}{L_{-1}} = \frac{\tilde{L}_{cj}}{\tilde{L}_{-1}}$. $\tilde{L}_{-1}$, $\tilde{L}_{0c}$ and $\tilde{L}_{cj}$ then equals (see Appendix for detailed derivation): □

$$\tilde{L}_{-1} = \sum_{c\in\mathbf{c}} \chi_{\boldsymbol{\phi}_c} \tilde{\gamma}_{t;c} \mathrm{H}_t(\tilde{\Theta}_t^c), \tag{5.14}$$

$$\begin{aligned}\tilde{L}_{0d} = \sum_{c\in\mathbf{c}} \chi_{\boldsymbol{\phi}_c} \tilde{\gamma}_{t;c} \Bigg[& \mathrm{H}_t(\tilde{\Theta}_t^c)\Big(\psi(\mathrm{v}_{t-1;d} + \delta_{dc}) - \psi\Big(\sum_{e\in\mathbf{c}} \mathrm{v}_{t-1;e} + 1\Big)\Big) \\ & + \frac{1}{\sum_{e\in\mathbf{c}} \mathrm{v}_{t-1;e} + 1} \sum_{\substack{e\in\mathbf{c} \\ e\neq d}} \tilde{\alpha}_{t;e}^c \mathrm{H}_t^2(\tilde{\Theta}_t^c) \sum_{O_t\in\mathbf{O}} \Big(\chi_{\phi_e} \prod_{j=1}^{m_e} \Big(\frac{\tilde{\omega}_{t;ej}^c}{K_{ej}} \Big)^{\Delta_{ej}} - \chi_{\phi_d} \prod_{j=1}^{m_d} \Big(\frac{\tilde{\omega}_{t;dj}^c}{K_{dj}} \Big)^{\Delta_{dj}} \Big) \Bigg], \quad d \in \mathbf{c},\end{aligned} \tag{5.15}$$

$$\begin{aligned}\tilde{L}_{di} = \sum_{c\in\mathbf{c}} \chi_{\boldsymbol{\phi}_c} \tilde{\gamma}_{t;c} \Bigg\{ & \mathrm{H}_t(\tilde{\Theta}_t^c)\Big(\psi(\mathrm{V}_{t-1;di} + \delta_{dc}\Delta_{di}) - \psi\Big(\sum_{j=1}^{m_d} (\mathrm{V}_{t-1;dj} + \Delta_{dj})\Big)\Big) \\ & + \frac{1}{\sum_{j=1}^{m_d} (\mathrm{V}_{t-1;dj} + \Delta_{dj})} \sum_{\substack{j=1 \\ j\neq i}}^{m_i} \Bigg[\tilde{\omega}_{t;dj}^c \tilde{\alpha}_{t;j}^c \mathrm{H}_t^2(\tilde{\Theta}_t^c) \sum_{O_t\in\mathbf{O}} \Big(\Big(\frac{1}{K_{dj}} \Big)^{\Delta_{dj}} 0^{1-\Delta_{dj}} - \Big(\frac{1}{K_{di}} \Big)^{\Delta_{di}} 0^{1-\Delta_{di}} \Big) \Bigg] \Bigg\}, \\ & d \in \mathbf{c}, i \in \{1, ..., m_d\},\end{aligned} \tag{5.16}$$

where $\tilde{\gamma}_{t;c} = \tilde{\tilde{\alpha}}_{t;c} \prod_{j=1}^{m_c} \Big(\frac{\tilde{\tilde{\omega}}_{cj}}{K_{cj}} \Big)^{\Delta_{cj}}$ $\qquad \tilde{\tilde{\alpha}}_{t;c} = \frac{\mathrm{v}_{t-1;c}}{\sum_{d\in\mathbf{c}} \mathrm{v}_{t-1;d}}$ $\qquad \tilde{\tilde{\omega}}_{t;cj} = \frac{\mathrm{V}_{t-1;cj}}{\sum_{i=1}^{m_c} \mathrm{V}_{t-1;ci}}.$

Finding the projection

The minimization tasks (5.8) and (5.9) are convex. The proof can be found e.g. in [28]. The convexity of tasks (5.8) and (5.9) implies its equivalence to finding the root of the derivatives of the minimized functions:

$$\hat{\mathrm{v}}_t = \underset{\mathrm{v}\in(0,+\infty)^{|\mathbf{c}|}}{\arg} \left(\begin{pmatrix} \psi(\mathrm{v}_1) - \psi\left(\sum_{c\in\mathbf{c}} \mathrm{v}_c\right) - \frac{\tilde{L}_{01}}{\tilde{L}_{-1}} \\ \vdots \\ \psi(\mathrm{v}_{|\mathbf{c}|}) - \psi\left(\sum_{c\in\mathbf{c}} \mathrm{v}_c\right) - \frac{\tilde{L}_{0|\mathbf{c}|}}{\tilde{L}_{-1}} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \right) \tag{5.17}$$

$$\hat{\mathrm{V}}_{t;c} = \underset{\mathrm{V}_c\in(0,+\infty)^{m_c}}{\arg} \left(\begin{pmatrix} \psi(\mathrm{V}_{c1}) - \psi\left(\sum_{j=1}^{m_c} \mathrm{V}_{cj}\right) - \frac{\tilde{L}_{c1}}{\tilde{L}_{-1}} \\ \vdots \\ \psi(\mathrm{V}_{cm_c}) - \psi\left(\sum_{j=1}^{m_c} \mathrm{V}_{cj}\right) - \frac{\tilde{L}_{cm_c}}{\tilde{L}_{-1}} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \right) \quad \forall c \in \mathbf{c}, \tag{5.18}$$

where $\psi(x)$ is the digamma function (see Definition 5), $\tilde{L}_{-1}, \tilde{L}_{0c}$ and $\tilde{L}_{cj}$, $c \in \mathbf{c}, j \in \{1, ..., m_c\}$, are computed via (5.14) - (5.16). The equations (5.17) and (5.18) are solved numerically, e.g. MATLAB function *fsolve* can be used to solve them.

Forgetting

The forgetting relies on the general formulas (4.25) and (4.26) applied on the Markov chain parametric model (5.3). The Markov chain parametric model (5.3) implies the following particular forms of needed functions for computing the forgetting factors (4.25) and (4.26):

- $\mathrm{N}_c(\mathrm{V}_c) = \mathrm{Be}(\mathrm{V}_c)$,
- $\gamma_{t;c} = \mathrm{Be}(\mathrm{v}_{t-1} + \delta(c)) \prod_{d \in \mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d} + \delta_{cd}\Delta_c)$,
- $\hat{\gamma}_{t;c} = \mathrm{Be}(\hat{\mathrm{v}}_t + \delta(c)) \prod_{d \in \mathbf{c}} \mathrm{Be}(\hat{\mathrm{V}}_{t;d} + \delta_{cd}\Delta_c)$,
- $\mathrm{P}_c\left(\psi^{\mathsf{c}}_{t;c}\right) = \frac{1}{|\boldsymbol{\psi}^{\mathsf{c}}_c|}$,
- $\mathrm{H}_c(\psi_t, \omega_c) = \frac{1}{\sum_{O_t \in \mathbf{O}} \chi_{\boldsymbol{\phi}_c} \prod_{j=1}^{m_c} \left(\frac{\omega_{cj}}{K_{cj}}\right)^{\Delta_{cj}}}$.

The forgetting factors $\lambda_t, \lambda_{t;c}$, $c \in \mathbf{c}$ then equal:

$$\lambda_t = \frac{\mathrm{Be}(\hat{\mathrm{v}}_t) \prod_{d \in \mathbf{c}} \mathrm{Be}(\hat{\mathrm{V}}_{t;d})}{\mathrm{Be}(\mathrm{v}_{t-1}) \prod_{d \in \mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d})} \frac{\sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{H}_t(\tilde{\Theta}^c_t)}{\sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \hat{\gamma}_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{H}_t(\hat{\Theta}^c_t)} \tag{5.19}$$

$$= \frac{\sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \frac{\mathrm{Be}(\mathrm{v}_{t-1} + \delta(c)) \prod_{d \in \mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d} + \delta_{cd}\Delta_c)}{\mathrm{Be}(\mathrm{v}_{t-1}) \prod_{d \in \mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d})} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{H}_t(\tilde{\Theta}^c_t)}{\sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \frac{\mathrm{Be}(\hat{\mathrm{v}}_t + \delta(c)) \prod_{d \in \mathbf{c}} \mathrm{Be}(\hat{\mathrm{V}}_{t;d} + \delta_{cd}\Delta_c)}{\mathrm{Be}(\hat{\mathrm{v}}_t) \prod_{d \in \mathbf{c}} \mathrm{Be}(\hat{\mathrm{V}}_{t;d})} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{H}_t(\hat{\Theta}^c_t)} = \frac{\sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \tilde{\tilde{\alpha}}_c \prod_{j=1}^{m_c} \left(\frac{\tilde{\tilde{\omega}}_{cj}}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{H}_t(\tilde{\Theta}^c_t)}{\sum_{c \in \mathbf{c}} \chi_{\boldsymbol{\phi}_c} \hat{\hat{\alpha}}_c \prod_{j=1}^{m_c} \left(\frac{\hat{\hat{\omega}}_{cj}}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{H}_t(\hat{\Theta}^c_t)}$$

$$\lambda_{t;c} = \frac{\mathrm{Be}(\mathrm{V}_{t-1;c} + \Delta_c)}{\mathrm{Be}(\mathrm{V}_{t-1;c})} \frac{\mathrm{Be}(\hat{\mathrm{V}}_{t;c})}{\mathrm{Be}(\hat{\mathrm{V}}_{t;c} + \Delta_c)} \frac{\mathrm{H}_c(\psi_t, \tilde{\omega}^c_c)}{\mathrm{H}_c(\psi_t, \hat{\omega}^c_c)} = \frac{\mathrm{H}_c(\psi_t, \tilde{\omega}^c_c) \prod_{j=1}^{m_c} (\tilde{\tilde{\omega}}_{cj})^{\Delta_{cj}}}{\mathrm{H}_c(\psi_t, \hat{\omega}^c_c) \prod_{j=1}^{m_c} (\hat{\hat{\omega}}_{cj})^{\Delta_{cj}}}, \tag{5.20}$$

where $\tilde{\Theta}^c_t = \left(\tilde{\alpha}^c_t, (\tilde{\omega}^c_{t;d})_{d \in \mathbf{c}}\right)$, $c \in \mathbf{c}$, are expected values of the parameter vector Θ computed in (5.10), $\hat{\Theta}^c_t = \left(\hat{\alpha}^c_t, (\hat{\omega}^c_{t;d})_{d \in \mathbf{c}}\right)$, $c \in \mathbf{c}$, are expected values of the parameter vector Θ computed analogically to (5.10), but with using the statistic $\hat{\mathrm{V}}_t = \left(\hat{\mathrm{v}}_t, (\hat{\mathrm{V}}_{t;c})_{c \in \mathbf{c}}\right)$ instead of the statistic $\mathrm{V}_{t-1} = (\mathrm{v}_{t-1}, (\mathrm{V}_{t-1;c})_{c \in \mathbf{c}})$,

$$\tilde{\tilde{\alpha}}_{t;c} = \frac{\mathrm{v}_{t-1;c}}{\sum_{d \in \mathbf{c}} \mathrm{v}_{t-1;d}} \qquad \tilde{\tilde{\omega}}_{t;cj} = \frac{\mathrm{V}_{t-1;cj}}{\sum_{i=1}^{m_c} \mathrm{V}_{t-1;ci}}$$

$$\hat{\hat{\alpha}}_{t;c} = \frac{\hat{\mathrm{v}}_{t;c}}{\sum_{d \in \mathbf{c}} \hat{\mathrm{v}}_{t;d}} \qquad \hat{\hat{\omega}}_{t;cj} = \frac{\hat{\mathrm{V}}_{t;cj}}{\sum_{i=1}^{m_c} \hat{\mathrm{V}}_{t;ci}}$$

Remark 8. Please note, that the dynamic forgetting factors (5.19) and (5.20) derived in this section will eventually not be used, instead the following algorithms will use fixed ones $\lambda_t = \lambda_{t;c} = 1$, $c \in \mathbf{c}$. The reason is the same as in Remark 3. □

Algorithm 3 (Learning algorithm for Markov Chain mixture ratio model)**.** One learning step of the learning algorithm converted for the Markov Chain mixture ratio model is summarized (similarly to Remark 2) as follows:

1. Compute the expected values of the parameter vector $\tilde{\Theta}^c_t, c \in \mathbf{c}$ via (5.10).
2. Evaluate L_{-1}, L_{0c}, L_{ci}, $c \in \mathbf{c}, i \in \{1, ..., m_c\}$ (5.5, 5.6, 5.7) via (5.14-5.16).
3. Perform minimization tasks, i.e. solve equations (5.17, 5.18).
4. Apply forgetting with factors $\lambda_t = \lambda_{t;c} = 1, c \in \mathbf{c}$.

□

Chapter 6

Mixture of Gaussian Components

Besides Markov chain mixtures, introduced in the Chapter 5, another important case of exponential-family mixture is a mixture of Gaussian (normal) components. They are used to model continuous valued observations $O \in \mathbf{O}$ depending on continuous valued regression vectors $\psi \in \boldsymbol{\psi}$.

This chapter will convert the general mixture ratio model (Chapter 3) as well as its learning (Chapter 4) for this specific form of components.

As described in the Chapter 3, the mixture-ratio model requires joint probability $\mathrm{J}(O_t, \psi_t|\Theta)$. To avoid using multi-dimensional Gaussian components and, therefore, avoid computational difficulties, the joint probability $\mathrm{J}(O_t, \psi_t|\Theta)$ is factorized as described in the following definition. Moreover, it opens the way to handling mixed (countinuous and discrete) data vectors.

Definition 13. (Joint probability of Gaussian mixture model) The joint probability (3.3) for the case of Gaussian components has the following form:

$$\mathrm{J}(O,\psi|\Theta) = \mathrm{J}(\phi|\Theta) = \sum_{c\in\mathbf{c}} \alpha_c \mathrm{J}_c(\phi_c|\Theta_c)\mathrm{P}_c(\psi_c^{\mathsf{c}}) = \sum_{c\in\mathbf{c}} \alpha_c \prod_{i=1}^{n_c} \mathrm{J}_c(\phi_{ci}|\psi_{ci}, r_{ci}, \Theta_{ci})\mathrm{P}_c(\psi_c^{\mathsf{c}}),$$

$$\mathrm{J}_c(\phi_{ci}|\psi_{ci}, r_{ci}, \Theta_{ci}) = \frac{1}{\sqrt{2\pi r_{ci}}}\exp\left(-\frac{\left(\phi_{ci} - \Theta_{ci}^T\psi_{ci}\right)^2}{2r_{ci}}\right) \qquad c \in \mathbf{c},\;\; i = 1,2,...,n_c, \tag{6.1}$$

where

- $\mathbf{c}$ is the set of component indices $c \in \mathbf{c}$, the symbol $|\mathbf{c}| < \infty$ stands for its cardinality, $\mathbf{c} = \{1,2,..,|\mathbf{c}|\}$,
- $\phi = (O,\psi)$ denotes a data vector; to simplify the notation, for $c \in \mathbf{c}$, the observation $O \in \mathbf{O}$ will be sometimes denoted as $O = \phi_{c1}$,
- α_c is the weight of c-th component; the set of all possible values of the weight-vectors $\alpha = \left(\alpha_1, \alpha_2, ..., \alpha_{|\mathbf{c}|}\right) \in \boldsymbol{\alpha}$ is defined as follows - $\boldsymbol{\alpha} = \left\{\alpha \subset [0,1]^{|\mathbf{c}|} \,|\, \sum_{c\in\mathbf{c}} \alpha_c = 1\right\}$,
- $\phi_c = (O, \psi_c)$, where ψ_c is a component-specific sub-vector of ψ,
- n_c denotes the length of the data vector ϕ_c, i.e. $n_c = |\boldsymbol{\psi}_c|$,
- ϕ_{ci} denotes i-th entry of the data vector ϕ_c, ψ_{ci} denotes the sub-vector of ϕ_c containing entries from $(i+1)$ to n_c, i.e. $\psi_{ci} = \left(\phi_{c(i+1)}, \phi_{c(i+2)}, ..., \phi_{cn_c}\right)$, for $i < n_c$, ψ_{cn_c} is an empty vector,

- $\mathrm{P}_c(\psi_c^{\mathsf{c}})$ is non-parametrized pd on the complement ψ_c^{c} of the sub-vector ψ_c with respect to ψ; it can be chosen e.g. as Gaussian distribution with zero mean and big variance as an expression of the lack information about these entries,
- $\Theta = \left(\alpha, (r_{ci}, \Theta_{ci})_{c\in\mathbf{c}, i=1}^{n_c}\right)$ denotes the parameter vector, the set of all possible values of the parameter vector, $\mathbf{\Theta}$, has the following form:

$$\mathbf{\Theta} = \boldsymbol{\alpha} \bigtimes_{c\in\mathbf{c}} \bigtimes_{i=1}^{n_c} \mathbf{r} \bigtimes \mathbf{\Theta}_{ci}, \qquad \text{where} \qquad \mathbf{r} = (0, +\infty), \qquad \mathbf{\Theta}_{ci} = \mathbb{R}^{n_c - i + 1}$$

□

Remark 9. The pd (6.1) can be rewritten into the following form:

$$\mathrm{J}(\phi_{ci}|\psi_{ci}, \Theta_{ci}) = \frac{1}{\sqrt{2\pi r_{ci}}} \exp\left\{ -\frac{1}{2r_{ci}} \mathrm{tr}\left([\phi_{ci}, \psi_{ci}][\phi_{ci}, \psi_{ci}]^T [-1, \Theta_{ci}^T]^T [-1, \Theta_{ci}^T] \right) \right\},$$

which corresponds with the form of an EF component (Definition 3), proof can be found in [32]. □

Remark 10. The corresponding parametric model of the observation $O \in \mathbf{O}$ (cf. Definition 10) has the following mixture ratio form:

$$\begin{aligned} \mathrm{M}(O|\psi, \Theta) &= \frac{\sum_{c\in\mathbf{c}} \alpha_c \prod_{i=1}^{n_c} \mathrm{J}_c(\phi_{ci}|\psi_{ci}, \Theta_{ci}) \mathrm{P}_c(\psi_c^{\mathsf{c}})}{\sum_{c\in\mathbf{c}} \alpha_c \mathrm{P}_c(\psi_c^{\mathsf{c}}) \prod_{i=2}^{n_c} \mathrm{J}_c(\phi_{ci}|\psi_{ci}, \Theta_{ci}) \underbrace{\int_{\mathbf{O}} \mathrm{J}_c(O|\phi_c, \Theta_{ci}) dO}_{=1}} \\ &= \mathrm{H}(\Theta, \psi) \sum_{c\in\mathbf{c}} \alpha_c \mathrm{P}_c(\psi_c^{\mathsf{c}}) \prod_{i=1}^{n_c} \frac{1}{\sqrt{2\pi r_{ci}}} \exp\left(-\frac{\left(\phi_{ci} - \Theta_{ci}^T \psi_{ci}\right)^2}{2r_{ci}} \right), \end{aligned} \tag{6.2}$$

$$\text{where} \qquad \mathrm{H}(\Theta, \psi) = \frac{1}{\sum_{c\in\mathbf{c}} \alpha_c \mathrm{P}_c(\psi_c^{\mathsf{c}}) \prod_{i=2}^{n_c} \frac{1}{\sqrt{2\pi r_{ci}}} \exp\left(-\frac{(\phi_{ci} - \Theta_{ci}^T \psi_{ci})^2}{2r_{ci}} \right)} \tag{6.3}$$

