Miroslav Kárný (Ed.)

Optimized Bayesian Dynamic Advising

Theory and Algorithms

CD-ROM

Springer

Advanced Information and Knowledge Processing

Series Editors
Professor Lakhmi Jain
Xindong Wu

Also in this series

Gregoris Mentzas, Dimitris Apostolou, Andreas Abecker and Ron Young
Knowledge Asset Management
1-85233-583-1

Michalis Vazirgiannis, Maria Halkidi and Dimitrios Gunopulos
Uncertainty Handling and Quality Assessment in Data Mining
1-85233-655-2

Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho
Ontological Engineering
1-85233-551-3

Arno Scharl (Ed.)
Environmental Online Communication
1-85233-783-4

Shichao Zhang, Chengqi Zhang and Xindong Wu
Knowledge Discovery in Multiple Databases
1-85233-703-6

Jason T.L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen and Dennis Shasha (Eds)
Data Mining in Bioinformatics
1-85233-671-4

C.C. Ko, Ben M. Chen and Jianping Chen
Creating Web-based Laboratories
1-85233-837-7

Manuel Graña, Richard Duro, Alicia d'Anjou and Paul P. Wang (Eds)
Information Processing with Evolutionary Algorithms
1-85233-886-0

Colin Fyfe
Hebbian Learning and Negative Feedback Networks
1-85233-883-0

Yun-Heh Chen-Burger and Dave Robertson
Automating Business Modelling
1-85233-835-0

Dirk Husmeier, Richard Dybowski and Stephen Roberts (Eds)
Probabilistic Modeling in Bioinformatics and Medical Informatics
1-85233-778-8

K.C. Tan, E.F. Khor and T.H. Lee
Multiobjective Evolutionary Algorithms and Applications
1-85233-836-9

Ajith Abraham, Lakhmi Jain and Robert Goldberg (Eds)
Evolutionary Multiobjective Optimization
1-85233-787-7

Miroslav Kárný (Ed.)
with Josef Böhm, Tatiana V. Guy, Ladislav Jirsa, Ivan Nagy, Petr Nedoma, Ludvík Tesař

Optimized Bayesian Dynamic Advising

Theory and Algorithms

Springer

Miroslav Kárný, Ing DrSc
Department of Adaptive Systems, Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, Prague, Czech Republic

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2005923319

AI&KP ISSN 1610-3947

ISBN-10: 1-85233-928-4
ISBN-13: 978-1-85233-928-9
Springer Science+Business Media
springeronline.com

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Electronic text files prepared by editor
Printed in the United States of America (MVY)
34-543210 Printed on acid-free paper

Miroslav Kárný, Josef Böhm, Tatiana V. Guy,
Ladislav Jirsa, Ivan Nagy, Petr Nedoma,
Ludvík Tesař

Optimized Bayesian Dynamic Advising

SPIN Springer's internal project number, if known

Theory and Algorithms

June 22, 2005

Springer
Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

This book compiles the results of three years of focused team work. Such a time span would be too short without a firm foundation. We therefore dedicate this text to

Václav Peterka, Alena Halousková and Rudolf Kulhavý

prominent representatives of the Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic who helped both professionally and personally with the basis of the presented solution.

Preface

This work summarizes the theoretical and algorithmic basis of *optimized probabilistic advising*. It developed from a series of targeted research projects supported both by the European Commission and Czech grant bodies.

The source text has served as a common basis of communication for the research team. When accumulating and refining the material we found that the text could also serve as

- a grand example of the strength of dynamic Bayesian decision making,
- a practical demonstration that computational aspects do matter,
- a reference to ready particular solutions in learning and optimization of decision-making strategies,
- a source of open and challenging problems for postgraduate students, young as well as experienced researchers,
- a departure point for a further systematic development of advanced optimized advisory systems, for instance, in multiple participant setting.

These observations have inspired us to prepare this book.

Prague, Czech Republic
October 2004

Miroslav Kárný
Josef Böhm
Tatiana V. Guy
Ladislav Jirsa
Ivan Nagy
Petr Nedoma
Ludvík Tesař

Contents

1

Introduction

This work summarizes the theoretical and algorithmic basis of *optimized probabilistic advising*. The proposed tool set will help the user to preserve and permanently improve the best practice in maintenance of complex systems.

The presented advisory system is intended to support operators of complex technological processes. The developed probabilistic mining of information hidden within the acquired data and the subsequent optimization of dynamic decision-making strategy are, however, applicable elsewhere, for instance, in medicine, traffic control, economy, society, etc.

This introductory chapter

- characterizes the motivating domain of operator control, Section 1.1,
- relates the proposed advisory system to the state of the art, Section 1.2,
- puts the advisory system into a decision support context, Section 1.3,
- classifies the target readers, usage of the book and its layout, Section 1.4.

1.1 Motivation

The following outline of the original target application clarifies the purpose and applicability domains of the advisory system.

Automation of a production line improves the achieved quality of production and decreases its costs. It is applied to increasingly complex production tasks while exploiting the ever-increasing power of contemporary computers. The complexity of real-life problems, however, makes optimal global solutions of these tasks unfeasible. This induces a hierarchically organized automation. The lowest hierarchical levels deal typically with low-dimensional, but highly dynamical phenomena. At these levels, the solutions are often used repeatedly. Thus, it pays to spend substantial energy on a careful modelling and control (dynamic decision-making) design. This area is nowadays well matured; see, e.g., [1, 2, 3, 4].

At the upper levels of the hierarchy,

- addressed design problems become multivariate,
- (almost) static relationships dominate,
- a detailed modelling is too expensive to be performed.

The automation of the upper hierarchic levels is far from being stabilized and substantial research and development effort is focused on it.

The book content falls into this research stream dealing with problems related to the optimized support of the upper level, where *management* (both *supervision* and *maintenance*) are carried out by an *operator*. The importance of such support follows from the strong influence of the operator's performance on the resulting product. In spite of the common awareness of this influence, the operators are still often insufficiently supported when facing complex situations that require their actions. The performance is predominantly determined then by the operator's experience, whose accumulation is costly, and his personal state, which may vary substantially. Thus, efficient computer support is badly needed. The decisive features of a good system are summarized below for reference purposes.

Desirable operator's support

The operator's support under consideration should

1. *help the operator to follow the best available practice and allow him to improve it gradually*,
2. *warn the operator against potential undesirable behavior modes of the managed system and direct his operations in order to achieve the desired state of the system*,
3. *enable the operator to cope with large-scale complex problems.*

Such support has to avoid excessive effort for development, tailoring and learning how to use it. Thus, it has to

4. *be as generic as possible while allowing simple tailoring to the specific problem at hand*,
5. *keep the knowledge and information load on the operator at a low level, i.e. provide easy-to-understand advice, ideally in a graphical form.*

Conditions for the design of the operator's support

The operator's support that guides his actions makes sense if the operator really influences the behavior of the managed system. As we have to rely predominantly on data-based information, this influence has to be sufficiently reflected in the available data records.

Each inspected multivariate data record contains a combination of sensor readings. Under similar working conditions, similar readings of the sensors can be expected. The different operating modes can be distinguished if they are reflected in observed data, if different *clusters* of data points arise when collecting them over a period of time.

We suppose that such clusters exist or, more generally, that the *closed loop* — formed by the managed system and operator — *exhibits multimodal behavior.*

1.2 State of the art

To our best knowledge *none of the existing systems meets the ambitious but highly desirable requirements* formulated above. At the same time, there are a lot of particular results and tools that have some of the required features. In fact, the amount of results related to the inspected topic is overwhelming [5]. Even yesterday's overview is necessarily incomplete and obsolete. The given reference samples are intended to justify this claim about the lack of such an advisory system. Moreover, scanning the available results confirms that consideration of multivariate and dynamic relationships is inevitable. It also indicates why the Bayesian decision-making paradigm [6] was chosen as the underlying methodological tool.

1.2.1 Operator supports

Attempts towards meeting industrial technological needs have ranged from simple expert systems based on acquiring and storing the knowledge and experiences of the operators [7] to a more systematic approach of knowledge extraction — labelled as *data mining* — from process data [8]. With the construction of expert systems often experiencing bottleneck problems [9] and with the availability of vast amounts of data, the data mining methods have started to dominate.

In the industrial sector, *statistical process control (SPC)* is widely used whereby control charts such as $\overline{X}$, Range and Cumsum charts are independently computed for each individual process quantity. Sometimes, the reliability of these charts is questioned [10]. However, their major drawback stems from the fact that the simultaneously displayed charts are mostly evaluated visually. This overloads the operators, puts a priori bounds on the feasible complexity and, to a significant extent, hides mutual relationships between the quantities.

While SPC software manufacturers admit that the multivariate SPC analysis is yet to be realized, both practical and theoretical attempts have been made towards it. Scanning of extensive databases reveals that the operator support is

- mostly connected with the chemical industry and nuclear power plants [11],

- bringing substantial improvements to production [12],
- oriented usually to highly qualified operators [13],
- prepared for applications, where the expensive tuning pays back [14],
- concerned significantly with the abilities of graphical user interface and software architecture [15], while standard algorithmic toolkits are taken from *neural networks* (*NN*) [16], fuzzy logic [12], multivariate statistics [17], filtering [18], and partially from artificial intelligence [19],
- intended to find a balance between centralized and distributed data processing [20],
- interpreted as a blend of
  - integrated information systems [21],
  - computerized operator's guides [22],
  - fault detection and isolation modules [23],
  - simple and multivariate statistical procedures [24],
  - specialized knowledge-based system [19],
  - tailored combination of grey-box [25] and black box [26] models,
  - multivariate predictors [16],
  - multivariate advanced controllers [27],
  - information filters and transformers allowing intelligent dialogue [28].

1.2.2 Mainstream multivariate techniques

The need to exploit mutual relationships of the inspected quantities leads us to the rich area of multivariate data analysis. This has a long tradition [29], a large number of results and strong persistent research. The latter statements are illustrated by an incomplete list of common techniques relevant to the addressed technical aim.

Principal component analysis

Principal component analysis (*PCA*) [17] dominates among multivariate statistical techniques used in the given context. Essentially, eigenvalues and eigenvectors of the covariance matrix made from data records are analyzed. PCA serves both for dimensionality reduction and recognition of nonstandard situations (faults). Novel variants try, for instance,

- to make data processing robust by eliminating outliers that are known to significantly spoil its results [30],
- to cope with dynamic character of data [31] by including lagged data into the data records whose covariance matrix is analyzed; moving-window PCA [32] is another approach to the same problem,
- to face multiscale processes requiring adaptive models [33],
- to combine NN with PCA to grasp nonlinear relationships among data [34],
- to support a decentralized treatment of local problems [35].

Bayesian versions of PCA are mostly based on functional approximation of parametric distributions [36] via minimization of the Kullback–Leibler divergence [37]. It provides a guaranteed lower bound on the predictive probability. For the treated topic, it is important that the idea also be applied in the context of probabilistic mixtures [38].

Neural networks

Modelling of multivariate and dynamic relationships is a specific case of modelling of multivariate, generally nonlinear, mappings. Neural networks serve as universal approximations of such mappings [39]. As such, they provide nonlinear black-box dynamic models used in various decision-supporting modules, for instance, as standards in fault detection or as predictors [16]. They are extensively used so that their advantages and limitations can be studied on real cases [40]. Conclusions are cautiously optimistic, reflecting the well-known, but rarely reported instability of results with NN. Specific tasks related to NN are permanently studied, for instance,

- coding of control charts for pattern recognition [22],
- facing reliability and safety issues under a time stress using blended, NN-centered, techniques [41].

Clustering, mixtures and local models

The assumed existence of multiple modes of behavior leads naturally to modelling and estimation tools that can cope with them. The NASA stimulated tool AutoClass [42] is still a prominent example of the implemented algorithm searching for multiple modes. The concern with initialization [43] together with structure estimation [44, 45] is the main research topic addressed at present. The achieved results are interesting but their use in high dimensions is still very limited.

The descriptive power of multivariate models increases substantially if at least short-term history of data record is taken into account. To cope with dynamics, nonlinear stochastic state-space models and related filtering [46] are applied to specialized supporting systems. It seems that the general “technology” labelled as hidden Markov chains [47] is the most promising direction in obtaining feasible estimates of models reflecting multimodal dynamics. It is not matured enough yet, but it is expected to bring widely applicable results after a sufficiently long development period.

Often, both static and dynamic clusters can be assumed to emerge from a mixture of normal (Gaussian) multivariate distributions. This observation has led to the study of normal mixtures [48, 49, 50] and probability density estimation [51, 52], to name but a few. The established methods of clustering differ widely across the diverse applications that utilize it, both in interpretation terms and algorithm design [53, 54, 55, 56, 57, 58].

Use of local models is another emerging area. It searches in extensive databases of historical records and fits local statistical models to past data similar to the current ones [59, 60].

1.2.3 Probabilistic dynamic optimized decision-making

The discussed multivariate techniques bring a substantial insight into mutual, possibly dynamic, relationships among inspected quantities. The extraction of knowledge from data definitely helps the operator but it remains an open-ended story without systematic guidance in the subsequent decision-making. Thus, the data mining is to be complemented by decision-making theory that can cope with uncertainty and incomplete knowledge inherent to the addressed problem. It singles out the framework of Bayesian decision-making [6], [61], [62] (see also Chapter 2) as the only available theory that has a potential to make a systematic step towards the optimized advising that meets the formulated requirements about it. Specifically (the numbering corresponds with that used for the particular requirement in Section 1.1):

1. The best available practice can be fixed by basing the optimized advising on the model learned from historical data. Improvements are reachable by repetitive batch, or even online, correction of this model followed by redesign of the advising strategy.
2. Warning against undesirable modes and guiding to desired states can be reached via dynamic optimization of optional elements in the closed loop formed by the operator and the managed system.
3. The scale of the feasible problems seems to be of a practical interest, in spite of the fact that this aspect is still the bottleneck of the approach. It is, however, partially removed by results of this book.
4. The generic nature is guaranteed when the probabilistic description is systematically used both in the learning and design part of the advisory system [63, 64]. The probabilistic description is rather universal and simple to tailor when the development of the relevant knowledge-transforming tools, like [65], is not neglected.
5. The knowledge and information load can be controlled by offering the operator suitable projections of high-dimensional models of the system and decision strategies. The basic projection operations with probabilities — conditioning and marginalization — open a straightforward way of presentation in a low-dimensional, graphically treatable way. The selection of the shown projections can be optimized, too.

1.3 Developed advising and its role in computer support

The need for an advisory system with the named features, the state of the art and our know-how resulted in the following ambitious aim.

Design a generic optimized dynamic advising based on black-box modelling and Bayesian decision-making theory. Simple tailoring to a specific instance of a managed system and a permanent adaptation possibility are to be supported. The offered advices have to be understood by operators who have no background on the advising principles.

Such a still incomplete but functioning system is described in the book. It covers both theory and algorithm of *probabilistic clustering* and subsequent *design of optimal strategies* within a unified framework of data mining and operator's guidance.

The adopted probabilistic interpretation of clusters and the probabilistic design provide a rich set of formal and algorithmic tools that help us to keep the overall solution internally consistent. The solution tries to capture and influence *dynamic properties* of a good operator. This makes it unique but rather difficult and specific. Thus, it is reasonable to exploit the significant overlap with achievements of our predecessors.

The adopted approach relies on *black-box modelling.* This focus on building universal data-based models is driven by the problem addressed: the modelled processes are so complex that grey-box or white-box modelling would be too expensive. Whenever possible, the grey-box approaches should, however, be used to complement the advocated approach. They can strengthen it, for instance, by providing prior "physical" information about parts of the overall model.

Black-box modelling and feedback control as well as advising rely on the availability of informative data. In this respect, we have to use information systems provided by leading firms specializing in them. Systems of this type usually care not only about sensor quality but also about the testing and checking of data integrity. They employ various signal processing and visualization techniques. There is a wide range of commercial products of this type. Their generality varies but some products integrate the information systems up to the plant level.

Many firms and products also support advanced multivariate analysis of data. Some of them even rely on the Bayesian paradigm adopted here. But to the best of our knowledge none of them covers unified data mining and dynamic optimized advising. Often, they provide tools inherently oriented to low-dimensional problems.

1.4 Presentation style, readership and layout

This text has arisen from the communication amongst the research team trying to create a unified, internally consistent, engineering product. The style of the presentation has been driven by the communication purpose of the text. It is structured so that developers of particular algorithms and their implementations can find all necessary information in a small subpart of this extensive text.

Moreover, real implementation requires solution of all minor steps including selection of defaults on various optional parameters. It makes us care about them too in spite of the fact they make the depth of the presentation a bit unbalanced. This concern with detail helps us also to decrease the number of manually tuned knobs. This aspect is often overlooked in academic solutions of complex problems and inhibits a practical use of otherwise sophisticated solutions. Simply put, it can hardly be expected that users will be willing and able to tune tens or more such knobs.

Ideally, the text should play its original role by serving as a *communication tool to a virtual research team of readers*. Experts on various aspects, including adjacent research domains, should be able to find relatively self-contained parts pertaining to their interest in order to complement and improve them.

The purpose, the text is expected to serve, dictates presentation style, which introduces a fine structure via "agreement/proposition/proof/remark" framework. This lets the reader focus on specific topics of interest and use the rest as a sort of reference.

The solution presented has many aspects that remain to be solved, inspected and improved. The most important ones are explicitly formulated throughout the text as open problems. Consequently, they will not be forgotten, and, moreover, they may serve as the *basis of typically postgraduate research projects.*

Obviously, this is <u>not a textbook</u>. It may, however, serve as an *advanced text on Bayesian dynamic decision-making and its algorithmic realization.* Chapter 5 provides an example of how to formalize a real-life nontrivial decision-making problem. General learning techniques are "illustrated" by non-trivial specific cases. The design methodology is illustrated similarly. Samples of real-life applications complete the overall picture.

Methodologically, the text is an example of a *successful way of solving an important complex engineering problem.* The solution

- is developed in a top-down way starting from abstract ideas and formulations — as is usual in advanced mathematics — and is elaborated on in detailed algorithms — as is usual in engineering;
- relies on the availability of a team of experts who are able to communicate with their "neighbors" in a common language.

This style goes *against that of traditional education* (i.e., bottom-up approach with individualistic attention on problem grasping and solving). Consequently, it can be painful reading but it reaches much farther in much shorter time than usual, and hopefully brings a new quality to the solution found.

We believe that patient readers may benefit much from the material organized in this way.

The *attached disk with examples illustrating the most important and the most complex aspects* of the material presented in the book tries to rectify at least partially the neglected educational features of the basic text.

Knowledge of standard university courses in matrix algebra, analysis and elementary probability theory should be sufficient for reading the text.

Knowledge of basic statistics and control theory simplifies the reading, but the text does not rely on it and tries to be as self-contained as possible.

The underlying Bayesian decision-making under uncertainty is discussed in **Chapter 2**. The chapter motivates our claim about the unique suitability of the adopted tool. Furthermore, the vast majority of subproblems solved is then formulated and solved within the simple structure of this methodology. Basically, the chapter serves as reference for the basic notions, operations and results exploited in the text. It can be just briefly scanned and consulted when necessary.

Chapters 3 and **4** summarize the established results related to algorithmic implementation of the Bayesian paradigm. Chapter 3 focuses on learning, and, Chapter 4 on the design of decision-making strategies. This separation has been found useful in the presentation and is adopted in subsequent chapters, too. Chapters 3, and 4 can be skipped during the first reading and consulted only when need be.

Chapter 5 provides a specific problem formulation, with extensions necessary for applying the Bayesian decision-making to the construction of the advisory system. Methodologically, this chapter is central to the book.

General techniques describing algorithmic solution of the learning part of the advisory system form the content of **Chapter 6. Chapters 8** and **10** specialize these techniques for normal (Gaussian) and Markov-chain mixtures, respectively. These chapters form the basis of the practically implemented advisory system [66]. They serve also as examples of general techniques described in Chapter 6.

The general solution of the design part of the advisory system is presented in **Chapter 7**. Its normal and Markov-chain counterparts are in **Chapters 9** and **11**. They form the basis of practically implemented advisory system [66]. At the same time, they serve as examples of general techniques proposed in Chapter 7.

Chapters 5–11 form the core of the book. Chapters dealing with specific aspects of the addressed problem complement this core.

Chapter 12 deepens the solution of the vital problem of learning initialization. At the same time, it improves substantially the so-called *mean tracking* (MT) algorithm [67] that was at the birth of the reported research.

Chapter 13 complements information on the treatment of mixed mixtures that blend models and data of different types.

Chapter 14 refers on applications we have been involved in. It

- illustrates the developed tool set on practical cases,
- confirms that the created tool is indeed of a generic nature.

Chapter 15 concludes by summarizing the status of the research achieved and names some problems to be addressed in future.

1.5 Acknowledgements

Any complex research relies on the support of various institutions and persons who are directly or indirectly involved. We want to confirm that without such support the present research would be impossible or, at least, its outcomes would be much weaker.

Institutions

European Commission sponsored the project ProDaCTool IST-99-12058.

GA ČR sponsored the projects RECiAS 102/99/1564 and further development of the theory within the project GA ČR 102/03/0049.

AV ČR supported applications of the advisory system by the grants AV ČR S1075351, 1ET100750401.

ÚTIA AV ČR, University of Reading, Trinity College Dublin, provided institutional support to the academic ProDaCTool partners and ÚTIA AV ČR supports the further development.

Compureg Plzeň s.r.o., Kovohutě Rokycany a.s., Department of Nuclear Medicine, FNM, and Eltodo, a.s. have participated as application partners. They invested their skills, energy, and served us as strong practical-feedback providers.

People

EC experts: Pekka Karp, Francois Arlabosse, Kauko Leiviskä, Colin Piddington, Åsbjorn Rolstadas gave us strong and useful feedback during the solution of ProDaCTool project.

ProDaCTool team members: (except authors) who contributed informally to this text are

Dean Parry, Matthew Maisey, Rak Pattel, Kevin Warwick (University of Reading)

Anthony Quinn, Václav Šmídl (Trinity College Dublin)

Josef Andrýsek, Jan Kracík, Bohumil Kovář, Markéta Valečková, Milan Tichý (ÚTIA AV ČR)

Pavel Ettler, František Jirkovský, Jaroslav Křen, Jiří Štika and Ivan Puchr (Compureg Plzeň, s.r.o.)

Václav Nejdl, František Šulc (Kovohutě Rokycany, a.s.)

Jindřiška Heřmanská, Nenad Medić, Petr Vlček, Jaroslav Zimák (Department of Nuclear Medicine, University Hospital Motol)

ProDaCTool project co-founders: Elmawatti L. Sutanto and Jiří Kadlec who started the research with us.

Experts: that helped us with specific problems are Jindřich Bůcha, Jiří Grim, Dani Juričić, Jozef Nagy, Andrej Rakar, Evgenia Suzdaleva, Jan Šindelář, Igor Vajda.

Technical support: was provided by Kate Darby, Jiří Fidler, Jaroslava Hradcová, Roman Kytka, Matthew Taylor.

2

Underlying theory

Decision-making theory should help the decision maker — typically, human being — to select one of the available options (actions $\equiv$ decisions). These options concern a description of a system (a part of the world) and (or) an influence on it.

This chapter summarizes design principles and tools exploited later in our discussion of a particular decision-making task, i.e., advising to operators of complex systems.

A similarly formulated design of controllers, which is a specific decision-making problem, is in [68]. For a detailed explanation of Bayesian learning see [69], and our overall view of the complete Bayesian decision-making can be found in [70].

The chapter starts with conventions and notions used throughout — Sections 2.1 and 2.2. The framework considered covers a broad range of problems. Inevitably, the adopted symbols and notions can have specific meanings in specific application fields. The reader is asked to be patient especially in this respect.

The adopted principle of optimal decision making under uncertainty, inspected in Section 2.3, implies that incomplete knowledge and randomness have the same operational consequences for decision-making. They should be treated in the same way, labelled as *Bayesian decision making.* In the same section, the design of optimal decision rules is presented.

In Section 2.4 the design of the optimal strategies is derived. The design works with models that are obtained through Bayesian learning described in Section 2.5.

2.1 General conventions

The conventions listed here are mostly followed in this work. If some exception is necessary it is introduced at the place of its validity. If some verbal notions are introduced within Propositions, Remarks, etc., then they are *emphasized.*

Moreover, they appear in the Index. Sometimes, important parts of sentences are stressed by underlining them.

f *is the letter reserved for probability (density) functions (p(d)f).*
The meaning of the p(d)f is given through the name of its argument.

x^* *denotes the range of* x, $x \in x^*$.