□

Learning with Ratio of Gaussian Mixtures

First of all, the set of feasible posterior pds (see Algorithm 1 and (4.3) $\mathbf{P}$ needs to be specified. It has the following product form:

$$\mathbf{P} = \left\{ \mathrm{P}_t(\alpha) \prod_{c\in\mathbf{c}} \mathrm{P}_t(\Theta_c) \right\} = \left\{ \mathrm{P}_t(\alpha) \prod_{c\in\mathbf{c}} \prod_{i=1}^{n_c} \mathrm{P}_t(\Theta_{ci}) \right\}. \tag{6.4}$$

Agreement. In this section, the indices determining the specific component and its part will be often dropped. The absence of this indices means the expression has the same form for each component and its part. For instance,

$$\mathrm{N}(\nu, V) = \frac{(2\pi)^{0.5|\psi|}}{|V_\psi|^{0.5}} \left(\frac{\Lambda}{2} \right)^{-0.5\nu} \Gamma(0.5\nu)$$

substitutes

$$\mathrm{N}(\nu_{ci}, V_{ci}) = \frac{(2\pi)^{0.5|\psi_{ci}|}}{|V_{\psi ci}|^{0.5}} \left(\frac{\Lambda_{ci}}{2}\right)^{-0.5\nu_{ci}} \Gamma(0.5\nu_{ci}) \qquad c \in \mathbf{c}, i \in \{1,2,...,n_c\}.$$

□

Pds $\mathrm{P}_t(\Theta_{ci})$ are chosen as conjugated pds to the respective components $\mathrm{J}_c(\phi_{ci}|\psi_{ci}, \Theta_{ci})$ (see Definition 13). The conjugated prior to a Gaussian component is called Gauss inverse-Wishart distribution, [32]. It has the following form (cf. 4.1):

$$\mathrm{P}(r, \Theta|\nu, V) = \frac{r^{-0.5(\nu+|\psi|+2)}}{\mathrm{N}(\nu, V)} \exp\left\{-\frac{1}{2r}\mathrm{tr}\Big(V[-1,\Theta^T]^T[-1,\Theta^T]\Big)\right\}. \tag{6.5}$$

For the practical computing, mainly because of numerical stability, which will be explained later on, the following alternative expression of Gauss inverse-Wishart pd is more suitable and will be used (conversion to this alternative expression is proved in [32]):

$$\mathrm{P}(r, \Theta|\nu, V) = \frac{r^{-0.5(\nu+|\psi|+2)}}{\mathrm{N}(\nu, V)} \exp\left\{-\frac{1}{2r}\Big((\Theta - \hat{\Theta})^T V_\psi (\Theta - \hat{\Theta}) + \Lambda\Big)\right\}, \tag{6.6}$$

$$\mathrm{N}(\nu, V) = \frac{(2\pi)^{0.5|\psi|}}{|V_\psi|^{0.5}} \left(\frac{\Lambda}{2}\right)^{-0.5\nu} \Gamma(0.5\nu), \tag{6.7}$$

where

$$V = \begin{bmatrix} V_{\phi_1} & V_{\phi_1\psi}^T \\ V_{\phi_1\psi} & V_\psi \end{bmatrix}, \qquad \hat{\Theta} = V_\psi^{-1} V_{\phi_1\psi},$$

$$\Lambda = V_{\phi_1} - V_{\phi_1\psi}^T V_\psi^{-1} V_{\phi_1\psi}.$$

The matrix V needs to be positive definite. Otherwise in (6.5), there are unbounded combinations of Θ_{ci} that prevent the discussed function be integrable. In practical situations, numerical implementation of the standard evolution $V_t = V_{t-1} + \psi\psi^T$ can easily cause the indefiniteness of V. This is overcomed by considering the Choleski decomposition of the inversion of V, [32], i.e.

$$V^{-1} = GG^T.$$

By partitioning G to

$$G = \begin{bmatrix} G_{\phi_1} & 0 \\ G_{\phi_1\psi} & G_\psi \end{bmatrix}, \tag{6.8}$$

the statistics V_ψ, $\hat{\Theta}$ and Λ used in (6.6) can be obtained from G as follows:

$$V_\psi = \left(G_\psi^T\right)^{-1} \left(G_\psi\right)^{-1} \tag{6.9}$$

$$\hat{\Theta} = -\frac{1}{G_{\phi_1}} G_{\phi_1\psi} \tag{6.10}$$

$$\Lambda = \frac{1}{(G_{\phi_1})^2}. \tag{6.11}$$

Hence, instead of updating the statistic V, it suffices to update the statistic G. This ensures that the statistic V remains positive definite even when computed on computers with bounded precision. The inversion in (6.9) is never directly computed as the updating runs directly on G. All of the formulas with

V_ψ contains its inversion V_ψ^{-1} or the determinant $|V_\psi|^{0.5}$. These can be computed directly using G_ψ by (6.9) and

$$|V_\psi|^{0.5} = \left|\left(G_\psi\right)^{-1}\right| = \frac{1}{\left|G_\psi\right|} = \frac{1}{\prod_i \left[G_\psi\right]_{ii}}, \tag{6.12}$$

respectively.

Also computing the determinant of the triangular matrix G_ψ is a trivial operation as seen in (6.12), which is another advantage of evolving G_t instead of V_t.

Updating step of Algorithm 1 The update of the statistic G_t to G_{t+1} is done by the following algorithm called REFIL, [32]. It takes G_t and ϕ_t as arguments and produces updated G_{t+1}.

Algorithm 4 REFIL

```
$G_{t+1}$ = REFIL($G_t, \phi_t$)
n = length($\phi_t$)
$\sigma_{n+1} = 1$
for $j = 1, ...n$ do
    $f_j = \sum_{i=1}^{j} G_{t;ij}\phi_i$
    $\sigma_j = \sqrt{\sigma^2_{[j+1]} + f^2_{[j]}}$
    $G_{t+1;jj} = \frac{\sigma_{j+1}}{\sigma_j} f_j$
    $g_j = G_{t+1;jj} f_j$
    for $i = j+1, ...n$ do
        $G_{t+1;jj} = \frac{\sigma_{j+1}}{\sigma_j}\left(G_{t;ij} - \frac{f_j}{\sigma^2_{j+1}} g_i\right)$
        $g_i = G_{t;ij} f_j + g_i$
    end
end
```

The update of the statistic G_t is computed in the first step of the general learning algorithm (Algorithm 1) to obtain the updated pd $\tilde{\mathrm{P}}_t(\Theta) = \tilde{\mathrm{P}}(\Theta|\tilde{V}_t)$ from the pd $\mathrm{P}_{t-1}(\Theta)$ available from the previous learning step, where

$$\begin{aligned}
\tilde{\mathrm{P}}(\Theta|\tilde{V}_t) &= \frac{\mathrm{H}_t(\Theta)\sum_{c\in\mathbf{c}} \gamma_{t;c}\mathrm{Q}_{t;c}(\Theta)}{L_{-1}} \\
\mathrm{Q}_{t;c}(\Theta) &= \mathrm{P}_t(\alpha|\tilde{\mathrm{v}}_t^c)\prod_{d\in\mathbf{c}}\prod_{i=1}^{n_d}\mathrm{P}(r_{di}, \Theta_{di}|\tilde{v}^c_{t;di}, \tilde{G}^c_{t;di}) \\
\gamma_{t;c} &= \mathrm{Be}(\tilde{\mathrm{v}}_t^c)\prod_{d\in\mathbf{c}}\prod_{i=1}^{n_d}\mathrm{N}(\tilde{v}^c_{t;di}, \tilde{G}^c_{t;di})/\mathrm{P}_c(\psi_c^{\mathbf{c}})
\end{aligned} \tag{6.13}$$

$$
\begin{aligned}
\mathrm{N}(\nu, G) &= (2\pi)^{0.5|\psi|}|G_{\psi}|\left(\frac{\Lambda}{2}\right)^{-0.5\nu}\Gamma(0.5\nu) \\
L_{-1} &= \int_{\mathbf{\Theta}} \mathrm{H}_t(\Theta)\sum_{c\in\mathbf{c}}\gamma_{t;c}\mathrm{Q}_{t;c}(\Theta)d\Theta \\
\tilde{V}_t &= \left(\left(\tilde{\mathrm{v}}^c_{t;d}\right)_{c,d\in\mathbf{c}}, \left(\tilde{\nu}^c_{t;di}, \tilde{G}^c_{t;di}\right)^{n_c}_{c,d\in\mathbf{c},i=1}\right).
\end{aligned}
$$

The statistics $\tilde{\mathrm{v}}^c_{t;d}$ and $\tilde{\nu}^c_{t;di}$, $c, d \in \mathbf{c}, i = 1, ..., n_c$ are obtained by the following recursion obtained from the Bayes' rule (1.7):

$$
\tilde{\mathrm{v}}^c_{t;d} = \mathrm{v}_{t-1;d} + \delta_{cd} \tag{6.14}
$$

$$
\tilde{\nu}^c_{t;di} = \nu_{t-1;di} + \delta_{cd}, \tag{6.15}
$$

which is the same for both $\tilde{\mathrm{v}}^c_{t;d}$ and $\tilde{\nu}^c_{t;di}$, but the initial conditions are generally (quite) different.

The recursion for the statistics $\tilde{G}^c_{t;di}$ is done by the mentioned REFIL algorithm (Algorithm 4)[1].

Projecting step of Algorithm 1 The second step in the learning algorithm (Algorithm 1) is the projection of $\tilde{\mathrm{P}}_t(\Theta)$ to the set of feasible pds $\mathbf{P}$ (6.4). As mentioned earlier in Algorithm 1, the projection is found as minimizer of the Kerridge inaccuracy (Definition 9), which is given as follows:

$$
\mathrm{K}(\tilde{\mathrm{P}}||\mathrm{P}) = -\int_{\mathbf{\Theta}} \tilde{\mathrm{P}}_t(\Theta)\ln\left(\mathrm{P}(\Theta)\right)d\Theta = \ln(\mathrm{Be}(\mathrm{v})) + \sum_{c\in\mathbf{c}}\sum_{j=1}^{n_c}\ln(\mathrm{N}(\nu_{cj}, G_{cj})) - \sum_{c\in\mathbf{c}}(\mathrm{v}_c - 1)\frac{L_{0c}}{L_{-1}} - \sum_{c\in\mathbf{c}}\sum_{i=1}^{n_c}\frac{L_{ci}}{L_{-1}}, \tag{6.16}
$$

where

$$
L_{-1} = \int_{\mathbf{\Theta}} \mathrm{H}_t(\Theta)\sum_{d\in\mathbf{c}}\gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta \tag{6.17}
$$

$$
L_{0c} = \int_{\mathbf{\Theta}} \ln(\alpha_c)\mathrm{H}_t(\Theta)\sum_{d}\gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta \tag{6.18}
$$

$$
\begin{aligned}
L_{ci} &= \int_{\mathbf{\Theta}}\left(-\frac{1}{2}\left(\nu_{ci} + |\psi_{ci}| + 2\right)\ln(r_{ci}) - \frac{1}{2r_{ci}}\left((\Theta_{ci} - \hat{\Theta}_{ci})^T V_{\psi ci}(\Theta_{ci} - \hat{\Theta}_{ci}) + \Lambda_{ci}\right)\right)\mathrm{H}_t(\Theta)\sum_{d\in\mathbf{c}}\gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta \\
&= -\frac{1}{2}(\nu_{ci} + |\psi_{ci}| + 2)L^1_{ci} - 0.5\Lambda_{ci}L^2_{ci} - 0.5L^3_{ci}
\end{aligned} \tag{6.19}
$$

$$
\begin{aligned}
L^1_{ci} &= \int_{\mathbf{\Theta}} \ln(r_{ci})\mathrm{H}_t(\Theta)\sum_{d\in\mathbf{c}}\gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta \\
L^2_{ci} &= \int_{\mathbf{\Theta}} \frac{1}{r_{ci}}\mathrm{H}_t(\Theta)\sum_{d\in\mathbf{c}}\gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta \\
L^3_{ci} &= \int_{\mathbf{\Theta}} \frac{1}{r_{ci}}\left((\Theta_{ci} - \hat{\Theta}_{ci})^T V_{\psi ci}(\Theta_{ci} - \hat{\Theta}_{ci})\right)\mathrm{H}_t(\Theta)\sum_{d\in\mathbf{c}}\gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta
\end{aligned}
$$

The projection $\hat{\mathrm{P}}_t(\Theta) \in \mathbf{P}$ is determined by the statistics $(\mathrm{v}_t, V_t) = \left(\mathrm{v}_t, (\nu_{t;ci}, G_{t;ci})^{n_c}_{c\in\mathbf{c},i=1}\right)$, thus, the Kerridge inaccuracy (6.16) can be viewed as a function of the statistics (v, V), i.e. $\mathrm{K}(\tilde{\mathrm{P}}||\mathrm{P}) = \mathrm{K}(\mathrm{v}, V)$. The

[1] In fact, only statistics $\tilde{G}^c_{t;ci}$ are updated, remaining statistics $\tilde{G}^c_{t;di}$, where $c \neq d$, remains the same, i.e. $\tilde{G}^c_{t;di} = G_{t-1;di}$

statistics determining the feasible projection $\hat{\mathrm{P}}_t(\Theta) \in \mathbf{P}$ is then found as follows:

$$(\mathrm{v}_t, V_t) = \underset{\mathrm{v},V}{\operatorname{argmin}}\, \mathrm{K}(\mathrm{v}, V). \tag{6.20}$$

Numerical Approximation of *L*'s

The projection is possible when knowing L's (6.17), (6.18), (6.19). Their calculation for the considered Markov-chain case is done here. The numerical approximation of them is based on linear Taylor expansions of the function $\mathrm{H}_t(\Theta) = \mathrm{H}(\psi_t, \Theta)$ (6.3) around expected values $\hat{\Theta}_t^d = \left(\hat{\alpha}_t^d, (\hat{r}_{t;ci}^d, \hat{\Theta}_{t;ci}^d)_{i=1,c\in\mathbf{c}}^{n_c}\right)$ according to pds $\mathrm{Q}_{t;d}(\Theta)$ (6.13), $\forall d \in \mathbf{c}$, i.e. (cf. Section 4.1):

$$\hat{\alpha}_{t;c}^d = \int_{\mathbf{\Theta}} \alpha_c \mathrm{Q}_{t;d}(\Theta) d\Theta = \int_{\boldsymbol{\alpha}} \alpha_c \frac{\prod_{e\in\mathbf{c}} \alpha_e^{\tilde{\mathrm{v}}_{t;e}-1}}{\mathrm{Be}(\tilde{\mathrm{v}}_{t-1})} d\alpha = \frac{\tilde{\mathrm{v}}_{t;c}}{\sum_{d\in\mathbf{c}} \tilde{\mathrm{v}}_{t;d}} \quad c \in \mathbf{c} \tag{6.21}$$

$$\hat{r}_{t;ci}^d = \int_{\mathbf{\Theta}} r_{ci} \mathrm{Q}_{t;d}(\Theta) d\Theta = \int_{\boldsymbol{r}_{ci}} r_{ci} \frac{r_{ci}^{-0.5(\tilde{\nu}_{t;ci}^d+2)}}{\mathrm{N}(\tilde{\Lambda}_{t;ci}^d, \tilde{\nu}_{t;ci}^d)} \exp\left(-\frac{\tilde{\Lambda}_{t;ci}^d}{2r_{ci}}\right) dr_{ci} = \frac{\tilde{\Lambda}_{t;ci}^d}{\tilde{\nu}_{t;ci}^d - 2} \tag{6.22}$$

$$\hat{\Theta}_{t;ci}^d = \hat{\hat{\Theta}}_{t;ci}^d, \tag{6.23}$$

where

$$\hat{\Theta}_{t;ci}^d = \left(\hat{\Theta}_{t;ci1}^d, \ldots, \hat{\Theta}_{t;ci|\psi_{ci}|}^d\right).$$

Hence, the function $\mathrm{H}_t(\Theta)$ is approximated as follows (depending on particular pd $\mathrm{Q}_{t;d}(\Theta)$, $d \in \mathbf{c}$, used in integration):

$$\mathrm{H}_t(\Theta) \approx \mathrm{H}_t(\hat{\Theta}_t^d) + \sum_{c\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \alpha_c} (\alpha_c - \hat{\alpha}_{t;c}^d) + \sum_{c\in\mathbf{c}} \sum_{j=1}^{n_c} \left(\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{cj}} \left(r_{cj} - \hat{r}_{t;cj}^d\right) + \frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{cj}} \left(\Theta_{cj} - \hat{\Theta}_{t;cj}^d\right) \right),$$

$$\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \alpha_c} = -\mathrm{H}_t^2(\hat{\Theta}_t^d) \mathrm{P}_c(\psi_c^{\mathsf{c}}) \prod_{i=2}^{n_c} \frac{1}{\sqrt{2\pi \hat{r}_{t;ci}^d}} \exp\left(-\frac{\left(\phi_{ci} - \left(\hat{\Theta}_{t;ci}^d\right)^T \psi_{ci}\right)^2}{2\hat{r}_{t;ci}^d} \right)$$

$$\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{cj}} = -\mathrm{H}_t^2(\hat{\Theta}_t^d) \frac{\hat{\alpha}_{t;c}}{\left(\hat{r}_{t;cj}^d\right)^2} \left(\left(\phi_{cj} - \left(\hat{\Theta}_{t;cj}^d\right)^T \psi_{cj}\right)^2 - \hat{r}_{t;cj}^d \right) \prod_{i=2}^{n_c} \frac{1}{\sqrt{2\pi \hat{r}_{t;ci}^d}} \exp\left(-\frac{\left(\phi_{ci} - \left(\hat{\Theta}_{t;ci}^d\right)^T \psi_{ci}\right)^2}{2\hat{r}_{t;ci}^d} \right)$$

$$\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{cjk}} = -\mathrm{H}_t^2(\hat{\Theta}_t^d) \frac{\hat{\alpha}_{t;c}}{\hat{r}_{t;cj}^d} \left(\phi_{cj} - \left(\hat{\Theta}_{t;cj}^d\right)^T \psi_{cj}\right) \psi_{cjk} \prod_{i=2}^{n_c} \frac{1}{\sqrt{2\pi \hat{r}_{t;ci}^d}} \exp\left(-\frac{\left(\phi_{ci} - \left(\hat{\Theta}_{t;ci}^d\right)^T \psi_{ci}\right)^2}{2\hat{r}_{t;ci}^d} \right)$$

$$\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{cj}} = \left(\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{cj1}}, \ldots, \frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{cj|\boldsymbol{\psi}_{cj}|}} \right).$$

The $L's$ are computed as follows (proofs of the following formulas are found in Appendix):

$$L_{-1} \approx \sum_{d \in \mathbf{c}} \mathrm{H}(\hat{\Theta}^d)\gamma_{t;d} \tag{6.24}$$

$$\begin{aligned} L_{0c} \approx \sum_{d \in \mathbf{c}} \gamma_{t;d} \Bigg[& \mathrm{H}(\hat{\Theta}^d)\left(\psi(\mathrm{v}_{t-1;d} + \delta_{cd}) - \psi\left(\sum_{e \in \mathbf{c}} \mathrm{v}_{t-1;e} + 1\right)\right) \\ & + \frac{1}{\sum_{e \in \mathbf{c}} \mathrm{v}_{t-1;f} + 1} \sum_{\substack{e \in \mathbf{c} \\ e \neq c}} \hat{\alpha}^d_{t;e} \left(\frac{\partial \mathrm{H}(\hat{\Theta}^d)}{\partial \alpha_c} - \frac{\partial \mathrm{H}(\hat{\Theta}^d)}{\partial \alpha_e}\right)\Bigg] \end{aligned} \tag{6.25}$$

$$L_{cj} = -0.5(\nu_{cj} + |\boldsymbol{\psi}_{cj}| + 2)L^1_{cj} - 0.5\Lambda_{cj}L^2_{cj} - 0.5L^3_{cj} \tag{6.26}$$

$$L^1_{cj} \approx \sum_{d \in \mathbf{c}} \gamma_{t;d}\left[\mathrm{H}(\hat{\Theta}^d_t)\left(\ln\left(0.5\tilde{\Lambda}^d_{t;cj}\right) - \psi\left(0.5\tilde{\nu}^d_{t;cj}\right)\right)\right] \tag{6.27}$$

$$L^2_{cj} \approx \sum_{d \in \mathbf{c}} \gamma_{t;d}\left(\mathrm{H}(\hat{\Theta}^d_t)\frac{\tilde{\nu}^d_{cj}}{\tilde{\Lambda}_{t;cj}} + \frac{\partial \mathrm{H}(\hat{\Theta}^d_t)}{\partial r_{cj}} \frac{2}{\tilde{\nu}^d_{cj} + 2}\right) \tag{6.28}$$

$$\begin{aligned} L^3_{cj} \approx \sum_{d \in \mathbf{c}} \gamma_{t;d} \Bigg[& \mathrm{H}(\hat{\Theta}^d_t)\mathrm{tr}\left(\left(\tilde{V}^d_{\psi cj}\right)^{-1} V_{\psi cj}\right) \\ & + \left(\hat{\hat{\Theta}}^d_{t;cj} - \hat{\Theta}_{t;cj}\right)^T V_{\psi cj}\left(\hat{\hat{\Theta}}^d_{t;cj} - \hat{\Theta}_{t;cj}\right)\left(\mathrm{H}(\hat{\Theta}^d_t)\frac{\tilde{\nu}^d_{cj}}{\tilde{\Lambda}^d_{t;cj}} + \frac{\partial \mathrm{H}(\hat{\Theta}^d_t)}{\partial r_{cj}} \frac{2}{\tilde{\nu}^d_{cj} + 2}\right) \\ & + \frac{\partial \mathrm{H}(\hat{\Theta}^d_t)}{\partial \Theta_{cj}} \left(\tilde{V}^d_{\psi cj}\right)^{-1}\left(V_{\psi cj} + V^T_{\psi cj}\right)\left(\hat{\hat{\Theta}}^d_{t;cj} - \hat{\Theta}_{t;cj}\right)\Bigg] \end{aligned} \tag{6.29}$$

Finding the projection

The minimization task (6.20) is convex and, therefore, it is equivalent to finding the root of the derivative of $\mathrm{K}(\mathrm{v}, V)$ (6.16), i.e. the statistics (v_t, V_t), determining the optimal projection, satisfies the following system of equations

$$\nabla \mathrm{K}(\mathrm{v}_t, V_t) = 0. \tag{6.30}$$

Some parts of the statistics $(\mathrm{v}_t, V_t) = \left(\mathrm{v}_t, (\nu_{t;ci}, G_{t;ci})^{n_c}_{c \in \mathbf{c}, i=1}\right)$ minimizing the Kerridge inaccuracy (6.16) can be found analytically. Note that the statistics $G_{t;ci}$ can be defined equivalently via $\left(\hat{\Theta}_{t;ci}, \Lambda_{t;ci}, V^{-1}_{t;\psi ci}\right)$, see (6.8), (6.10), (6.11) and (6.9).

Minimizing statistics $\hat{\Theta}_{t;ci}$ First of all, $\left(\hat{\Theta}_{t;ci}\right)_{c\in\mathbf{c},i=1}^{n_c}$ are found as follows:

$$\begin{aligned}\frac{\partial \mathrm{K}(\mathrm{v},V)}{\partial\hat{\Theta}_{ci}} &= \sum_{d\in\mathbf{c}} a_d\left(V_{\psi cj}+V_{\psi cj}^T\right)\left(\hat{\hat{\Theta}}_{t;ci}^d-\hat{\Theta}_{ci}\right)-\sum_{d\in\mathbf{c}} b_d\left(V_{\psi cj}+V_{\psi cj}^T\right)[1,...,1]^T\\ &= -\left(V_{\psi cj}+V_{\psi cj}^T\right)\hat{\Theta}_{ci}\sum_{d\in\mathbf{c}} a_d+\left(V_{\psi cj}+V_{\psi cj}^T\right)\sum_{d\in\mathbf{c}}\left(a_d\hat{\hat{\Theta}}_{t;ci}^d-b_d[1,...,1]^T\right),\end{aligned}$$

where
$$\begin{aligned}a_d &= \mathrm{H}(\hat{\Theta}_t^d)\frac{\tilde{\nu}_{cj}^d}{\tilde{\Lambda}_{t;cj}^d}+\frac{\partial\mathrm{H}(\hat{\Theta}_t^d)}{\partial r_{cj}}\frac{2}{\tilde{\nu}_{cj}^d+2}\\ b_d &= \left[b_{d1},...,b_{d(n_c-j+1)}\right]=\frac{\partial\mathrm{H}(\hat{\Theta}_t^d)}{\partial\Theta_{cj}}\left(\tilde{V}_{\psi cj}^d\right)^{-1}.\end{aligned}$$

The statistics $\left(\hat{\Theta}_{t;ci}\right)_{c\in\mathbf{c},i=1}^{n_c}$ satisfying (6.30) are given as follows:

$$\hat{\Theta}_{t;ci}=\frac{1}{\sum_{d\in\mathbf{c}}a_d}\sum_{d\in\mathbf{c}} b_d[1,...,1]^T+\frac{1}{\sum_{d\in\mathbf{c}}a_d}\sum_{d\in\mathbf{c}} a_d\hat{\hat{\Theta}}_{t;ci}^d \qquad (6.31)$$

Minimizing statistics $V_{\psi ci}^{-1}$ To find the analytical formula for the statistics $\left(V_{\psi ci}^{-1}\right)_{c\in\mathbf{c},i=1}^{n_c}$, the formula for L_{cj}^3 (6.29) is rewritten as follows[2]:

$$\begin{aligned}L_{cj}^3 = \sum_{k,l}\left[V_{\psi cj}\right]_{kl}\sum_d\gamma_{t;d}\Bigg\{&\mathrm{H}(\hat{\Theta}_t^d)\left[\tilde{V}_{\psi cj}^d\right]_{kl}^{-1}\\ &+\left(\mathrm{H}(\hat{\Theta}_t^d)\frac{\tilde{\nu}_{cj}^d}{\tilde{\Lambda}_{t;cj}^d}+\frac{\partial\mathrm{H}(\hat{\Theta}_t^d)}{\partial r_{cj}}\frac{2}{\tilde{\nu}_{cj}^d+2}\right)\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_k-\left[\hat{\Theta}_{cj}\right]_k\right)\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_l-\left[\hat{\Theta}_{cj}\right]_l\right)\\ &+\sum_m\frac{\partial\mathrm{H}(\hat{\Theta}_t^d)}{\partial\left[\Theta_{cj}\right]_m}\left(\left[\tilde{V}_{\psi cj}^d\right]_{mk}^{-1}\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_l-\left[\hat{\Theta}_{cj}\right]_l\right)+\left[\tilde{V}_{\psi cj}^d\right]_{ml}^{-1}\right)\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_k-\left[\hat{\Theta}_{cj}\right]_k\right)\Bigg\}\\ = \sum_{k,l}\left[V_{\psi cj}\right]_{kl}\left[L_{cj}^3\right]_{kl},&\end{aligned}$$

where
$$\begin{aligned}\left[L_{cj}^3\right]_{kl} = \sum_d\gamma_{t;d}\Bigg\{&\mathrm{H}(\hat{\Theta}_t^d)\left[\tilde{V}_{\psi cj}^d\right]_{kl}^{-1}\\ &+\left(\mathrm{H}(\hat{\Theta}_t^d)\frac{\tilde{\nu}_{cj}^d}{\tilde{\Lambda}_{t;cj}^d}+\frac{\partial\mathrm{H}(\hat{\Theta}_t^d)}{\partial r_{cj}}\frac{2}{\tilde{\nu}_{cj}^d+2}\right)\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_k-\left[\hat{\Theta}_{cj}\right]_k\right)\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_l-\left[\hat{\Theta}_{cj}\right]_l\right)\\ &+\sum_m\frac{\partial\mathrm{H}(\hat{\Theta}_t^d)}{\partial\left[\Theta_{cj}\right]_m}\left(\left[\tilde{V}_{\psi cj}^d\right]_{mk}^{-1}\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_l-\left[\hat{\Theta}_{cj}\right]_l\right)+\left[\tilde{V}_{\psi cj}^d\right]_{ml}^{-1}\right)\left(\left[\hat{\hat{\Theta}}_{cj}^d\right]_k-\left[\hat{\Theta}_{cj}\right]_k\right)\Bigg\}.\end{aligned}$$

[2] The symbol $\left[\tilde{V}_{\psi cj}^d\right]_{kl}^{-1}$ denotes the k,l-th entry of the matrix $\left(\tilde{V}_{\psi cj}^d\right)^{-1}$, similarly $\left[\hat{\Theta}_{cj}\right]_k$ denotes k-th entry of the vector $\hat{\Theta}_{cj}$. The alternative expression of L_{cj}^3 is obtained from (6.29) by using basic matrix and vector multiplication and by appropriate repositioning all of the elements.

Then, $\frac{\partial \mathrm{K}(\mathrm{v},V)}{\partial \left[V_{\psi ci}\right]_{kl}}$ equals as follows:

$$\frac{\partial \mathrm{K}(\mathrm{v},V)}{\partial \left[V_{\psi ci}\right]_{kl}} = -0.5\frac{1}{|V_{\psi ci}|}\frac{\partial}{\partial \left[V_{\psi ci}\right]_{kl}}|V_{\psi ci}| + 0.5\frac{\left[L^2_{ci}\right]_{kl}}{L_{-1}} = -0.5\left[V^{-1}_{\psi ci}\right]_{kl} + 0.5\frac{\left[L^2_{ci}\right]_{kl}}{L_{-1}}.$$

The matrix statistics $\left(V^{-1}_{\psi ci}\right)^{n_c}_{c\in\mathbf{c},i=1}$ satisfying (6.30) are given by the following formula:

$$\left[V^{-1}_{\psi ci}\right]_{kl} = \frac{\left[L^2_{ci}\right]_{kl}}{L_{-1}} \tag{6.32}$$

Minimizing statistics $\Lambda_{t;ci}$ The statistics $(\Lambda_{t;ci})^{n_c}_{c\in\mathbf{c},i=1}$ are found by computing the following derivative:

$$\frac{\partial \mathrm{K}(\mathrm{v},V)}{\partial \Lambda_{ci}} = -0.5\frac{\nu_{t;ci}}{\Lambda_{ci}} + 0.5\frac{L^2_{ci}}{L_{-1}}.$$

The statistics $(\Lambda_{t;ci})^{n_c}_{c\in\mathbf{c},i=1}$ satisfying (6.30) then equals:

$$\Lambda_{t;ci} = \nu_{t;ci}\frac{L_{-1}}{L^2_{ci}} \tag{6.33}$$

Minimizing statistics $\mathrm{v}_{t;c}, \nu_{t;ci}$ The remaining statistics $\left((\mathrm{v}_{t;c})_{c\in\mathbf{c}}, (\nu_{t;ci})^{n_c}_{c\in\mathbf{c},i=1}\right)$ are computed numerically and the specific equation for numerical solving are obtained from the following derivatives:

$$\begin{aligned}
\frac{\partial \mathrm{K}(\mathrm{v},V)}{\partial \nu_{ci}} &= -0.5 - 0.5\ln\left(\nu_{t;ci}\frac{L_{-1}}{L^2_{cj}}\right) + 0.5\ln(2) + 0.5\Gamma(0.5\nu_{ci})\psi(0.5\nu_{ci}) + 0.5\frac{L^1_{ci}}{L_{-1}}\\
\frac{\partial \mathrm{K}(\mathrm{v},V)}{\partial \mathrm{v}_c} &= \psi(\mathrm{v}_c) - \psi\left(\sum_{d\in\mathbf{c}}\mathrm{v}_d\right) - \frac{L_{0c}}{L_{-1}}.
\end{aligned}$$

The statistics $\left((\mathrm{v}_{t;c})_{c\in\mathbf{c}}, (\nu_{t;ci})^{n_c}_{c\in\mathbf{c},i=1}\right)$ are then found as a numerical solution to the following equations:

$$\Gamma(0.5\nu_{ci})\psi(0.5\nu_{ci}) - \ln(\nu_{ci}) + \frac{L^1_{ci}}{L_{-1}} + \ln\left(\frac{2L^2_{ci}}{L_{-1}}\right) - 1 = 0 \tag{6.34}$$

$$\begin{pmatrix} \psi(\mathrm{v}_1) - \psi\left(\sum_{c\in\mathbf{c}}\mathrm{v}_c\right) - \frac{L_{01}}{L_{-1}} \\ \vdots \\ \psi(\mathrm{v}_{|\mathbf{c}|}) - \psi\left(\sum_{c\in\mathbf{c}}\mathrm{v}_c\right) - \frac{L_{0|\mathbf{c}|}}{L_{-1}} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{6.35}$$

The tasks (6.34) and (6.35) can be easily solved, e.g. by MATLAB function *fsolve*.

Proposition 1. The statistics $\left(\mathrm{v}_t, (\nu_{t;ci}, G_{t;ci})^{n_c}_{c\in\mathbf{c},i=1}\right) = \left(\mathrm{v}_t, \left(\nu_{t;ci}, \hat{\Theta}_{ci}, \Lambda_{ci}, V^{-1}_{\psi ci}\right)^{n_c}_{c\in\mathbf{c},i=1}\right)$ determining the projection to the feasible set of pds $\mathrm{P}_t(\Theta) \in \mathbf{P}$ (6.4) are obtained via formulas (6.31), (6.32), (6.33) and by solving equations (6.34) and (6.35).

□

Algorithm 5 (Learning algorithm for Gaussian mixture ratio model)**.** One learning step of the learning algorithm converted for the Gaussian mixture ratio model is summarized (similarly to Remark 2) as follows:

1. Compute the expected values of the parameter vector $\hat{\Theta}_t^d = \left(\hat{\alpha}_t^d, (\hat{r}_{t;ci}^d, \hat{\Theta}_{t;ci}^d)_{i=1,c\in\mathbf{c}}^{n_c}\right)$, $d \in \mathbf{c}$, via (6.21 - 6.23).

2. Evaluate L_{-1}, L_{0c}, L_{ci}, $c \in \mathbf{c}, i \in \{1,...,n_c\}$ (5.5, 5.6, 5.7) via (6.24 -6.26).

3. Find the updated statistics $\left(\mathrm{v}_t, (\nu_{t;ci}, G_{t;ci})_{c\in\mathbf{c},i=1}^{n_c}\right) = \left(\mathrm{v}_t, \left(\nu_{t;ci}, \hat{\Theta}_{ci}, \Lambda_{ci}, V_{\psi ci}^{-1}\right)_{c\in\mathbf{c},i=1}^{n_c}\right)$ via (6.31), (6.32), (6.33) and by solving equations (6.34) and (6.35).

4. Apply forgetting with factors $\lambda_t = \lambda_{t;c} = 1, c \in \mathbf{c}$ (see Remark 3).

□

The way how the mixture ratio model is constructed, i.e. by modelling joint probability on output and regression vector, allows to seamlessly model also mixed discrete-continuous systems. The practical modelling of such systems is complicated mainly if the dependence of discrete variable on continuous regression vector is present. However, if the joint probability is modelled, it can always be factorized and modelled in the way that overcomes this. It is shown in the following remark.

Remark 11 (General mixed component)**.** Let $\mathrm{J}_c(\phi_t|\Theta), \phi_t = (O_t, \psi_t)$, be one component of joint pd from mixture ratio model (3.3). Without loss of generality, suppose that the data vector ϕ_t contain two entries $\phi_t = (\phi_{t;C_1}, \phi_{t;C_2}, \phi_{t;D_1}, \phi_{t;D_2})$, where $\phi_{t;C_1}, \phi_{t;C_2}$ are continuous variables and $\phi_{t;D_1}, \phi_{t;D_2}$ are discrete variables. Such component has then the following form

$$\mathrm{J}_c(O_t, \psi_t|\Theta) = (\mathrm{J}_c(\phi_{t;C_1}|\phi_{t;C_2}\Theta_1)\mathrm{J}_c(\phi_{t;C_2}|\Theta_2))^{\delta(\phi_{t;D_1}, i_{t;1})\delta(\phi_{t;D_2}, i_{2;t})}\,\mathrm{J}_c(\phi_{t;D_1}, \phi_{t;D_2}|\Theta_3),$$

where

- $\mathrm{J}_c(\phi_{t;C_1}|\phi_{t;C_2}\Theta_1), \mathrm{J}_c(\phi_{t;C_2}|\Theta_2)$ are Gaussian pd from Chapter 6,
- $\mathrm{J}_c(\phi_{t;D_1}, \phi_{t;D_2}|\Theta_3)$ is Markov chain pf from Chapter 5,
- $\delta(\phi_{t;D_1}, i_{t;1})$, $\delta(\phi_{t;D_2}, i_{t;2})$ are Kronecker delta functions,
- $i_{t;1} \in \{1, 2, ..., |\boldsymbol{\phi}_{t;D_1}|\}$, $i_{t;2} \in \{1, 2, ..., |\boldsymbol{\phi}_{t;D_2}|\}$ are observed values of the variables $\phi_{t;D_1}, \phi_{t;D_2}$, respectively.

□

Chapter 7

Experimental Part

This chapter experimentally examines properties of mixture ratio model on several examples. It is compared with standard mixture model, [18], in three scenarios: i) Modelling strengths of both models are compared by simulating a dynamic systems by these two models. ii) Monte Carlo study on learning quality of both models on simulation data are made. iii) Both models is tested on real futures trading data and their ability to serve for creating DM strategies is compared.

All of the experiments are made for the case of Markov chain components. Similar experiments with Gaussian and mixed components are under preparation and will be subject to the future research.

7.1 Mixture Ratio Modelling Strength

A dynamic system with system model M, generating real scalar observations O_t with no actions and regression vector $\psi_t = O_{t-1}$ was simulated. The joint probability was the mixture of joint Gaussian pds $\mathcal{N}_c(\mu_c, \omega_c^{0.5})$ with expectations μ_c and square roots $\omega_c^{0.5}$ of precision matrices, $c = \{1,2\}$, which had the following form

$$\mathrm{J}(O_t, O_{t-1}) = \underbrace{0.5}_{\alpha_1} \mathcal{N}_1 \underbrace{O_t, O_{t-1}}_{\phi_t} \left(\underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{\mu_1}, \underbrace{\begin{bmatrix} 1 & 3 \\ 0 & 0.5 \end{bmatrix}}_{\omega_1^{0.5}} \right) + \underbrace{0.5}_{\alpha_2} \mathcal{N}_2 \underbrace{O_t, O_{t-1}}_{\phi_t} \left(\underbrace{\begin{bmatrix} -1 \\ -1 \end{bmatrix}}_{\mu_2}, \underbrace{\begin{bmatrix} 1 & -2 \\ 0 & 0.5 \end{bmatrix}}_{\omega_2^{0.5}} \right). \tag{7.1}$$

The system model then equalled as follows:

$$\mathrm{M}(O_t|O_{t-1}) = \frac{0.5\mathcal{N}_{1 O_t, O_{t-1}} + 0.5\mathcal{N}_{2 O_t, O_{t-1}}}{\underbrace{0.5 \int_{\mathbf{O}} \mathcal{N}_{1 O_t, O_{t-1}} dO_t}_{\mathrm{W}_1(O_{t-1})} + \underbrace{0.5 \int_{\mathbf{O}} \mathcal{N}_{2 O_t, O_{t-1}} dO_t}_{\mathrm{W}_2(O_{t-1})}} = \mathrm{w}_1(O_{t-1}) \frac{\mathcal{N}_{1 O_t, O_{t-1}}}{\mathrm{W}_1(O_{t-1})} + \mathrm{w}_2(O_{t-1}) \frac{\mathcal{N}_{2 O_t, O_{t-1}}}{\mathrm{W}_2(O_{t-1})}$$

$$\mathrm{w}_1(O_{t-1}) = \frac{\mathrm{W}_1(O_{t-1})}{\mathrm{W}_1(O_{t-1}) + \mathrm{W}_2(O_{t-1})} \qquad \mathrm{w}_2(O_{t-1}) = \frac{\mathrm{W}_2(O_{t-1})}{\mathrm{W}_1(O_{t-1}) + \mathrm{W}_2(O_{t-1})} \tag{7.2}$$

The typical simulation results, differing only in initial values $O_0 \in \{-0.8, 0, 3\}$, are shown in Figure 7.1.