$\mathring{x}$ *denotes the number of members in the countable set* x^* or the *number of entries in the vector* x.

$\equiv$ *means the equality by definition.*

x_t *is a quantity* x *at the discrete time instant labelled by* $t \in t^* \equiv \{1, \ldots, \mathring{t}\}$.

$\mathring{t} \leq \infty$ *is called (decision, learning, prediction, control, advising) horizon.*

$x_{i;t}$ *is an* i*th entry of the array* x *at time* t.
The semicolon in the subscript stresses that the symbol following it is a time index.

$x(k \cdots l)$ *denotes the sequence* $x_k, \ldots, x_l$, *i.e.*, $x(k \cdots l) \equiv x_k, \ldots, x_l$ *for* $k \leq l$.

$x(k \cdots l)$ *is an empty sequence and reflects just the prior information if* $l < k$

$x(t) \equiv x(1 \cdots t) \equiv x_1, \ldots, x_t$ is the sequence from the initial time moment till time instance t.

$x_{k \cdots lc} \equiv x_{kc \cdots lc}$ *denotes the sequence* $x_{kc}, \ldots, x_{lc}$.

$\mathrm{supp}\,[\,f(x)]$ *is the support of the pdf* $f : x^* \to [0, \infty]$, *i.e., the subset of* x^* *on which* $f(x) > 0$.

$\setminus$ is the set subtraction or an omission of a term from a sequence.

Agreement 2.1 (Interface of decision making and reality)
Physical connections of the decision-making elements to the real world — sensors, transmission lines, actuators, etc. — are taken here as a part of the physical system dealt with. Consequently, all considered quantities and mappings are mathematical entities living in an abstract calculating machine.

2.2 Basic notions and notations

The notions introduced here specify elements occurring in dynamic decision-making; see Fig. 2.1.

A brief characterization of the introduced notion is complemented by explanatory comments.

Quantity *is a multivariate mapping.*
The domain and form of the quantity are mostly unused and unspecified. The introduced notion corresponds with *random variable* used in probability theory. The use of the alternative term should stress that probability serves us as the tool adopted for decision-making under uncertainty. The term “quantity” stresses our orientation on numerical values that arise mostly by observing physical quantities. However, quantities with a discrete range that do not need numerical meaning are also considered.

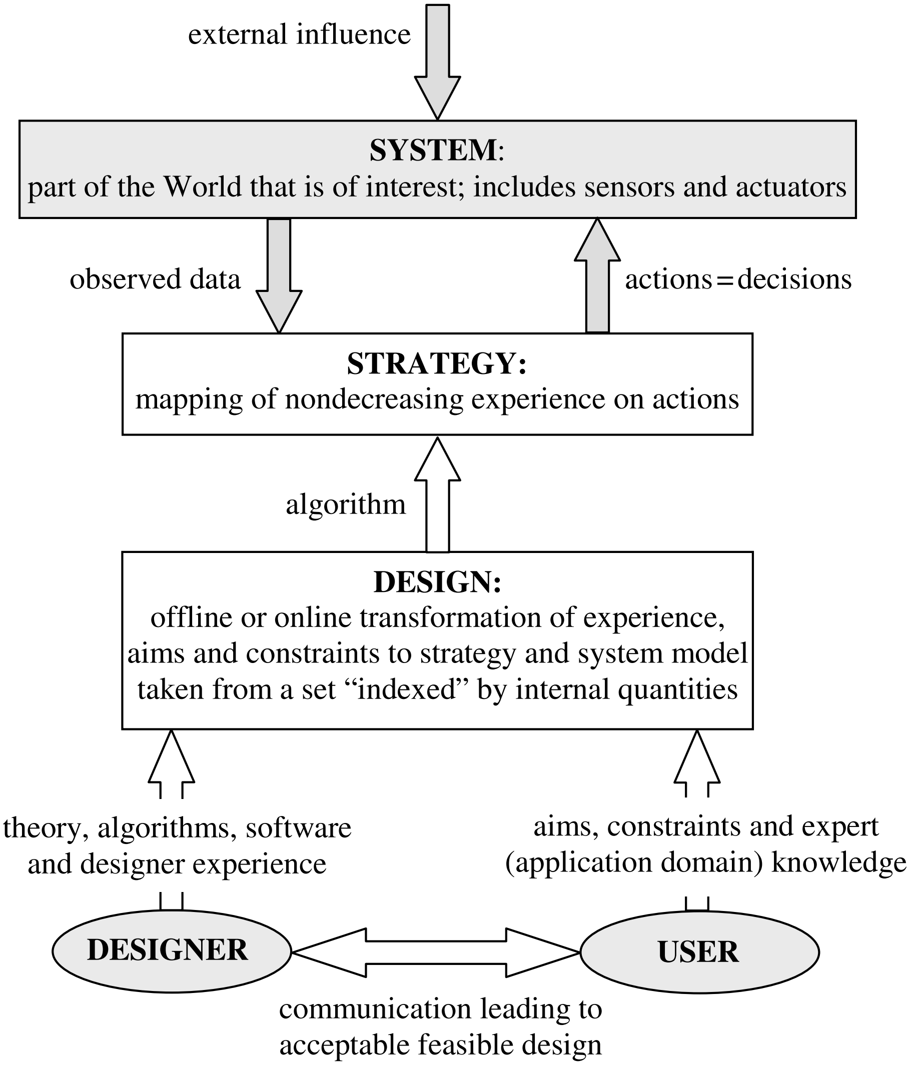

external influence
SYSTEM:
part of the World that is of interest; includes sensors and actuators
observed data
actions = decisions
STRATEGY:
mapping of nondecreasing experience on actions
algorithm
DESIGN:
offline or online transformation of experience,
aims and constraints to strategy and system model
taken from a set "indexed" by internal quantities
theory, algorithms, software
and designer experience
aims, constraints and expert
(application domain) knowledge
DESIGNER
USER
communication leading to
acceptable feasible design

Fig. 2.1. Basic decision-making scenario and notions used.

Realization *is a value of the quantity for its fixed argument.*

Often, the quantity and its realization are not distinguished. The proper meaning is implied by the context.

Decision maker *is a person or mechanism who has to select among several options called decisions or actions.*

A group of persons or mechanisms may form a single decision maker.

System *is the part of the world that is of interest for a decision maker who should either describe or influence it.*

The system is specified with respect to the aim that the decision maker wants to reach and with respect to the tools the decision maker has available. In other words, the penetrable boundaries of the system are implied by the decision task.

Decision ≡ Action $a \in a^*$ *is the value of a quantity that can be directly chosen by the decision maker for reaching decision-maker's aims.*

The terms "decision" and "action" are here used as synonyms.

A decision task arises iff there are several options available, i.e., iff $\mathring{a} > 1$.

(Decision) experience $\mathcal{P}_{a^*} \in \mathcal{P}^*_{a^*}$ *is knowledge about the system available to the decision maker for the selecting the decision* $a \in a^*$.

For example, if just data values D are available for constructing the estimate $\hat{\Theta}$ of an unknown quantity $\Theta \in \Theta^*$ then the experience is $\mathcal{P}_{\hat{\Theta}^*} \equiv D$. Often, experience includes the past data observed.

(Decision) ignorance $\mathcal{F}_{a^*} \in \mathcal{F}^*_{a^*}$ *is knowledge about the system unavailable to the decision maker for the choice of the decision* $a \in a^*$.

An estimated quantity Θ belongs to the ignorance $\mathcal{F}_{\hat{\Theta}^*}$ of the estimate $\hat{\Theta}$. Often, ignorance contains future, still unobserved data.

(System) behavior $\mathcal{Q}^*$ *consists of all possible realizations (of trajectories)* $\mathcal{Q}$, *i.e., values of all quantities considered by the decision maker within the time span determined by the horizon of interest and related to the system.*

Any decision $a \in a^*$ splits each realization $\mathcal{Q}$ into the corresponding experience $\mathcal{P}_{a^*}$ and ignorance $\mathcal{F}_{a^*}$. Formally, $\mathcal{Q} = (\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*})$. A single realization $\mathcal{Q}$ splits differently with respect to decisions $a \in a^*$, $\tilde{a} \in \tilde{a}^*$ with different experience $\mathcal{P}_{a^*} \neq \mathcal{P}_{\tilde{a}^*}$ and, consequently, different ignorance $\mathcal{F}_{a^*} \neq \mathcal{F}_{\tilde{a}^*}$. $\mathcal{Q} = (\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*}) = (\mathcal{P}_{\tilde{a}^*}, \tilde{a}, \mathcal{F}_{\tilde{a}^*})$.

(System) input $u \in u^*$ *is a decision, which is supposed to* $\underline{\textit{influence}}$ *the ignorance part* $\mathcal{F}_{u^*}$ *of the (system) behavior.*

For instance, a manipulated valve position influencing a fluid flow is the (system) input. On the other hand, an estimate $\hat{\Theta}$ of an unknown (realization of) quantity Θ is an instance of the decision that is not the input. The estimate describes the system but has no direct influence on it.

(System) output $y \in y^*$ *is an observable quantity that provides the decision maker information about the (system) behavior.*

To be or not to be output or input is a relative property. The input is always directly manipulated. For instance, a pressure measured in a heated system is an instance of the output. A pressure applied to the system is an instance of the input.

Innovation $\Delta_t \in \Delta^*_t$ *contains quantities included in the ignorance* $\mathcal{F}_{a^*_t}$ $\underline{\textit{and}}$ *in* $\mathcal{P}_{a^*_{t+1}} \setminus a_t$.

Often, $\Delta_t = y_t =$ the system output at time t.

Decision rule $\mathcal{R} : \mathcal{Q}^* \rightarrow a^*$ *is a mapping that assigns a decision* $a \in a^*$ *to the behavior* $\mathcal{Q} \in \mathcal{Q}^*$.

Causal decision rule $\mathcal{R} : \mathcal{P}^*_{a^*} \rightarrow a^*$ *is a mapping that assigns the decision* $a \in a^*$ *to its experience* $\mathcal{P}_{a^*} \in \mathcal{P}^*_{a^*}$.

In other words, the decision a made by a causal decision rule is uninfluenced by the related ignorance $\mathcal{F}_{a^*}$. We deal with the causal decision rules so that the term "causal" is mostly dropped. Estimator is an instance of the causal decision rule $\mathcal{R} : \mathcal{P}^*_{\hat{\Theta}^*} \to \hat{\Theta}^*$ that assigns an estimate $\hat{\Theta}$ of the unknown quantity $\Theta \in \Theta^*$ to the available experience $\mathcal{P}_{\hat{\Theta}^*}$.

Strategy *is a sequence of decision rules* $\{\mathcal{R}_t : \mathcal{Q}^* \to a_t^*\}_{t \in t^*}$.

Causal strategy $\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}_{t \in t^*}$ *is a sequence made of causal decision rules.*

Again, we deal with causal strategies so that the term "causal" is mostly dropped.

Controller *is a causal strategy assigning inputs* u_t *to experience* $\mathcal{P}_{u_t^*}$, $\forall t \in t^*$.

For instance, the controller given by the proportionality constant C is an example of the causal control strategy $\{y_{t-1}^* \to u_t^* : u_t = -Cy_{t-1}\}_{t \in t^*}$ if $y_{t-1}^* \subset \mathcal{P}_{u_t^*}$. The same controller is not causal if, for instance, $\mathcal{P}_{u_t^*} = \emptyset$.

Design *selects the decision rule or strategy.*

The design selecting a single rule is called *static design.* The choice of the strategy is called *dynamic design.* The person (group) who makes the selection is the *designer.* Authors and readers of this text are supposed to be designers. In that sense, the term *we* used within the text should mostly be read: *we designers.* The designers work for the *users* whose aims should be reached by using the strategy designed.

Uncertain behavior *(related to static design) is a behavior whose realizations* $\mathcal{Q}$ *can be decomposed into*

- $\mathcal{Q}_{\mathcal{R}} \equiv$ *the part that is unambiguously determined by the considered decision rule* $\mathcal{R} \in \mathcal{R}^*$,
- uncertainty Υ *that is defined as the part of the behavior that belongs to the ignorance* $\mathcal{F}_{\mathcal{R}^*(\mathcal{P})}$ *of decisions* $\mathcal{R}(\mathcal{P})$ *generated by the admissible rules* $\mathcal{R} \in \mathcal{R}^*$ *and uninfluenced by them, even indirectly.*

With an abuse of notation, we write the corresponding decomposition of the realization $\mathcal{Q} = (\mathcal{Q}_{\mathcal{R}}, \Upsilon)$. By definition, incomplete knowledge of (the realization of) a considered quantity $\Theta \in \Theta^*$ makes the behavior uncertain. Also, external unobserved noise influencing the system makes its behavior uncertain.

Uncertainty expresses both incomplete knowledge and randomness. Uncertain behavior related to dynamic design is encountered if any of its rules faces uncertainty.

Decision-making *means design and application of a decision rule (strategy).*

Admissible strategy *is a strategy* $\{\mathcal{R}_t\}_{t \in t^*}$ *that*

- is *causal*, i.e., $\{\mathcal{R}_t\}_{t \in t^*} \equiv \{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}_{t \in t^*}$ *and*
- meets *physical constraints*, i.e., *the ranges of its decision rules are in prespecified subsets of respective sets of decisions.*

Loss function, $\mathcal{Z} : \mathcal{Q}^* \to [0, \infty]$, *quantifies the degree of achievement of the design aim.*

The loss function measures the quality of the realizations $\mathcal{Q}$. The smaller the value of $\mathcal{Z}(\mathcal{Q})$ is, the better. The loss function orders indirectly, but only partially, admissible strategies influencing the behavior. Those leading to the smaller loss are taken as better ones. The important case of multivalued loss function [71] is beyond the scope of this text, but the approach discussed in Section 2.3 could be relatively simply extended to the case by embedding the index of entries of the loss values into uncertainty.

"Expected" loss $\tilde{\mathcal{E}}(\mathcal{Z}) \equiv \tilde{\mathcal{E}}_{\mathcal{R}}(\mathcal{Z})$ *assigns to the considered loss function* $\mathcal{Z}$ *and strategy* $\mathcal{R}$ *a value in* $[0, \infty]$. *The value is to be independent of the realization of the involved uncertainty.*

The quotation marks and the sign $\tilde{}$ are used temporarily. They serve us in the discussion, which shows that, under widely acceptable conditions, we have to deal with expectation in a mathematical sense. Then they are not used any more.

Optimal design *selects an admissible strategy that leads to the smallest value of the "expected" loss function.*

Practically admissible strategy *is an admissible strategy that respects constraints limiting the complexity of the decision-making.*

The complexity is considered with respect to the computational resources available at the design and application stages. The majority of discussed problems in which the complexity constraints play a role are computationally hard in terms of computer sciences. An intuitive understanding of the computational complexity is sufficient to our purposes.

Practically optimal design *selects a practically admissible strategy giving the smallest values of the "expected" loss.*

The presented optimal design provides optimal admissible strategies and can be simply adapted to provide strategies of a prespecified complexity by optimizing over a set of simple decision rules, for instance, over proportional controllers only. Operational formal tools for practically optimal design are not available. It is not known how to make the optimal design of a prespecified complexity.

We never know whether the selection of the constant determining proportional controller made with use of, say, ten algebraic operations is really the best one possible among all selections that are allowed to perform ten algebraic operations. This is the *main barrier of the applicability* of the theory describing the optimal design. The optimal design becomes a practical tool by employing sound engineering heuristics. The practical optimum is not guaranteed. This fact is stressed by using the term *suboptimal design* giving *suboptimal strategy.*

2.3 Decision making under uncertainty

Here, we describe a general way how to understand and face uncertainty that causes incomplete ordering of strategies. In order to avoid a cumbersome notation, we formulate the adopted design principle, related requirements and their consequences for the static design, i.e., the design of a single, not necessarily causal, decision rule. The obtained conclusions apply also to the dynamic design, i.e., to the choice of decision strategies.

Agreement 2.2 (Uncertainty in decision making) *Decision making under uncertainty arises if the optimal decision-making is to be performed and*

- *at least a pair of different decisions can be made,* $\mathring{a} > 1$,
- *the considered loss function* $\mathcal{Z}(\mathcal{Q}) \equiv \mathcal{Z}(\mathcal{Q}_{\mathcal{R}}, \Upsilon)$ *depends on a non-void set* Υ^* *of uncertainties.*

The loss $\mathcal{Z}_{\mathcal{R}}(\Upsilon) \equiv \mathcal{Z}(\mathcal{Q}_{\mathcal{R}}, \Upsilon)$ *is a function of uncertainty* Υ, *i.e., that part of the realization belonging to ignorance and being uninfluenced by the rule* $\mathcal{R}$. *The function* $\mathcal{Z}_{\mathcal{R}}(\cdot)$ *is assigned to each considered decision rule* $\mathcal{R} \in \mathcal{R}^* \equiv$ *set of admissible rules. The set of such functions is denoted* $\mathcal{Z}_{\mathcal{R}^*}$

$$\mathcal{Z}_{\mathcal{R}^*} \equiv \{\mathcal{Z}_{\mathcal{R}} : \Upsilon^* \to [0, \infty],\ \mathcal{Z}_{\mathcal{R}}(\Upsilon) \equiv \mathcal{Z}(\mathcal{Q}_{\mathcal{R}}, \Upsilon)\}_{\mathcal{R} \in \mathcal{R}^*}. \tag{2.1}$$

Under uncertainty, the loss function is insufficient for a complete ordering (comparing) of admissible rules in spite of the fact that its values are in the fully ordered interval: due to the uncertainty, not a single number but a function on the set (2.1) is assigned to each decision rule $\mathcal{R}$.

For instance, let two estimators give a pair of estimates $\hat{\Theta}_1 \neq \hat{\Theta}_2$ of an unknown scalar quantity $\Theta \in \Theta^* \equiv (-\infty, \infty)$. The better one cannot be unambiguously chosen using the quadratic loss function $(\Theta - \hat{\Theta})^2$: we do not know whether $\Theta \in \mathcal{F}_{\hat{\Theta}^*}$ is in that part of Θ^* where $(\Theta - \hat{\Theta}_1)^2 \leq (\Theta - \hat{\Theta}_2)^2$ or in its complement.

2.3.1 Complete ordering of decision rules

This section inspects conditions under which the compared decision rules can be completely ordered. The adopted conditions try to make the ordering as objective as possible, i.e., as little dependent as possible on the subject ordering them.

Any systematic choice of an optimal decision rule can be reduced to the following principle.

Agreement 2.3 ("Expectation"-minimization design principle)

- *A functional* $\tilde{\mathcal{E}}_{\mathcal{R}}$, *called "*expectation*", is selected by the designer. It assigns to functions in (2.1) — determined by the loss function* $\mathcal{Z}$ *and compared decision rules* $\mathcal{R} \in \mathcal{R}^*$ *— an "*expected loss*"* $\tilde{\mathcal{E}}_{\mathcal{R}}[\mathcal{Z}]$

$$\tilde{\mathcal{E}}_{\mathcal{R}} : \mathcal{Z}_{\mathcal{R}^*} \to [0, \infty]. \tag{2.2}$$

- *The minimizer of* $\tilde{\mathcal{E}}[\mathcal{Z}_{\mathcal{R}}] \equiv \tilde{\mathcal{E}}_{\mathcal{R}}[\mathcal{Z}(\mathcal{Q}_{\mathcal{R}}, \Upsilon)]$ *found in* $\mathcal{R}^*$ *is taken as the optimal decision rule.*

The outcome of this design depends on the "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$. Its choice has to at least guarantee that unequivocally bad rules are avoided. Such bad rules are identified here.

Agreement 2.4 (Dominated rules; strictly isotonic expectation) *Let a loss function $\mathcal{Z}$ measure the quality of the behavior. The decision rule $\mathcal{R}: \mathcal{Q}^* \rightarrow a^*$ is called* dominated iff (if and only if) *there is another decision rule $\tilde{\mathcal{R}}: \mathcal{Q}^* \rightarrow a^*$ such that*

$$\mathcal{Z}_{\mathcal{R}}(\Upsilon) \geq \mathcal{Z}_{\tilde{\mathcal{R}}}(\Upsilon) \ \Leftrightarrow \ \mathcal{Z}(\mathcal{Q}_{\mathcal{R}}, \Upsilon) \geq \mathcal{Z}(\mathcal{Q}_{\tilde{\mathcal{R}}}, \Upsilon), \ \forall \Upsilon \in \Upsilon^*. \tag{2.3}$$

The decision rule is called strictly dominated *iff there is a nontrivial subset of Υ^* on which the inequality (2.3) is strict.*

The "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$ (2.2) is said to be strictly isotonic if for a decision rule $\mathcal{R}$, strictly dominated by a decision rule $\tilde{\mathcal{R}}$, it holds

$$\tilde{\mathcal{E}}_{\mathcal{R}}[\mathcal{Z}_{\mathcal{R}}] > \tilde{\mathcal{E}}_{\tilde{\mathcal{R}}}[\mathcal{Z}_{\tilde{\mathcal{R}}}].$$

We take the dominated decision rules as those to be surely avoided.

Requirement 2.1 (Inadmissibility of strictly dominated rules) *The considered "expectation"-minimization design, Agreement 2.3, must not lead to a strictly dominated decision rule.*

The "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$ allows us to order the decision rules in spite of the influence of uncertainty. Thus, it characterizes uncertainty and neither the ordered set of decision rules nor the specific loss function $\mathcal{Z}$. Consequently, in the quest for objectivity of the constructed complete ordering, the "natural" requirement on the "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$ can be formulated as follows.

Requirement 2.2 (Independence of $\mathcal{R}^*$) *The design, Agreement 2.3, with the chosen "expectation" must not take a strictly dominated rule as the optimal one (i.e., must meet Requirement 2.1) even if the set of possible decision rules $\mathcal{R}^*$ is reduced to its nontrivial subset.*

A subset of $\mathcal{R}^*$ is taken as nontrivial if it contains at least two different rules while at least one of them gives a finite "expected" loss.

Proposition 2.1 (Isotonic ordering) *Assume that there is a rule in $\mathcal{R}^*$ for which the "expected" loss is finite. Then, Requirement 2.2 is fulfilled iff the "expectation" is strictly isotonic; see Agreement 2.4.*

Proof.

1. We prove by contradiction that — with strictly isotonic "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$ — the minimizer cannot be strictly dominated. Let $\tilde{\mathcal{E}}_{\mathcal{R}}[\mathcal{Z}]$ be strictly isotonic on its domain $\mathcal{Z}_{\mathcal{R}^*}$ (2.2) and $\mathcal{R}_o \in \mathcal{R}^*$ be a minimizer of the "expected" loss. The minimizer gives necessarily a finite value of the corresponding $\tilde{\mathcal{E}}_{\mathcal{R}}[\mathcal{Z}]$. Let $\mathcal{R}_d \in \mathcal{R}^*$ dominate it strictly. Then, because of the

construction of $\mathcal{R}_o$, the strict dominance and strictly isotonic nature of $\tilde{\mathcal{E}}_{\mathcal{R}}$, we get the following contradictory inequality

$$\tilde{\mathcal{E}}_{\mathcal{R}_o}[\mathcal{Z}(\mathcal{Q}_{\mathcal{R}_o},\Upsilon)] \underbrace{\leq}_{\text{minimum}} \tilde{\mathcal{E}}_{\mathcal{R}_d}[\mathcal{Z}(\mathcal{Q}_{\mathcal{R}_d},\Upsilon)] \underbrace{<}_{\text{strictly isotonic}} \tilde{\mathcal{E}}_{\mathcal{R}_o}[\mathcal{Z}(\mathcal{Q}_{\mathcal{R}_o},\Upsilon)].$$

2. We prove by contradiction that use of an "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$ that is not strictly isotonic leads to violation of Requirement 2.1 when Requirement 2.2 holds. If $\tilde{\mathcal{E}}_{\mathcal{R}}[\mathcal{Z}]$ is not strictly isotonic on its domain $\mathcal{Z}_{\mathcal{R}^*}$ (2.2) then there is a rule $\mathcal{R}_1 \in \mathcal{R}^*$ strictly dominated by the decision rule $\mathcal{R}_d \in \mathcal{R}^*$ such that
$$\tilde{\mathcal{E}}_{\mathcal{R}_d}[\mathcal{Z}(\mathcal{Q}_{\mathcal{R}_d},\Upsilon)] \geq \tilde{\mathcal{E}}_{\mathcal{R}_1}[\mathcal{Z}(\mathcal{Q}_{\mathcal{R}_1},\Upsilon)].$$
If we restrict the set of decision rules $\mathcal{R}^*$ to the pair $\{\mathcal{R}_d, \mathcal{R}_1\}$ then $\mathcal{R}_1$ can always be taken as the optimal decision rule. Thus, under Requirement 2.2, Requirement 2.1 is not met with the "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$. □

Requirement 2.2 guarantees suitability of the "expectation" to a wide range of decision rules. It remains to guarantee that the "expectation" $\tilde{\mathcal{E}}_{\mathcal{R}}$ serves as an objective tool for ordering strategies for any loss function $\mathcal{Z}$ from a rich set $\mathcal{Z}^*$.