Discussion The mixture ratio model 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.

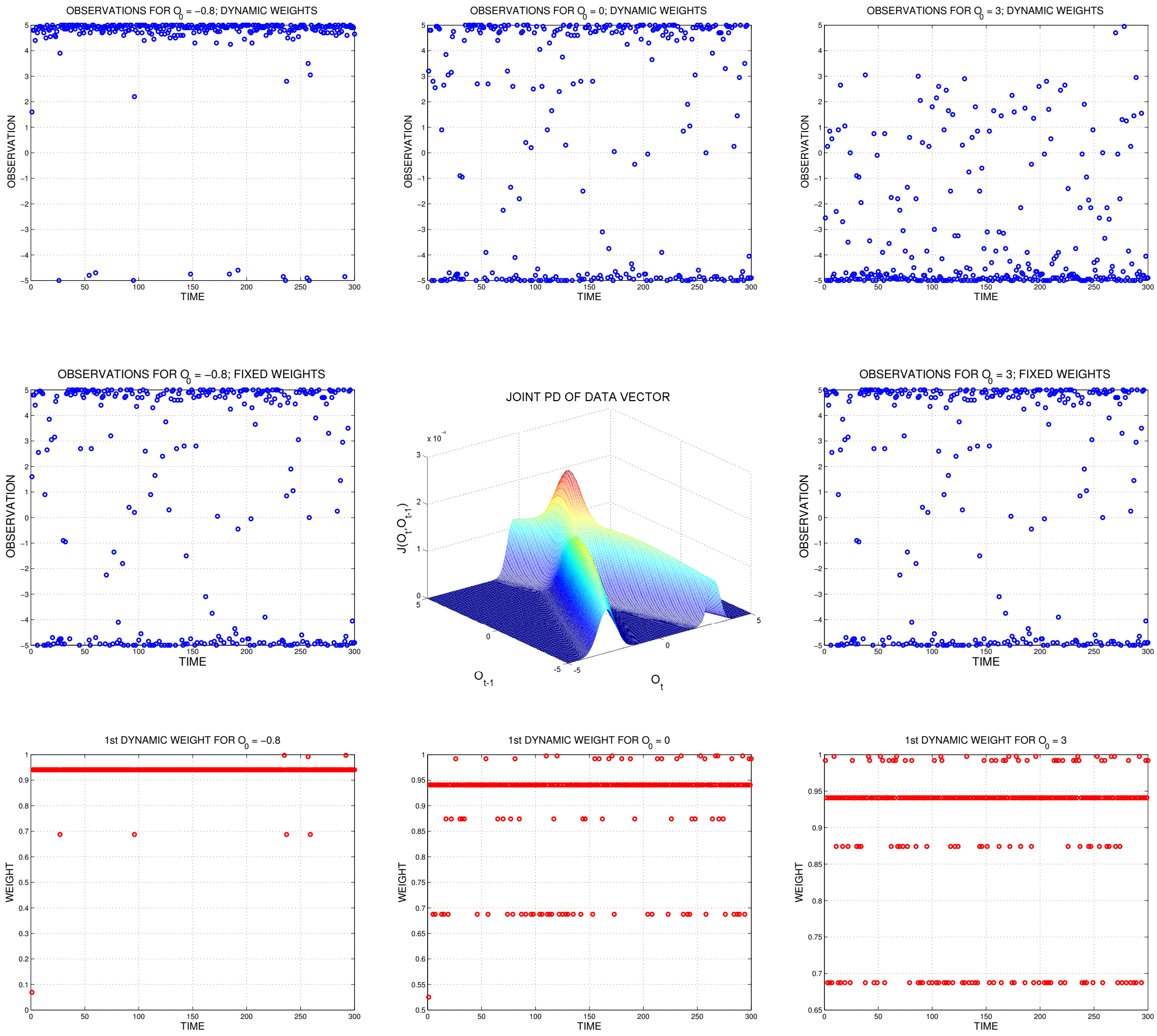

Figure 7.1: Simulation of system (7.1). Columns differ in the initial value of regression vector $\psi_1 = O_0 \in \{-0.8,\ 0,\ 3\}$. Pseudo-random realisations are the same in all cases. The 1st row shows the realized observations. The 2nd row shows realized observations when the same components as in (7.1) 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 $\mathrm{w}_1(O_{t-1})$ (7.2) corresponding to the 1st-row.

7.2 Learning Simulation with Mixture Ratio of Markov Chains

The simulation comparison of the Markov chain *mixture ratio model* discussed in Chapter 5 and the *standard* Markov chain *mixture model*, where the parametric conditional pf $\mathrm{M}_C(O_t|\psi_t,\bar{\Theta})$ (compared to the joint pf $\mathrm{J}(O_t,\psi_t|\Theta)$) is modelled, see [18], was made. The comparison was made as follows:

- A sequence of observations $(O_t)_{t=1}^{|\mathbf{t}|}$ was generated by the pf (5.3) with known parameters Θ_R (i.e. $\mathrm{P}_R(O_t|\psi_t) = \mathrm{M}(O_t|\psi_t,\Theta_R)$; $\mathrm{P}_R(O_t|\psi_t)$ is referred as the *real model*), actions were not present in the system, $\psi_t = (O_{t-1},...,O_{t-n-1})$,
- both, the mixture ratio model and the standard mixture model were learnt recursively,
- the quality of both models during the simulation was compared using Kullback–Leibler divergence of pairs[1] of joint pfs $\mathrm{P}(O_1,...,O_t|\psi_0)$, $t \in \mathbf{t}$.

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

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

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

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

$$\mathrm{P}_N(O_1,...,O_t|\psi_0) = \prod_{\tau=1}^{t} \mathrm{P}_N(O_\tau|\psi_\tau)\mathrm{P}(\psi_0), \qquad N \in \{R,J,C\}, \tag{7.3}$$

with known $\mathrm{P}(\psi_0)$. Kullback–Leibler divergences of the joint pfs then equal[2]:

$$\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_N) = \sum_{O_1,...O_t \in \mathbf{O}} \prod_{\tau=1}^{t} \mathrm{P}_R(O_\tau|\psi_\tau) \ln\left(\prod_{\tau=1}^{t} \frac{\mathrm{P}_R(O_\tau|\psi_\tau)}{\mathrm{P}_N(O_\tau|\psi_\tau)}\right), \qquad N \in \{J,C\}, t \in \mathbf{t}. \tag{7.4}$$

(7.4) can be rewritten as follows:

$$\begin{aligned}\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_L) &= \sum_{\tau=1}^{t} \sum_{O_1,...O_t \in \mathbf{O}} \prod_{\tau=1}^{t} \mathrm{P}_R(O_\tau|\psi_\tau) \ln\left(\frac{\mathrm{P}_R(O_\tau|\psi_\tau)}{\mathrm{P}_N(O_\tau|\psi_\tau)}\right) \\ &= \sum_{\tau=1}^{t} \sum_{\substack{O_\tau \in \mathbf{O} \\ \psi_\tau \in \boldsymbol{\psi}}} \mathrm{P}_R(O_\tau|\psi_\tau)\mathrm{P}(\psi_\tau) \ln\left(\frac{\mathrm{P}_R(O_\tau|\psi_\tau)}{\mathrm{P}_N(O_\tau|\psi_\tau)}\right), \qquad N \in \{J,C\}. \end{aligned} \tag{7.5}$$

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

$$\mathrm{P}(\psi_t) = \sum_{O_{t-n-1} \in \mathbf{O}} \mathrm{P}(\psi_t, O_{t-n-1}) = \sum_{O_{t-n-1} \in \mathbf{O}} \mathrm{P}(O_{t-1}, \psi_{t-1}) = \sum_{O_{t-n-1} \in \mathbf{O}} \mathrm{P}_R(O_{t-1}|\psi_{t-1})\mathrm{P}(\psi_{t-1}),$$

[1] In particular, Kullback–Leibler divergence was computed for two pairs of the “real“ and “learnt“ pfs (one learnt with mixture ratio model and the other with standard mixture model), the details follow.

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

with

$$
\begin{aligned}
\psi_t &= O_{t-1}, O_{t-2}, ..., O_{t-n} \\
\psi_{t-1} &= O_{t-2}, O_{t-3}, ..., O_{t-n-1}.
\end{aligned}
$$

As stated above, the quality of both models is compared by comparing the values of $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_J)$ and $\mathrm{KL}^t(\mathrm{P}_R||\mathrm{P}_C)$ computed via (7.5), where the smaller value of the Kullback-Leibler divergence indicates better predicting and modelling quality of the respective model.

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

7.2.1 Compared Models

In simulation, the system with 5 possible observations $O_t \in \mathbf{O} = \{1, 2, ..., |\mathbf{O}|\}$, where $|\mathbf{O}| = 5$ was modelled. Both mixture ratio model (denoted as $\mathrm{M}_J(O_t|\psi_t, \Theta))$ and standard mixture model (denoted as $\mathrm{M}_C(O_t|\psi_t, \bar{\Theta}))$ were 2-component and used the same regression vector ψ_t and respective components operated on the same subsets of data-vector $\boldsymbol{\phi}$ ($\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, see Definition 5.1), which are specified as follows:

- $\boldsymbol{\Theta} = \alpha \times \omega_1 \times \omega_2$,
- $\psi_t = (O_{t-1}, O_{t-2}) \in \boldsymbol{\psi} = \mathbf{O} \times \mathbf{O}$,
- $\phi_t = (O_t, \psi_t) \in \boldsymbol{\phi} = \mathbf{O} \times \mathbf{O} \times \mathbf{O}$,
- $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \mathbf{O} \times \mathbf{O}$,
- $\phi_{t;1} = (O_t, O_{t-1}) \qquad \phi_{t;2} = (O_t, O_{t-2})$.

Mixture ratio model The joint pf of the mixture ratio model equals (cf. (5.1)):

$$
\mathrm{J}(\phi_t|\Theta) = \mathrm{J}(O_t, \psi_t|\Theta) = \alpha \mathrm{J}(O_t, O_{t-1}|\Theta) + (1-\alpha)\mathrm{J}(O_t, O_{t-2}|\Theta) = \alpha \frac{1}{25} \prod_{i=1}^{25} \omega_{1i}^{\Delta_{1i}} + (1-\alpha) \frac{1}{25} \prod_{i=1}^{25} \omega_{2i}^{\Delta_{2i}},
$$

where

- $\Theta = \left(\alpha, (\omega_{1i})_{i=1}^{25}, (\omega_{2i})_{i=1}^{25}\right) \in \boldsymbol{\Theta}, \quad \sum_{i=1}^{25} \omega_{1i} = \sum_{i=1}^{25} \omega_{2i} = 1, \quad \alpha, \omega_{1i}, \omega_{2i} \in [0,1], \quad i \in \{1, 2, ..., 25\}$,
- $\Delta_{1i} = \Delta_{1i}(O_t, O_{t-1})$ and $\Delta_{2i} = \Delta_{2i}(O_t, O_{t-2})$ are indicators for all combinations (O_t, O_{t-1}) and (O_t, O_{t-2}), respectively.

The corresponding parametric system model then reads (cf. (3.4), (5.3)):

$$
\begin{aligned}
\mathrm{M}_J(O_t|\psi_t, \Theta) &= \frac{\alpha \prod_{i=1}^{25} \omega_{1i}^{\Delta_{1i}} + (1-\alpha) \prod_{i=1}^{25} \omega_{2i}^{\Delta_{2i}}}{\alpha \underbrace{\sum_{O_t \in \mathbf{O}} \prod_{i=1}^{25} \omega_{1i}^{\Delta_{1i}}}_{\mathrm{W}_1(\psi_t, \omega_1)} + (1-\alpha) \underbrace{\sum_{O_t \in \mathbf{O}} \prod_{i=1}^{25} \omega_{2i}^{\Delta_{2i}}}_{\mathrm{W}_2(\psi_t, \omega_2)}} \\
&= \mathrm{w}_1(\psi_t, \Theta) \frac{\prod_{i=1}^{25} \omega_{1i}^{\Delta_{1i}}}{\mathrm{W}_1(\psi_t, \omega_1)} + \mathrm{w}_2(\psi_t, \Theta) \frac{\prod_{i=1}^{25} \omega_{2i}^{\Delta_{2i}}}{\mathrm{W}_2(\psi_t, \omega_2)},
\end{aligned}
\tag{7.6}
$$

where

$$\mathrm{w}_1(\psi_t,\Theta) = \frac{\alpha\,\mathrm{W}_1(\psi_t,\omega_1)}{\alpha\mathrm{W}_1(\psi_t,\omega_1) + (1-\alpha)\mathrm{W}_2(\psi_t,\omega_2)} \qquad \mathrm{w}_2(\psi_t,\Theta) = \frac{(1-\alpha)\,\mathrm{W}_2(\psi_t,\omega_2)}{\alpha\mathrm{W}_1(\psi_t,\omega_1) + (1-\alpha)\mathrm{W}_2(\psi_t,\omega_2)}. \tag{7.7}$$

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

$$\mathrm{M}_C(O_t|\psi_t,\bar\Theta) = \bar\alpha \prod_{i=1}^{25} \bar\omega_{1i}^{\Delta_{1i}} + (1-\bar\alpha)\prod_{i=1}^{25} \bar\omega_{2i}^{\Delta_{2i}}, \tag{7.8}$$

where

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

7.2.2 Results of Comparison

Compared quality of models The quality of models was compared by a MC study. Three sets of 200 stochastic simulations (with identical initial conditions) of the dynamic system were made. In each simulation, both, the mixture ratio and the standard mixture models were learnt recursively on 500 observations. Those were generated by the real model $\mathrm{P}_{R_i}(O_t|\psi_t)$, $i \in \{1,2,3\}$, which was different for each set of simulations:

1. The real model was the mixture ratio model (7.6) with parameters $\Theta_R \in \mathbf{\Theta}$ generated randomly for each simulation, i.e.
$$\mathrm{P}_{R_1}(O_t|\psi_t) = \mathrm{M}_J(O_t|\psi_t,\Theta_R). \tag{7.9}$$
2. The real model was the standard mixture model (7.8) with parameters $\Theta_S \in \bar{\mathbf{\Theta}}$ generated randomly for each simulation, i.e.
$$\mathrm{P}_{R_2}(O_t|\psi_t) = \mathrm{M}_C(O_t|\psi_t,\Theta_S). \tag{7.10}$$
3. The real model was the mixture ratio model (7.6) with parameters $\Theta_F \in \mathbf{\Theta}$ fixed for all simulations giving true dynamic weights $\mathrm{w}_c(\psi_t)$, i.e.
$$\mathrm{P}_{R_3}(O_t|\psi_t) = \mathrm{M}_J(O_t|\psi_t,\Theta_F). \tag{7.11}$$

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

$$\mathrm{v}_0 = 1 \qquad \mathrm{V}_{0;c} = 1 \ \text{(entrywise)} \qquad c \in \mathbf{c}$$

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

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

The values of Kullback-Leibler divergence for both models (at the end of the simulation), $\mathrm{KL}^{500}(\mathrm{P}_R||\mathrm{P}_J)$ and $\mathrm{KL}^{500}(\mathrm{P}_{R_i}||\mathrm{P}_C)$, were studied (see (7.5)). They were compared by their differences

$$\varDelta\mathrm{KL}_i = \mathrm{KL}^{500}(\mathrm{P}_{R_i}\|\mathrm{P}_J) - \mathrm{KL}^{500}(\mathrm{P}_{R_i}\|\mathrm{P}_C), \qquad i \in \{1,2,3\}, \tag{7.12}$$

where negative values of $\varDelta\mathrm{KL}_i$ indicates better predicting performance of mixture ratio model. $\mathrm{P}_{R_i}(O_1,...,O_t|\psi_0)$, $\mathrm{P}_J(O_1,...,O_t|\psi_0)$ and $\mathrm{P}_C(O_1,...,O_t|\psi_0)$, $i \in \{1,2,3\}$, denote the joint pfs for the real models $\mathrm{P}_{R_1}, \mathrm{P}_{R_2}$ and P_{R_3}, the mixture ratio model and the standard mixture model, respectively (cf. (7.3)).

The results are summarized in Table 7.1. For the illustration, Figure 7.2 displays time-evolution of Kullback-Leibler divergence for both models, $\mathrm{KL}^t(\mathrm{P}_{R_i}\|\mathrm{P}_J)$ and $\mathrm{KL}^t(\mathrm{P}_{R_i}\|\mathrm{P}_C)$, $i \in \{1,2,3\}$ (7.5), in one of the simulations.

Simulated Case	P_{R_1}	P_{R_2}	P_{R_3}
Mean	-0.7001	-0.6029	-9.1191
Median	-0.7818	-0.5957	-9.0598
Minimum	-3.2138	-4.0519	-16.1047
Maximum	4.9534	1.9341	-4.5774
Standard Deviation	1.0583	0.9377	2.0317

Table 7.1: Sample statistics of $\varDelta\mathrm{KL}_i$ (7.12), $i \in \{1,2,3\}$. The columns belong to the simulations carried with different real models $\mathrm{P}_{R_1}, \mathrm{P}_{R_2}, \mathrm{P}_{R_3}$ (7.9-7.11).

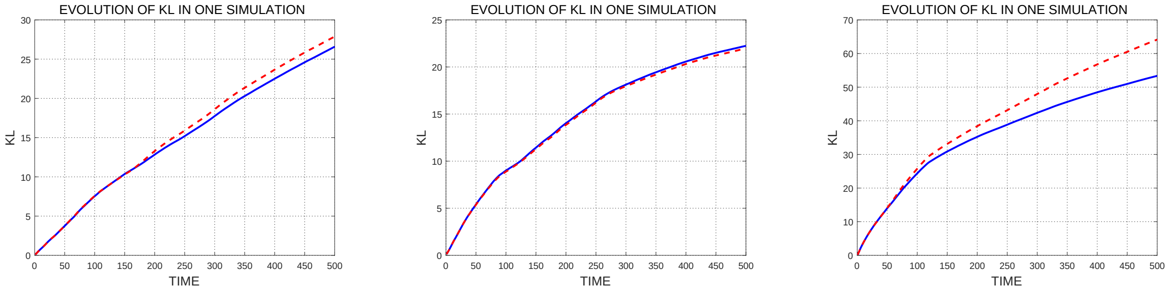

Figure 7.2: Time evolutions of $\mathrm{KL}^t(\mathrm{P}_R\|\mathrm{P}_J)$ (blue line —) and $\mathrm{KL}^t(\mathrm{P}_R\|\mathrm{P}_C)$ (red line - -) (7.5) in one of the simulations. Left to right: P_{R_1} with randomly chosen Θ_R (7.9); P_{R_2} with randomly chosen Θ_S (7.10); P_{R_3} with fixed Θ_F (7.11) causing truly dynamic weights $\mathrm{w}_c(\psi_t)$ (7.7).