We adopt rather technical conditions on the set $\mathcal{Z}^*$. Essentially, applicability to a very smooth functions and a restricted version of "linearity" of $\tilde{\mathcal{E}}_{\mathcal{R}}$ are required.

Requirement 2.3 (Independence of loss function) *Let us consider various loss functions $\mathcal{Z} \in \mathcal{Z}^*$. The "expectation" $\tilde{\mathcal{E}}$ acts on the union $\mathcal{Z}^*_{\mathcal{R}^*}$ of the sets of functions $\mathcal{Z}_{\mathcal{R}^*}$ (2.1) with a common uncertainty set Υ^**

$$\mathcal{Z}^*_{\mathcal{R}^*} \equiv \cup_{\mathcal{Z}\in\mathcal{Z}^*}\mathcal{Z}_{\mathcal{R}^*}. \tag{2.4}$$

The set $\mathcal{Z}^*_{\mathcal{R}^*}$ is required to contain a subset of test loss functions. *The test functions are zero out of a compact nonempty subset Ω of Υ^* and continuous on Ω (supremum norm defines the corresponding topology).*

The "expectation" is assumed to be an isotonic, sequentially continuous, and uniformly continuous functional on $\mathcal{Z}^*_{\mathcal{R}^*}$. It is, moreover, additive on loss functions with nonoverlapping supports

$$\tilde{\mathcal{E}}[\mathcal{Z}_1 + \mathcal{Z}_2] = \tilde{\mathcal{E}}[\mathcal{Z}_1] + \tilde{\mathcal{E}}[\mathcal{Z}_2] \text{ if } \mathcal{Z}_1\mathcal{Z}_2 = 0,\ \mathcal{Z}_1,\, \mathcal{Z}_2 \in \mathcal{Z}^*_{\mathcal{R}^*}.$$

Technical Requirement 2.3 allows us to get an integral representation of the "expectation" searched for. Its proof, as well as definitions of the adopted noncommon terms, can be found in Chapter 9 of the book [72]; see Theorem 5 there.

Proposition 2.2 (Integral form of "expectation") *Under Requirement 2.3, the "expectation" $\tilde{\mathcal{E}}$ has the form*

$$\tilde{\mathcal{E}}[\mathcal{Z}] = \int_{\Omega} \mathcal{U}(\mathcal{Z}(\Upsilon), \Upsilon)\, \mu(d\Upsilon), \quad \textit{where} \tag{2.5}$$

μ is a finite regular nonnegative Borel measure on Ω. The utility function *$\mathcal{U}$ satisfies $\mathcal{U}(0, \Upsilon) = 0$. It is continuous in values of $\mathcal{Z}(\cdot)$* almost everywhere *(*a.e.*) on Ω, bounded a.e. on Ω for each $\mathcal{Z}$ in the set of the test loss functions.*

Remark(s) 2.1

1. *The test loss functions are widely applicable and their consideration implies no practical restriction. The continuity requirements on $\tilde{\mathcal{E}}$ are also widely acceptable.*
2. *The linearity of $\tilde{\mathcal{E}}$ on functions with nonoverlapping support seems to be sound. Any loss function $\mathcal{Z} \in \mathcal{Z}^*_{\mathcal{R}^*}$ can be written as $\mathcal{Z} = \mathcal{Z}\chi_\omega + \mathcal{Z}(1 - \chi_\omega) \equiv \mathcal{Z}_1 + \mathcal{Z}_2$, $\mathcal{Z}_1\mathcal{Z}_2 = 0$ with χ_ω denoting an* indicator *of a set $\omega \subset \Omega \subset \Upsilon^*$. The indicator χ_ω equals 1 on ω and it is zero outside of it.*
   The loss "expected" on the set ω and its complement should sum to the loss "expected" on the whole set of arguments.
3. *The* utility function *$\mathcal{U}$ allows the designer to express his/her attitude toward the design consequences: the decision maker might be risk aware, risk prone, or risk indifferent [71].*
4. *The utility function $\mathcal{U}$ and the nonnegative measure μ are universal for the whole set of test functions. $\mathcal{U}$ and μ are (almost) "objective", i.e., suitable for a wide range of decision tasks facing the same uncertainty.*

We formulate now our final objectivity-oriented requirement. Hereafter, we use the fact that the behavior $\mathcal{Q}$ is uniquely determined by the decision rule $\mathcal{R}$ and uncertainty Υ. Thus, we can work with the nonreduced behavior $\mathcal{Q}$ without explicit separation of the uncertainty Υ.

Requirement 2.4 (Indifference of the designer)

- *The designer is risk indifferent, which means that $\mathcal{U}(\mathcal{Z}(\cdot), \cdot) = \mathcal{Z}(\cdot)$.*
- *The "expectation" preserves any constant loss $\tilde{\mathcal{E}}[constant] = constant$.*
- *The involved measure μ has Radon–Nikodým derivative $f(\mathcal{Q})$ with respect to a dominating measure denoted $d\mathcal{Q}$, [72]. In the treated cases, $d\mathcal{Q}$ is either Lebèsgue or counting measure.*

Adopting Requirement 2.4, we get the basic representation Proposition that introduces *objective expectation*.

Proposition 2.3 (Objective expectation) *Under Requirement 2.4, the "expectation" $\tilde{\mathcal{E}}$ (2.5) is formally identical with a* mathematical expectation*. The Radon–Nikodým's derivative f has all the properties of the* joint probability (density) function *(*p(d)f*) on $\mathcal{Q}^*$.*

Proof. It is sufficient to observe that the preservation of constants implies that μ is a probabilistic measure, i.e., $\mu \geq 0,\ \mu(\mathcal{Q}^*) = 1$. □

Remark(s) 2.2

1. *The (mathematical) "expectation" is singled out by dropping the sign* ˜ *as well as the quotation symbols* " "
$$\mathcal{E}[\mathcal{Z}] \equiv \int \mathcal{Z}(\mathcal{Q}) f(\mathcal{Q})\, d\mathcal{Q}. \tag{2.6}$$
2. *The first item in Requirement 2.4 has clear meaning: objective, emotionally indifferent, designers are supported here.*
3. *The last item in Requirement 2.4 is unnecessary but it helps us to deal with simpler objects, namely, with probability density functions (pdf) or probability functions (pf).*
4. *The pdf defining the objective expectation is referred to as the* objective pdf.
5. *Mostly, we use notation related to pdfs even to pfs. Only when necessary, we underline that we deal with a pf and write integrals as sums.*
6. *We have arrived to the* unconditional expectation. *Its values are independent of the uncertainty realization and are determined by our prior experience only. When dealing with dynamic design, the observed part of the realization becomes a part of experience. The uncertainty changes with changing experience. Consequently, we always deal with* conditional expectation $\mathcal{E}[\bullet|\mathit{available\ experience}]$. *Rigorous definition of the conditional expectation can be found in [72]. Here, it is treated in a naive way as the integral* $\int \bullet(\alpha) f(\alpha|\mathit{available\ experience})\, d\alpha$ *weighted by the* conditional pdf $f(\alpha|\mathit{available\ experience})$.

2.3.2 Calculus with pdfs

The joint pdf f on $\mathcal{Q} \equiv (\alpha, \beta, \gamma)$ is analyzed and synthesized using several pdfs related to it. Let us recall the meaning of pdfs derived from $f(\mathcal{Q})$.

Agreement 2.5 (Nomenclature of pdfs; Independence)
Basic pdfs dealt with are

Name	Meaning
joint pdf $f(\alpha,\beta|\gamma)$ *of* α,β conditioned on γ	*a pdf on* $(\alpha,\beta)^*$ *restricting* $f(\mathcal{Q})$ *on the cross-section of* $\mathcal{Q}^*$ *given by a fixed* γ
marginal pdf $f(\alpha|\gamma)$ *of* α *conditioned on* γ	*a pdf on* α^* *restricting* $f(\mathcal{Q})$ *on the cross-section of* $\mathcal{Q}^*$ *given by a fixed* γ *with no information on* β
marginal pdf $f(\beta|\alpha,\gamma)$ *of* β *conditioned on* α,γ	*a pdf on* β^* *restricting* $f(\mathcal{Q})$ *on the cross-section of* $\mathcal{Q}^*$ *given by a fixed* α,γ

The conditioning symbol | *is dropped if just trivial conditions are considered.*

The pdf $f(\alpha, \beta)$ *is the* lower dimensional joint pdf *of the pdf* $f(\alpha, \beta, \gamma)$ *and* $f(\beta)$ *is its* marginal pdf. *Quantities* α *and* β *are* conditionally independent *under the condition* γ *iff*

$$f(\alpha, \beta|\gamma) = f(\alpha|\gamma) f(\beta|\gamma). \tag{2.7}$$

Our manipulations with the introduced pdfs rely on the following calculus.

Proposition 2.4 (Calculus with pdfs) *For any* $(\alpha, \beta, \gamma) \in (\alpha, \beta, \gamma)^*$, *the following relationships between pdfs hold.*

Non-negativity $f(\alpha, \beta|\gamma),\ f(\alpha|\beta, \gamma),\ f(\beta|\alpha, \gamma),\ f(\beta|\gamma) \geq 0.$

Normalization $\int f(\alpha, \beta|\gamma)\, d\alpha d\beta = \int f(\alpha|\beta, \gamma)\, d\alpha = \int f(\beta|\alpha, \gamma)\, d\beta = 1.$

Chain rule $f(\alpha, \beta|\gamma) = f(\alpha|\beta, \gamma) f(\beta|\gamma) = f(\beta|\alpha, \gamma) f(\alpha|\gamma).$

Marginalization $f(\beta|\gamma) = \int f(\alpha, \beta|\gamma)\, d\alpha,\ f(\alpha|\gamma) = \int f(\alpha, \beta|\gamma)\, d\beta.$

Bayes rule $f(\beta|\alpha, \gamma) =$

$$= \frac{f(\alpha|\beta, \gamma) f(\beta|\gamma)}{f(\alpha|\gamma)} = \frac{f(\alpha|\beta, \gamma) f(\beta|\gamma)}{\int f(\alpha|\beta, \gamma) f(\beta|\gamma)\, d\beta} \propto f(\alpha|\beta, \gamma) f(\beta|\gamma). \tag{2.8}$$

The proportion sign, $\propto$, *means that the factor, independent of* β *and uniquely determined by the normalization, is not explicitly written in the equality represented.*

The conditional independence (2.7) can be expressed equivalently

$$f(\alpha, \beta|\gamma) = f(\alpha|\gamma) f(\beta|\gamma) \Leftrightarrow f(\alpha|\beta, \gamma) = f(\alpha|\gamma) \text{ or } f(\beta|\alpha, \gamma) = f(\beta|\gamma). \tag{2.9}$$

Proof. For motivation see [69], a more precise and more technical treatment exploits the measure theory [72]. Technically, an intermediate insight can be gained by considering loss functions dependent only on a part of $\mathcal{Q}$ or with some parts of $\mathcal{Q}$ “fixed by the condition”, [68]. □

Remark(s) 2.3

1. *Alternative presentations of formulas stress their symmetry.*
2. *The technically correct statements that the identities, like (2.9), are valid only almost everywhere is mostly omitted in the subsequent explanations.*
3. *The Bayes rule (2.8) is a simple consequence of previous formulas. Its importance in this text cannot be exaggerated, cf. Propositions 2.13, 2.14.*
4. *The symmetric identities (2.9) say that* β *does not influence the description of* α *(and vice versa) if* α *and* β *are conditionally independent for a given* γ.

Often, we need the pdf of a quantity β that is the image of other multivariate quantity α, i.e., $T : \alpha^* \to \beta^* \equiv T(\alpha^*)$, whose pdf is known. The desired pdf is found by simply respecting the need to preserve the expectation.

Proposition 2.5 (Pdfs of transformed quantities) *Let the expectation $\mathcal{E}_T$, acting on functions $\mathcal{B} : \beta^* \to (-\infty, \infty)$, be specified by the pdf $f_T(\beta)$, i.e.*

$$\mathcal{E}_T[\mathcal{B}] = \int \mathcal{B}(\beta) f_T(\beta)\, d\beta.$$

Then, this functional expresses the same expectation as

$$\mathcal{E}[\mathcal{B}] = \int \mathcal{B}(T(\alpha)) f(\alpha)\, d\alpha$$

iff

$$\int_{T(A)} f_T(T(\alpha))\, dT(\alpha) = \int_A f(\alpha)\, d\alpha, \tag{2.10}$$

for all measurable sets $A \subset \alpha^*$.

Let α be a real vector , $\alpha \equiv [\alpha_1, \ldots, \alpha_{\mathring{\alpha}}]$ and $T = [T_1, \ldots, T_{\mathring{\alpha}}]$ bijection (one-to-one mapping) with finite continuous partial derivatives a.e. on α^*

$$J_{ij}(\alpha) \equiv \frac{\partial T_i(\alpha)}{\partial \alpha_j},\ \ i, j = 1, \ldots, \mathring{\alpha}, \tag{2.11}$$

for all entries T_i of T and entries α_j of α. Then,

$$f_T(T(\alpha))|J(\alpha)| = f(\alpha), \quad \textit{where} \tag{2.12}$$

$|\cdot|$ *denotes absolute value of the* determinant *of the matrix in its argument.*

Proof. Proposition describes substitutions in multivariate integrals; see, for instance, [72, 73]. □

It is useful to summarize basic properties of expectation, which help us to simplify formal manipulations. Its conditional version $\mathcal{E}[\cdot|\gamma]$ is considered. For this text, it is sufficient to take it in a naive way as an integral weighted by the conditional pdf $f(\cdot|\gamma)$. The textbook [72] can be consulted for a rigorous treatment.

Proposition 2.6 (Basic properties of $\mathcal{E}$) *For arbitrary functions $\mathcal{Z}_1(\cdot)$, $\mathcal{Z}_2(\cdot)$ on which the conditional expectation $\mathcal{E}[\cdot|\gamma]$ is well defined, $\mathcal{E}[\cdot|\gamma]$ has the following properties.*

Isotonic nature of $\mathcal{E}[\cdot|\gamma]$: $\mathcal{Z}_1 \leq \mathcal{Z}_2, cf.(2.3), \Rightarrow \mathcal{E}[\mathcal{Z}_1|\gamma] \leq \mathcal{E}[\mathcal{Z}_2|\gamma]$.

Linearity of $\mathcal{E}[\cdot|\gamma]$: $\mathcal{E}[A(\gamma)\mathcal{Z}_1 + B(\gamma)\mathcal{Z}_2|\gamma] = A(\gamma)\mathcal{E}[\mathcal{Z}_1|\gamma] + B(\gamma)\mathcal{E}[\mathcal{Z}_2|\gamma]$ *for arbitrary coefficients A, B depending at most on the condition γ.*

Chain rule for expectation: $\mathcal{E}\left[\mathcal{E}[\cdot|\gamma, \zeta]|\gamma\right] = \mathcal{E}[\cdot|\gamma]$ for an arbitrary <u>*additional condition*</u> *ζ.*

Conditional covariance of a vector α $\mathrm{cov}[\alpha|\gamma] \equiv \mathcal{E}\left[(\alpha - \mathcal{E}[\alpha|\gamma])(\alpha - \mathcal{E}[\alpha|\gamma])'|\gamma\right]$ *is related to the noncentral moments through the formula*

$$\mathrm{cov}[\alpha|\gamma] = \mathcal{E}[\alpha\alpha'|\gamma] - \mathcal{E}[\alpha|\gamma]\mathcal{E}[\alpha'|\gamma], \quad ' \textit{ is } \text{transposition.} \tag{2.13}$$

Jensen inequality bounds expectation of a convex function $T_\gamma : \alpha^* \to (-\infty, \infty)$

$$\mathcal{E}[T_\gamma(\alpha)|\gamma] \geq T_\gamma\left(\mathcal{E}[\alpha|\gamma]\right). \tag{2.14}$$

Proof. All statements can be verified by using the integral expression (2.6) of the expectation. Proof of the Jensen inequality can be found, e.g., in [74]. □

Remark(s) 2.4

1. *The proposition is formulated for the conditional expectation. The unconditional case is formally obtained by omitting the condition used.*
2. *Note that whenever the expectation is applied to an array function* V *it should be understood as the array of expectations* $[\mathcal{E}(V)]_i \equiv \mathcal{E}(V_i)$.

2.3.3 Basic decision-making lemma

The optimal choice of *admissible decision rules* relies on the key proposition that reduces minimization over mappings to an "ordinary" minimization.

Proposition 2.7 (Basic decision-making lemma) *The* optimal admissible decision rule ${}^{\llcorner o}\mathcal{R}$

$$^{\llcorner o}\mathcal{R}(\mathcal{P}_{a^*}) \equiv {}^{\llcorner o}a(\mathcal{P}_{a^*}),\ \forall \mathcal{P}_{a^*} \in \mathcal{P}^*_{a^*}$$

minimizing the expected loss (2.6) can be constructed valuewise as follows.

To each $\mathcal{P}_{a^*} \in \mathcal{P}^*_{a^*}$, *a minimizing argument* ${}^{\llcorner o}a(\mathcal{P}_{a^*})$ *in*

$$\min_{a \in a^*} \mathcal{E}[\mathcal{Z}(\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*})|a, \mathcal{P}_{a^*}] \tag{2.15}$$

is assigned as the value of the optimal decision rule corresponding to the considered argument. The minimum reached is

$$\min_{\{\mathcal{R}:\, \mathcal{P}^*_{a^*} \to a^*\}} \mathcal{E}[\mathcal{Z}(\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*})] = \mathcal{E}\left\{\min_{a \in a^*} \mathcal{E}[\mathcal{Z}(\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*})|a, \mathcal{P}_{a^*}]\right\}. \tag{2.16}$$

Proof. Let us fix an arbitrary $\mathcal{P}_{a^*} \in \mathcal{P}^*_{a^*}$. The definition of minimum implies that for all $a \in a^*$

$$\mathcal{E}\left[\mathcal{Z}\left(\mathcal{P}_{a^*}, {}^{\llcorner o}\mathcal{R}(\mathcal{P}_{a^*}), \mathcal{F}_{a^*}\right) | {}^{\llcorner o}\mathcal{R}(\mathcal{P}_{a^*}), \mathcal{P}_{a^*}\right] \leq \mathcal{E}[\mathcal{Z}(\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*})|a, \mathcal{P}_{a^*}].$$

Let an admissible rule $\mathcal{R} : \mathcal{P}^*_{a^*} \to a^*$ assign a decision $a \in a^*$ to the considered $\mathcal{P}_{a^*}$. Then, the previous inequality can be written

$$\begin{aligned}\mathcal{E}[\mathcal{Z}(\mathcal{P}_{a^*}, {}^{\llcorner o}\mathcal{R}(\mathcal{P}_{a^*}), \mathcal{F}_{a^*})| {}^{\llcorner o}\mathcal{R}(\mathcal{P}_{a^*}), \mathcal{P}_{a^*}]& \\ \leq \mathcal{E}[\mathcal{Z}(\mathcal{P}_{a^*}, \mathcal{R}(\mathcal{P}_{a^*}), \mathcal{F}_{a^*})|\mathcal{R}(\mathcal{P}_{a^*}), \mathcal{P}_{a^*}].&\end{aligned}$$

Let us apply unconditional expectation $\mathcal{E}[\cdot]$ acting on functions of $\mathcal{P}_{a^*}$ to this inequality. Due to the isotonic nature of $\mathcal{E}[\cdot]$, the inequality is preserved. The

the chain rule for expectations — see Proposition 2.6 — implies that on the right-hand side of the resulting inequality we get the unconditional expected loss corresponding to an arbitrarily chosen $\mathcal{R} : \mathcal{P}^*_{a^*} \to a^*$. On the left-hand side the unconditional expected loss for ${}^{\lfloor o}\mathcal{R}$ arises. Thus, ${}^{\lfloor o}\mathcal{R}$ that assigns to each $\mathcal{P}_{a^*} \in \mathcal{P}^*_{a^*}$ the decision ${}^{\lfloor o}a(\mathcal{P}_{a^*})$ is the optimal decision rule. □

The proposition and its proof imply no preferences if there are several globally minimizing arguments ${}^{\lfloor o}a(\mathcal{P}_{a^*})$. We can use any of them or switch between them in a random manner whenever the pdf $f(a_t|\mathcal{P}_{a_t^*})$ has its support concentrated on them. This is an example where a *randomized causal strategy* may occur. We specify it formally as it is extensively used later on.

Agreement 2.6 (Outer model of randomized strategy) *The pdf* $f(a|\mathcal{P}_{a^*})$ *is called the* outer model of the decision rule*. The pdfs* $\left\{f(a_t|\mathcal{P}_{a_t^*})\right\}_{t\in t^*}$ *form the* outer model of the decision strategy*.*

A decision rule $f(a|\mathcal{P}_{a^*})$ *is called a* randomized decision rule *if its support contains at least two different values of* a_t*. The strategy is called a* randomized strategy *if some of its rules are randomized.*

Remark(s) 2.5

1. *We do not enter the technical game with* ε*-optimum: the existence of the various minimizing arguments is implicitly supposed.*
2. *It is worth repeating that the optimal decision rule is constructed valuewise. Formally, the minimization should be performed for all possible instances of experience* $\mathcal{P}_{a^*} \in \mathcal{P}^*_{a^*}$ *in order to get the decision rule. Often, we are interested in the optimal decision for a given fixed, say observed, experience. Then, just a single minimization is necessary. This is typically the case of the estimation problem. This possibility makes the main distinction from the dynamic design, when optimal strategy, a sequence of decision rules, is searched for. In this case, discussed in next section, the construction of decision rules is necessary. This makes the dynamic design substantially harder and, mostly, exactly infeasible [75, 76].*

2.4 Dynamic design

We are searching for the optimal admissible strategy assuming that each rule has at least the same experience as its predecessor. This *extending experience* models an increasing number of data available for the decision-making.

2.4.1 Dynamic programming

The optimal admissible strategy can be found by using a stochastic version of celebrated *dynamic programming* [77]. It is nothing but a repetitive application of Proposition 2.7 evolving an auxiliary function $\mathcal{V}(\mathcal{P}_{a_t^*})$ and determining actions of the constructed optimal strategy.

Proposition 2.8 (Stochastic dynamic programming) *Optimal causal strategy ${}^{\lfloor o}\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}_{t\in t^*} \in \{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}^*_{t\in t^*}$ with extending experience $\mathcal{P}_{a_t^*} \subset \mathcal{P}_{a_{t+1}^*}$ and minimizing the expected loss function $\mathcal{E}[\mathcal{Z}(\mathcal{Q})]$ can be constructed in a valuewise way.*

For every $t \in t^*$ and $\mathcal{P}_{a_t^*} \in \mathcal{P}^*_{a_t^*}$, it is sufficient to take a minimizing argument ${}^{\lfloor o}a(\mathcal{P}_{a_t^*})$ of

$$\mathcal{V}(\mathcal{P}_{a_t^*}) = \min_{a_t \in a_t^*} \mathcal{E}[\mathcal{V}(\mathcal{P}_{a_{t+1}^*})|a_t, \mathcal{P}_{a_t^*}],\ t \in t^* \tag{2.17}$$

as the action generated by the tth rule of the optimal strategy, i.e., ${}^{\lfloor o}a(\mathcal{P}_{a_t^*}) = {}^{\lfloor o}\mathcal{R}_t(\mathcal{P}_{a_t^*})$.

The functional recursion (2.17) is evaluated in the backward manner against the course given by the extending experience. It starts with

$$\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}+1}}) \equiv \mathcal{E}[\mathcal{Z}(\mathcal{Q})|\mathcal{P}_{a^*_{\mathring{t}+1}}], \tag{2.18}$$

where $\mathcal{P}_{a^*_{\mathring{t}+1}}$ contains all information available up to and including time $\mathring{t}$. The reached minimum has the value

$$\mathcal{E}[\mathcal{V}(\mathcal{P}_{a_1^*})] = \min_{\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}^*_{t\in t^*}} \mathcal{E}[\mathcal{Z}(\mathcal{Q})].$$

Proof. Let us define $\mathcal{P}_{a^*_{\mathring{t}+1}}$ as all information available at time $\mathring{t}$. The the chain rule for expectations and definition (2.18) imply

$$\mathcal{E}[\mathcal{Z}(\mathcal{Q})] = \mathcal{E}[\mathcal{E}[\mathcal{Z}(\mathcal{Q})|\mathcal{P}_{a^*_{\mathring{t}+1}}]] \equiv \mathcal{E}[\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}+1}})].$$

This identity allows us to get a uniform notation. Note that the definition of $\mathcal{P}_{a^*_{\mathring{t}+1}}$ is legitimate as $a_{\mathring{t}+1}$ is not optimized.

The definition of minimum and Proposition 2.7 imply

$$\begin{aligned}
&\min_{\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}^*_{t\in t^*}} \mathcal{E}[\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}+1}})] \\
&= \min_{\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}^*_{t<\mathring{t}}} \left\{ \min_{\{\mathcal{R}_{\mathring{t}} : \mathcal{P}^*_{a^*_{\mathring{t}}} \to a^*_{\mathring{t}}\}} \mathcal{E}[\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}+1}})] \right\} \\
&\underbrace{=}_{(2.16)} \min_{\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}^*_{t<\mathring{t}}} \mathcal{E}\left[\min_{a_{\mathring{t}} \in a^*_{\mathring{t}}} \mathcal{E}[\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}+1}})|a_{\mathring{t}}, \mathcal{P}_{a^*_{\mathring{t}}}] \right].
\end{aligned}$$

Denoting $\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}}}) \equiv \min_{a_{\mathring{t}} \in a^*_{\mathring{t}}} \mathcal{E}[\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}+1}})|a_{\mathring{t}}, \mathcal{P}_{a^*_{\mathring{t}}}]$, we proved the first step of the recursion and specified the start (2.18). The following step becomes

$$\min_{\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}^*_{t<\mathring{t}}} \mathcal{E}\left[\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}}})\right].$$

We face the identical situation as above with the horizon decreased by one. Thus, the procedure can be repeated until the initial optimal rule ${}^{\lfloor o}\mathcal{R}_1$ is constructed. □

The optimization relies on our ability to evaluate the expectations

$$\mathcal{E}[\mathcal{V}(\mathcal{P}_{a^*_{t+1}})|a_t, \mathcal{P}_{a^*_t}] = \int \mathcal{V}(\mathcal{P}_{a^*_t}, a_t, \Delta_t) f(\Delta_t|a_t, \mathcal{P}_{a^*_t})\, d\Delta_t, \ \forall t \in t^*.$$

The introduced *innovation* Δ_t contains those observable quantities that can be used for the choice a_{t+1} but not for the choice of a_t. They belong to $\mathcal{P}_{a^*_{t+1}} \setminus a_t$ but not to $\mathcal{P}_{a^*_t}$. The pdfs $\left\{f(\Delta_t|a_t, \mathcal{P}_{a^*_t})\right\}_{t\in t^*}$ model the relationships of Δ_t to a_t and $\mathcal{P}_{a^*_t}$.

Agreement 2.7 (Outer model of the system) *The collection of pdfs*

$$\left\{f(\Delta_t|a_t, \mathcal{P}_{a^*_t})\right\}_{t\in t^*}, \tag{2.19}$$

needed for the optimal design, is called the outer model of the system*.*

Remark(s) 2.6

1. *The term outer model of the system is shortened to the model of the system or even to the model. The exact meaning is clear from the specific context.*
2. *Often, the innovation $\Delta_t = y_t =$ observable output of the system.*
3. *The set-point s_t to which the output should be driven by the chosen input u_t has to be included into Δ_{u_t} if its values are uncertain, i.e., if $s_t \in \mathcal{P}_{u^*_{t+1}} \setminus \mathcal{P}_{u^*_t} \setminus u_t$.*

The following agreement is used in a presentation of the most common version of dynamic programming.

Agreement 2.8 (Internal quantities; data-driven design) *The behavior $\mathcal{Q}$ consists generally of potentially observable innovations $\Delta(\mathring{t})$, optional actions $a(\mathring{t})$, and* internal quantities $\Theta(\mathring{t})$ *in $\mathcal{Q}$ that are never observed directly, i.e., $\Theta(\mathring{t}) \in \mathcal{F}_{a^*_\tau}$, $\tau \in t^*$.*

The design is called data driven *iff the involved loss function depends on optional and potentially observable quantities, i.e., with ignorance consisting only of unobserved data*

$$\mathcal{Z}(\mathcal{Q}) \equiv \mathcal{Z}(\Delta(\mathring{t}), a(\mathring{t})) \equiv \mathcal{Z}(\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*}). \tag{2.20}$$

Note that in the general case the loss function may depend also on internal quantities that are never observed by decision maker. Then, the evaluation of the conditional expectation of the terminal condition (2.18) requires the additional pdf $f(\Theta(\mathring{t})|\mathcal{P}_{a^*_{\mathring{t}+1}})$. The discussion how to construct this pdf is postponed to Section 2.5 after presenting specialized data-driven versions of dynamic programming.

Proposition 2.9 (Dynamic programming for additive loss) *Let us consider data-driven design and search for the optimal admissible strategy* $\left\{ {}^{\lfloor o}\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^* \right\}_{t\in t^*}$ *acting on an extending experience* $\{\mathcal{P}_{a_t^*}\}_{t\in t^*}$, $\mathcal{P}_{a_t^*} \subset \mathcal{P}_{a_{t+1}^*}$. *Then, the optimal strategy* $\left\{ {}^{\lfloor o}\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^* \right\}_{t\in t^*}$ *minimizing the expected additive loss function, determined by the* partial loss $z(\Delta(t), a(t)) \geq 0$,

$$\mathcal{E}\left[\mathcal{Z}\left(\Delta(\mathring{t}), a(\mathring{t})\right)\right] \equiv \mathcal{E}\left[\sum_{t\in t^*} z(\Delta(t), a(t))\right], \tag{2.21}$$

can be constructed in the following valuewise way.

For all $t \in t^*$ *and* $\mathcal{P}_{a_t^*} \in \mathcal{P}^*_{a_t^*}$, *a minimizing argument* ${}^{\lfloor o}a(\mathcal{P}_{a_t^*})$ *of*

$$\mathcal{V}(\mathcal{P}_{a_t^*}) = \min_{a_t \in a_t^*} \mathcal{E}[z(\Delta(t), a(t)) + \mathcal{V}(\mathcal{P}_{a_{t+1}^*})|a_t, \mathcal{P}_{a_t^*}],\ t \in t^* \tag{2.22}$$

is taken as the optimal decision, ${}^{\lfloor o}a(\mathcal{P}_{a_t^*}) = {}^{\lfloor o}\mathcal{R}_t(\mathcal{P}_{a_t^*})$. *The recursion (2.22) is performed in the backward manner against the course given by the extending experience, starting from*

$$\mathcal{V}(\mathcal{P}_{a^*_{\mathring{t}+1}}) \equiv 0. \tag{2.23}$$

The minimum reached has the value

$$\min_{\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}^*_{t\in t^*}} \mathcal{E}[\mathcal{Z}(\mathcal{P}_{a^*_{\mathring{t}+1}})] = \mathcal{E}[\mathcal{V}(\mathcal{P}_{a_1^*})].$$

Proof. It follows exactly the line of Proposition 2.8 with a modified definition of the function $\mathcal{V}(\cdot)$

$$\mathcal{V}(\mathcal{P}_{a_t^*}) \equiv \min_{\{\mathcal{R}_\tau : \mathcal{P}^*_{a_\tau^*} \to a_\tau^*\}^*_{\tau\geq t}} \mathcal{E}\sum_{\tau\geq t}\left[z(\Delta(\tau), a(\tau))|\, a_t = \mathcal{R}_t(\mathcal{P}_{a_t^*}), \mathcal{P}_{a_t^*}\right]. \tag{2.24}$$

□

For reference purposes, we formulate the following agreement.

Agreement 2.9 (Bellman function; loss-to-go) *The function* $\mathcal{V}(\cdot)$ *occurring in dynamic programming is called the* Bellman function. *The Bellman function* $\mathcal{V}(\cdot)$ *in (2.24) is also called the* optimal loss-to-go.

2.4.2 Fully probabilistic design

A specific design that expresses losses fully in probabilistic terms is formulated and solved here. It is systematically used in the body of the text. Moreover, it is believed to form a bridge between optimal and practically optimal designs.

The notion of the *Kullback–Leibler divergence* [37] that measures well proximity of a pair of pdfs is widely used.

Agreement 2.10 (Kullback–Leibler divergence) *Let f, g be a pair of pdfs acting on a common set x^*. Then, the Kullback–Leibler divergence $\mathcal{D}(f||g)$ is defined by the formula*

$$\mathcal{D}(f||g) \equiv \int f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx. \qquad (2.25)$$

For conciseness, the Kullback–Leibler divergence is referred to as the KL divergence.

Proposition 2.10 (Basic properties of KL divergence) *Let f, g be a pair of pdfs acting on a same set. It holds*

1. $\mathcal{D}(f||g) \geq 0$*,*
2. $\mathcal{D}(f||g) = 0$ *iff* $f = g$ *(a.e.),*
3. $\mathcal{D}(f||g) = \infty$ *iff on a set of a positive dominating measure* $f > 0$ *and* $g = 0$*,*
4. $\mathcal{D}(f||g) \neq \mathcal{D}(g||f)$ *and the KL divergence does not obey triangle inequality.*

Proof. See, for instance, [74]. □

Now, the fully probabilistic design problem can be formulated and solved. A simple version considering the data-driven design is presented here. In this case, the joint pdf $f(\mathcal{Q}) \equiv f(\Delta(\mathring{t}), a(\mathring{t}))$ describing observable quantities of interest can be factorized by a repetitive use of the chain rule

$$f(\Delta(\mathring{t}), a(\mathring{t})) = \prod_{t \in t^*} f(\Delta_t | a_t, \mathcal{P}_{a_t^*}) f(a_t | \mathcal{P}_{a_t^*}). \qquad (2.26)$$

The first factors $\left\{ f(\Delta_t | a_t, \mathcal{P}_{a_t^*}) \right\}_{t \in t^*}$ under the product sign describe possible reactions of the system on the decision a_t under the experience $\mathcal{P}_{a_t^*}$. These pdfs form the outer model of the system; see Agreement 2.7. Similarly, the pdfs $\left\{ f(a_t | \mathcal{P}_{a_t^*}) \right\}_{t \in t^*}$ represent an outer model of the randomized decision strategy to be chosen; see Agreement 2.6. Looking at the joint pdf (2.26), it seems to be natural to formulate the design as the selection of such a decision strategy that make this pdf as close as possible to some “ideal” joint pdf.

Agreement 2.11 (Fully probabilistic design) *The fully probabilistic, data-driven design, $\mathcal{P}_{a_t^*} \equiv d(t-1) \equiv (\Delta(t-1), a(t-1)))$, specifies its loss function through an* ideal pdf

$${}^{\lfloor I}f(\mathcal{Q}) = \prod_{t \in t^*} {}^{\lfloor I}f(\Delta_t | a_t, d(t-1)) \, {}^{\lfloor I}f(a_t | d(t-1)).$$

The optimal admissible, possibly randomized, decision strategy is defined as a minimizer of the KL divergence (2.25) of $f(d(\mathring{t})) = f(\Delta(\mathring{t}), a(\mathring{t}))$ and ${}^{\lfloor I}f(d(\mathring{t})) = {}^{\lfloor I}f(\Delta(\mathring{t}), a(\mathring{t}))$

$$\mathcal{D}\left(f \,\middle|\middle|\, {}^{\lfloor I}f\right) \equiv \int f(\Delta(\mathring{t}), a(\mathring{t})) \ln\left(\frac{f(\Delta(\mathring{t}), a(\mathring{t}))}{{}^{\lfloor I}f(\Delta(\mathring{t}), a(\mathring{t}))}\right) d(\Delta(\mathring{t}), a(\mathring{t})). \qquad (2.27)$$

Proposition 2.11 (Solution of fully probabilistic design) *The optimal strategy minimizing the KL divergence (2.27) has the form*

$$f(a_t|d(t-1)) = {}^{\lfloor I}f(a_t|d(t-1))\frac{\exp[-\omega_\gamma(a_t,d(t-1))]}{\gamma(d(t-1))}, \tag{2.28}$$

$$\gamma(d(t-1)) \equiv \int {}^{\lfloor I}f(a_t|d(t-1))\exp[-\omega(a_t,d(t-1))]\,da_t,\ \textit{for } t<\mathring{t},$$

$$\gamma(d(\mathring{t})) = 1, \tag{2.29}$$

$$\omega_\gamma(a_t,d(t-1)) \equiv \int f(\Delta_t|a_t,d(t-1))\ln\left(\frac{f(\Delta_t|a_t,d(t-1))}{\gamma(d(t))\,{}^{\lfloor I}f(\Delta_t|a_t,d(t-1))}\right)d\,\Delta_t.$$

The solution is performed against the time course, starting at $t=\mathring{t}$.

Proof. With the chain rule, the KL divergence gets the form $\mathcal{D}\left(f||\,{}^{\lfloor I}f\right)$

$$= \mathcal{E}\left\{\sum_{t\in t^*}\int f(a_t|d(t-1))\left[\ln\left(\frac{f(a_t|d(t-1))}{{}^{\lfloor I}f(a_t|d(t-1))}\right)+\omega(a_t,d(t-1))\right]da_t\right\},$$

$$\omega(a_t,d(t-1)) \equiv \int f(\Delta_t|a_t,d(t-1))\ln\left(\frac{f(\Delta_t|a_t,d(t-1))}{{}^{\lfloor I}f(\Delta_t|a_t,d(t-1))}\right)d\Delta_t.$$

Let us denote

$$-\ln(\gamma(d(t))) \equiv \min_{\{f(a_{\tau+1}|d(\tau))\}_{\tau=t}^{\mathring{t}}} \mathcal{E}\left\{\sum_{\tau=t+1}^{\mathring{t}}\int f(a_\tau|d(\tau-1))\right.$$
$$\left.\times\left[\ln\left(\frac{f(a_\tau|d(\tau-1))}{{}^{\lfloor I}f(a_\tau|d(\tau-1))}\right)+\omega(a_\tau,d(\tau-1))\right]da_\tau\right|d(t)\Bigg\}.$$

Then, this definition implies that $\gamma(d(\mathring{t}))=1$ and $-\ln(\gamma(d(t)))$

$$\equiv \min_{f(a_{t+1}|d(t))}\int f(a_{t+1}|d(t))\left[\ln\left(\frac{f(a_{t+1}|d(t))}{{}^{\lfloor I}f(a_{t+1}|d(t))}\right)+\omega_\gamma(a_{t+1},d(t))\right]da_{t+1},$$

$$\omega_\gamma(a_{t+1},d(t)) \equiv \int f(\Delta_{t+1}|a_{t+1},d(t))\ln\left(\frac{f(\Delta_{t+1}|a_{t+1},d(t))}{\gamma(d(t+1))\,{}^{\lfloor I}f(\Delta_{t+1}|a_{t+1},d(t))}\right)d\Delta_{t+1}.$$

It gives $-\ln(\gamma(d(t)))$

$$\equiv \min_{f(a_{t+1}|d(t))}\int f(a_{t+1}|d(t))\left[\ln\left(\frac{f(a_{t+1}|d(t))}{\frac{{}^{\lfloor I}f(a_{t+1}|d(t))\exp[-\omega_\gamma(a_{t+1},d(t))]}{\int {}^{\lfloor I}f(\tilde{a}_{t+1}|d(t))\exp[-\omega_\gamma(\tilde{a}_{t+1},d(t))]\,d\tilde{a}_{t+1}}}\right)da_{t+1}\right.$$
$$\left.-\ln\left(\int {}^{\lfloor I}f(a_{t+1}|d(t))\exp\left[-\omega_\gamma(a_{t+1},d(t))\right]\,da_{t+1}\right)\right].$$

The first term in the above identity is the KL divergence that reaches its smallest zero value for the claimed pdf. It also defines the form of the minima reached. □

Remark(s) 2.7

1. *For an alternative derivation with more details see [63, 78].*
2. *At a descriptive level, the stochastic dynamic programming consists of a sequence of the evaluation pairs*

   $$(\textit{conditional expectation, minimization}).$$

   Except of a few numerically solvable cases, some approximation techniques have to be employed. The complexity of the approximated optimum prevents a systematic use of the standard approximation theory. Consequently, various ad hoc schemes are adopted. The fully probabilistic design finds minimizers explicitly and thus reduces the design to a sequence of conceptually feasible multivariate integrations.
3. *The found optimal strategy is randomized and obviously causal one. The physical constraints are met trivially if the chosen ideal strategy respects them, i.e., if* $\mathrm{supp}\left[\ ^{\lfloor I}f(a_t|\mathcal{P}_{a_t^*})\right] \subset a_t^*$, *cf. (2.28).*

2.4.3 Asymptotic of the design

The asymptotic of the dynamic programming is analyzed for horizon $\mathring{t} \to \infty$ within this section. The outlined analysis serves us only as a motivation for approximate design; for instance, see Algorithm 4.2. Thus, all technicalities are suppressed as much as possible.

The data-driven case with an additive loss function (2.21) is considered. The general, data-dependent loss function can always be converted into the additive form by defining the partial loss

$$z(\mathit{\Delta}(t), a(t))) = \left\{ \begin{array}{cl} \mathcal{Z}(\mathit{\Delta}(\mathring{t}), a(\mathring{t})) & \text{if } t = \mathring{t}, \\ 0 & \text{otherwise} \end{array} \right. . \tag{2.30}$$

We deal, however, with a simpler but still useful case by assuming that

- there is a finite-dimensional *information state* x_{t-1}, i.e., $\mathcal{P}_{a_t^*} \equiv x_{t-1} \equiv$ an observed finite-dimensional vector,
- the partial loss depends on the information state x_t and the action a_t only $z(\mathit{\Delta}(t), a(t)) \equiv z(x_t, a_t)$, i.e., the considered loss is

$$\mathcal{Z}(\mathit{\Delta}(\mathring{t}), a(\mathring{t})) = \sum_{t \in t^*} z(x_t, a_t). \tag{2.31}$$

Agreement 2.12 (Stabilizing strategy) *Let us consider sequence of decision-making problems with the growing horizon* $\mathring{t} \to \infty$, *i.e., with extending sets* $^{\lfloor \mathring{t}}t^* \equiv \{1, \ldots, \mathring{t}\}$. *The infinite sequence of decision rules*

$$\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\}_{t \in\ ^{\lfloor \infty}t^* \equiv \{1,2,\ldots,\}}$$

is called the stabilizing strategy *if there is a finite constant c such that*

$$\mathcal{E}[z(x_t, a_t)|a_t, \mathcal{P}_{a_t^*}] \le c < \infty,\ t \in\ ^{\lfloor \infty}t^* \equiv \{1, 2, 3, \ldots\}. \tag{2.32}$$

Intuitively, it is obvious that with growing decision horizon the expected loss, as a sum of positive terms, grows to infinity. Consequently, the influence of individual rules on it decreases. The following proposition shows that in this practically important case the optimal strategy can be chosen as stationary one. The *stationary strategy* is formed by a repetitive use of the same rule whose (approximate) evaluation is simpler than that of a general strategy with time-varying rules.

Proposition 2.12 (Asymptotic design) *Let a stabilizing strategy exist. Then, for $\mathring{t} \to \infty$, the optimal strategy can be chosen as stationary one. Decisions generated by the rule defining it are minimizing arguments in the formal analogy of (2.22)*

$$ {}^{\lfloor\infty}\mathcal{V}(x_{t-1}) + {}^{\lfloor\infty}C = \min_{a_t \in a_t^*} \mathcal{E}\left[z(x_t, a_t) + {}^{\lfloor\infty}\mathcal{V}(x_t) \middle| a_t, x_{t-1} \right] \qquad (2.33) $$

with a constant ${}^{\lfloor\infty}C \le c$ and a time-invariant Bellman function ${}^{\lfloor\infty}\mathcal{V}(x)$.

Proof. Let us take any finite horizon $\mathring{t}$ and, within this horizon, denote ${}^{\lfloor\mathring{t}}\tilde{\mathcal{V}}(\mathcal{P}_{a_t^*}) \equiv {}^{\lfloor\mathring{t}}\tilde{\mathcal{V}}(x_{t-1})$ the optimal loss-to-go; see Agreement 2.9.

Let us define ${}^{\lfloor\mathring{t}}C$ as the smallest value such that

$$ {}^{\lfloor\mathring{t}}\mathcal{V}(x_t) \equiv {}^{\lfloor\mathring{t}}\tilde{\mathcal{V}}(x_t) - (\mathring{t} - t)\, {}^{\lfloor\mathring{t}}C $$

is bounded from above for $\mathring{t} \to \infty$ and any fixed t, x_t. Obviously, the optimal strategy cannot lead to a higher expected loss than any stabilizing strategy. Thus, the optimal strategy has to also be a stabilizing strategy. Thus, ${}^{\lfloor\mathring{t}}C \le c$ and $\overline{\lim}_{\mathring{t}\to\infty} {}^{\lfloor\mathring{t}}C = {}^{\lfloor\infty}C$ exists.

The optimization is uninfluenced if we subtract the value ${}^{\lfloor\mathring{t}}C$ from all partial losses. For arbitrary fixed t, x_t, the corresponding modified Bellman function ${}^{\lfloor\mathring{t}}\mathcal{V}(x_t)$, is bounded from above and ${}^{\lfloor\mathring{t}}\mathcal{V}(x_t) = {}^{\lfloor\mathring{t}}\tilde{\mathcal{V}}(x_t) - (\mathring{t} - t)\, {}^{\lfloor\mathring{t}}C$ is the difference between a pair of monotonous sequences (indexed by $\mathring{t}$). Thus, a finite limit ${}^{\lfloor\infty}\mathcal{V}(x_t) = \lim_{\mathring{t}\to\infty} {}^{\lfloor\mathring{t}}\mathcal{V}(x_t)$ exists.

The modified Bellman function fulfills the equation

$$ {}^{\lfloor\mathring{t}}\mathcal{V}(x_{t-1}) + {}^{\lfloor\mathring{t}}C = \min_{a_t \in a_t^*} \mathcal{E}\left[z(x_t, a_t) + {}^{\lfloor\mathring{t}}\mathcal{V}(x_t) \middle| a_t, x_{t-1} \right]. $$

Existence and finiteness of the involved limits imply that the asymptotic version of the Bellman equation is fulfilled, too,

$$ {}^{\lfloor\infty}\mathcal{V}(x_{t-1}) + \overline{\lim}_{\mathring{t}\to\infty} {}^{\lfloor\mathring{t}}C = \min_{a_t \in a_t^*} \mathcal{E}\left[z(x_t, a_t) + {}^{\lfloor\infty}\mathcal{V}(x_t) \middle| a_t, x_{t-1} \right]. $$

Limits of ${}^{\lfloor\mathring{t}}\mathcal{V}(x_t)$ exist and, thus, $\overline{\lim}_{\mathring{t}\to\infty} {}^{\lfloor\mathring{t}}C = \lim_{\mathring{t}\to\infty} {}^{\lfloor\mathring{t}}C = {}^{\lfloor\infty}C$.

The identical optimization is performed for each $t < \infty$. Thus, it provides the same decision rule for each t: the optimal strategy is a stationary one. □

Remark(s) 2.8

1. *The same proof is directly applicable to the fully probabilistic design as it can be seen as an instance of the additive loss function.*
2. *Solutions of the Bellman equation obtained for a growing finite horizon* $\mathring{t}$ *can be interpreted as successive approximations for solving its stationary counterpart (2.33).*
3. *So-called* iterations in strategy space *[79] are an alternative and efficient way of finding the asymptotic solution. Essentially, a stabilizing stationary strategy* $\{\mathcal{R}\}$ *is selected and the linear equation*

$$\mathcal{V}(x) + C = \mathcal{E}[z(\tilde{x}, \mathcal{R}(x)) + \mathcal{V}(\tilde{x})|\mathcal{R}(x), x]$$

is solved for the function $\mathcal{V}(\cdot)$ *and constant* C. *Then, a new approximating strategy is found valuewise* $\mathcal{R}(x) \in \mathrm{Arg\,min}_{a \in a^*} \mathcal{E}[z(\tilde{x}, a) + \mathcal{V}(\tilde{x})|a, x]$ *with such a* $\mathcal{V}(\cdot)$ *(*Arg min *denotes a* set of minimizing arguments*). Under general conditions, the newly found strategy is stabilizing and iterations may be repeated until the guaranteed convergence. Details of this procedure are beyond the scope of this work but it should be considered when searching for efficient numerical procedures.*

2.5 Learning

Considered behavior $\mathcal{Q}^*$ contains generally internal quantities $\Theta(\mathring{t})$, Agreement 2.8, that are never observed directly. Despite this, we want to describe or influence them. Still, the optimal decision-making needs the outer model (2.19); see Proposition 2.8. If, moreover, the loss function depends on $\Theta(\mathring{t})$, the general dynamic programming needs the pdf $f(\Theta(\mathring{t})|\mathcal{P}_{a^*_{\mathring{t}+1}})$ for evaluation of the initial condition (2.18).