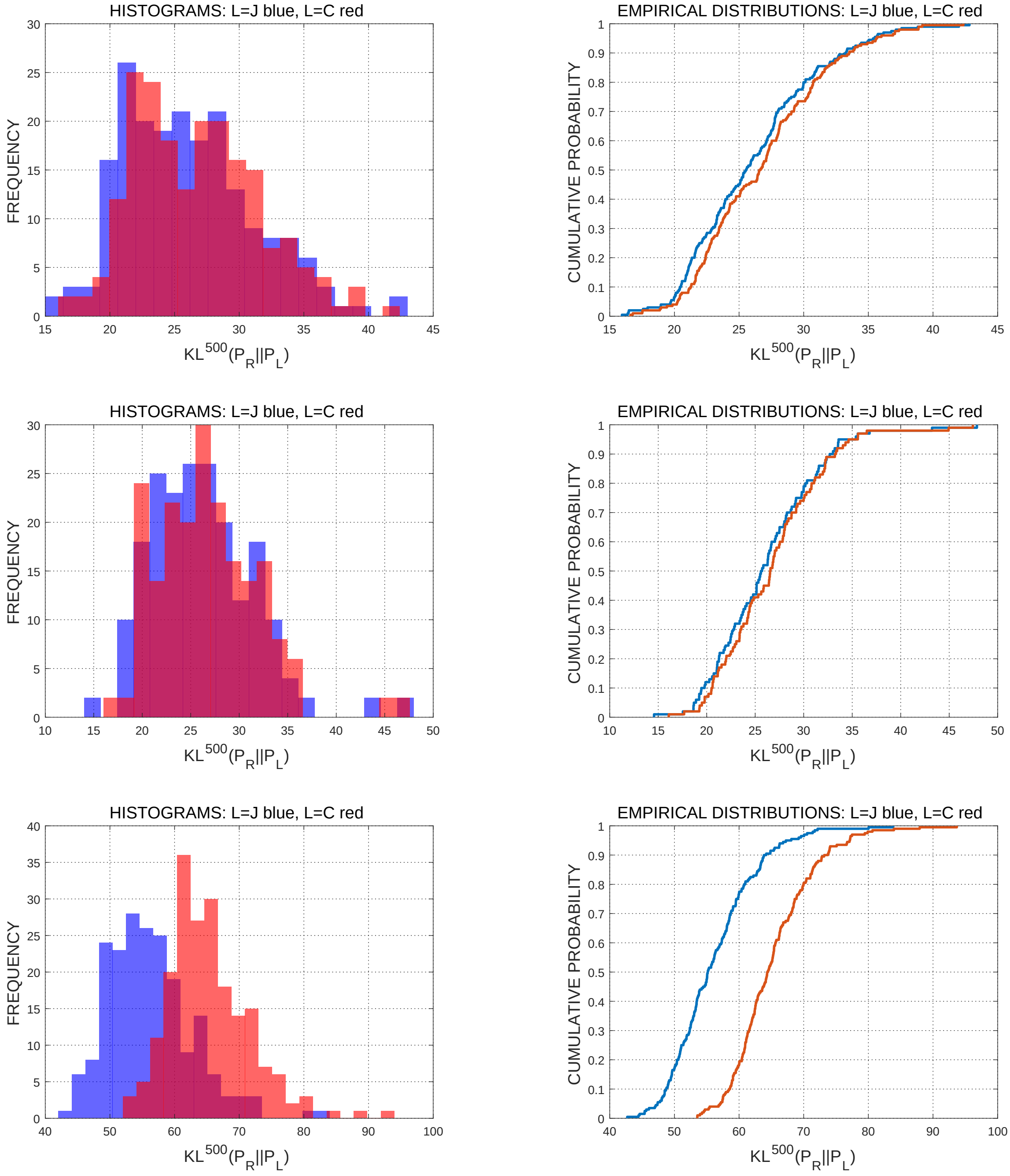

Figure 7.3: Histograms of KL^{500}(P$_{R_i}$||P$_J$) (blue) and KL^{500}(P$_{R_i}$||P$_C$) (7.5) (left) and their empirical distribution function (right). From the top row to the bottom one: P$_{R_1}$ with randomly chosen Θ_R (7.9); P$_{R_2}$ with randomly chosen Θ_S (7.10); P$_{R_3}$ with fixed Θ_F (7.11) causing truly dynamic weights w$_c(\psi_t)$ (7.7)..

Discussion The carried MC study shows two important results, which are implied by Table 7.1 and Figure 7.1. The first is that the mixture ratio model serves well as a parametric system model even if the system is simulated by a standard mixture model. It is even slightly better, in terms of the studied Kullback-Leibler divergence 7.5, then the standard mixture model as it (probably) copes better with errors caused by approximate learning (Algorithm 2).

The second result is that the expectation that the mixture ratio outperforms the standard mixture if the mixture ratio is simulated was confirmed. The numerical approximations used in learning of this model allowed feasible and approximate sufficiently-precise learning of it. The simulation results suggest that it can be used for learning and predicting real-life stochastic systems.

7.3 Trading with Futures

The mixture ratio and the standard mixture models were also compared on a real data containing daily closing prices of various futures contracts traded in public market for 20 years. The prices was discretized into 15 equally spaced segments between their minimum and maximum. Both models were used to create an automatic adaptive agent who optimizes her trading strategy.

The key ingredient of a successful trading strategy is the price prediction. Traders with futures enters a so-called long position when they expect the futures price to rise and to short position when they expect it to fall. Their gain is then the differences in prices between the price when they entered a position and the price when they closed their position. Importing aspect for modelling is also that the agent do not influence the future price by her actions. More details about trading with futures contracts can be found e.g. in [34].

7.3.1 General strategy design

Let $S_t \in \mathbf{S}$ denote a state of the system after the t-th decision epoch. To quantify the agent's aim, a real-valued reward function $\mathrm{R} = \mathrm{R}(S_t, S_{t-1}, A_t)$ is introduced. In general, it depends on the actual state S_{t-1}, the agent's action A_t and on the next state S_t.

The aim of the agent is to optimize her expected reward $\mathrm{E}[\mathrm{R}(S_t, S_{t-1}, A_t)]$ accumulated through $h \in \mathbb{N}$ days, it equals as follows:

$$\mathrm{E}[\mathrm{R}(S_\tau, S_{\tau-1}, A_\tau)] = \sum_{\substack{S_\tau \in \mathbf{S} \\ S_{\tau-1} \in \mathbf{S}}} \sum_{A_\tau \in \mathbf{A}} \mathrm{R}(S_\tau, S_{\tau-1}, A_\tau)\mathrm{P}(S_\tau | S_{\tau-1}, A_\tau)\pi(A_\tau | S_{\tau-1})\mathrm{P}(S_{\tau-1}), \tag{7.13}$$

where t is the initial decision epoch, $\mathrm{P}(S_\tau | S_{\tau-1}, A_\tau)$ is the transition pf, $\pi(A_\tau | S_{\tau-1}) \in \boldsymbol{\pi}$ is the decision rule and $\mathrm{P}(S_{\tau-1})$ is the pf of the state $S_{\tau-1}$.

The agent does so by choosing the set of optimal decision rules $(\pi^*(A_\tau | S_{\tau-1}))_{\tau=t}^{t+h}$ (optimal policy) maximizing the accumulated reward, i.e.

$$(\pi^*(A_\tau | S_{\tau-1}))_{\tau=t}^{t+h} = \operatorname*{argmax}_{(\pi(A_\tau | S_{\tau-1}))_{\tau=t}^{t+h}} \sum_{\tau=t}^{t+h} \mathrm{E}[\mathrm{R}(S_\tau, S_{\tau-1}, A_\tau)]. \tag{7.14}$$

The optimal policy $(\pi^*(A_\tau | S_{\tau-1}))_{\tau=t}^{t+h}$ is found by dynamic programming, [8].

Algorithm 6 (Dynamic programming)**.** Let $S_t \in \mathbf{S}$ denotes the state of the system. The solution $(A_\tau^*(S_{t-1}))_{\tau=t}^{t+h}$ to the following optimization problem

$$(\pi^*(A_\tau | S_{\tau-1}))_{\tau=t}^{t+h} = \operatorname*{argmax}_{(\pi(A_\tau | S_{\tau-1}))_{\tau=t}^{t+h}} \sum_{\tau=t}^{t+h} \mathrm{E}[\mathrm{R}(S_\tau, S_{\tau-1}, A_\tau)] \tag{7.15}$$

is given by the following algorithm. Starting with $\varphi_{t+h}(S_{t+h}) = 0$. For $\tau = t+h, t+h-1, ..., t$

- $A^*_\tau(S_{\tau-1}) = \underset{A_\tau(S_{t-1})}{\operatorname{argmax}} \mathrm{E}[\mathrm{R}(S_\tau, S_{\tau-1}, A_\tau) + \varphi_\tau(S_\tau)|S_{\tau-1}, A_\tau]$
  $= \underset{A_\tau(S_{t-1})}{\operatorname{argmax}} \sum_{S_\tau \in \mathbf{S}} (\mathrm{R}(S_\tau, S_{\tau-1}, A_\tau) + \varphi_\tau(S_\tau))\mathrm{P}(S_\tau|S_{\tau-1}, A_\tau)$
- $\pi^*(A_\tau|S_{\tau-1}) = \delta_{A_\tau A^*_\tau(S_{\tau-1})}$
- $\varphi_{\tau-1}(S_{\tau-1}) = \mathrm{E}[\mathrm{R}(S_\tau, S_{\tau-1}, A_\tau) + \varphi_\tau(S_\tau)|S_{\tau-1}, A^*_\tau(S_\tau)]$

The predictive pf $\mathrm{P}(S_\tau|S_{\tau-1}, A_\tau)$ are assumed to be constant during optimization. Generally the pf $\mathrm{P}(S_\tau|S_{\tau-1}, A_\tau)$ are learned from data on-line, thus it is not constant. This approximation is included to assure feasibility of the optimisation. More details and proof of optimality can be found in [8]. □

7.3.2 Trading strategy

For the trading problem, the reward function, respecting the trading described in the beginning of this section, is specified as follows.

Definition 14 (Reward function)**.** The agent's reward function is defined as follows

$$\mathrm{R}(O_{t-1}, C_{t-1}, A_t) = -\delta_{A_t 1} 2^{\delta_{C_{t-1} 2}} O_{t-1} + \delta_{A_t 2} 2^{\delta_{C_{t-1} 1}} O_{t-1} + \delta_{A_t 3} (-1)^{\delta_{C_{t-1} 2}} O_{t-1}, \tag{7.16}$$

where

- times indices t denote days,
- δ denotes Kronecker delta function (Definition 7),
- $O_t \in \mathbf{O}$ denotes closing price after t-th day,
- $A_t \in \mathbf{A} = \{1, 2, 3, 4\}$ denotes action at the beginning of the t-th day,
  - $A_t = 1$ means entering long position,
  - $A_t = 2$ denotes entering short position,
  - $A_t = 3$ stands for closing actual position,
  - $A_t = 4$ means doing nothing,
- $C_t \in \mathbf{C} = \{1, 2, 3\}$ denotes opened position in the t-th day,
  - $C_t = 1$ means long position,
  - $C_t = 2$ denotes short position,
  - $C_t = 3$ stands for no open position.

□

As mentioned in the beginning of this section, the key element in the successful trading strategy is the prediction of the future price. Here it will be expressed by a transition pf $\mathrm{P}(O_t|\psi_t)$, where $O_t \in \mathbf{O}$

is the future price and ψ_t is the regression vector[3], cf. (3.1). It will contain the actual price O_{t-1}, the previous price O_{t-2} and the average price for the past 7 days denotes as $\overline{O}_{t-1}$, i.e.

$$\psi_t = \left(O_{t-1}, O_{t-2}, \overline{O}_{t-1}\right). \tag{7.17}$$

The transition pf $\mathrm{P}(O_t|\psi_t)$ will be learned from the data by using the mixture ratio model (Chapter 5) as well as the standard mixture model, [18]. It is discussed below.

For the specific reward function (7.16) quantifying trading goals, Algorithm 6 is used with the following elements:

- the state S_t from the algorithm corresponds to (ψ_{t+1}, C_t),
- the transition pf $\mathrm{P}(S_t|S_{t-1}, A_t) = \mathrm{P}(\psi_{t+1}, C_t|\psi_t, C_{t-1}, A_t)$, respecting the structure and dependencies of the trading problem, has the following form[4]
  $\mathrm{P}(\psi_{t+1}, C_t|\psi_t, C_{t-1}, A_t) = \mathrm{P}(O_t|\psi_t)\mathrm{d}_1(C_t|A_t, C_{t-1})\mathrm{d}_2(\overline{O}_t|O_t, \overline{O}_{t-1})\delta_{O_{t-1}\psi_{t;1}}$ with
  - $\mathrm{d}_1(S_t|A_t, S_{t-1}) = \delta_{S_t S_{new}}$
  $$S_{new} = \begin{cases} 1, & A_t = 1 \\ 2, & A_t = 2 \\ 3, & A_t = 3 \\ S_{t-1} & A_t = 4 \end{cases}$$
  - $\mathrm{d}_2(\overline{O}_t|\overline{O}_{t-1}) = \delta_{\overline{O}_t \overline{O}_{new}}$ $\qquad \overline{O}_{new} = \overline{O}_{t-1} + \frac{O_t - O_{t-7}}{7}$

Hence, it remains to specify the transition pf $\mathrm{P}(O_t|\psi_t)$. It is learned on-line by both agents using the mixture ratio and the standard mixture models.

Mixture ratio model The joint probability in mixture ratio model $\mathrm{J}(O_t, \psi_t|\Theta)$ (5.1) has the following form:

$$\mathrm{J}(O_t, \psi_t|\Theta) = \alpha_1 \frac{1}{225} \prod_{i=1}^{225} \omega_{1i}^{\Delta_{1i}} + \alpha_2 \frac{1}{225} \prod_{i=1}^{225} \omega_{2i}^{\Delta_{2i}} + \alpha_3 \frac{1}{225} \prod_{i=1}^{225} \omega_{3i}^{\Delta_{3i}} \tag{7.18}$$

where

- $\Theta = \left((\alpha_i)_{i=1}^{3}, (\omega_{1i})_{i=1}^{225}, (\omega_{2i})_{i=1}^{225}, (\omega_{3i})_{i=1}^{225}\right) \in \boldsymbol{\Theta}$, $\quad \sum_{i=1}^{3} \alpha_i = \sum_{i=1}^{225} \omega_{1i} = \sum_{i=1}^{225} \omega_{2i} = 1$,
- $\Delta_{1i} = \Delta_{1i}(O_t, O_{t-1})$, $\Delta_{2i} = \Delta_{2i}(O_t, O_{t-2})$ and $\Delta_{3i} = \Delta_{3i}(O_t, \overline{O}_{t-1})$ are indicators for all combinations $(O_t,\ O_{t-1})$,$(O_t,\ O_{t-2})$ and $(O_t, \overline{O}_{t-1})$, respectively.

The parametric model $\mathrm{M}_J(O_t|\psi_t, \Theta)$ is then given as follows:

$$\mathrm{M}_J(O_t|\psi_t, \Theta) = \frac{\sum_{c=1}^{3} \alpha_c \prod_{i=1}^{225} \omega_{ci}^{\Delta_{ci}}}{\sum_{O_t \in \mathbf{O}} \sum_{c=1}^{3} \alpha_c \prod_{i=1}^{225} \omega_{ci}^{\Delta_{ci}}} = \mathrm{H}_t(\Theta) \sum_{c=1}^{3} \alpha_c \prod_{i=1}^{225} \omega_{ci}^{\Delta_{ci}} \tag{7.19}$$

$$\mathrm{H}_t(\Theta) = \frac{1}{\sum_{O_t \in \mathbf{O}} \sum_{c=1}^{3} \alpha_c \prod_{i=1}^{225} \omega_{ci}^{\Delta_{ci}}}$$

[3] The regression vector will not contain action A_t as in (cf. 3.1). The action does not influence the predicting price, thus omitting A_t from regression vector simplifies the notation.

[4] The functions $\mathrm{d}_1(C_t|A_t, C_{t-1},)\mathrm{d}_2(\overline{O}_t|O_t, \overline{O}_{t-1}), \delta_{O_{t-1}\psi_{t;1}}$ describes deterministic evolution of C_t, $\overline{O}_t$ and O_{t-1}, only the evolution of the future price O_t is assumed to be stochastic and is described by the transition pf $\mathrm{P}(O_t|\psi_t)$.

The key transition probability $\mathrm{P}_J(O_t|\psi_t)$ is then obtained from the parametric model $\mathrm{M}_J(O_t|\psi_t,\Theta)$ (7.19) as follows:

$$\mathrm{P}_J(O_t|\psi_t) = \int_{\boldsymbol{\Theta}} \mathrm{M}_J(O_t|\psi_t,\Theta)\mathrm{P}_{Jt}(\Theta)d\Theta, \tag{7.20}$$

where pd $\mathrm{P}_{Jt}(\Theta) = \mathrm{P}_J(\Theta|\mathrm{v}_t,\mathrm{V}_t)$ is obtained recursively as described in Section 6. The integration in (7.20) is approximated similarly as the integral L_{-1} (5.5), i.e. the function $\mathrm{H}_t(\Theta)$ is approximated by its linear Taylor expansion around the expected values of parameter $\hat{\Theta}_t^c$, $c \in \{1,2,3\}$ computed by (5.10). The transition probability $\mathrm{P}_J(O_t|\psi_t)$ is then computed as follows:

$$\begin{aligned}
\mathrm{P}_J(O_t|\psi_t) &\approx \sum_{c=1}^{3} \mathrm{H}_t(\tilde{\Theta}_t)\frac{\int_{\boldsymbol{\Theta}} \alpha_c \prod_{i=1}^{225}\omega_{ci}^{\Delta_{ci}} \prod_{d=1}^{3}\alpha_d^{\mathrm{v}_d-1}\prod_{i=1}^{225}\omega_{di}^{V_{di}-1}d\Theta}{\mathrm{Be}(\mathrm{v})\prod_{d=1}^{3}\mathrm{Be}(\mathrm{V}_c)} \\
&= \sum_{c=1}^{3} \mathrm{H}_t(\tilde{\Theta}_t)\frac{\mathrm{Be}(\mathrm{v}+\delta_c)\prod_{d=1}^{3}\mathrm{Be}(\mathrm{V}_d+\delta_c\Delta_c)}{\mathrm{Be}(\mathrm{v})\prod_{d=1}^{3}\mathrm{Be}(\mathrm{V}_d)} \\
&= \sum_{c=1}^{3} \mathrm{H}_t(\tilde{\Theta}_t)\frac{\mathrm{Be}(\mathrm{v}+\delta_c)\prod_{d=1}^{3}\mathrm{Be}(\mathrm{V}_d+\delta_c\Delta_c)}{\mathrm{Be}(\mathrm{v})\prod_{d=1}^{3}\mathrm{Be}(\mathrm{V}_d)} = \sum_{c=1}^{3}\mathrm{H}_t(\tilde{\Theta}_t)\tilde{\alpha}_{t;c}\prod_{i=1}^{225}\tilde{\omega}_{t;ci}^{\Delta_{ci}}
\end{aligned} \tag{7.21}$$

$$\tilde{\alpha}_{t;c} = \frac{\mathrm{v}_{t-1;c}}{\sum_{d\in\mathbf{c}}\mathrm{v}_{t-1;d}} \qquad \tilde{\omega}_{t;cj} = \frac{\mathrm{V}_{t-1;cj}}{\sum_{i=1}^{m_c}\mathrm{V}_{t-1;ci}} \qquad \tilde{\Theta}_t = (\tilde{\alpha}_t,(\tilde{\omega}_{t;c})_{c\in\mathbf{c}}) \tag{7.22}$$

Standard mixture model The conditional probability $\mathrm{M}_C(O_t|\psi_t,\Theta)$ modelled in standard mixture model has the following form:

$$\mathrm{M}_C(O_t|\psi_t,\Theta) = \bar{\alpha}_1\prod_{i=1}^{225}\bar{\omega}_{1i}^{\Delta_{1i}} + \bar{\alpha}_2\prod_{i=1}^{225}\bar{\omega}_{2i}^{\Delta_{2i}} + \bar{\alpha}_3\prod_{i=1}^{225}\bar{\omega}_{3i}^{\Delta_{3i}} \tag{7.23}$$

where

- $\bar{\Theta} = \left(\bar{\alpha},(\bar{\omega}_{1i})_{i=1}^{225},(\bar{\omega}_{2i})_{i=1}^{225},(\bar{\omega}_{3i})_{i=1}^{225}\right) \in \boldsymbol{\Theta}$, $\sum_{j=1}^{15}\bar{\omega}_{1(15i+j)} = \sum_{j=1}^{15}\bar{\omega}_{2(15i+j)} = \sum_{j=1}^{15}\bar{\omega}_{3(15i+j)} = 1$ $i \in \{0,1,...,14\}$,
- $\Delta_{1i} = \Delta_{1i}(O_t,O_{t-1})$, $\Delta_{2i} = \Delta_{2i}(O_t,O_{t-2})$ and $\Delta_{3i} = \Delta_{3i}(O_t,\overline{O}_{t-1})$ are indicators for all combinations $(O_t,\ O_{t-1})$,$(O_t,\ O_{t-2})$ and $(O_t,\overline{O}_{t-1})$, respectively.