Here we describe how to get both outer model and this estimate of internal quantities. The solved problem, known as nonlinear filtering [46], is of independent interest as its solution provides a consistent *formal model of learning.*

2.5.1 Bayesian filtration

The joint pdf $f(\mathcal{Q})$ describing both observed and internal quantities is constructed from the following elements.

Requirement 2.5 (Models; natural conditions of decision making)

1. *The innovations* Δ_t *are related to experience* $\mathcal{P}_{a^*_t}$ *and decisions* a_t *through the* observation model

$$\{f(\Delta_t|a_t, \mathcal{P}_{a^*_t}, \Theta_t) \equiv f(\Delta_t|a_t, \mathcal{P}_{a^*_t}, \Theta(t))\}_{t \in t^*} \quad (2.34)$$

that is given up to unknown internal quantities $\Theta_t \in \Theta^*_t \subset \mathcal{F}_{a^*_\tau}$, $\forall \tau \in t^*$.

2. *The internal quantities $\Theta(\mathring{t}) \in \Theta^*(\mathring{t})$ are described by a known collection of pdfs called the* time evolution model
$$\left\{f(\Theta_t|a_t, \mathcal{P}_{a_t^*}, \Theta_{t-1}) \equiv f(\Theta_t|a_t, \mathcal{P}_{a_t^*}, \Theta(t-1))\right\}_{t\in t^*}. \tag{2.35}$$
3. *The quantities $\Theta(\mathring{t})$ are unknown to the strategies considered. The* natural conditions of decision making *(a slight generalization of natural conditions of control [69]) express it formally. They postulate independence of a_t and Θ_t when conditioned on $\mathcal{P}_{a_t^*}$*
$$f(a_t|\mathcal{P}_{a_t^*}, \Theta_t) = f(a_t|\mathcal{P}_{a_t^*}) \quad \underbrace{\Leftrightarrow}_{Proposition\ 2.4} \quad f(\Theta_t|a_t, \mathcal{P}_{a_t^*}) = f(\Theta_t|\mathcal{P}_{a_t^*}). \tag{2.36}$$
4. *The initial experience $\mathcal{P}_{a_1^*}$ coincides with the prior information about the initial internal quantity Θ_0 so that the* prior pdf $f(\Theta_0)$ *fulfills*
$$f(\Theta_0) \equiv f(\Theta_0|\mathcal{P}_{a_1^*}) \underbrace{=}_{(2.36)} f(\Theta_0|a_1, \mathcal{P}_{a_1^*}). \tag{2.37}$$

Remark(s) 2.9

1. *The conditional independence, required by (2.34) for observations and by (2.35) for time evolution, is used to simplify notation. Generally, it is unnecessary.*
2. *Often, the unknown quantities Θ_t together with the decision a_t are assumed to describe the involved conditional pdfs fully. Then, $\mathcal{P}_{a_t^*}$ can be omitted and Θ_t can be identified with the* information state.
3. *The natural conditions of decision making express the assumption that $\Theta_t \notin \mathcal{P}_{a_\tau^*}$ $\forall\tau$, $\forall t \in t^*$. Thus, values of Θ_t cannot be used by the decision rules forming the admissible strategy. Alternatively, we cannot gain information about Θ_t from the decision a_t if the corresponding innovation Δ_t (the corresponding reaction of the system) is not available.*
   The natural conditions of decision making are "naturally" fulfilled by strategies we are designing. They have to be checked when the data influenced by an "externally chosen" strategy are processed.

Proposition 2.13 (Generalized Bayesian filtering) *Let Requirement 2.5 be met. Then, the outer model of the system (2.19) is given by the formula*
$$f(\Delta_t|a_t, \mathcal{P}_{a_t^*}) = \int f(\Delta_t|a_t, \mathcal{P}_{a_t^*}, \Theta_t) f(\Theta_t|\mathcal{P}_{a_t^*})\, d\Theta_t. \tag{2.38}$$

The evolution of the pdf $f(\Theta_t|\mathcal{P}_{a_t^*})$, called (generalized Bayesian) filtration of unknown quantities Θ_t, *is described by the following recursion that starts from the prior pdf $f(\Theta_0)$.*

- Time updating

$$f(\Theta_{t+1}|\mathcal{P}_{a^*_{t+1}}) = \int f(\Theta_{t+1}|a_{t+1}, \mathcal{P}_{a^*_{t+1}}, \Theta_t) f(\Theta_t|\mathcal{P}_{a^*_{t+1}})\, d\Theta_t \tag{2.39}$$

that reflects the time evolution $\Theta_t \to \Theta_{t+1}$.

- Data updating

$$f(\Theta_t|\mathcal{P}_{a^*_{t+1}}) = \frac{f(\Delta_t|a_t, \mathcal{P}_{a^*_t}, \Theta_t) f(\Theta_t|\mathcal{P}_{a^*_t})}{f(\Delta_t|a_t, \mathcal{P}_{a^*_t})} \propto f(\Delta_t|a_t, \mathcal{P}_{a^*_t}, \Theta_t) f(\Theta_t|\mathcal{P}_{a^*_t}) \tag{2.40}$$

that incorporates the innovation Δ_t *and the decision* a_t.

Proof. Sequential use of marginalization, the chain rule, Proposition 2.4, and the natural conditions of decision making (2.36) imply (2.38)

$$\begin{aligned} f(\Delta_t|a_t, \mathcal{P}_{a^*_t}) &= \int f(\Delta_t, \Theta_t|a_t, \mathcal{P}_{a^*_t})\, d\Theta_t \\ &= \int f(\Delta_t|a_t, \mathcal{P}_{a^*_t}, \Theta_t) f(\Theta_t|a_t, \mathcal{P}_{a^*_t})\, d\Theta_t \\ &= \int f(\Delta_t|a_t, \mathcal{P}_{a^*_t}, \Theta_t) f(\Theta_t|\mathcal{P}_{a^*_t})\, d\Theta_t. \end{aligned}$$

Marginalization, the chain rule, and natural conditions of decision making also imply the formula for time updating.

Data updating coincides with the Bayes rule in which the outer model of the strategy cancels as it does not depend on Θ_t due to the natural conditions of decision making (2.36). □

Agreement 2.13 (Filtering; predictive pdf) *The process of generating filtration is called (generalized Bayesian)* filtering. *The outer model of the system obtained by filtering is called* predictive pdf.

Remark(s) 2.10

1. *The term generalized distinguishes a nonstandard use of the terms Bayesian filtering and predictions. Without this adjective, they are understood as specific decision-making problems. The "generalization" means that the conditional pdfs needed for these tasks are evaluated only. They serve for solving a whole class of decision-making problems.*
2. *The term* predictive pdf *reflects how the outer model of the system has been obtained. It uses the observed experience and extrapolates it into ignorance assuming that the* mechanism of generating Θ_t does not change.
   This accumulation of experience and its extrapolation represent a good formal model of learning.
3. *It has to be stressed that the accumulation of experience can take place only when the rules governing the behavior are not changed during it, i.e., when we can rely on the validity of the underlying models.*

4. *The filtering results are often of independent interest. The construction of the predictive pdf and of the pdf needed in (2.18) are our key motivation for filtering. Under the adopted conditions stated in Requirement 2.5, the latter pdf can be evaluated recursively as follows.*

$$f(\Theta(\mathring{t})|\mathcal{P}_{a^*_{\mathring{t}+1}}) = \frac{f(\Theta(\mathring{t}), d_{\mathring{t}}|\mathcal{P}_{a^*_{\mathring{t}}})}{f(d_{\mathring{t}}|\mathcal{P}_{a^*_{\mathring{t}}})}$$

$$= \frac{f(d_{\mathring{t}}|\mathcal{P}_{a^*_{\mathring{t}}}, \Theta(\mathring{t}))f(\Theta_{\mathring{t}}|\mathcal{P}_{a^*_{\mathring{t}}}, \Theta(\mathring{t}-1))}{f(d_{\mathring{t}}|\mathcal{P}_{a^*_{\mathring{t}}})} f(\Theta(\mathring{t}-1)|\mathcal{P}_{a^*_{\mathring{t}}}) \qquad \underbrace{=}_{(2.34),(2.35),(2.36)}$$

$$= \frac{f(\mathit{\Delta}_{\mathring{t}}|a_{\mathring{t}}, \mathcal{P}_{a^*_{\mathring{t}}}, \Theta_t)f(\Theta_{\mathring{t}}|\mathcal{P}_{a^*_{\mathring{t}}}, \Theta_{\mathring{t}-1})}{f(\mathit{\Delta}_{\mathring{t}}|a_{\mathring{t}}, \mathcal{P}_{a^*_{\mathring{t}}})} f(\Theta(\mathring{t}-1)|\mathcal{P}_{a^*_{\mathring{t}}}).$$

   This recursion uses the observation model, the time-evolution model, and the predictive pdf. It can be formally repeated up to reaching prior pdf as starting point $f(\Theta(0)|\mathcal{P}_{a^*_1}) \equiv f(\Theta_0)$.
5. *Under the natural conditions of decision making, filtering relies on the knowledge of decisions and not on the knowledge of rules* $\mathcal{R} : \mathcal{P}^*_{a^*} \to a^*$ *generating them. It is important practically when we learn while decision loop is closed, especially, by a human decision maker.*
6. *The time evolution model* $f(\Theta_t|a_t, \mathcal{P}_{a^*_t}, \Theta_{t-1})$ *as well as the observation model* $f(\mathit{\Delta}_t|a_t, \mathcal{P}_{a^*_t}, \Theta_t)$ *have to result from a theoretical modelling of the system in question. Such modelling uses both field knowledge, like laws of conservation, and approximation capabilities of the selected family of models involved. Often, deterministic relationships are modelled and then the "deviations" from an "expected" trajectory are described.*
7. *The prior pdf* $f(\Theta_0)$ *allows us to introduce information based on expert knowledge or analogy to situations observed previously.*
8. *The observed data, the only bridge to reality, enter the evaluations in the data-updating step only when the newest innovation-decision pair is processed. This simple fact is important for approximation of the time evolution model; see Section 3.1.*
9. *In summary, the described Bayesian filtering combines prior information in* $f(\Theta_0)$*, theoretical knowledge of the specific fields described by* $f(\mathit{\Delta}_t|a_t, \mathcal{P}_{a^*_t}, \Theta_t)$, $f(\Theta_t|a_t, \mathcal{P}_{a^*_t}, \Theta_{t-1})$ *and data* $d(\mathring{t}) = (\mathit{\Delta}(\mathring{t}),\ a(\mathring{t}))$ *by using coherent deductive calculus with pdfs. This combination of information sources is a powerful internally consistent framework describing the essence of learning. Due to its deductive structure, an important assurance is gained: the incorrect modelling or non-informative data can only be blamed for a failure of the specific learning process.*

2.5.2 Bayesian estimation

This section deals with a special version of filtering called *estimation*. It arises when the internal quantities Θ_t are time invariant

$$\Theta_t = \Theta, \ \forall t \in t^*. \tag{2.41}$$

The common value Θ is called a *parameter*. In this case, the time evolution model is $f(\Theta_t|a_t, \mathcal{P}_{a_t^*}, \Theta_{t-1}) = \delta(\Theta_t - \Theta_{t-1})$. The employed Dirac delta function $\delta(\cdot)$ is a formal pdf of the measure fully concentrated on zero.

Proposition 2.14 (Generalized Bayesian estimation) *Let Requirement 2.5 be met with time invariant $\Theta_t = \Theta \in \Theta^* \subset \mathcal{F}_{a_\tau^*}$, $\forall \tau \in t^*$. Then, the outer model of the system (2.19) is given by the formula*

$$f(\Delta_t|a_t, \mathcal{P}_{a_t^*}) = \int f(\Delta_t|a_t, \mathcal{P}_{a_t^*}, \Theta) f(\Theta|\mathcal{P}_{a_t^*})\, d\Theta. \tag{2.42}$$

The evolution of the pdf $f(\Theta|\mathcal{P}_{a_t^*})$, called (generalized Bayesian) parameter estimation, *generates the parameter estimate coinciding with the* posterior pdf *of the* unknown parameter*. It is described by the recursion identical with the data updating (2.40)*

$$f(\Theta|\mathcal{P}_{a_{t+1}^*}) = \frac{f(\Delta_t|a_t, \mathcal{P}_{a_t^*}, \Theta) f(\Theta|\mathcal{P}_{a_t^*})}{f(\Delta_t|a_t, \mathcal{P}_{a_t^*})} \propto f(\Delta_t|a_t, \mathcal{P}_{a_t^*}, \Theta) f(\Theta|\mathcal{P}_{a_t^*}). \tag{2.43}$$

It starts from the prior pdf $f(\Theta) \equiv f(\Theta|a_1, \mathcal{P}_{a_1^*}) = f(\Theta|\mathcal{P}_{a_1^*})$.

The simplicity of the estimation formula allows us to write down its (non-recursive) batch variant

$$f(\Theta|\mathcal{P}_{a_{t+1}^*}) = \frac{\prod_{\tau \le t} f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \Theta) f(\Theta)}{\int \prod_{\tau \le t} f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \Theta) f(\Theta)\, d\Theta} \equiv \frac{\mathcal{L}(\Theta, \mathcal{P}_{a_{t+1}^*}) f(\Theta)}{\mathcal{I}(\mathcal{P}_{a_{t+1}^*})}. \tag{2.44}$$

The introduced likelihood function

$$\mathcal{L}(\Theta, \mathcal{P}_{a_{t+1}^*}) \equiv \prod_{\tau \le t} f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \Theta) \tag{2.45}$$

evolves according to the recursion identical with that for the non-normalized posterior pdf (2.43). It starts, however, from the $\mathcal{L}(\Theta, \mathcal{P}_{a_1^*})$ identically equal to 1. The normalization factor $\mathcal{I}(\cdot)$ is defined by the formula

$$\mathcal{I}(\mathcal{P}_{a_{t+1}^*}) = \int \mathcal{L}(\Theta, \mathcal{P}_{a_{t+1}^*}) f(\Theta)\, d\Theta \propto f(\Delta_t|a_t, \mathcal{P}_{a_t^*}). \tag{2.46}$$

With it, the outer model of the system (2.19) can alternatively be expressed as

$$f(\Delta_t|a_t, \mathcal{P}_{a_t^*}) = \frac{\mathcal{I}(\mathcal{P}_{a_{t+1}^*})}{\mathcal{I}(\mathcal{P}_{a_t^*})}. \tag{2.47}$$

Proof. It is again a simple exercise in calculus with pdfs, marginalization, the chain rule, and the Bayes rule, Proposition 2.4, under the natural conditions of decision making (2.36). □

Remark(s) 2.11

1. *The observation model $f(\Delta_t|a_t, \mathcal{P}_{a_t^*}, \Theta)$ is called a* parameterized model *whenever the estimation problem is considered. We respect this tradition.*
2. *Note that the recursive evolution of the pdf $f(\Theta|\mathcal{P}_{a_t^*})$ allows us to interpret the posterior pdf as the prior one before processing new observations.*
3. *The data inserted into the "objective" parameterized (observation) model gradually correct the subjectively chosen prior pdf $f(\Theta)$. The posterior pdf $f(\Theta|\mathcal{P}_{a_t^*})$ always reflects both objective and subjective information pieces. If the data are informative enough, the relative contribution of the single subjective factor $f(\Theta)$ to the posterior pdf is decreasing with increasing t as the likelihood function $\mathcal{L}(\Theta, \mathcal{P}_{a_{t+1}^*})$ contains t "objective" factors (2.45).*
4. *Zero values are preserved by multiplication. Thus, the posterior pdf redistributes the probability mass only within the support of the prior pdf, i.e., within the set* $\mathrm{supp}\,[\, f(\Theta)] \equiv \{\Theta \in \Theta^* : f(\Theta) > 0\}$. *This fact allows us to introduce hard bounds on possible parameter values but does not allow us to "learn" about parameters Θ out of the support* $\mathrm{supp}\,[\, f(\Theta)]$.
5. *Remarks 2.10, related to filtering, apply mostly to estimation, too. The parameter estimation is a task on its own; unknown parameters are always in the ignorance of the decision to be chosen; under the natural conditions of decision making (2.36), decisions values are only needed and the strategy $\left\{\mathcal{R}_t : \mathcal{P}^*_{a_t^*} \to a_t^*\right\}_{t\in t^*}$ generating them need not to be known.*
6. *The parameters Θ_t are usually assumed to be finite-dimensional in order to avoid technicalities related to measure theory. In exceptional cases, like description of the so-called equivalence approach (see Section 3.4), we deal with* potentially infinite-dimensional parameter. *It means that the number of unknown quantities is finite but increases without limitations. This case is often called* nonparametric estimation.

2.5.3 Asymptotic of estimation

The analysis outlined here serves us primarily for interpretation of estimation results when none of the considered parameterized models describes reality exactly. This interpretation can be directly used for constructions of approximate estimation, for instance; see Section 6.4.8. Similarly as for design, all technicalities are suppressed as much as possible.

The "objective" pdf $f(\mathcal{Q})$ (see Section 2.3.1) describing the system behavior is denoted here ${}^{\llcorner o}f(\mathcal{Q})$. The corresponding outer model of the system $f(\Delta_t|a_t, \mathcal{P}_{a_t^*})$ is denoted ${}^{\llcorner o}f(\Delta_t|a_t, \mathcal{P}_{a_t^*})$. Its relationship to the predictive pdf $f(\Delta_t|a_t, \mathcal{P}_{a_t^*})$ — obtained through the parameter estimation; see Proposition 2.14 — is inspected here.

For the analysis, the notion of a (relative) *entropy rate* $\mathcal{H}_\infty\left({}^{\llcorner o}f\middle|\middle|\,\Theta\right)$ is needed; see the discrete-valued analogy in [80]. For a given realization of behavior, it measures the distance of a parameterized model to the objective pdf and determines asymptotic behavior of Bayesian estimation.

For each $\Theta \in \Theta^*$, the entropy rate is defined by the formula

$$\mathcal{H}_\infty\left({}^{\lfloor o}f \middle|\middle| \Theta \right) \equiv \overline{\lim}_{t\to\infty} \mathcal{H}_t\left({}^{\lfloor o}f \middle|\middle| \Theta \right) \tag{2.48}$$

$$\equiv \overline{\lim}_{t\to\infty} \frac{1}{t} \sum_{\tau \le t} \int {}^{\lfloor o}f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}) \ln\left(\frac{{}^{\lfloor o}f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*})}{f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \Theta)} \right) d\Delta_\tau.$$

Proposition 2.15 (Asymptotic of estimation) *Let the natural conditions of decision-making (2.36) hold. For almost all $\Theta \in \Theta^*$, let there exist positive $\underline{C}_\Theta, \overline{C}_\Theta$ uniformly bounded by a finite c, i.e., $0 < \underline{C}_\Theta \le \overline{C}_\Theta \le c < \infty$, and a finite time moment $\bar{t}_\Theta \in \{1, 2, \ldots\}$, such that $\forall t > \bar{t}_\Theta$, $\forall \mathcal{P}_{a_{t+1}^*} \in \mathcal{P}^*_{a_{t+1}^*}$*

$$\underline{C}_\Theta f(\Delta_t|a_t, \mathcal{P}_{a_t^*}, \Theta) \le {}^{\lfloor o}f(\Delta_t|a_t, \mathcal{P}_{a_t^*}) \le \overline{C}_\Theta f(\Delta_t|a_t, \mathcal{P}_{a_t^*}, \Theta). \tag{2.49}$$

Then, the posterior pdf $f(\Theta|\mathcal{P}_{a_t^*})$ (2.43) converges almost surely to a pdf $f(\Theta|\mathcal{P}_{a_\infty^*})$. Its support coincides with the set of minimizing arguments in

$$\text{supp}\left[\, f(\Theta|\mathcal{P}_{a_\infty^*})\right] = \text{Arg} \min_{\Theta \in \text{supp}[\, f(\Theta)] \cap \Theta^*} \mathcal{H}_\infty\left({}^{\lfloor o}f \middle|\middle| \Theta \right). \tag{2.50}$$

Proof. Under the natural conditions of decision making (2.36), the posterior pdf (2.43) can be written in the form

$$f(\Theta|\mathcal{P}_{a_{t+1}^*}) \propto f(\Theta) \exp[-t\mathcal{H}(\mathcal{P}_{a_{t+1}^*}, \Theta)], \tag{2.51}$$

$$\mathcal{H}(\mathcal{P}_{a_{t+1}^*}, \Theta) = \frac{1}{t}\sum_{\tau \le t} \ln[\eta(\mathcal{P}_{a_\tau^*}, \Theta)], \quad \eta(\mathcal{P}_{a_{\tau+1}^*}, \Theta) \equiv \frac{{}^{\lfloor o}f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*})}{f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \Theta)}. \tag{2.52}$$

This form exploits the fact that the non-normalized posterior pdf can be multiplied by any factor independent of Θ.

Let us fix the argument $\Theta \in \Theta^*$ and define

$$e_{\Theta;\tau} \equiv \ln(\eta(\mathcal{P}_{a_{\tau+1}^*}, \Theta)) - {}^{\lfloor o}\mathcal{E}\left[\ln(\eta(\mathcal{P}_{a_{\tau+1}^*}, \Theta)) \middle| a_\tau, \mathcal{P}_{a_\tau^*}\right]$$

$$\equiv \ln(\eta(\mathcal{P}_{a_{\tau+1}^*}, \Theta)) - \int {}^{\lfloor o}f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}) \ln(\eta(\mathcal{P}_{a_{\tau+1}^*}, \Theta))\, d\Delta_\tau.$$

A direct check reveals that the introduced deviations $e_{\Theta;\tau}$ of $\ln(\eta(\mathcal{P}_{a_{\tau+1}^*}, \Theta))$ from their conditional expectations ${}^{\lfloor o}\mathcal{E}[\ln(\eta(\mathcal{P}_{a_{\tau+1}^*}, \Theta))|a_\tau, \mathcal{P}_{a_\tau^*}]$, given by ${}^{\lfloor o}f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*})$, are zero mean and mutually uncorrelated. With them,

$$\mathcal{H}(\mathcal{P}_{a_{t+1}^*}, \Theta) = \mathcal{H}_t\left({}^{\lfloor o}f \middle|\middle| \Theta \right) + \frac{1}{t}\sum_{\tau \le t} e_{\Theta;\tau}.$$

The assumption (2.49) implies that the variance of $e_{\Theta;\tau}$ is bounded. Consequently, the last term in the above expression converges to zero almost surely;

see [81], page 417. The first term on the right-hand side of the last equality is nonnegative as it can be viewed as a sum of Kullback–Leibler divergences; see Proposition 2.10. Due to (2.49), it is also finite. Thus, (2.52) converges a.s. to the nonnegative value $\mathcal{H}_\infty\left({}^{\llcorner o}f||\Theta\right)$. The posterior pdf remains unchanged if we subtract $t\min_{\Theta\in\text{supp}[f(\Theta)]\cap\Theta^*}\mathcal{H}_\infty\left({}^{\llcorner o}f||\Theta\right)$ from the exponent of its non-normalized version (2.51). Then, the exponent contains ($-t\times$ an asymptotically nonnegative factor). Thus, the posterior pdf $f(\Theta|\mathcal{P}_{a^*_\infty})$ may be asymptotically nonzero on minimizing arguments (2.50) only. □

Remark(s) 2.12

1. *The entropy rate can be seen as an extension of the Kullback–Leibler divergence (2.25) that covers well asymptotic and controlled cases. It coincides with the Kullback–Leibler divergence in a range of particular cases.*
2. *The assumption (2.49) can be weakened. It is, however, intuitively acceptable. It excludes parameterized models, which assign no confidence to data generated by the system with a nonzero probability, and vice versa.*
3. *The Bayesian estimation chooses among candidates $f(\Delta_t|a_t,\mathcal{P}_{a^*_t},\Theta)$, $\Theta\in\Theta^*$, the pdf that minimizes the asymptotic entropy rate from the objective pdf ${}^{\llcorner o}f(\Delta_t|a_t,\mathcal{P}_{a^*_t})$. In other words, a* best projection *of the objective pdf to the considered parameterized models is asymptotically found. The prior pdf can be interpreted as a prior belief assigned to the individual parameters $\Theta\in\Theta^*$ that the corresponding parameterized model is the best projection of the objective pdf [82]: not knowing the reality we do not know the best projection we finally arrive at.*
4. *The posterior pdf concentrates on a point if there is a unique minimizer of the entropy rate. In this case, the model is called* identifiable*. The possibility to identify the model can be influenced by*
   - *the considered class of the parameterized models,*
   - *the decisions chosen, for instance, by the controller used: e.g., the controller generating constant inputs prevents us from learning their dynamic influence on outputs.*
5. *If the objective pdf ${}^{\llcorner o}f(\Delta_t|a_t,\mathcal{P}_{a^*_t})$ coincides with $f(\Delta_t|a_t,\mathcal{P}_{a^*_t},\Theta)$ for some $\Theta={}^{\llcorner o}\Theta$ with $f\left({}^{\llcorner o}\Theta\right)>0$ then ${}^{\llcorner o}\Theta$ is in the support of the asymptotic posterior pdf $f(\Theta|\mathcal{P}_{a^*_\infty})$. If, moreover, the model is identifiable, then the objective pdf is asymptotically identified by the adopted Bayesian approach. This fact can be expressed in a more appealing form:*
   The Bayesian estimate is consistent whenever there is a consistent estimator.
6. *Often, a similar analysis is performed by measuring the distance of parameterized models to the empirical pdf of data [83]. It gives similar answers if the empirical pdf converges to the objective pdf. Moreover, it provides hints of how to approximate the posterior pdf [84]; see also Section 3.4. On the other hand, the known conditions of such convergence are more restrictive. For instance, an analysis of the controlled case is much harder.*

Problem 2.1 (How to unify statistics?) *Asymptotic analysis and the finite-data-oriented Bayesian approach are often perceived in an antagonistic way. Their harmonized use still waits for its full exploitation.*

3

Approximate and feasible learning

The operator support we deal with relies on the ability to describe different operating modes. The finite probabilistic mixture [49] is the black-box model we use for this purpose. It is a convex combination of unimodal pdfs called *component*s; see Agreement 5.4. The set of mixtures among which a specific model of a specific process is searched for is parameterized by an unknown finite-dimensional parameter Θ. Formally, it can be estimated using Proposition 2.13. In the considered high-dimensional, data-intensive applications, the formal solution is useless as the exact likelihood function cannot be practically handled. The inevitable approximate mixture estimation (see Section 6.5) relies on our ability to solve estimation and prediction tasks for variety of parameterized components. This preliminary task is addressed here with the aim of preparing common tools used for the mixture estimation and consequently for estimation of its normal, Markov-chain and uniform variants.

Specifically, estimation with forgetting allows us to track slow changes of parameters and thus to make the advisory system adaptive; see Section 3.1.

The estimation within an *exponential family* (EF) of parameterized models is recalled in Section 3.2. It covers the majority of components whose exact estimation is feasible. Its most important special instances — *normal parameterized models* and *Markov-chain parameterized models* — are treated in detail Chapters 8 and 10.

The the chain rule allows us to decompose any parameterized component

$$f(d_t|d(t-1),\Theta) = \prod_{i=1}^{\mathring{d}} f(d_{i;t}|d_{i+1;t},\ldots,d_{\mathring{d};t},d(t-1),\Theta). \qquad (3.1)$$

The individual pdfs describing scalars $d_{i;t}$ are called *factors*; see Agreement 5.4. Starting in Chapter 5, factor-based modelling will be systematically used as it describes data records with mixed — discrete and continuous — entries. Moreover, factors reveal a fine structure of dependencies that spares parameters and contributes significantly to identifiability; cf. Remark 4, 2.12. Basic principles of their structure estimation are outlined in Section 3.3.

The descriptive power of a finite mixture depends strongly on the richness of the set of parameterized factors we are able to handle efficiently. This makes us to present in Section 3.4 so-called equivalence approach to parameter estimation [84]. It provides a promising tool for handling nonlinear and/or non-normal factors. At the same time, it points to an important but neglected problem of approximate recursive Bayesian estimation.

3.1 Estimation with forgetting

The advisory system we are constructing relies on the validity of the model learned. We have to grasp all significant time-varying relationships. At the same time, we cannot exclude slow changes caused, for instance, by aging of the maintained system. Thus, it is reasonable to inspect the case of *slowly varying parameters* as a widely met intermediate case bridging estimation and filtering. It admits time variations of Θ_t, but it assumes that $\Theta_{t+1} \approx \Theta_t$ as the only information related to missing time evolution model. This incompletely formulated filtering problem is mostly addressed by various forgetting techniques, e.g., [85, 86, 87].

We summarize here an approach called *stabilized forgetting*, [88]. It is based on a flexible problem formulation of this formally ill-posed problem.

Let $f(\Theta_{t+1} = \Theta|d(t))$ be the pdf that assumes no parameter changes happened after measuring d_t and before processing d_{t+1}, i.e., $\Theta_{t+1} = \Theta_t$. Let ${}^{\lfloor A}f(\Theta_{t+1} = \Theta|d(t))$ be an *alternative pdf* that describes parameters after expected changes within time interval $(t, t+1)$.

Let $\lambda \in [0,1]$ be the probability that the "correct" unknown pdf $f(\Theta_{t+1} = \Theta|d(t))$ has the best projection equal to the former pdf and $1-\lambda$ the probability that the latter one is relevant. We are searching for the best compromise $\hat{f}(\Theta_{t+1} = \Theta|d(t))$ minimizing its expected KL divergence to the unknown pdf $f(\Theta_{t+1} = \Theta|d(t))$. Solution of this decision-making task is described by the following proposition.

Proposition 3.1 (Geometric representation of a pair of pdfs) *Let an unknown pdf $f \in f^* \equiv \{f_1, f_2\}$ be equal to f_1 with a probability $\lambda \in \lambda^* \equiv [0,1]$ and equal to f_2 with the complementary probability $1-\lambda$. The pdfs f_1, f_2 are supposed to have a common support x^*. Then, the pdf*

$$\hat{f}(x) \propto [f_1(x)]^{\lambda}[f_2(x)]^{1-\lambda}, \ x \in x^* \quad \textit{is the estimate of } f(\cdot) \textit{ that} \tag{3.2}$$

- *uses the experience $\mathcal{P}_{\hat{f}^*} \equiv \{\lambda, f_1, f_2\}$ and*
- *minimizes the expected KL divergence (2.25), i.e., the functional*

$$\mathcal{E}\left[\mathcal{D}\left(\hat{f}\middle|\middle| f\right)\right] \equiv \lambda \mathcal{D}\left(\hat{f}\middle|\middle| f_1\right) + (1-\lambda)\mathcal{D}\left(\hat{f}\middle|\middle| f_2\right). \tag{3.3}$$

The reached minimum is

$$\omega(\lambda) \equiv \mathcal{E}\left[\mathcal{D}\left(\hat{f}\middle|\middle| f\right)\right] = -\ln \int f_1^{\lambda}(x) f_2^{1-\lambda}(x)\, dx. \tag{3.4}$$

If $f_1 \neq f_2$, the function $\omega(\lambda)$ reaches its maximum on $(0,1)$.

Proof. The minimized functional (3.3) can be rewritten into the form

$$\mathcal{D}\left(\hat{f}(x)\middle|\middle|\frac{[f_1(x)]^{\lambda} f_2(x)]^{1-\lambda}}{\int f_1^{\lambda}(\tilde{x}) f_2^{1-\lambda}(\tilde{x})\, d\tilde{x}}\right) - \ln \int f_1^{\lambda}(x) f_2^{1-\lambda}(x)\, dx.$$

This form and properties of the KL divergence imply the form of the minimizer (3.2) as well as the attained minimum value (3.4).

The last statement of the proposition is implied by the assumption $f_1 \neq f_2$, closeness of the set λ^*, and continuity of $\omega(\lambda)$ guaranteed by the common support of f_1, f_2. Non-negativity of the KL divergence and obvious equalities $\omega(0) = \omega(1) = 0$ imply that the extreme in (0,1) must be the maximum. □

The direct application of Proposition 3.1 provides the best correction of the posterior pdf to expected changes of parameters

$$\hat{f}(\Theta_{t+1} = \Theta_t = \Theta | d(t)) \propto [f(\Theta_{t+1} = \Theta | d(t))]^{\lambda} \left[{}^{\llcorner A}f(\Theta_{t+1} = \Theta | d(t))\right]^{1-\lambda}. \tag{3.5}$$

By using this formula, we approximate the time-updating step (2.39) in filtering without explicitly specifying the model of time evolution (2.35).

Algorithm 3.1 (Stabilized forgetting)
Initial (offline) mode

- *Specify the prior pdf $f(\Theta_1 = \Theta) \equiv f(\Theta_1 = \Theta | d(0))$ corresponding to the treated parameterized model $f(\Delta_t | a_t, d(t-1), \Theta_t)$.*
- *Select the probability $\lambda \in [0,1]$ that parameters do not change.*

Sequential (online) mode, *running for $t \in t^*$,*

1. *Collect the newest data d_t.*
2. *Perform data updating*

$$f(\Theta_t = \Theta | d(t)) \propto f(\Delta_t | a_t, d(t-1), \Theta) f(\Theta_t = \Theta | d(t-1)).$$

3. *Select or update the alternative pdf ${}^{\llcorner A}f(\Theta_{t+1} = \Theta | d(t))$.*
4. *Approximate* <u>time updating</u> *(*forget*)*

$$f(\Theta_{t+1} = \Theta | d(t)) \propto [f(\Theta_{t+1} = \Theta_t = \Theta | d(t))]^{\lambda} \left[{}^{\llcorner A}f(\Theta_{t+1} = \Theta | d(t))\right]^{1-\lambda}.$$

Remark(s) 3.1

1. *The pdf (3.5) after the time updating is a compromise between the posterior pdf obtained under the hypothesis that Θ_t is time invariant and an externally supplied alternative ${}^{\lfloor A}f$. The closer λ is to unity, the slower changes are expected, i.e., the higher weight the posterior pdf corresponding to the time invariant case gets.*
2. *The forgetting operation (3.5) preserves the basic property of time updating: the posterior pdf on parameters propagates without obtaining any new measured information.*
3. *Let us assume that ${}^{\lfloor A}f \propto 1$ and $\lambda < 1$. Then, $f(\Theta_{t+1} = \Theta|d(t)) \propto [f(\Theta_{t+1} = \Theta_t = \Theta|d(t))]^{\lambda} \equiv [f(\Delta_t|a_t, d(t-1))f(\Theta|d(t-1))]^{\lambda}$, i.e., the pdf after time updating is a* flattened version *of the pdf obtained after data updating. It is intuitively appealing as our uncertainty about parameters can hardly decrease without knowing a good time evolution model (2.35) and with no new information processed.*
4. *It is instructive to inspect the influence of forgetting on old data built in through the observation model. The older data are, the stronger flattening is applied to the values of the corresponding parameterized model. Consequently, the older data influence the estimation results less than new ones. Data are gradually "forgotten". This explains why the probability λ is called the* forgetting factor*.*
5. *The alternative pdf ${}^{\lfloor A}f(\cdot)$ expresses our belief where the parameters might move within the time interval $(t, t+1)$ while we have no new observable information. Often, the pessimistic uniform alternative pdf ($\propto 1$) has been used. This special case of stabilized forgetting is called* exponential forgetting. *It allows us to follow relatively fast parameter changes but it forgets the accumulated information with an often too high exponential rate. For this reason, it is worth preserving what we feel as a guaranteed information. The prior pdf $f(\Theta_{t+1} = \Theta)$ is a typical, reasonably conservative, choice of the alternative pdf ${}^{\lfloor A}f(\Theta_{t+1} = \Theta|d(t))$.*
6. *The nontrivial alternative pdf prevents us to forget the "guaranteed" information as it is always incorporated after flattening (exponential forgetting). This stabilizes whole learning and reflects very positively in its numerical implementations. Without this, the posterior pdf may become too flat whenever the information brought by new data is poor. Note that lack of information brought by new data is more rule than exception. It is true especially in the so-called regulation problem [89]. In it, the controller tries to make the closed control loop as quiet as possible; it tries to suppress any new information brought by data.*
7. *The forgetting factor λ can be either taken as a tuning knob or estimated. The predictive pdf parameterized by it, however, depends on it in a very complex way so that a partitioned estimation has to be applied when its posterior pdf is estimated on a prespecified grid [90].*

Alternatively, the forgetting factor can be chosen in a pessimistic way as the maximizer of the reached minima; see Proposition 3.2.

8. *The practical importance of this particular case of estimating slowly varying parameters cannot be overstressed: the vast majority of adaptive systems rely on a version of forgetting.*

3.2 Exponential family

A majority of the factors we deal with are taken from exponential family.

Agreement 3.1 (Exponential family) *The ith parameterized factor in (3.1) belongs to the dynamic exponential family iff it can be written in the form*

$$\begin{aligned} f(d_{i;t}|d_{i+1;t},\ldots,d_{\mathring{d};t},d(t-1),\Theta) &\equiv f(d_{i;t}|\psi_{i;t},\Theta) \\ &\equiv A(\Theta)\exp[\langle B(\Psi_{i;t}),C(\Theta)\rangle + D(\Psi_{i;t})], \quad \textit{where} \end{aligned} \tag{3.6}$$

$\psi_{i;t}$ *is the finite-dimensional* regression vector *determined by* $d_{i+1;t},\ldots,d_{\mathring{d};t}$ *and* $a_t, d(t-1)$,

$\Psi'_{i;t} \equiv [d_{i;t},\psi'_{i;t}]$ *is a finite-dimensional* data vector, *whose values can be updated recursively according to a known rule* $(\Psi_{i;t-1},d_t)^* \to \Psi^*_{i;t}$, *where*

$'$ *denotes* transposition

$\langle\cdot,\cdot\rangle$ *is the functional, linear in the first argument. Within this text it is defined*

$$\langle x,y\rangle = \begin{cases} x'y & \textit{if } x,\ y \textit{ are vectors} \\ \mathrm{tr}[xy] & \textit{if } x,\ y \textit{ are matrices}, \ \mathrm{tr} \textit{ is trace} \\ \sum_{i\in i^*} x_i y_i & \textit{if } x,\ y \textit{ are arrays with a multi-index } i, \end{cases} \tag{3.7}$$

$A(\cdot)$ *is a nonnegative scalar function defined on* Θ^*,

$B(\cdot)$, $C(\cdot)$ *are either vector or matrix functions of compatible, finite and fixed dimensions; they are defined on respective arguments in* $\Psi^*_{i;t}$ *and* Θ^*,

$D(\cdot)$ *is a nonnegative scalar function defined on* Ψ^*_i.

Remark(s) 3.2

1. *Our definition of the exponential family contains the nonstandard requirement on the recursive updating of the data vector* $\Psi_{i;t}$. *It is practically important for dynamic cases we deal with.*
2. *Notice that equality is used in (3.6). The normalization of this pdf must not spoil this form. It makes the allowed form rather restrictive. In the dynamic case with a nonempty regression vector* ψ, *normal (Gaussian) linear-in-parameters factors and Markov chains almost cover exponential family.*
3. *The scalar function* D *entering (3.6) does not influence estimation. The factor* $\exp(D(\Psi_{i;t}))$ *enters prediction unchanged. We use this property only in Section 8.1.6. Otherwise,* $D(\Psi_{i;t})$ *is omitted.*

4. *In the rest of this chapter, we consider a factor related to a predicted data entry $d_{i;t}$ with a fixed i and the index i is dropped.*

The practical significance of the exponential family becomes obvious when we apply Proposition 2.14, which describes the corresponding estimation and prediction.

Proposition 3.2 (Estimation and prediction in exponential family) *Let natural conditions of decision making, Requirement 2.5, be met with the time-invariant parameter $\Theta_t = \Theta \in \Theta^*$. Let the parameterized model belong to the exponential family (3.6). Then, the* predictive pdf, *the outer model of the system, is given by the formula*

$$f(\Delta_t|a_t, d(t-1)) = \frac{\mathcal{I}(V_{t-1} + B(\Psi_t), \nu_{t-1} + 1)}{\mathcal{I}(V_{t-1}, \nu_{t-1})} \exp[D(\Psi_t)], \tag{3.8}$$

$$V_t = V_{t-1} + B(\Psi_t),\ V_0 = 0; \quad \nu_t = \nu_{t-1} + 1,\ \nu_0 = 0, \tag{3.9}$$

$$\mathcal{I}(V, \nu) = \int A^{\nu}(\Theta) \exp[\langle V, C(\Theta)\rangle] f(\Theta) \chi_{\Theta^*}(\Theta)\, d\Theta, \tag{3.10}$$

where $f(\Theta)$ is a prior pdf. Its support is restricted by the indicator $\chi_{\Theta^*}(\cdot)$ of the set Θ^* .

The Bayesian parameter estimate (posterior pdf) is

$$f(\Theta|d(t)) = \frac{A^{\nu_t}(\Theta) \exp[\langle V_t, C(\Theta)\rangle] \chi_{\Theta^*}(\Theta) f(\Theta)}{\mathcal{I}(V_t, \nu_t)}, \tag{3.11}$$

i.e., the likelihood function is

$$\mathcal{L}(\Theta, d(t)) \equiv \mathcal{L}(\Theta, V_t, \nu_t) = A^{\nu_t}(\Theta) \exp[\langle V_t, C(\Theta)\rangle]. \tag{3.12}$$

Let us consider the conjugate prior *pdf*

$$f(\Theta) \propto A^{\nu_0}(\Theta) \exp[\langle V_0, C(\Theta)\rangle] \chi_{\Theta^*}(\Theta), \tag{3.13}$$

determined by the "prior statistics" V_0, ν_0. Then, the prediction and estimation formulas (3.8) and (3.11) are valid if

- *V_0, ν_0 replace the zero initial conditions in (3.9),*
- *the indicator $\chi_{\Theta^*}(\cdot)$ is formally used as the prior pdf.*

Let us allow slow parameter changes with the forgetting factor $\lambda \in [0, 1]$ and the alternative pdf given in the conjugate form determined by the pair of sufficient statistics ${}^{\lfloor A}V_t$, ${}^{\lfloor A}\nu_t$. Then, the prediction and estimation formulas remain unchanged with statistics evolving according to the recursion

$$\begin{aligned} V_t &= \lambda(V_{t-1} + B(\Psi_t)) + (1-\lambda)\, {}^{\lfloor A}V_t,\ V_0 \textit{ given},\\ \nu_t &= \lambda(\nu_{t-1} + 1) + (1-\lambda)\, {}^{\lfloor A}\nu_t,\quad \nu_0 \textit{ given}. \end{aligned} \tag{3.14}$$

Remark(s) 3.3

1. *Estimation and prediction within the exponential family is very simple, especially with the conjugate prior pdf. The updating of functions (pdfs) converts into the algebraic recursive updating of the finite-dimensional sufficient statistic consisting of V_t and the* sample counter ν_t. *Moreover, a single type of the normalization integral $\mathcal{I}(V,\nu)$ has to be evaluated. The need to have the complete recursion explains the requirement for a possibility to update Ψ_t recursively; see Agreement 3.1.*
2. *An inspection whether there is a wider set of parameterized models with advantageous properties of the exponential family opens just a narrow space [91]. Essentially, the exponential family coincides with all parameterized models that are sufficiently smooth functions of Θ and with supports independent of Θ. Uniform distribution with unknown constant boundaries represents one of a few feasible examples of pdfs out of the exponential family.*
3. *The class of models that lead to a finite-dimensional characterization of pdfs occurring in filtering is even more restrictive. Its discussion can be found in [92].*

3.3 Structure estimation in the nested exponential family

Often, specification of the parameterized model can be done by selecting discrete-valued pointers to a finite set of alternative parameterized models. These pointers define *model structure*; see [93] for a detailed discussion. Formally, estimation of the best structure can be performed fully within a Bayesian set-up by assigning the prior pf to competitive structures and computing the posterior one. The problem becomes specific and difficult due to the usual extremely large cardinality of the set of possible structures. Then, special measures have to be taken. Often, we use nested models with sufficient statistics contained in a single one corresponding to the *richest structure.*

Proposition 3.3 (Nesting in the exponential family) *Let the parameterized model with the* richest structure *belong to the exponential family (3.6) $f(d|\psi_r,\Theta_r) = A_r(\Theta_r)\exp[\langle B_r(\Psi_r), C_r(\Theta_r)\rangle]$. Let us consider another model $f(d|\psi,\Theta) = A(\Theta)\exp[\langle B(\Psi), C(\Theta)\rangle]$ describing the same data $d(\mathring{t})$.*

Let N be a time invariant linear nesting operator *such that*

$$B(\Psi(d(t))) \equiv N[B_r(\Psi_r(d(t)))].$$

Let us consider a pair of conjugate prior pdfs given by statistics $V_{r;0}, \nu_{r;0}$ and $V_0 \equiv N[V_{r;0}], \nu_0$. Then, the V statistics of the posterior pdfs of both models are *related by the nesting mapping $V_{\mathring{t}} = N[V_{r;\mathring{t}}]$. The predictive pdf of the nested model is*

$$f(d(\mathring{t})) \propto \frac{\mathcal{I}(V_{\mathring{t}}, \nu_t)}{\mathcal{I}(V_0, \nu_0)} = \frac{\mathcal{I}\left(N[V_{r;\mathring{t}}], \nu_t\right)}{\mathcal{I}\left(N[V_{r;0}], \nu_0\right)}. \tag{3.15}$$

Proof. It is implied directly by Proposition 3.2 and linearity of N. □

For structures $s \in s^*$, that can be obtained by a nesting operator from the richest structure, we are able to evaluate the pdf $f(d(\mathring{t})|s)$ for any specific structure s. Let $f(s)$ be prior pf on s^*, then the Bayes rule, Proposition 2.4, gives us <u>formally</u> the posterior pf on structures, i.e., full information needed for its point estimation.

If ψ_r is the *richest regression vector* then the number $\mathring{s}$ of possible competitive structures nested in it is $2^{\mathring{\psi}_r}$. This is mostly an excessive number that prevents us from evaluating completely posterior probabilities of nested structures. Instead, we are searching for the *maximum a posteriori probability* (MAP) estimate. Of course, we can inform also about highly probable structures met during the search for MAP estimate. The following conceptual algorithm is used for it.

Algorithm 3.2 (MAP estimate of a factor structure)
Do while the prespecified number of restarts is not exceeded.

```
1. Select an initial guess of the structure.
2. Do while the value of the likelihood increases and a prespecified number of
   searches is not exceeded.
   a) Make a full search for the best structure within a “neighborhood” of
      the current guess of the structure and find the structure maximizing
      the posterior likelihood $f(d(\mathring{t})|s)f(s)$ within it.
   b) Take the maximizer as a new guess of the structure.
```

The algorithm is “parameterized” by the following:

- The generator of the initial guesses: a specified number of random choices, empty, richest and user-specified regression vectors are mostly used. Specification of the number of random draws has to balance computational demands and probability that global maximum is found. A solution of this problem exploiting the Bayesian sequential stopping rule was developed for this purpose [94].
- Definition of neighborhood: a good choice depends on the specific parameterized model considered. A detailed solution tailored to normal regression models is described in [95].

3.4 Equivalence approach

The exponential family and special uniform distributions provide us with a basic supply of factors. Sometimes, however, we are forced to go out of this family. Under natural conditions of decision making (2.36), the generalized Bayesian estimation, Proposition 2.14, updates the posterior pdfs according to the Bayes rule (2.43)

$$f(\Theta|d(t)) = \frac{f(\Delta_t|a_t, d(t-1), \Theta) f(\Theta|d(t-1))}{f(\Delta_t|a_t, d(t-1))}, \ t \in t^*.$$

The complexity of these pdfs increases quickly with an increasing number of data, with increasing t. The exponential family (3.6) is essentially the only exception from this rule. This section tries to cope with *recursive estimation applicable out of the exponential family*. The equivalence approach presented in this section allows us to find a well-justified approximation even in these cases.

3.4.1 Recursively feasible representation

The limited capabilities of computers call for a reduced representation of the propagated posterior pdfs. It is a difficult task as the posterior pdfs concentrate quickly on a very narrow support at a priori unknown position in Θ^*. Thus, a representation on a sufficiently fine grid that does not miss the final position becomes soon computationally prohibitive. The way out has been elaborated in a sequence of papers and summarized in [84]. Here, we just outline the essence of this *equivalence approach*.