The objective transition probability $\mathrm{P}_C(O_t|\psi_t)$ is obtained as follows:

$$\mathrm{P}_C(O_t|\psi_t) = \int_{\boldsymbol{\Theta}}\mathrm{M}_C(O_t|\psi_t,\Theta)\mathrm{P}_{Ct}(\Theta)d\Theta = \sum_{c=1}^{3}\tilde{\bar{\alpha}}_{t;c}\prod_{i=1}^{225}\tilde{\bar{\omega}}_{t;ci}^{\Delta_{ci}}, \tag{7.24}$$

where pd $\mathrm{P}_{Ct}(\Theta) = \mathrm{P}_C(\Theta|\bar{\mathrm{v}}_t,\bar{\mathrm{V}}_t)$ is learned recursively, $\bar{\mathrm{v}}_t,\bar{\mathrm{V}}_t$ are sufficient statistics determining this pd and

$$\tilde{\bar{\alpha}}_{t;c} = \frac{\bar{\mathrm{v}}_{t-1;c}}{\sum_{d\in\mathbf{c}}\bar{\mathrm{v}}_{t-1;d}} \qquad \tilde{\bar{\omega}}_{t;cj} = \frac{\bar{\mathrm{V}}_{t-1;cj}}{\sum_{i=1}^{m_c}\bar{\mathrm{V}}_{t-1;ci}}.$$

Learning of standard mixture model is described in detail in [18].

7.3.3 Results of Trading

Both agents traded with 16 different futures contracts. Each of the time series was discretized (see Figure 7.4) into 15 values and divided into two halves. The first half of the data was used to learning only, while in the second half, both agents also traded.

The accumulated reward (7.16) of both agents was compared. The results are summarized in Table 7.2 and Figure 7.5. Trading with three picked futures can be seen on Figure 7.6.

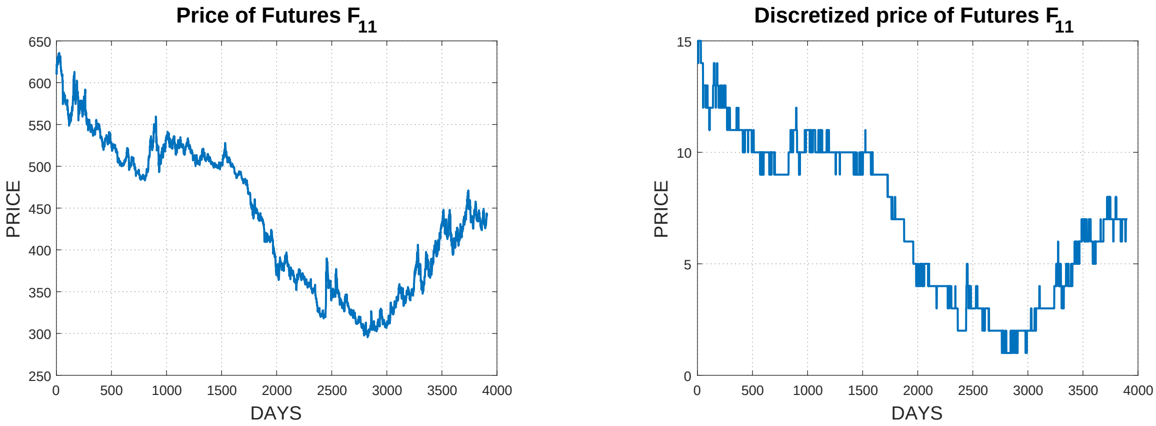

Figure 7.4: Price evolution of futures F_{11} for 3910 trading days; the real prices on the left, the discretized ones on the right.

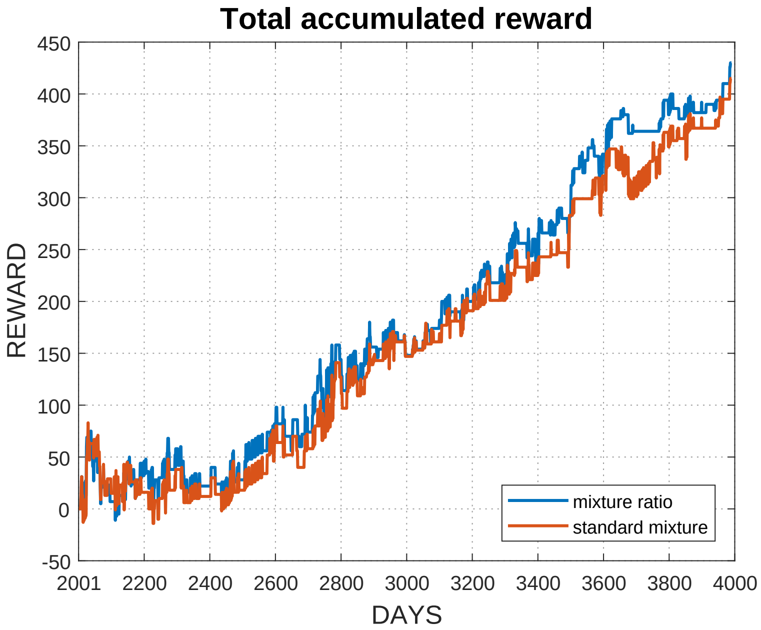

Figure 7.5: Total accumulated reward (7.16) while trading, i.e. for the past 1989 days.

futures	**mixture ratio reward**	**standard mixture reward**
F_1	32	18
F_2	-1	-3
F_3	0	0
F_4	20	10
F_5	-6	-2
F_6	17	25
F_7	37	55
F_8	70	74
F_9	42	38
F_{10}	14	10
F_{11}	22	22
F_{12}	13	15
F_{13}	42	37
F_{14}	60	56
F_{15}	-6	-6
F_{16}	74	66
TOTAL	430	415
MEAN	26.875	25.938
MEDIAN	21	20
MIN	-6	-6
MAX	74	74

Table 7.2: The results from trading with all of the traded futures in terms of accumulated reward (7.16).

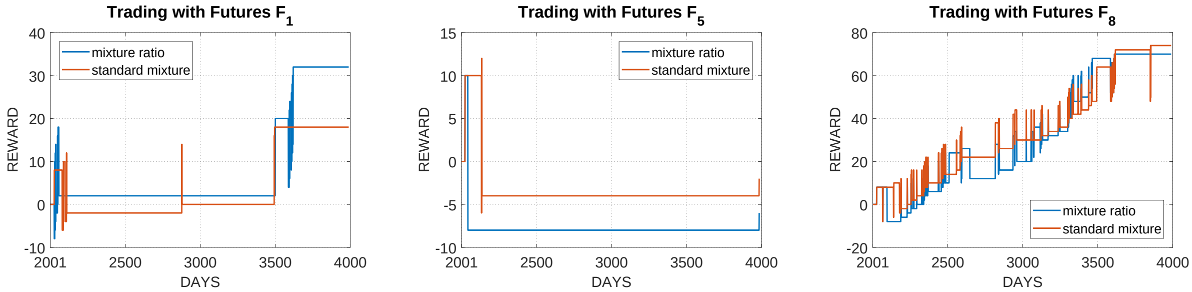

Figure 7.6: Accumulated reward while trading with three futures - from the left F_1, F_5 and F_8.

Discussion As Figure 7.5 and Table 7.2 show, the performance of agent with mixture ratio model is similar to the agent with standard mixture model. Predicting time series of futures prices is in general complicated task and it is very hard to beat the useless prediction that the price will stay the same the next day. Both agents ended in positive numbers in terms of their reward, however this does not mean that the agents would earn proportional amount of real money if they traded on a futures market. Our reward function (7.16) does not include transaction costs, which complicate real algorithmic trading even more.

Despite this, the results on the futures data leave an important message on mixture ratio model. As in simulation in Section 7.2, the mixture ratio model is at least as good as the standard mixture model even when real data is predicted and it is suitable for real decision making.

Conclusions

In this work, a very flexible, but yet unconsidered model set of mixture ratios with components from exponential family was build. Commonly used dynamic mixture models mostly uses fixed weights of components. This work overcame this limitation and provided universally approximating parametric model for a general dynamic system. Also, the approximate Bayesian learning, presented in [3] and summarized in Chapter 2, was specified for the mixture ratio models.

Two important types of components from EF were closely studied - Markov chains components (Chapter 5) and Gaussian components (Chapter 6). The first type serves as a model for discrete variables dependent on discrete regression vector while the latter one models continuous variables dependent on continuous regression vector. For mixtures with these components, the general learning algorithm described in Chapter 4 was converted into particular formulas and summarized in Algorithm 3 and Algorithm 5. The formulas include necessary numerical approximations of integrals occurring in Bayes' theorem (Theorem 1) and in Kullback-Leibler divergence (Definition 8) used for projection of pds.

The way how the mixture ratio model was constructed, i.e. by modelling joint pd on output and regression vector, allowed to model also mixed discrete-continuous systems. The practical modelling of such systems is complicated mainly if the dependence of discrete variable on continuous regression vector is present. The mixture ratio model can cope easily with such cases, because the joint pd can always be factorized and modelled in the way that overcomes this as it shown in Remark 11.

Chapter 7 provided simulation as well as real-life examples demonstrating simulation and learning strength of mixture ratio model. It was compared with standard mixtures, introduced in [18], in three set-ups - 1) basic simulation of mixture of two Gaussian components was made, 2) Monte Carlo simulation study of learning and predicting quality of both mixtures with Markov chain components, 3) both models was compared in real-life scenario involving trading with futures on real data. The main outcomes are following:

- When both models are simulated, mixture ratio model shows true dynamic behaviour compared to the standard mixture, see Figure 7.3.
- Mixture ratio model showed similar predictive quality (even slightly better) to standard mixture model when the system was simulated from standard mixture model.
- When the system was simulated from mixture ratio model with parameters giving true dynamic weights, the standard mixture model showed significantly worse predictive quality compared with the mixture ratio model.
- Mixture ratio model showed its applicability for decision making in real-life scenario - trading with futures. It produced slightly better results in terms of accumulated reward then the standard mixture model.

The conversion of the general approximate learning algorithm from Chapter 2 for these specific components required a lot of detailed technical work present throughout the thesis and in Appendix.

Although, the theoretical and experimental parts of this thesis show that it was worthwhile as the mixture ratio models provide a universal approach for modelling general dynamic systems and designing DM strategies.

Many challenging tasks remains to be studied, in particular: i) simulation and real-life testing mixture ratio model with Gaussian and mixed components, ii) examining the suitability of the proposed numerical approximations for high-dimensional models, iii) refining the data-dependent choice of forgetting factors, iv) including structure estimation of the mixture ratio model as well as its use for feature extraction similarly as in [33].

Appendix

Derivation of L's (5.11 - 5.13)

In this section, the formulas (5.11 - 5.13) are proved. The derivation starts with recalling definitions of treated objects:

$$L_{-1} = \sum_{c \in \mathbf{c}} \chi_{\phi_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} \int_{\mathbf{\Theta}} \mathrm{H}(\psi_t, \Theta) \mathrm{Q}_{t;c}(\Theta) d\Theta \tag{7.25}$$

$$L_{0c} = \sum_{c \in \mathbf{c}} \chi_{\phi_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} \int_{\mathbf{\Theta}} \ln(\alpha_c) \mathrm{H}(\psi_t, \Theta) \mathrm{Q}_{t;c}(\Theta) d\Theta \tag{7.26}$$

$$L_{cj} = \sum_{c \in \mathbf{c}} \chi_{\phi_c} \gamma_{t;c} \prod_{j=1}^{m_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} \int_{\mathbf{\Theta}} \ln(\omega_{cj}) \mathrm{H}(\psi_t, \Theta) \mathrm{Q}_{t;c}(\Theta) d\Theta \quad \forall j \in \{1, 2..., m_c\} \tag{7.27}$$

$$\mathrm{H}_t(\Theta) \approx \mathrm{H}_t(\tilde{\Theta}_t^d) + \sum_{c \in \mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c} (\alpha_c - \tilde{\alpha}_{t;c}^d) + \sum_{c \in \mathbf{c}} \sum_{j=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{cj}} (\omega_{cj} - \tilde{\omega}_{t;cj}^d), \tag{7.28}$$

$$\text{where} \qquad \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c} = -\mathrm{H}_t^2(\tilde{\Theta}_t^d) \sum_{O_t \in \mathbf{O}} \chi_{\phi_c} \prod_{j=1}^{m_c} \left(\frac{\tilde{\omega}_{cj}^d}{K_{cj}} \right)^{\Delta_{cj}}$$

$$\frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{cj}} = -\mathrm{H}_t^2(\tilde{\Theta}_t^d) \tilde{\alpha}_{t;c}^d \sum_{O_t \in \mathbf{O}} \chi_{\phi_c} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} 0^{1-\Delta_{cj}}$$

To prove the formulas (5.11 - 5.13), the following integrals will be used (see Definition 6 and Definition 5):

$$\int_{\boldsymbol{x}} \prod_{i=1}^{n} x_i^{p_i - 1} dx = \mathrm{Be}(p), \tag{7.29}$$

$$\int_{\boldsymbol{x}} \ln(x_j) \prod_{i=1}^{n} x_i^{p_i - 1} dx = \frac{\partial}{\partial p_j} \int_{\boldsymbol{x}} \prod_{i=1}^{n} x_i^{p_i - 1} dx = \frac{\partial}{\partial p_j} \mathrm{Be}(p) = \frac{\partial}{\partial p_j} \frac{\prod_{i=1}^{n} \Gamma(p_i)}{\Gamma(\sum_{i=1}^{m} p_i)}$$

$$= \frac{\Gamma'(p_i) \prod_{\substack{i=1 \\ i \neq j}}^{n} \Gamma(p_i)}{\Gamma(\sum_{i=1}^{m} p_i)} - \frac{\prod_{i=1}^{n} \Gamma(p_i) \Gamma'(\sum_{i=1}^{m} p_i)}{\Gamma^2(\sum_{i=1}^{m} p_i)} = \mathrm{Be}(p) \left(\psi(p_j) - \psi\left(\sum_{i=1}^{n} p_i \right) \right), \tag{7.30}$$

where

$$p \in (\mathbb{R}^+)^n, \quad x \in \boldsymbol{x} = \left\{ x \in [0,1]^n, \sum_{i=1}^{n} x_i = 1 \right\}$$

By inserting (7.28) into (7.26) it is obtained:

$$\begin{aligned} L_{-1}(\phi_t, \mathrm{V}_{t-1}) &= \sum_{d \in \mathbf{c}} \chi_{\boldsymbol{\phi}_d} \gamma_{t;d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}} \right)^{\Delta_{dj}} \int_{\boldsymbol{\Theta}} \Big[\Big(\mathrm{H}_t(\tilde{\Theta}_t^d) + \sum_{e \in \mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} (\alpha_e - \tilde{\alpha}_{t;e}^d) \\ &\qquad + \sum_{e \in \mathbf{c}} \sum_{i=1}^{m_e} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ei}} (\omega_{ei} - \tilde{\omega}_{t;ei}^d) \Big) \mathrm{Q}_{t;d}(\Theta) d\Theta \Big] \\ &= \sum_{d \in \mathbf{c}} \chi_{\boldsymbol{\phi}_d} \gamma_{t;d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}} \right)^{\Delta_{dj}} \mathrm{H}_t(\tilde{\Theta}_t^d), \end{aligned}$$

which proves (5.11).

Subsequently, insertion of (7.28) into (7.26) gives

$$\begin{aligned} L_{0c} &= \sum_{d \in \mathbf{c}} \chi_{\boldsymbol{\phi}_d} \gamma_{t;d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}} \right)^{\Delta_{dj}} \int_{\boldsymbol{\Theta}} \Big[\ln(\alpha_c) \Big(\mathrm{H}_t(\tilde{\Theta}_t^d) + \sum_{e \in \mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} (\alpha_e - \tilde{\alpha}_{t;e}^d) \\ &\qquad + \sum_{e \in \mathbf{c}} \sum_{i=1}^{m_e} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ei}} (\omega_{ei} - \tilde{\omega}_{t;ei}^d) \Big) \mathrm{Q}_{t;d}(\Theta) d\Theta \Big] \\ &= \sum_{d \in \mathbf{c}} \chi_{\boldsymbol{\phi}_d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}} \right)^{\Delta_{dj}} \Bigg(\mathrm{H}_t(\tilde{\Theta}_t^d) \prod_{f \in \mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;f} + \delta_{df} \Delta_d) \int_{\boldsymbol{\alpha}} \ln(\alpha_c) \prod_{f \in \mathbf{c}} \alpha_f^{\mathrm{v}_{t-1;f} - 1 + \delta_{fd}} d\alpha \\ &\qquad \underbrace{+ \sum_{e \in \mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} \prod_{f \in \mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;f} + \delta_{df} \Delta_d) \int_{\boldsymbol{\alpha}} \ln(\alpha_c)(\alpha_e - \tilde{\alpha}_{t;e}^d) \prod_{f \in \mathbf{c}} \alpha_f^{\mathrm{v}_{t-1;f} - 1 + \delta_{fd}} d\alpha}_{\mathrm{I}} \Bigg) \\ &= \sum_{d \in \mathbf{c}} \chi_{\boldsymbol{\phi}_d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}} \right)^{\Delta_{dj}} \Big(\gamma_{t;d} \mathrm{H}_t(\tilde{\Theta}_t^d) (\psi(\mathrm{v}_c + \delta_{cd}) - \psi(\Sigma_{e \in \mathbf{c}} \mathrm{v}_e + 1)) + \mathrm{I} \Big). \end{aligned} \tag{7.31}$$

The remaining part of L_{0c}, I, is computed as follows:

$$
\begin{aligned}
\mathrm{I} &= \sum_{e\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} \prod_{f\in\mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;f}+\delta_{df}\Delta_d)\Big[\mathrm{Be}(\upsilon+\delta_d+\delta_e)\Big(\psi(\mathrm{v}_c+\delta_{cd}+\delta_{ce})-\psi(\Sigma_{f\in\mathbf{c}}\mathrm{v}_f+2)\Big) \\
&\qquad\qquad\qquad - \Big(\tilde{\alpha}_{t;e}^d \mathrm{Be}(\upsilon+\delta_d)\Big(\psi(\mathrm{v}_c+\delta_{cd})-\psi(\Sigma_{f\in\mathbf{c}}\mathrm{v}_f+1)\Big)\Big)\Big] \\
&= \sum_{e\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} \tilde{\alpha}_{t;e}^d \gamma_{t;d} [\frac{\delta_{ce}}{\mathrm{v}_c+\delta_{cd}} - \frac{1}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1}] \\
&= \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c} \tilde{\alpha}_{t;c}^d \gamma_{t;d} [\frac{1}{\mathrm{v}_c+\delta_{cd}} - \frac{1}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1}] - \sum_{\substack{e\in\mathbf{c}\\ e\neq c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} \tilde{\alpha}_{t;e}^d \gamma_{t;d} [\frac{1}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1}] \\
&= \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c} \gamma_{t;d} \frac{\mathrm{v}_c+\delta_{cd}}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1} \frac{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1-(\mathrm{v}_c+\delta_{cd})}{(\mathrm{v}_c+\delta_{cd})(\sum_{f\in\mathbf{c}}\mathrm{v}_f+1)} - \sum_{\substack{e\in\mathbf{c}\\ e\neq c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} \tilde{\alpha}_{t;e}^d \gamma_{t;d} [\frac{1}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1}] \\
&= \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c} \gamma_{t;d} \left(\frac{1}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1}\right) \sum_{\substack{e\in\mathbf{c}\\ e\neq c}} \tilde{\alpha}_{t;e}^d - \sum_{\substack{e\in\mathbf{c}\\ e\neq c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} \tilde{\alpha}_{t;e}^d \gamma_{t;d} \left(\frac{1}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1}\right) \\
&= \frac{\gamma_{t;d}}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1} \sum_{\substack{e\in\mathbf{c}\\ e\neq c}} \tilde{\alpha}_{t;e}^d \left(\frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_c} - \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e}\right) \\
&= \frac{\gamma_{t;d}}{\sum_{f\in\mathbf{c}}\mathrm{v}_f+1} \sum_{\substack{e\in\mathbf{c}\\ e\neq c}} \tilde{\alpha}_{t;e}^d \mathrm{H}_t^2(\tilde{\Theta}_t^d) \left(\sum_{O_t\in\mathbf{O}} \chi_{\boldsymbol{\phi}_e} \prod_{j=1}^{m_e} \left(\frac{\tilde{\omega}_{ej}^d}{K_{ej}}\right)^{\Delta_{ej}} - \sum_{O_t\in\mathbf{O}} \chi_{\boldsymbol{\phi}_c} \prod_{j=1}^{m_c} \left(\frac{\tilde{\omega}_{cj}^d}{K_{cj}}\right)^{\Delta_{cj}}\right). \qquad (7.32)
\end{aligned}
$$

Hence, (7.31) and (7.32) prove the formula (5.12).