Proposition 3.4 (Equivalence-preserving mapping) *Let $f^*(\Theta|d(t-1))$ be a set of posterior pdfs $f(\Theta|d(t-1))$ with a common, time, data, and parameter invariant support. Let the mapping*

$$\mathcal{G}_{t-1} : \ f^*(\Theta|d(t-1)) \to g^*_{t-1} \tag{3.16}$$

assign to each pdf $f(\Theta|d(t-1))$ from $f^*(\Theta|d(t-1))$ a finite-dimensional vector statistics $g_{t-1} \equiv g(d(t-1))$ representing it. Then, the value of g_{t-1} can be exactly recursively updated *using only its previous value and the current parameterized model $f(\Delta_t|a_t, d(t-1), \Theta)$ iff $\mathcal{G}_t$ is a* time-invariant linear mapping $\mathcal{G}_t \equiv \mathcal{G}$, $t \in t^*$, *acting on logarithms of the pdfs involved. The logarithmic pdfs are treated as functions of Θ.*

$\mathcal{G}_t$ has to map Θ-independent elements to zero.

Proof. To demonstrate necessity is rather difficult, and the interested reader is referred to [96, 97]. To show that the conditions on $\mathcal{G}_t \equiv \mathcal{G}$, $t \in t^*$, are sufficient is simple and instructive. They become obvious if we apply $\mathcal{G}$ to the logarithmic version of the Bayes rule (2.43) and use both time invariance and linearity of G. The normalizing term $\ln(f(\Delta_t|a_t, d(t-1))$ is independent of Θ and as such mapped to zero. The recursion for g_t is then

$$\begin{aligned} g_t &= \mathcal{G}\left[\ln\left(f(\Delta_t|a_t, d(t-1), \Theta)\right)\right] + g_{t-1}, \quad \text{with} \\ g_0 &= \mathcal{G}(\ln(f(\Theta))) \equiv \mathcal{G}(\ln(\text{prior pdf})). \end{aligned} \tag{3.17}$$

□

Note that (3.17) becomes the true recursion if we need not store complete past observed data for evaluating $f(\varDelta_t|a_t, d(t-1), \varTheta)$. Thus, similarly to the case of the exponential family (see Agreement 3.1) we restrict ourselves to such models. Together with their formal description we introduce other notions useful in the subsequent discussion.

Agreement 3.2 (Finite memory; Riezs representation) *The parameterized model is said to have* finite memory *iff*

$$f(\varDelta_t|a_t, d(t-1), \varTheta) \equiv M(\varTheta, \varPsi_t), \tag{3.18}$$

where $M(\cdot, \cdot)$ is a known function of $\varTheta$ and of a finite-dimensional data vector $\varPsi_t$ that can be updated recursively

$$(\varPsi_{t-1}, d_t)^* \to \varPsi_t^*.$$

Let $\{\varPsi_\tau\}_{\tau=1}^t$ be measured data vectors. Then,

$$f_t(\varPsi) \equiv \frac{1}{t} \sum_{\tau=1}^{t} \delta(\varPsi - \varPsi_\tau), \;\; \varPsi \in \varPsi^* \equiv \bigcup_{t \in t^*} \varPsi_t^* \tag{3.19}$$

is the formal empirical pdf *of $\varPsi$. The Dirac delta function $\delta(\cdot)$ used is the linear functional assigning to (reasonable) functions $B(\varPsi)$ their values at zero argument. It has a Riezs integral representation, [98],*

$$\int B(\varPsi)\delta(\varPsi)\, d\varPsi = B(0), \tag{3.20}$$

where $\delta(\varPsi)$should be formally taken as a generalized function [99].

We assume that the same representation exists for the linear mapping $\mathcal{G}$ introduced in Proposition 3.4, i.e., there is a, possibly generalized, vector function $G(\varTheta)$ such that

$$\mathcal{G}(C) \equiv \int C(\varTheta)G(\varTheta)\, d\varTheta \tag{3.21}$$

for any considered function $C: \varTheta^* \to (-\infty, \infty)$.

Assuming (3.18), using the empirical pdf (3.19) and Riezs representation of G (3.21), we see that the representation g_t of the posterior pdf $f(\varTheta|d(t))$ can be written in the form

$$g_t = t \int \left[\int \ln[M(\varTheta, \varPsi)]G(\varTheta)\, d\varTheta \right] f_t(\varPsi)\, d\varPsi + g_0. \tag{3.22}$$

The integral in brackets defines the vector function

$$h(\Psi) \equiv \int \ln[M(\Theta, \Psi)] G(\Theta) \, d\Theta. \tag{3.23}$$

With it, we get the equivalent form of (3.22)

$$g_t = t \int h(\Psi) f_t(\Psi) \, d\Psi + g_0. \tag{3.24}$$

Recall that left-hand side of (3.24) is known as it can be updated recursively according to (3.17), which has the equivalent Riezs representation

$$\begin{aligned} g_t &\underbrace{=}_{(3.17),(3.18),(3.21)} \int \ln[M(\Theta, \Psi_t)] G(\Theta) \, d\Theta + g_{t-1} \underbrace{\equiv}_{(3.23)} h(\Psi_t) + g_{t-1} \\ g_0 &= \mathcal{G}(\ln(f(\Theta))). \end{aligned} \tag{3.25}$$

Also, for the chosen parameterized model (3.18) and functions $G(\Theta)$ representing the admissible projections to g_t^*, the vector function $h(\Psi)$ (3.23) is known as well as its values in measured data vectors Ψ_t.

3.4.2 Approximation as a point estimation

The posterior pdf we are interested in can be expressed in terms of the empirical pdf $f_t(\Psi)$ as follows, cf. the similar transformation in the proof of Proposition 2.15,

$$f(\Theta|d(t)) \equiv f(\Theta|f_t(\cdot)) \propto f(\Theta) \exp\left[t \int \ln[M(\Theta, \Psi)] f_t(\Psi) \, d\Psi \right]. \tag{3.26}$$

The empirical pdf is unknown to us as we are not able to store it and map it without information loss on a finite-dimensional statistic. Thus, this posterior pdf is unknown, too. As we want to estimate it, we are facing the decision-making problem.

The selection of a suitable estimate $\hat{f}(\Theta|g_t)$($\equiv$ a pdf acting on Θ^*) fits our general decision-making framework as follows.

- $\mathcal{Q} \equiv (\mathcal{P}_{a^*}, a, \mathcal{F}_{a^*}) \equiv \left(g_t, \hat{f}(\Theta|g_t), f(\Theta|f_t(\cdot)) \right) \equiv$ (stored statistic, posterior-pdf estimate, posterior pdf given by the unknown empirical pdf $f_t(\Psi)$, $\Psi \in \Psi^*$).
- Admissible decision rules are of the form
$$\mathcal{R}_t : g_t^* \to \hat{f}^* \equiv \left\{ \hat{f}(\Theta) : \int \hat{f}(\Theta) \, d\Theta = 1, \; \hat{f}(\Theta) \geq 0, \; \forall \Theta \in \Theta^* \right\}.$$
Note that g_0 is a known fixed representation of the known prior pdf $f(\Theta)$ and as such neither of them is explicitly mentioned as a part of the domain of $\mathcal{R}_t$.
- The loss function is selected as the KL divergence (2.25) $\mathcal{D}(\hat{f}(\Theta)||f(\Theta|f_t))$.

Formally, the optimal point estimate of the posterior pdf is

$$\hat{f}(\Theta|g_t) \in \operatorname{Arg}\min_{\hat{f}\in\hat{f}^*} \mathcal{E}\left[\mathcal{D}\left(\hat{f}(\Theta)\middle|\middle| f(\Theta|f_t)\right)\middle| g_t\right] \tag{3.27}$$

$$= \operatorname{Arg}\min_{\hat{f}\in\hat{f}^*}\left[\mathcal{D}\left(\hat{f}(\Theta)||f(\Theta)\right) - t\int_{\Theta^*}\int_{\Psi^*}\hat{f}(\Theta)\ln[M(\Theta,\Psi)]\mathcal{E}[f_t(\Psi)|g_t]\,d\Psi\,d\Theta\right].$$

The conditional expectation $\mathcal{E}[\cdot|g_t]$ is taken with respect to the uncertain empirical pdf $f_t(\Psi)$, $\Psi \in \Psi^*$.

The minimizing argument can be found explicitly as

$$\hat{f}(\Theta|g_t) \propto f(\Theta)\exp\left\{t\int \ln[M(\Theta,\Psi)]\mathcal{E}[f_t(\Psi)|g_t]\,d\Psi\right\}. \tag{3.28}$$

3.4.3 Specification of $\mathcal{E}[f_t(\Psi)|g_t]$

The estimate (3.28) depends only on the conditional expectation

$$\hat{f}_t(\Psi) \equiv \mathcal{E}\left[f_t(\Psi)|\, g_t\right],\ \Psi \in \Psi^*. \tag{3.29}$$

Its choice and the resulting estimation are described in this subsection.

First we notice that the empirical pdf $f_t(\Psi)$ obeys the identity (3.24) that is linear in it. Consequently, its conditional expectation $\hat{f}_t(\Psi) = \mathcal{E}[f_t(\Psi)|g_t]$ has to fulfill it, too (recall, g_0 is known)

$$g_t = \int h(\Psi)\mathcal{E}\left[f_t(\Psi)|\, g_t\right]\,d\Psi + g_0 \underbrace{\Leftrightarrow}_{(3.29)} g_t - g_0 = \int h(\Psi)\hat{f}_t(\Psi)\,d\Psi. \tag{3.30}$$

Obviously, $\hat{f}_t(\Psi)$ has to be pdf. The identity (3.30) is the only objective knowledge we have about it. Thus, it is reasonable to select it as close as possible to the pdf describing prior information about $f_t(\Psi)$ while respecting the constraint (3.30).

Marginalization and the chain rule, Proposition 2.4, together with the assumption (3.18) imply that the prior distribution on Ψ is

$$f(\Psi) = \int f(\Psi|\Theta)f(\Theta)\,d\Theta = f(\psi)\int M(\Theta,\Psi)f(\Theta)\,d\Theta.$$

There the pdf $f(\psi)$ describes our prior information on possible regression vectors. It can be taken as a flat pdf with its support on ψ^*. The pdf $M(\Theta,\Psi)$ is the model in question and $f(\Theta)$ is the chosen prior pdf on its parameters. As usual, we measure the distance of $\hat{f}_t(\Psi)$ to $f(\Psi)$ with the help of the KL divergence. Taking into account the constraint (3.30), we minimize

$$\min_{\hat{f}(\Psi)}\int \hat{f}(\Psi)\left[\ln\left(\frac{\hat{f}(\Psi)}{f(\Psi)}\right) - v_t'h(\Psi)\right]\,d\Psi \tag{3.31}$$

$$= \min_{\hat{f}(\Psi)}\mathcal{D}\left(\hat{f}(\Psi)\middle|\middle|\frac{f(\Psi)\exp\left(v_t'h(\Psi)\right)}{\int f(\tilde{\Psi})\exp\left(v_t'h(\tilde{\Psi})\right)\,d\tilde{\Psi}}\right) - \ln\left(\int f(\Psi)\exp\left(v_t'h(\Psi)\right)\,d\Psi\right),$$

where the vector v_t of Lagrangian multipliers is to be selected so that (3.30) is met. This requirement, the second form of the minimized functional and properties of the KL divergence give the optimal $\hat{f}(\Psi)$

$$\hat{f}(\Psi) = \frac{f(\Psi)\exp\left(v_t' h(\Psi)\right)}{\int f(\Psi)\exp\left(v_t' h(\Psi)\right)\, d\Psi}, \tag{3.32}$$

with the vector v_t solving the equation

$$g_t - g_0 = \int h(\Psi)\hat{f}_t(\Psi)\, d\Psi = \frac{\partial}{\partial v_t}\ln\left(\int f(\Psi)\exp\left(v_t' h(\Psi)\right)\, d\Psi\right).$$

This option completes the overall algorithm.

Algorithm 3.3 (Recursive equivalence estimation)
Initial (offline) mode

- *Select the parameterized model* $f(\Delta_t|a_t, d(t-1), \Theta) \equiv M(\Theta, \Psi_t)$.
- *Select the prior pdfs* $f(\Theta)$ *and* $f(\psi)$.
- *Evaluate the prior pdf* $f(\Psi) = f(\psi)\int M(\Theta, \Psi) f(\Theta)\, d\Theta$ *on* Ψ^*.
- *Select the generalized vector function* $G(\Theta)$ *such that* $\int G(\Theta)\, d\Theta = 0$.
- *Prepare the evaluation of the function*

$$h(\Psi) \equiv \int \ln(M(\Theta, \Psi)) G(\Theta)\, d\Theta. \tag{3.33}$$

- *Set* $g_0 = 0$.

Sequential (online) mode, *running for* $t \in t^*$,

1. *Measure data vector* Ψ_t.
2. *Evaluate* $h_t \equiv h(\Psi_t) \equiv \int \ln(M(\Theta, \Psi_t)) G(\Theta)\, d\Theta$.
3. *Update the weights* $g_t = g_{t-1} + h_t$.
4. *Find the vector* v_t *solving the equation* (g_t *is known fixed vector*)

$$g_t = \frac{\partial}{\partial v_t}\ln\left(\int f(\Psi)\exp\left(v_t' h(\Psi)\right)\, d\Psi\right)$$

5. *Exploit the obtained approximation of the posterior pdf*

$$\hat{f}(\Theta|g_t) \propto f(\Theta)\exp\left[\frac{\int \ln(M(\Theta, \Psi)) f(\Psi)\exp\left(v_t' h(\Psi)\right)\, d\Psi}{\int f(\Psi)\exp\left(v_t' h(\Psi)\right)\, d\Psi}\right]. \tag{3.34}$$

Remark(s) 3.4

1. *Zero value of the initial weights* g_0 *reflects that the weights are always determined by the increment* $g_t - g_0$*; see (3.32).*
2. *The name "equivalence approach" stresses the fact that the set of posterior pdfs is reduced to equivalence classes. The pdfs with the same representation* g *cannot be distinguished.*

3. *The required* commutative updating and projecting *of the posterior pdfs is crucial. The recursion for gs is exact and the approximation errors caused by the use of* $\hat{f}(\Theta|g_t)$ *instead of* $f(\Theta|d(t)) \equiv f(\Theta|f_t(\cdot))$ *do not accumulate! Use of a noncommutative projection* $\mathcal{G}_t : f^*(\Theta|d(t)) \to g_t^*$ *is always endangered by a divergence as the estimation described by the Bayes rule can be viewed as a dynamic system evolving* $f(\Theta|d(t))$ *at the stability boundary.*
4. *The indicated integrations represent the computationally most demanding part of the algorithm. They can be performed in offline mode if their results can be efficiently stored (the resulting functions interpolated). The solution of the nonlinear equation for* v_t *is also hard, but is a standard problem.*
5. *We would like to get the exact posterior pdf if the model belongs to the exponential family (3.6). This dictates the choice of the mapping* $\mathcal{G}$ *that should make* $h(\Psi) = [B'(\Psi), 1]'$. *It is sufficient, to introduce the prior initial moments of the vector function* $\tilde{C}(\Theta) \equiv [C'(\Theta), \ln(A(\Theta))]'$

$$\bar{C} \equiv \int \tilde{C}(\Theta) f(\Theta)\, d\Theta,\ \mathcal{C} \equiv \mathrm{cov}\left[\tilde{C}\right] = \int \tilde{C}(\Theta)\tilde{C}'(\Theta) f(\Theta)\, d\Theta - \bar{C}\bar{C}'$$

and define the weighting function

$$G(\Theta) \equiv [\tilde{C}(\Theta) - \bar{C}]'\mathcal{C}^{-1} f(\Theta).$$

Problem 3.1 (How to choose the mapping $\mathcal{G}$?) *The generalized functions G represent the key tuning knobs of the approach. Options leading to discrete versions of the function and/or its derivatives, or* $M(\Theta, \Psi_i)$ *on a grid of* Ψ_i *have been tried with a success, but a deeper insight is needed in order to arrive at a cookbook.*

A possible direction can be found by approximating $M(\Psi, \Theta)$ *by a member from the exponential family and by using the recommendation given in step 5 in Remarks 3.4.*

Problem 3.2 (Application of equivalence approach to mixtures) *This section describes how a correct approximate recursive estimation should look like. As it will be seen in Chapters 6, 8, 10 and 12, we found no practical way how to apply it to mixtures. Any progress in this respect would be invaluable.*

4

Approximate design

Formally, the fully probabilistic design adopted for the design of advising strategies is described by Proposition 2.11. The corresponding functional equation is rarely solvable in the high-dimensional multimodal cases considered. Thus, specific approximation techniques are needed to get feasible strategies. This chapter prepares such techniques that are exploited in Chapters 7 and 9 for the design of advising strategies.

The dynamic design essentially predicts possible behavior of the system interacting with the judged strategy and selects the most favorable one. The design complexity is significantly influenced by the richness of the inspected space. Its reduction is behind the majority of available approximation schemes including those discussed in this chapter.

First, we relate an approximate design with the notion of adaptive systems, Section 4.1. It provides us a common perspective on established approximate designs that look otherwise as a “bag of tricks”. Section 4.2 lists the most common techniques used. It will serve us a reference point in the design body of the text (Chapters 7, 9, 11). The reader is referred to classical references [1, 68, 89, 100] for a detailed presentation of adaptive systems.

Adaptive control has a relatively long history within which substantial experience has been accumulated. It guides designers on how to split this complex decision-making task into a sequence of meaningful and solvable subtasks. This “prototype line” is generalized in Section 4.3.

4.1 Adaptive systems

The ideal solution of the decision-making under uncertainty is described by the combination of Bayesian filtering (see Proposition 2.13) and dynamic programming; see Propositions 2.9 and 2.11. The functional equations describing them are mostly computationally infeasible and are solved approximately.

The involved multivariate functions should be represented in a computer, i.e., in the device that can operate on a high but finite number of values. Thus, a sort of approximation is needed. The global approximation of functions of many variables that we are dealing with is known to be computationally hard. On the other hand, the application of the strategies resulting from the design requires knowledge of the discussed solutions only for the recorded experience. Thus, it is sufficient to know them locally around the actual experience. Such a *local approximation* can be identified with an *adaptive system* [101]. The mixture model we use as well as its quasi-Bayes estimation (see Chapter 6) follow exactly this direction. In this way, we get practically feasible learning.

Below, the localization principles used in the design are discussed.

4.2 Suboptimal design

The design complexity is the key issue addressed repeatedly at various places of this text. At the design stage, the complexity stems from

- richness of the space of the ignorance part of the behavior that has to be inspected for the optimal selection of decisions;
- complexity of models describing relationships of the experience and optional decisions to ignorance part of the behavior.

The suboptimal design tries to reduce the influence of one or both of these sources of complexity. The selected techniques described below are suitable to the design of the adaptive advisory system.

4.2.1 Strategies examining reduced space

Here, we outline common techniques oriented on simplification of the space searched for.

Receding horizon

The reduction of the design horizon is the most direct way to a simplified (suboptimal) design. The reduction obtained by planning just one-step-ahead has been popular for a long time [102]. Dynamic decision-making, however, means that consequences of a decision are encountered far behind the time moment of its application. Consequently, the decision that is optimal when judged from a short-sighted perspective might be quite bad from the long-term viewpoint [68].

This observation has stimulated the search for a compromise between the ideal planning over the whole horizon of interest and short-sighted, locally optimizing strategies.

A little-steps-ahead planning provides just an approximation of the optimal design. Thus, it is reasonable to apply just the initial planned decisions and redesign strategy whenever a new information about the system and its state is processed. This is the essence of the design technique called *receding-horizon strategy*. Let us describe its algorithm in the case of additive loss function (2.31) with the finite-dimensional observed information state x_t and for a prespecified value T of the *receding horizon* $T < \mathring{t}$. Note that the information state is generally composed of the observed state of the system and statistics describing results of learning [103, 104].

Algorithm 4.1 (Receding horizon strategy)
Repeat for $t \in t^* \equiv \{1, \ldots, \mathring{t}\}$,

1. *Find the rules* $\mathcal{R}_t, \ldots, \mathcal{R}_{t+T}$ *approximately minimizing*
$$\mathcal{E}\left[\sum_{\tau=t}^{t+T} z(x_\tau, a_\tau)\middle| \mathcal{P}_{a_t^*}\right]$$
using the available experience $\mathcal{P}_{a_t^*} \equiv x_{t-1}$ *and the outer model of the system, Agreement 2.7.*
2. *Apply* $a_t \equiv \mathcal{R}_t(\mathcal{P}_{a_t^*})$ *for the available* $\mathcal{P}_{a_t^*}$.
3. *Extend the experience by the new data* Δ_t, a_t *and use them in learning — perform filtration or estimation, Propositions 2.13 or 2.14 — so that an improved outer model of the system is obtained.*

Iterations spread in time

It is intuitively clear and practically verified that the receding horizon T guaranteeing good approximation of the optimal design can be too large if the decision horizon $\mathring{t}$ is large. In the case of additive loss, the asymptotic analysis of the design, Proposition 2.12, shows that for $\mathring{t} \to \infty$ the optimal decision is the minimizer in

$$^{\lfloor\infty}\mathcal{V}(x_{t-1}) + {}^{\lfloor\infty}C = \min_{a_t \in a_t^*} \mathcal{E}\left[z(x_t, a_t) + {}^{\lfloor\infty}\mathcal{V}(x_t)\middle| a_t, x_{t-1}\right].$$

This stationary form can be interpreted as one-step-ahead design for the partial loss $z(x_t, a_t)$ increased by the stationary Bellman function $^{\lfloor\infty}\mathcal{V}(x_t)$. Knowing the function $^{\lfloor\infty}\mathcal{V}(x_t)$, the design with receding horizon $T = 1$ is the optimal one. At the same time, dynamic programming has been interpreted as an iterative search for the Bellman function $^{\lfloor\infty}\mathcal{V}(x_t)$, Proposition 2.12. Thus, we can use the Bellman function resulting from an approximate design at time moment $t-1$ as an estimate of $^{\lfloor\infty}\mathcal{V}(x_t)$ at time moment t. Then, we perform the receding-horizon design that has this estimate as its terminal condition, i.e., increases the terminal partial loss. This allows us to use a short design horizon T . This reasoning leads to the following approximate strategy.

Algorithm 4.2 (Strategy with iterations spread in time)
Initial (offline) phase
Select an initial guess of the stationary Bellman function $\mathcal{V}_0(x)$.
Iterative (online) phase *repeated for* $t \in t^* \equiv \{1, \ldots, \mathring{t}\}$

1. *Find the rules* $\mathcal{R}_t, \ldots, \mathcal{R}_{t+T}$ *approximately minimizing*

$$\mathcal{E}\left[\sum_{\tau=t}^{t+T} z(x_\tau, a_\tau) + \mathcal{V}_{t-1}(x_{t+T-1}) \middle| x_{t-1}\right]$$

   using the available experience $\mathcal{P}_{a_t^*}$ *stored in information state* x_{t-1} *and the outer model of the system, Agreement 2.7. Evaluate these rules and achieved minima* $\mathcal{V}_t(x)$ *for all possible states* $x \in x^*$ *in the role of* x_{t-1}.
2. *Take the final Bellman function* $\mathcal{V}_t(x)$ *resulting from this design as an updated approximation of the stationary Bellman function.*
3. *Apply* $a_t \equiv \mathcal{R}_t(x_{t-1})$ *for the measured information state* x_{t-1}.
4. *Extend the experience by the new data* Δ_t, a_t *and use them in learning — perform filtration or estimation, Propositions 2.13 or 2.14 — so that an improved outer model of the system is obtained.*

Predictive strategies

There is a whole set of strategies that drastically reduce the space of inspected behaviors by working only with point predictions of uncertain quantities to be influenced. Sometimes, they deal with set predictions using credibility sets of a very simplified form. Such strategies are usually labelled as *predictive strategies* [89, 105, 106, 107].