Finally, by inserting (7.28) into (7.27), it is obtained:

$$
\begin{aligned}
L_{cj} &= \sum_{d\in\mathbf{c}} \chi_{\boldsymbol{\phi}_d} \gamma_{t;d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}}\right)^{\Delta_{dj}} \int_{\boldsymbol{\Theta}} \Big[\ln(\omega_{cj})\Big(\mathrm{H}_t(\tilde{\Theta}_t^d) + \sum_{e\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e}(\alpha_e - \tilde{\alpha}_{t;e}^d) \\
&\qquad\qquad + \sum_{e\in\mathbf{c}} \sum_{i=1}^{m_e} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ei}}(\omega_{ei} - \tilde{\omega}_{t;ei}^d)\Big)\mathrm{Q}_{t;d}(\Theta) d\Theta\Big] \\
&= \sum_{d\in\mathbf{c}} \chi_{\boldsymbol{\phi}_d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}}\right)^{\Delta_{dj}} \Bigg(\mathrm{H}_t(\tilde{\Theta}_t^d)\mathrm{Be}(\mathrm{v}_{t-1}+\delta_d) \prod_{\substack{f\in\mathbf{c}\\ f\neq c}} \mathrm{Be}(\mathrm{V}_{t-1;f}+\delta_{df}\Delta_d) \int_{\boldsymbol{\omega}_c} \ln(\omega_{cj}) \prod_{i=1}^{m_c} \omega_{ci}^{\mathrm{V}_{t-1;ci}-1+\delta_{cd}\Delta_{ci}} d\omega \\
&\qquad \underbrace{+ \sum_{i=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ci}} \mathrm{Be}(\mathrm{v}_{t-1}+\delta_d) \prod_{\substack{f\in\mathbf{c}\\ f\neq c}} \mathrm{Be}(\mathrm{V}_{t-1;f}+\delta_{df}\Delta_d) \int_{\boldsymbol{\omega}_c} \ln(\omega_{cj})(\omega_{ci} - \tilde{\omega}_{t;ci}^d) \prod_{i=1}^{m_c} \omega_{ci}^{\mathrm{V}_{t-1;ci}-1+\delta_{cd}\Delta_{ci}} d\omega \Bigg)}_{\mathrm{II}} \\
&= \sum_{d\in\mathbf{c}} \chi_{\boldsymbol{\phi}_d} \prod_{j=1}^{m_d} \left(\frac{1}{K_{dj}}\right)^{\Delta_{dj}} \Big(\gamma_{t;d}\mathrm{H}_t(\tilde{\Theta}_t^d)\Big(\psi(\mathrm{V}_{t-1;cj}+\delta_{cd}\Delta_{cj}) - \psi\Big(\Sigma_{i=1}^{m_c}(\mathrm{V}_{t-1;ci}+\delta_{cd}\Delta_{ci})\Big)\Big) + \mathrm{II}\Big), \qquad (7.33)
\end{aligned}
$$

where the remaining part II is computed as follows:

$$
\begin{aligned}
\mathrm{II} = & \sum_{i=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ci}} \mathrm{Be}(\mathrm{v}_{t-1} + \delta_d) \prod_{\substack{f \in \mathbf{c} \\ f \neq c}} \mathrm{Be}(\mathrm{V}_{t-1;f} + \delta_{df}\Delta_d) \Big[\mathrm{Be}(\mathrm{V}_{t-1;c} + \delta_{cd}\Delta_c + \delta_i) \\
& \times \Big(\psi(\mathrm{V}_{cj} + \delta_{cd}\Delta_{cj} + \delta_{ij}) - \psi\Big(\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci}) + 1 \Big) \Big) \\
& - \tilde{\omega}_{t;ci}^d \mathrm{Be}(\mathrm{V}_{t-1;c} + \delta_{cd}\Delta_c) \Big(\psi(\mathrm{V}_{cj} + \delta_{cd}\Delta_{cj}) - \psi\Big(\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci}) \Big) \Big) \Big] \\
= & \sum_{i=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ci}} \gamma_{t;d} \tilde{\omega}_{t;ci}^d \Big[\psi(\mathrm{V}_{cj} + \delta_{cd}\Delta_{cj} + \delta_{ij}) - \psi\Big(\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci}) + 1 \Big) \\
& - \Big(\psi(\mathrm{V}_{cj} + \delta_{cd}\Delta_{cj}) - \psi\Big(\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci}) \Big) \Big] \\
= & \sum_{i=1}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \alpha_e} \gamma_{t;d} \tilde{\omega}_{t;ci}^d \Big[\frac{\delta_{ij}}{\mathrm{V}_{t-1;cj} + \delta_{cd}\Delta_{cj}} - \frac{1}{\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci})} \Big] \\
= & \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{cj}} \gamma_{t;d} \frac{\mathrm{V}_{t-1;cj} + \delta_{cd}\Delta_{cj}}{\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci})} \Big[\frac{\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci}) - (\mathrm{V}_{t-1;cj} + \delta_{cd}\Delta_{cj})}{(\mathrm{V}_{t-1;cj} + \delta_{cd}\Delta_{cj}) \Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci})} \Big] \\
& - \sum_{\substack{i=1 \\ i \neq j}}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ci}} \gamma_{t;d} \tilde{\omega}_{t;ci}^d \Big[\frac{1}{\Sigma_{i=1}^{m_c} (\mathrm{V}_{ci} + \delta_{cd}\Delta_{ci})} \Big] \\
= & \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{cj}} \gamma_{t;d} \sum_{\substack{i=1 \\ i \neq j}}^{m_c} \tilde{\omega}_{t;ci}^d \Big[\frac{1}{\Sigma_{i=1}^{m_c} (V_{ci} + \delta_{cd}\Delta_{ci})} \Big] - \sum_{\substack{i=1 \\ i \neq j}}^{m_c} \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ci}} \gamma_{t;d} \tilde{\omega}_{t;ci}^d \Big[\frac{1}{\Sigma_{i=1}^{m_c} (V_{ci} + \delta_{cd}\Delta_{ci})} \Big] \\
= & \gamma_{t;d} \Big[\frac{1}{\Sigma_{i=1}^{m_c} (V_{ci} + \delta_{cd}\Delta_{ci})} \Big] \sum_{\substack{i=1 \\ i \neq j}}^{m_c} \left(\frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{cj}} - \frac{\partial \mathrm{H}_t(\tilde{\Theta}_t^d)}{\partial \omega_{ci}} \right) \tilde{\omega}_{t;ci}^d \\
= & \frac{\chi_{\boldsymbol{\phi}_c} \gamma_{t;d} \mathrm{H}_t^2(\tilde{\Theta}_t^d)}{\Sigma_{i=1}^{m_c} (V_{ci} + \delta_{cd}\Delta_{ci})} \sum_{\substack{i=1 \\ i \neq j}}^{m_c} \tilde{\omega}_{t;ci}^d \tilde{\alpha}_{t;c}^d \left(\sum_{O_t \in \mathbf{O}} \left(\frac{1}{K_{ci}} \right)^{\Delta_{ci}} 0^{1-\Delta_{ci}} - \sum_{O_t \in \mathbf{O}} \left(\frac{1}{K_{cj}} \right)^{\Delta_{cj}} 0^{1-\Delta_{cj}} \right),
\end{aligned}
\tag{7.34}
$$

which, along with (7.33), proves (5.13).

Derivation of (5.14- 5.16)

The formulas (5.14 - 5.16) (with $\tilde{\gamma}_{t;c}$) is derived by applying the properties of beta function (see Definition 6). Schematically, the derivation progresses as follows:

$$\frac{\sum_{c\in\mathbf{c}} C_c \prod_{j=1}^{m_c}\left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \gamma_{t;c}}{\sum_{c\in\mathbf{c}} D_c \prod_{j=1}^{m_c}\left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \gamma_{t;c}} = \frac{\sum_{c\in\mathbf{c}} C_c \prod_{j=1}^{m_c}\left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{Be}(\mathrm{v}_{t-1}+\delta(c)) \prod_{d\in\mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d}+\delta_{cd}\Delta_c)}{\sum_{c\in\mathbf{c}} D_c \prod_{j=1}^{m_c}\left(\frac{1}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{Be}(\mathrm{v}_{t-1}+\delta(c)) \prod_{d\in\mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d}+\delta_{cd}\Delta_c)}$$

$$= \frac{\sum_{c\in\mathbf{c}} C_c \tilde{\tilde{\alpha}}_{t;c} \prod_{j=1}^{m_c}\left(\frac{\tilde{\tilde{\omega}}_{t;cj}}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{Be}(\mathrm{v}_{t-1}) \prod_{d\in\mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d})}{\sum_{c\in\mathbf{c}} D_c \tilde{\tilde{\alpha}}_{t;c} \prod_{j=1}^{m_c}\left(\frac{\tilde{\tilde{\omega}}_{t;cj}}{K_{cj}}\right)^{\Delta_{cj}} \mathrm{Be}(\mathrm{v}_{t-1}) \prod_{d\in\mathbf{c}} \mathrm{Be}(\mathrm{V}_{t-1;d})}$$

$$= \frac{\sum_{c\in\mathbf{c}} C_c \tilde{\tilde{\alpha}}_{t;c} \prod_{j=1}^{m_c}\left(\frac{\tilde{\tilde{\omega}}_{t;cj}}{K_{cj}}\right)^{\Delta_{cj}}}{\sum_{c\in\mathbf{c}} D_c \underbrace{\tilde{\tilde{\alpha}}_{t;c} \prod_{j=1}^{m_c}\left(\frac{\tilde{\tilde{\omega}}_{t;cj}}{K_{cj}}\right)^{\Delta_{cj}}}_{\tilde{\gamma}_{t;c}}},$$

$$\text{where} \qquad \tilde{\tilde{\alpha}}_{t;c} = \frac{\mathrm{v}_{t-1;c}}{\sum_{d\in\mathbf{c}} \mathrm{v}_{t-1;d}} \qquad \tilde{\tilde{\omega}}_{t;cj} = \frac{\mathrm{V}_{t-1;cj}}{\sum_{i=1}^{m_c} \mathrm{V}_{t-1;ci}}.$$

The form of $C_c, D_c, c \in \mathbf{c}$ depends on the particular formula (5.14 - 5.16), but the derivation remains unchanged for all of formulas (5.14 - 5.16).

Derivation of L_{ci} (6.26)

In this section the derivation of L_{ci} (6.26) will be provided. See Chapter 6 for the meaning and explanation of all of the used symbols here. It is computed from the following general formulas (6.19):

$$L_{ci} = -\frac{1}{2}(\nu_{ci} + |\psi_{ci}| + 2)L^1_{ci} - 0.5\Lambda_{ci}L^2_{ci} - 0.5L^3_{ci} \tag{7.35}$$

$$L^1_{ci} = \int_{\mathbf{\Theta}} \ln(r_{ci})\mathrm{H}_t(\Theta) \sum_{d\in\mathbf{c}} \gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta \tag{7.36}$$

$$L^2_{ci} = \int_{\mathbf{\Theta}} \frac{1}{r_{ci}}\mathrm{H}_t(\Theta) \sum_{d\in\mathbf{c}} \gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta \tag{7.37}$$

$$L^3_{ci} = \int_{\mathbf{\Theta}} \frac{1}{r_{ci}}\left((\Theta_{ci} - \hat{\Theta}_{ci})^T V_{\psi ci}(\Theta_{ci} - \hat{\Theta}_{ci})\right)\mathrm{H}_t(\Theta) \sum_{d\in\mathbf{c}} \gamma_{t;d}\mathrm{Q}_{t;d}(\Theta)d\Theta, \tag{7.38}$$

where the function $\mathrm{H}_t(\Theta)$ is approximated via linear Taylor expansion around the expected values of parameter Θ (6.21) - (6.23) as follows

$$\mathrm{H}_t(\Theta) \approx \mathrm{H}_t(\hat{\Theta}^d_t) + \sum_{c\in\mathbf{c}} \frac{\partial \mathrm{H}_t(\hat{\Theta}^d_t)}{\partial \alpha_c}(\alpha_c - \hat{\alpha}^d_{t;c}) + \sum_{c\in\mathbf{c}} \sum_{j=1}^{n_c} \left(\frac{\partial \mathrm{H}_t(\hat{\Theta}^d_t)}{\partial r_{cj}}\left(r_{cj} - \hat{r}^d_{t;cj}\right) + \frac{\partial \mathrm{H}_t(\hat{\Theta}^d_t)}{\partial \Theta_{cj}}\left(\Theta_{cj} - \hat{\Theta}^d_{t;cj}\right)\right). \tag{7.39}$$

Following integrals will be useful and the derivation will refer to them:

$$\int_{\boldsymbol{r}_{ci}} r_{ci}^{-0.5(\tilde{v}^d_{t;ci}+2)} \exp\left\{-\frac{\tilde{\Lambda}^d_{t;ci}}{2r_{ci}}\right\} dr_{ci} = \left| x = \frac{\tilde{\Lambda}^d_{t;ci}}{2r_{ci}},\ dr_{ci} = -\frac{\tilde{\Lambda}^d_{t;ci}}{2x^2} dx \right| = \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\tilde{v}^d_{t;ci}} \int_0^\infty x^{0.5(\tilde{v}^d_{t;ci}-2)} e^{-x} dx$$
$$= \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\tilde{v}^d_{t;ci}} \Gamma\left(\frac{\tilde{v}^d_{t;ci}}{2}\right) \qquad (7.40)$$

$$\int_{\boldsymbol{r}_{ci}} r_{ci}^{-0.5(\tilde{v}^d_{t;ci})} \exp\left\{-\frac{\tilde{\Lambda}^d_{t;ci}}{2r_{ci}}\right\} dr_{ci} = \left| x = \frac{\tilde{\Lambda}^d_{t;ci}}{2r_{ci}},\ dr_{ci} = -\frac{\tilde{\Lambda}^d_{t;ci}}{2x^2} dx \right| = \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\left(\tilde{v}^d_{t;ci}-2\right)} \int_0^\infty x^{0.5(\tilde{v}^d_{t;ci}-4)} e^{-x} dx$$
$$= \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\left(\tilde{v}^d_{t;ci}-2\right)} \Gamma\left(\frac{\tilde{v}^d_{t;ci}}{2} - 1\right) = \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\left(\tilde{v}^d_{t;ci}-2\right)} \Gamma\left(\frac{\tilde{v}^d_{t;ci}}{2}\right) \frac{2}{\tilde{v}^d_{t;ci} - 2} \qquad (7.41)$$

$$\int_{\boldsymbol{r}_{ci}} r_{ci}^{-0.5(\tilde{v}^d_{t;ci}+4)} \exp\left\{-\frac{\tilde{\Lambda}^d_{t;ci}}{2r_{ci}}\right\} dr_{ci} = \left| x = \frac{\tilde{\Lambda}^d_{t;ci}}{2r_{ci}},\ dr_{ci} = -\frac{\tilde{\Lambda}^d_{t;ci}}{2x^2} dx \right| = \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\left(\tilde{v}^d_{t;ci}-2\right)} \int_0^\infty x^{0.5(\tilde{v}^d_{t;ci}-4)} e^{-x} dx$$
$$= \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\left(\tilde{v}^d_{t;ci}+2\right)} \Gamma\left(\frac{\tilde{v}^d_{t;ci}}{2} + 1\right) = \left(\frac{\tilde{\Lambda}^d_{t;ci}}{2}\right)^{-0.5\left(\tilde{v}^d_{t;ci}+2\right)} \Gamma\left(\frac{\tilde{v}^d_{t;ci}}{2}\right) \frac{\tilde{v}^d_{t;ci}}{2} \qquad (7.42)$$

L^1_{ci} **6.27**

By inserting (7.39) into (7.36) it is obtained:

$$L^1_{ci} \approx \sum_{d \in \mathbf{c}} \mathrm{Be}(\tilde{\mathrm{v}}^d_t) \prod_{\substack{e \in \mathbf{c} \\ e \neq c}} \prod_{j=1}^{n_e} \mathrm{N}(\tilde{v}^d_{t;ej}, \tilde{G}^d_{t;ej}) / \mathrm{P}_e(\psi^{\mathsf{c}}_e)$$
$$\times \int_{\boldsymbol{r}_{ci}} \int_{\boldsymbol{\Theta}_{ci}} \ln(r_{ci}) \left(\mathrm{H}_t(\hat{\Theta}^d_t) + \frac{\partial \mathrm{H}_t(\hat{\Theta}^d_t)}{\partial r_{ci}} \left(r_{ci} - \hat{r}^d_{t;ci}\right) \right) r_{ci}^{-0.5(\nu_{ci}+|\boldsymbol{\psi}_{ci}|+2)} \mathrm{P}(r_{ci}, \Theta_{ci} | \tilde{v}^d_{t;ci}, \tilde{V}^d_{t;ci}) d\Theta_{ci} dr_{ci} \qquad (7.43)$$

The integral in (7.43) is computed in the following steps:

$$\begin{aligned}
&\int_{\boldsymbol{r}_{ci}}\int_{\boldsymbol{\Theta}_{ci}}\Bigg[\ln(r_{ci})\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial\mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)\right)r_{ci}^{-0.5(\nu_{ci}+|\boldsymbol{\psi}_{ci}|+2)}\\
&\qquad\times\exp\left\{-\frac{1}{2r_{ci}}\left((\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d)^T\tilde{V}_{t;\psi ci}^d(\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d)+\tilde{\Lambda}_{t;ci}^d\right)\right\}d\Theta_{ci}dr_{ci}\Bigg]\\
&=\left|x=r_{ci}^{-0.5}(\tilde{V}_{t;\psi ci}^d)^{0.5}(\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d),\;\; dx=r_{ci}^{-0.5|\boldsymbol{\psi}_{ci}|}\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}d\Theta_{ci}\right|\\
&=\frac{1}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\int_{\boldsymbol{r}_{ci}}\int_{\boldsymbol{\Theta}_{ci}}\ln(r_{ci})\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial\mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)\right)r_{ci}^{-0.5(\tilde{\nu}_{t;ci}^d+2)}\exp\left\{-\frac{\tilde{\Lambda}_{t;ci}^d}{2r_{ci}}\right\}\exp\left\{-\frac{x^Tx}{2}\right\}dxdr_{ci}\\
&=\frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\int_{\boldsymbol{r}_{ci}}\ln(r_{ci})\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial\mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)\right)r_{ci}^{-0.5(\tilde{\nu}_{t;ci}^d+2)}\exp\left\{-\frac{\tilde{\Lambda}_{t;ci}^d}{2r_{ci}}\right\}dr_{ci}\\
&=-2\frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\frac{\partial}{\partial\tilde{\nu}_{t;ci}^d}\int_{\boldsymbol{r}_{ci}}\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial\mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)\right)r_{ci}^{-0.5(\tilde{\nu}_{t;ci}^d+2)}\exp\left\{-\frac{\tilde{\Lambda}_{t;ci}^d}{2r_{ci}}\right\}dr_{ci}\\
&=-2\frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\mathrm{H}_t(\hat{\Theta}_t^d)\frac{\partial}{\partial\tilde{\nu}_{t;ci}^d}\int_0^\infty r_{ci}^{-0.5(\tilde{\nu}_{t;ci}^d+2)}\exp\left\{-\frac{\tilde{\Lambda}_{t;ci}^d}{2r_{ci}}\right\}dr_{ci}\\
&\overbrace{=}^{(7.40)}-2\frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\mathrm{H}_t(\hat{\Theta}_t^d)\frac{\partial}{\partial\tilde{\nu}_{t;ci}^d}\left(\frac{\tilde{\Lambda}_{t;ci}^d}{2}\right)^{-0.5\tilde{\nu}_{t;ci}^d}\Gamma\left(\frac{\tilde{\nu}_{t;ci}^d}{2}\right)\\
&=\frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\mathrm{H}_t(\hat{\Theta}_t^d)\left(\frac{\tilde{\Lambda}_{t;ci}^d}{2}\right)^{-0.5\tilde{\nu}_{t;ci}^d}\Gamma\left(\frac{\tilde{\nu}_{t;ci}^d}{2}\right)\left(\ln\left(0.5\tilde{\Lambda}_{t;ci}^d\right)-\psi\left(0.5\tilde{\nu}_{t;ci}^d\right)\right)
\end{aligned}\tag{7.44}$$

By using (7.44), the integral L_{ci}^1 (7.43) is computed as follows:

$$L_{ci}^1\approx\sum_{d\in\mathbf{c}}\gamma_{t;d}\left[\mathrm{H}(\hat{\Theta}_t^d)\left(\ln\left(0.5\tilde{\Lambda}_{t;cj}^d\right)-\psi\left(0.5\tilde{\nu}_{t;cj}^d\right)\right)\right],$$

which proves (6.27).

L_{ci}^2 (6.28)

By inserting (7.39) into (7.37) it is obtained:

$$\begin{aligned}
L_{ci}^2&\approx\sum_{d\in\mathbf{c}}\mathrm{Be}(\tilde{\mathrm{v}}_t^d)\prod_{\substack{e\in\mathbf{c}\\ e\neq c}}\prod_{j=1}^{n_e}\mathrm{N}(\tilde{\nu}_{t;ej}^d,\tilde{G}_{t;ej}^d)/\mathrm{P}_e(\psi_e^{\mathsf{c}})\\
&\quad\times\int_{\boldsymbol{r}_{ci}}\int_{\boldsymbol{\Theta}_{ci}}\frac{1}{r_{ci}}\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial\mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)\right)r_{ci}^{-0.5(\nu_{ci}+|\boldsymbol{\psi}_{ci}|+2)}\mathrm{P}(r_{ci},\Theta_{ci}|\tilde{\nu}_{t;ci}^d,\tilde{V}_{t;ci}^d)d\Theta_{ci}dr_{ci}
\end{aligned}\tag{7.45}$$

Similarly to (7.44), the integral in (7.45) is derived as follows:

$$\int_{\boldsymbol{r}_{ci}}\int_{\boldsymbol{\Theta}_{ci}}\Bigg[\frac{1}{r_{ci}}\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)\right)r_{ci}^{-0.5(\nu_{ci}+|\boldsymbol{\psi}_{ci}|+2)}$$

$$\times\exp\left\{-\frac{1}{2r_{ci}}\left((\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d)^T\tilde{V}_{t;\psi ci}^d(\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d)+\tilde{\Lambda}_{t;ci}^d\right)\right\}d\Theta_{ci}dr_{ci}\Bigg]$$

$$=\left|x=r_{ci}^{-0.5}(\tilde{V}_{t;\psi ci}^d)^{0.5}(\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d),\ \ dx=r_{ci}^{-0.5|\boldsymbol{\psi}_{ci}|}\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}d\Theta_{ci}\right|$$

$$=\frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\int_{\boldsymbol{r}_{ci}}\frac{1}{r_{ci}}\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)\right)r_{ci}^{-0.5(\tilde{\nu}_{t;ci}^d+2)}\exp\left\{-\frac{\tilde{\Lambda}_{t;ci}^d}{2r_{ci}}\right\}dr_{ci} \tag{7.46}$$

L_{ci}^2 is then computed by using (7.46), (7.40) and (7.42) as follows:

$$L_{ci}^2\approx\sum_{d\in\mathbf{c}}\gamma_{t;d}\left(\mathrm{H}(\hat{\Theta}_t^d)\frac{\tilde{\nu}_{ci}^d}{\tilde{\Lambda}_{t;ci}}+\frac{\partial \mathrm{H}(\hat{\Theta}_t^d)}{\partial r_{cj}}\frac{2}{\tilde{\nu}_{ci}^d+2}\right),$$

which proves (6.28).

L_{ci}^3 (6.29)

By inserting (7.39) into (7.38) it is obtained:

$$L_{ci}^3\approx\sum_{d\in\mathbf{c}}\mathrm{Be}(\tilde{\mathrm{v}}_t^d)\prod_{\substack{e\in\mathbf{c}\\e\neq c}}\prod_{j=1}^{n_e}\mathrm{N}(\tilde{\nu}_{t;ej}^d,\tilde{G}_{t;ej}^d)/\mathrm{P}_e(\psi_e^{\mathsf{c}})\int_{\boldsymbol{r}_{ci}}\int_{\boldsymbol{\Theta}_{ci}}\Bigg[\frac{1}{r_{ci}}\left((\Theta_{ci}-\hat{\Theta}_{ci})^TV_{\psi ci}(\Theta_{ci}-\hat{\Theta}_{ci})\right)$$

$$\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{ci}}\left(\Theta_{ci}-\hat{\Theta}_{t;ci}^d\right)\right)\mathrm{P}(r_{ci},\Theta_{ci}|\tilde{\nu}_{t;ci}^d,\tilde{V}_{t;ci}^d)d\Theta_{ci}dr_{ci}\Bigg] \tag{7.47}$$

The integration in (7.47) is done in the following steps:

$$\int_{\boldsymbol{r}_{ci}}\int_{\boldsymbol{\Theta}_{ci}}\Bigg[\frac{1}{r_{ci}}\left((\Theta_{ci}-\hat{\Theta}_{ci})^TV_{\psi ci}(\Theta_{ci}-\hat{\Theta}_{ci})\right)\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{ci}}\left(\Theta_{ci}-\hat{\Theta}_{t;ci}^d\right)\right)$$

$$r_{ci}^{-0.5(\nu_{ci}+|\boldsymbol{\psi}_{ci}|+2)}\exp\left\{-\frac{1}{2r_{ci}}\left((\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d)^T\tilde{V}_{t;\psi ci}^d(\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d)+\tilde{\Lambda}_{t;ci}^d\right)\right\}d\Theta_{ci}dr_{ci}\Bigg]$$

$$=\left|x=r_{ci}^{-0.5}(\tilde{V}_{t;\psi ci}^d)^{0.5}(\Theta_{ci}-\hat{\hat{\Theta}}_{t;ci}^d),\ \ dx=r_{ci}^{-0.5|\boldsymbol{\psi}_{ci}|}\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}d\Theta_{ci}\right|$$

$$=\frac{1}{\left|\tilde{V}_{t;\psi ci}^d\right|^{0.5}}\int_{\boldsymbol{r}_{ci}}\int_{\boldsymbol{\Theta}_{ci}}\Bigg[\frac{1}{r_{ci}}\left(\left(\hat{\hat{\Theta}}_{t;ci}^d-\hat{\Theta}_{ci}+r_{ci}^{0.5}\left(\tilde{V}_{t;\psi ci}^d\right)^{-0.5}x\right)^TV_{\psi ci}\left(\hat{\hat{\Theta}}_{t;ci}^d-\hat{\Theta}_{ci}+r_{ci}^{0.5}\left(\tilde{V}_{t;\psi ci}^d\right)^{-0.5}x\right)\right)$$

$$\left(\mathrm{H}_t(\hat{\Theta}_t^d)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial r_{ci}}\left(r_{ci}-\hat{r}_{t;ci}^d\right)+\frac{\partial \mathrm{H}_t(\hat{\Theta}_t^d)}{\partial \Theta_{ci}}\left(r_{ci}^{0.5}\left(\tilde{V}_{t;\psi ci}^d\right)^{-0.5}x\right)\right)$$

$$r_{ci}^{-0.5(\nu_{ci}+2)}\exp\left\{-\frac{x^Tx}{2}\right\}\exp\left\{-\frac{\tilde{\Lambda}_{t;ci}^d}{2r_{ci}}\right\}dxdr_{ci}\Bigg]=\mathrm{I}+\mathrm{II}. \tag{7.48}$$

The integrals I and II are computed as follows:

$$\begin{aligned}
\mathrm{I} &= \frac{1}{\left|\tilde{V}^{d}_{t;\psi ci}\right|^{0.5}} \int_{\boldsymbol{r}_{ci}} \int_{\boldsymbol{\Theta}_{ci}} \Bigg[\frac{1}{r_{ci}} \left(\left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} + r_{ci}^{0.5} \left(\tilde{V}^{d}_{t;\psi ci}\right)^{-0.5} x \right)^{T} V_{\psi ci} \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} + r_{ci}^{0.5} \left(\tilde{V}^{d}_{t;\psi ci}\right)^{-0.5} x \right) \right) \\
&\qquad \left(\mathrm{H}_t(\hat{\Theta}^{d}_{t}) + \frac{\partial \mathrm{H}_t(\hat{\Theta}^{d}_{t})}{\partial r_{ci}} \left(r_{ci} - \hat{r}^{d}_{t;ci} \right) \right) r_{ci}^{-0.5(\nu_{ci}+2)} \exp\left\{ -\frac{x^{T} x}{2} \right\} \exp\left\{ -\frac{\tilde{\Lambda}^{d}_{t;ci}}{2 r_{ci}} \right\} dx dr_{ci} \Bigg] \\
&= \frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}^{d}_{t;\psi ci}\right|^{0.5}} \int_{\boldsymbol{r}_{ci}} \Bigg[\left(\frac{1}{r_{ci}} \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} \right)^{T} V_{\psi ci} \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} \right) + \mathrm{tr}\left(\left(\tilde{V}^{d}_{t;\psi ci} \right)^{-1} V_{\psi ci} \right) \right) \\
&\qquad \left(\mathrm{H}_t(\hat{\Theta}^{d}_{t}) + \frac{\partial \mathrm{H}_t(\hat{\Theta}^{d}_{t})}{\partial r_{ci}} \left(r_{ci} - \hat{r}^{d}_{t;ci} \right) \right) r_{ci}^{-0.5(\nu_{ci}+2)} \exp\left\{ -\frac{\tilde{\Lambda}^{d}_{t;ci}}{2 r_{ci}} \right\} dr_{ci} \Bigg] \\
&\overbrace{=}^{(7.40),(7.42)} \frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}^{d}_{t;\psi ci}\right|^{0.5}} \left(\frac{\tilde{\Lambda}^{d}_{t;ci}}{2} \right)^{-0.5\tilde{\nu}^{d}_{t;ci}} \Gamma\left(\frac{\tilde{\nu}^{d}_{t;ci}}{2} \right) \Bigg\{ \mathrm{H}(\hat{\Theta}^{d}_{t}) \mathrm{tr}\left(\left(\tilde{V}^{d}_{\psi cj} \right)^{-1} V_{\psi cj} \right) \\
&\qquad + \left(\hat{\hat{\Theta}}^{d}_{t;cj} - \hat{\Theta}_{t;cj} \right)^{T} V_{\psi cj} \left(\hat{\hat{\Theta}}^{d}_{t;cj} - \hat{\Theta}_{t;cj} \right) \left(\mathrm{H}(\hat{\Theta}^{d}_{t}) \frac{\tilde{\nu}^{d}_{cj}}{\tilde{\Lambda}^{d}_{t;cj}} + \frac{\partial \mathrm{H}(\hat{\Theta}^{d}_{t})}{\partial r_{cj}} \frac{2}{\tilde{\nu}^{d}_{cj} + 2} \right) \Bigg\}, \qquad (7.49)
\end{aligned}$$

$$\begin{aligned}
\mathrm{II} &= \frac{1}{\left|\tilde{V}^{d}_{t;\psi ci}\right|^{0.5}} \int_{\boldsymbol{r}_{ci}} \int_{\boldsymbol{\Theta}_{ci}} \Bigg[\frac{1}{r_{ci}} \left(\left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} + r_{ci}^{0.5} \left(\tilde{V}^{d}_{t;\psi ci}\right)^{-0.5} x \right)^{T} V_{\psi ci} \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} + r_{ci}^{0.5} \left(\tilde{V}^{d}_{t;\psi ci}\right)^{-0.5} x \right) \right) \\
&\qquad \frac{\partial \mathrm{H}_t(\hat{\Theta}^{d}_{t})}{\partial \Theta_{ci}} \left(r_{ci}^{0.5} \left(\tilde{V}^{d}_{t;\psi ci} \right)^{-0.5} x \right) r_{ci}^{-0.5(\nu_{ci}+2)} \exp\left\{ -\frac{x^{T} x}{2} \right\} \exp\left\{ -\frac{\tilde{\Lambda}^{d}_{t;ci}}{2 r_{ci}} \right\} dx dr_{ci} \Bigg] \\
&= \frac{1}{\left|\tilde{V}^{d}_{t;\psi ci}\right|^{0.5}} \frac{\partial \mathrm{H}_t(\hat{\Theta}^{d}_{t})}{\partial \Theta_{ci}} \int_{\boldsymbol{r}_{ci}} \int_{\boldsymbol{\Theta}_{ci}} \Bigg[\left(\tilde{V}^{d}_{t;\psi ci} \right)^{-0.5} x \left(x^{T} \left(\left(\tilde{V}^{d}_{t;\psi ci} \right)^{-0.5} \right)^{T} V_{\psi ci} \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} \right) + \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} \right)^{T} V_{\psi ci} \left(\tilde{V}^{d}_{t;\psi ci} \right)^{-0.5} x \right) \\
&\qquad r_{ci}^{-0.5(\nu_{ci}+2)} \exp\left\{ -\frac{x^{T} x}{2} \right\} \exp\left\{ -\frac{\tilde{\Lambda}^{d}_{t;ci}}{2 r_{ci}} \right\} dx dr_{ci} \Bigg] \\
&= \frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}^{d}_{t;\psi ci}\right|^{0.5}} \frac{\partial \mathrm{H}_t(\hat{\Theta}^{d}_{t})}{\partial \Theta_{ci}} \left(\left(\tilde{V}^{d}_{t;\psi ci} \right)^{-1} \left(V_{\psi ci} + V^{T}_{\psi ci} \right) \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} \right) \right) \int_{\boldsymbol{r}_{ci}} r_{ci}^{-0.5(\nu_{ci}+2)} \exp\left\{ -\frac{\tilde{\Lambda}^{d}_{t;ci}}{2 r_{ci}} \right\} dr_{ci} \\
&\overbrace{=}^{(7.40)} \frac{(2\pi)^{0.5|\boldsymbol{\psi}_{ci}|}}{\left|\tilde{V}^{d}_{t;\psi ci}\right|^{0.5}} \left(\frac{\tilde{\Lambda}^{d}_{t;ci}}{2} \right)^{-0.5\tilde{\nu}^{d}_{t;ci}} \Gamma\left(\frac{\tilde{\nu}^{d}_{t;ci}}{2} \right) \frac{\partial \mathrm{H}_t(\hat{\Theta}^{d}_{t})}{\partial \Theta_{ci}} \left(\left(\tilde{V}^{d}_{t;\psi ci} \right)^{-1} \left(V_{\psi ci} + V^{T}_{\psi ci} \right) \left(\hat{\hat{\Theta}}^{d}_{t;ci} - \hat{\Theta}_{ci} \right) \right). \qquad (7.50)
\end{aligned}$$

By combining (7.47), (7.48), (7.49) and (7.50), L_{ci}^{3} is obtained as follows:

$$
\begin{aligned}
L_{cj}^{3} \approx \sum_{d\in\mathbf{c}} \gamma_{t;d} \Bigg[& \mathrm{H}(\hat{\Theta}_t^d)\mathrm{tr}\left(\left(\tilde{V}_{\psi cj}^{d}\right)^{-1} V_{\psi cj}\right) \\
& + \left(\hat{\tilde{\Theta}}_{t;cj}^{d} - \hat{\Theta}_{t;cj}\right)^{T} V_{\psi cj}\left(\hat{\tilde{\Theta}}_{t;cj}^{d} - \hat{\Theta}_{t;cj}\right)\left(\mathrm{H}(\hat{\Theta}_t^d)\frac{\tilde{\nu}_{cj}^{d}}{\tilde{\Lambda}_{t;cj}^{d}} + \frac{\partial \mathrm{H}(\hat{\Theta}_t^d)}{\partial r_{cj}}\frac{2}{\tilde{\nu}_{cj}^{d}+2}\right) \\
& + \frac{\partial \mathrm{H}(\hat{\Theta}_t^d)}{\partial \Theta_{cj}}\left(\tilde{V}_{\psi cj}^{d}\right)^{-1}\left(V_{\psi cj} + V_{\psi cj}^{T}\right)\left(\hat{\tilde{\Theta}}_{t;cj}^{d} - \hat{\Theta}_{t;cj}\right)\Bigg],
\end{aligned}
$$

which proves (6.29).

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[28] I. Mező. *Some infinite sums arising from the Weierstrass Product Theorem*. Applied Mathematics and Computation. 219: 9838–9846, 2013.

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

[30] M. Kárný, M.Ruman *Mixture-Ratio Modeling of Dynamic Environments*. Submitted to ECML-PKDD-2018.

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

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

[33] Kárný M., R. Kulhavý (1988). *Structure determination of regression-type models for adaptive prediction and control*. In: Spall J. C. (ed.): Bayesian Analysis of Time Series and Dynamic Models. Marcel Dekker, New York.

[34] Colin A. Carter (1999). *Commodity futures markets: a survey*. In Australian Journal of Agricultural and Resource Economics, vol. 43, pages 209-247.