Let us consider an additive loss function with receding-horizon T. It serves us an example demonstrating essence of these strategies. For minimization purposes, the predictive strategies use the approximation

$$\mathcal{E}\left[\sum_{\tau=t}^{t+T} z\left(x_\tau, a_\tau\right) \middle| \mathcal{P}_{a_t^*}\right] \approx \sum_{\tau=t}^{t+T} z\left(\tilde{x}_{\tau|\mathcal{P}_{a_t^*}}, a_\tau\right),$$

with $\tilde{x}_{\tau|\mathcal{P}_{a_t^*}}$ obtained from the uncertain future information state x_τ by approximating

$$\Delta_\tau \approx \mathcal{E}[\Delta_\tau|\mathcal{P}_{a_t^*}] \equiv \tilde{\Delta}_{\tau|\mathcal{P}_{a_t^*}}.$$

Obviously, such an approximation may be quite rough when there is a non-negligible uncertainty caused either by lack of knowledge or by noise influence. On the other hand, the gained simplification allows the user to concentrate on respecting hard bounds on actions or other quantities. The practical significance of physical constraints justifies this simplification and explains the popularity of predictive strategies and its permanent use. Papers [108, 109] serve as samples of such applications.

4.2.2 Strategies simplifying models

Let us consider the receding-horizon strategy applied at time t. Then, a substantial degree of the design complexity is caused by the use of predictive pdfs

$\left\{f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*})\right\}_{\tau=t}^{t+T}$ obtained through the Bayesian estimation. They have the form (see Proposition 2.14)

$$f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}) = \int f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \Theta) f(\Theta|\mathcal{P}_{a_\tau^*})\, d\Theta. \tag{4.1}$$

For a relatively short planning horizon T, the predictors (4.1) can be approximated by

$$f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}) \approx \int f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \Theta) \hat{f}_\tau(\Theta)\, d\Theta,\ \tau = t, \ldots, t+T, \tag{4.2}$$

with a simpler pdf $\hat{f}_\tau(\Theta) \approx f(\Theta|\mathcal{P}_{a_\tau^*})$. Below we outline widespread versions of $\hat{f}_\tau(\Theta)$ used.

Supercautious strategy

If the posterior pdf $f(\Theta|\mathcal{P}_{a_t^*})$ is reasonably concentrated then it makes sense to exploit the approximate equality

$$f(\Theta|\mathcal{P}_{a_\tau^*}) \approx f(\Theta|\mathcal{P}_{a_t^*}) \equiv \hat{f}_t(\Theta),\ \tau = t, \ldots, t+T. \tag{4.3}$$

The receding-horizon strategy combined with the approximation of (4.2) and (4.3) is called *supercautious* as it is based on the assumption that we learn nothing about unknown parameters until the receding horizon. The uncertainty of the parameter estimates (4.3) projected into the predictors through the formula (4.2) inhibits excessive decision values. It makes the strategy supercautious.

Cautious strategy

The assumption (4.3) is often too strong. Then, it is reasonable to assume that

$$f(\Theta|\mathcal{P}_{a_\tau^*}) \approx F(\tau, f(\Theta|\mathcal{P}_{a_t^*})) \equiv \hat{f}_\tau(\Theta),\ \tau = t, \ldots, t+T,\ \Theta \in \Theta^*. \tag{4.4}$$

The mapping $F(\tau, \cdot)$ modifies the posterior pdf $f(\Theta|\mathcal{P}_{a_t^*})$ to another, more concentrated, pdf. The mapping F is chosen beforehand and does not use data belonging to the ignorance $\mathcal{F}_{a_t^*}$. Typically, the generated pdf $\hat{f}_\tau(\Theta)$ has the same first moment (expectation) as $f(\Theta|\mathcal{P}_{a_t^*})$, but its covariance $C_{\tau|t}$ decreases with increasing distance $\tau - t$ in a prespecified way. Often,

$$C_{\tau|t} = b(\tau - t) \times C_t \equiv b(\tau - t) \times \text{ covariance determined by } f(\Theta|\mathcal{P}_{a_t^*}).$$

The following nonnegative functions $b(j)$ serve as representative examples.

$$b(j) = \frac{1}{j+1} \text{ or } b(j) = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4.5}$$

Such a choice can be seen as *sharpening* of the pdf $f(\Theta|\mathcal{P}_{a_t^*})$ by taking its power $f^{\lambda_\tau^{-1}}(\Theta|\mathcal{P}_{a_t^*})$ using a prespecified sequence $\{\lambda_\tau \in (0,1)\}_{\tau=t}^{t+T}$.

Certainty-equivalence strategy

This strategy is the most widespread one. It gets the approximate predictive pdf by replacing an unknown parameter in the parameterized model by its current point estimate $\hat{\Theta}_t$

$$f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}) \approx f(\Delta_\tau|a_\tau, \mathcal{P}_{a_\tau^*}, \hat{\Theta}_t), \ \tau = t, \ldots, t+T. \tag{4.6}$$

It corresponds to the approximate generalized Bayesian parameter estimate

$$f(\Theta|\mathcal{P}_{a_\tau^*}) \approx \hat{f}_t(\Theta) \equiv \delta(\Theta - \hat{\Theta}_t), \ \tau = t, \ldots, t+T, \text{ where} \tag{4.7}$$

$\delta(\cdot)$ is Dirac delta function, a formal pdf of the measure with its support equal to $\{0\}$.

Active strategies

All outlined strategies are *passive strategies* in the sense that the planned decisions do not care about learning. At the same time, it is known (see [75, 76]) that the optimal strategy is *dual strategy.* It cares both about immediate decisions and learning process in a balanced way. This stimulates a search for measures that support learning. The resulting strategies are known as *active strategies.* Two basic suboptimal ways are used for their design.

- An external stimulating signal is fed into the closed loop. It is added to optional quantities like inputs or set points. It improves learning conditions at the cost of deteriorating the achievable quality.
- A term reflecting learning quality even under a passive-type design is added to the original loss function [104]. It is usually numerically demanding and sensitive to the relative weight of the added term.

Problem 4.1 (Development of active strategies) *Practical experience indicates that the active strategies improve the overall performance just a little for normal linear models. It can be shown, however, that passivity may result in completely bad performance in the case of controlled Markov-chain models [110]. The latter fact indicates possible problems in the context of mixture estimation as well as in hidden Markov-chains area [47]. Systematic attempts to solve this difficult problem are rare; see reference in [111]. The problem is less pronounced in signal-processing applications where the discussed models have been predominantly used. Their foreseen wider use for feedback advising and control [64, 112] calls for a change of this state.*

4.3 Decomposition of decision-making

Splitting of the decision-making task in subtasks is the known way to get an approximation of the practically optimal design but the splitting violates *golden decision-making rule* — try to solve the problem at hand in its entirety — and drives the solution from its optimum. Thus, a compromise has to be searched for.

Lack of the formal tools for the decomposition leaves us with empirical rules in this area. This makes us summarize here the experience we have collected in this respect in a long-term project *DESIGNER* [65, 113, 114, 115]. The project tries to decrease the burden on designers of adaptive controllers by creating algorithmic support for their prior tuning. It has led to "natural" decomposition of the design we list below while generalizing it to wider set of decision-making tasks. Each item in the list has been found as a relatively self-containing decision subproblem. The majority of them have a good solution for adaptive control based on normal linear controlled regression models. The majority of them are newly addressed in this book in order to cope with the mixture-based advising.

4.3.1 Offline phase

The design, as any human activity, is iterative. The iterations should be mostly made in the offline phase of the design in order to minimize expenses related to the commission of the proposed strategy.

The following indicative list of subtasks is solved, often with internal iterations, until the user is satisfied.

Formulate the addressed problem

The formulation specifies

- the technical decision-making aims,
- the system,
- the available data, decisions, and innovations,
- the physical and complexity constraints,
- the knowledge available.

Perform experimental design and collect data for offline analysis

This important subtask [116] is still weakly supported [117].

Preprocess data

Data preprocessing is an extensive area [118] with a lot of significant results; see e.g., [118, 119] and Sections 6.2, 8.2, 10.2.

Select the used class of parameterized models

The choice is mostly dictated by our ability to handle them. This text provides tools for normal mixtures (Chapters 8 and 9) and Markov chains (Chapters 10 and 11). Chapter 13 opens a way to treatment of their mixed versions.

Quantify prior knowledge on unknown parameters

This is still an underestimated area with particular results in [65, 120] and Section 6.3.

Estimate model structure

There are a lot of significant result in this respect, e.g., [26, 65, 95, 121], but the problem is far from being completely solved; see Section 6.6.

Estimate the period with which the actions are applied

The choice of the period for adaptive control is discussed in [122, 123] but a general solution for a general multivariate case is missing.

Perform generalized Bayesian estimation

It combines theory behind the adopted models, prior knowledge, and available data; cf. [69], Proposition 2.14, and the mixture-tailored versions in Sections 6.5, 8.5, 10.5. The results are used as the prior and/or alternative pdf in online phase.

Estimate forgetting factor

The maximum a posteriori probability estimate on a discrete grid is used; for instance, see Algorithm 6.2.

Validate estimation results.

Commonly, the achieved modelling quality is compared with that obtained when using a wider model set. The quality is also compared with learning and validation data [124]. Implementation of these ideas has no generally accepted methodology. Thus, a whole range of variants is to be used; see Section 6.7.

Transform the decision aims into the loss function

The class of loss functions is mostly determined by our ability to handle it. The approximate expression of the aims is then achieved by the tuning of optional knobs determining the loss function. Typically, weights of quadratic loss and, generally, the parameters determining the ideal pdf in fully probabilistic design have to be chosen. The iterative choice runs as follows.

1. *Select the test values parameters of the tuned parameters and design the corresponding strategy.*
   The optimization-based selection of the test values seems to be adequate [125].
2. *Predict closed loop behavior.*
   The closed loop is formed by the coupling system with the considered strategy. The prediction consists formally of transformation of all uncertain quantities into the user-defined quality indicators. Its formal solution provides Proposition 2.5 but practically a Monte Carlo evaluation has to be employed [126]. It becomes feasible using sequential stopping rules [94, 127, 128].
3. *Compare predictions with user's specifications.*
   Stop if the user's specifications are met. Otherwise, go to Step 1 while there are observable changes of indicators; otherwise, demand changes in the problem specification.

Validate the design results

Again, no complete established formal methodology is available but a range of particular tests can be made, Section 7.4, often based on extensive simulations combined with Monte Carlo techniques.

4.3.2 Online phase

The decision-making subtasks listed here are solved in real time for $t \in t^*$. Thus, there is almost no freedom for iterative trial-and-error solutions. Of course, it is wise to store the data collected during the online phase and use them for an improved offline design. The processing proceeds as follows.

- Collect and preprocess data.
- Generate reference signals.
- Estimate the parameters with stabilized forgetting; see [88] and Section 3.1.
- Use receding-horizon or iterations-spread-in-time design, in an appropriate version; Section 4.2;
- Generate the action using the designed strategy and measured data.
- Check possible discrepancies and make a finer tuning of optional parameters of the design; this supervision can be based on the theory discussed in this book.

Problem 4.2 (Completion of the design support) *The presented outline shows the complexity of the overall design chain. Some steps are weakly supported by theory and algorithms. Others are solved for linear normal models only. Some of them are not supported at all. Our experience indicates that the systematic solution of the design support pays back. The completion and generalization of this support represent a research challenge with a significant potential impact on the resulting quality of decision-making.*

5

Problem formulation

This chapter formulates the problem of the design of an advisory system and outlines its conceptual solution. The interconnection of the managed system with its operator is modelled by the finite mixture model. Its optional parts are optimized so that the resulting mixture model is as close as possible to management aims expressed by a user's ideal pdf in a fully probabilistic sense, cf. 2.4.2. The resulting mixture is offered to the operator as the recommended ideal pdf to be followed. Unlike the user's ideal pdf, the state described by the recommended ideal pdf can be practically achieved.

The application of this conceptually simple design requires formulation and solution of a sequence of particular technical steps. Their description forms the content of this chapter.

Design conditions and adopted design principles are clarified in Section 5.1. Special attention is paid to relationships of data spaces accessible to the operator and to the advisory system. Also, necessary reductions and extensions of involved data spaces are proposed there. The learning conditions we assume are specified in Section 5.2. Dynamic predictions, forming a bridge between learning and advising parts, are discussed in Section 5.3. The design of advices offered to the operator is described in Section 5.4. Basic types of advisory systems differing in the extent of optional ingredients of the modified model are classified there. The design of the presentation part of the advisory system is in Section 5.5. Learning and design steps, which form the backbone of detailed solutions presented in Chapters 6 and 7, are summarized in Section 5.6.2.

5.1 Design principle and design conditions

This section introduces notions needed for formalization and solution of the addressed design problem.

5.1.1 Systems and data spaces

We start with inspection of relationships of the managed system, its operator, and the constructed advisory system. The following agreement introduces fixes notions.

Agreement 5.1 (Nomenclature of systems) *The system managed by the operator is called the* m-system*. The closed loop formed by the m-system and its operator (supervisor) is called the* o-system*. The* advisory system *transforming the data measured on the o-system into advices to the operator is called the* p-system*; see Fig. 5.1.*

If need be, the quantities related to m-, o-, and p-systems are distinguished by prefixes m-, o-, and p- or by subscripts m, o, and p.

Out of this chapter, we deal mostly with the data handled by the p-system. Then, their subscript p is dropped if there is no danger of misunderstanding.

Formulation and solution of the overall design need clarification of relationships among quantities dealt with by the o- and p-systems.

Agreement 5.2 (Data, observation, and action spaces) *Values of the quantities available to the operator form the* data space of the operator $d_o^*(\mathring{t})$*. This space is the Cartesian product of the* observation space of the operator $\mathit{\Delta}_o^*(\mathring{t})$ *— formed by the innovations $\mathit{\Delta}_o(\mathring{t})$ available to the operator; see Section 2.2 — and of the* action space of the operator $a_o^*(\mathring{t})$*, i.e., $d_o^*(\mathring{t}) \equiv (\mathit{\Delta}_o^*(\mathring{t}), a_o^*(\mathring{t}))$.*

Values of the quantities available to the advisory system form the data space of the advisory system $d_p^*(\mathring{t})$*. This space is the Cartesian product of the* observation space of the advisory system $\mathit{\Delta}_p^*(\mathring{t})$ *and of the* action space of the advisory system $a_p^*(\mathring{t})$*, i.e., $d_p^*(\mathring{t}) \equiv (\mathit{\Delta}_p^*(\mathring{t}), a_p^*(\mathring{t}))$.*

Values of the quantities available to the operator but unavailable to the advisory system form the surplus data space of the operator $d_{o+}^*(\mathring{t}) \equiv d_p^*(\mathring{t}) \cup d_o^*(\mathring{t}) \setminus d_p^*(\mathring{t})$*.*

Values of the quantities available to the advisory system but unavailable to the operator form the surplus data space of the advisory system $d_{p+}^*(\mathring{t}) \equiv d_p^*(\mathring{t}) \cup d_o^*(\mathring{t}) \setminus d_o^*(\mathring{t})$*; see Fig. 5.1.*

Note that the possible nonemptiness of the surplus data spaces singles out the addressed problem from the standard formulation of decision-making. Obviously, the design of the advisory system is meaningful only when the p- and o-data spaces overlap.

Requirement 5.1 (Overlap of data spaces) *The data spaces available to the operator $d_o^*(\mathring{t})$ and to the advisory system $d_p^*(\mathring{t})$ have a nonempty intersection $d_{op}^*(\mathring{t}) \neq \emptyset$. Thus,*

$$\begin{aligned} d_o^*(\mathring{t}) &= d_{op}^*(\mathring{t}) \cup d_{o+}^*(\mathring{t}), \quad d_{op}^*(\mathring{t}) \cap d_{o+}^*(\mathring{t}) = \emptyset, \\ d_p^*(\mathring{t}) &= d_{op}^*(\mathring{t}) \cup d_{p+}^*(\mathring{t}), \quad d_{op}^*(\mathring{t}) \cap d_{p+}^*(\mathring{t}) = \emptyset, \quad \textit{where} \end{aligned} \tag{5.1}$$

$d_{o+}^*(\mathring{t})$ is the surplus data space of the operator and $d_{p+}^*(\mathring{t})$ is the surplus data space of the advisory system.

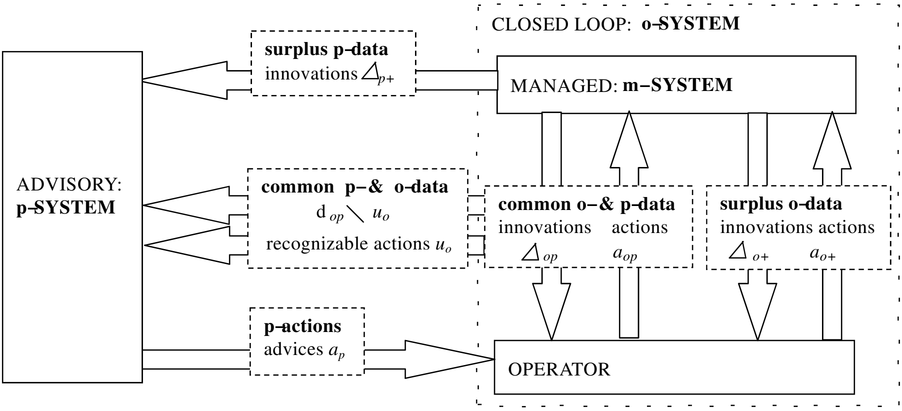


Fig. 5.1. Relationships of systems and data spaces: Agreements 5.1, 5.2, 5.7.

5.1.2 Basic scenario and design principle

The following basic scenario is considered within the time span determined by a *horizon* $\mathring{t} \leq \infty$.

The operator handling the m-system deals with a sequence $d_o(\mathring{t})$ of the o-data. A nonempty part of the data record $d_{o;t}$ is formed by operator's actions $a_{o;t} \in a_o^*$. The rest consists of innovations $\Delta_{o;t}$, i.e.,

$$d_{o;t} = (\Delta_{o;t}, a_{o;t}) \equiv (\text{o-innovations, o-actions}).$$

Formally, the operator implements the causal strategy

$$\left\{ d_o^*(t-1) \to a_{o;t}^* \right\}_{t \in t^*}.$$

The strategy of the operator is to be judged according to the expected value $\mathcal{E}$ of a loss function

$$\mathcal{Z} : \, d_o^*(\mathring{t}) \to [0, \infty] \tag{5.2}$$

that reflects the management aims. The real operator's strategy is usually chosen informally, but the joint pdf $f(d_o(\mathring{t}))$ would be the adequate description of the o-system needed for the formal evaluation of the operator's strategy; see Chapter 2.

The p-system works on a sequence of the p-data $d_p(\mathring{t})$. Each record $d_{p;t}$ contains data items $d_{pi;t}, \; i \in i^* \equiv \{1, \ldots, \mathring{d}_p\}$ with either real or discrete values. The record $d_{p;t} \equiv (\Delta_{p;t}, a_{p;t})$ includes p-innovations $\Delta_{p;t}$, observed by the p-system on the o-system and actions, of the p-system $a_{p;t}$, called *advices*.

The p-system implements its causal strategy $\left\{ d_p^*(0), d_p^*(t-1) \to a_{p;t}^* \right\}_{t \in t^*}$, where $d_p(0)$ denotes experience collected before using the p-system.

The p-system strategy is designed so that the o-system accepting advices achieves the smallest expected loss (5.2).

Agreement 5.3 (Guided and unguided o-system) *The interconnection of the o-system with the p-system creates a new* guided o-system*. We shall use also the mirror term* unguided o-system *for the o-system working without a p-system.*

The adjectives guided and unguided *are also used in connection with the corresponding models, behaviors, situations, etc. For instance,* unguided model *and* guided model *mean the model of the unguided and guided o-system, respectively.*

The complete outer description of the guided o-system is given by the joint pdf $f(d_p(\mathring{t}), d_{o+}(\mathring{t}))$ of all involved data $(d_p(\mathring{t}), d_{o+}(\mathring{t}))$, cf. (5.1). The structure of the discussed interconnection is predominantly determined by the communication ways of the managed system, the operator and the advisory system. The m-system

- provides its o-innovations to the operator as responses to the o-actions,
- offers some data to the p-system that generally differ from $d_{o;t}$,
- is indirectly influenced by the p-system through the o-actions that are stimulated by advices, i.e., by actions of the p-system.

We would like to design the p-system that guides the operator to make the expectation of the loss function (5.2) as small as possible. The advisory system cannot force the operator to follow its advices. The p-system can only present a “target” to be reached. The advices will be useful if the operator is able to respect them, i.e., if the advised target is reachable by the cooperating operator. Both the design of the p-system and presentation of advices can be done in a systematic way if we adopt the fully probabilistic design; see Section 2.4.2. It determines the way we intend to stimulate the operator.

> The *ideal pdf* in the sense of the fully probabilistic design, Agreement 2.7, is adopted as the optimized target offered to the operator.

The corresponding conceptual algorithm implementing this *design principle* looks as follows.

Algorithm 5.1 (Design principle of the advisory system)

1. *Express managing aims as the* user's ideal pdf ${}^{\lfloor U}f\left(d_p(\mathring{t}), d_{o+}(\mathring{t})\right)$.
2. *Estimate an outer* multimodal *model $f(d_p(\mathring{t}), d_{o+}(\mathring{t}))$ describing relationships among the considered data $(d_p(\mathring{t}), d_{o+}(\mathring{t}))$.*
3. *Create the ideal pdf ${}^{\lfloor I}f\left(d_p(\mathring{t}), d_{o+}(\mathring{t})\right)$ that*
   - *is as close as possible to the user's ideal pdf ${}^{\lfloor U}f\left(d_p(\mathring{t}), d_{o+}(\mathring{t})\right)$; the KL divergence is used as the proximity measure,*
   - *inherits those constituents of the pdf $f\left(d_p(\mathring{t}), d_{o+}(\mathring{t})\right)$ describing an unguided o-system that cannot be changed even by the* fully cooperating operator *that implements the* randomized *strategy recommended to him.*

The left superscript ^{I} marks the resulting ideal *pdf and its optimized constituents.*

4. *Present low-dimensional projections of the ideal pdf* ${}^{\lfloor I}f\left(d_p(\mathring{t}), d_{o+}(\mathring{t})\right)$ *as the "target" to be followed by the operator.*

Remark(s) 5.1

1. *The applicability of Algorithm 5.1 depends on the possibility of creating the involved elements practically. This nontrivial task is solved gradually in subsequent sections.*
2. *The focus on a fully probabilistic design allows us to deal with a uniform probabilistic description of learning, design, and advising.*
3. *The optimization with properly preserved elements of the m-system guarantees that a good past practice can be followed. The ideal pdf, constructed in the outlined way,*
   - *is reachable,*
   - *relates all observed consequences to their observed causes unlike human being can.*

5.1.3 Reduction of surplus data of the operator

Algorithm 5.1 deals formally with all data occurring in the guided system, i.e., with the union of p- and o-data. This subsection shows that, in the design of the p-system, we need not model the surplus o-data $d_{o+}(\mathring{t})$.

The p-system has no information on $d^*_{o+}(\mathring{t})$. Thus, it has to leave this part of the o-behavior to its fate. In other words, the strategy of the p-system, designed without considering the surplus o-data, implies that their distribution has to be accepted as the ideal one. Formally, it restricts the constructed ideal pdf ${}^{\lfloor I}f(\cdot)$ by the requirement

$$
{}^{\lfloor I}f\left(d_{o+}(\mathring{t})|d_p(\mathring{t})\right) \equiv f\left(d_{o+}(\mathring{t})|d_p(\mathring{t})\right). \tag{5.3}
$$

The objective pdf $f(d_{o+}(\mathring{t})|d_p(\mathring{t}))$ (see Chapter 2) describes the data $d_{o+}(\mathring{t})$ unavailable to the p-system.

According to the concept of the fully probabilistic design, the distance of the pdf describing the inspected behavior to its ideal pdf is measured through the KL divergence (2.25). In this context, the requirement (5.3) implies a simple but important consequence.

Proposition 5.1 (Ideal pdf offered by the advisory system) *Let the ideal pdf* ${}^{\lfloor I}f(d_p(\mathring{t}), d_{o+}(\mathring{t}))$ *offered by the advisory system meet the requirement (5.3). Then,*

$$
\mathcal{D}\left(f \middle|\middle| {}^{\lfloor I}f\right) \equiv \int f(d_p(\mathring{t}), d_{o+}(\mathring{t})) \ln\left(\frac{f(d_p(\mathring{t}), d_{o+}(\mathring{t}))}{{}^{\lfloor I}f(d_p(\mathring{t}), d_{o+}(\mathring{t}))}\right) d(d_p(\mathring{t}), d_{o+}(\mathring{t}))
$$

$$
\begin{aligned}
&= \int f(d_p(\mathring{t})) \ln\left(\frac{f(d_p(\mathring{t}))}{{}^{\lfloor I}f(d_p(\mathring{t}))} \right) dd_p(\mathring{t}) \quad \textit{and} \qquad (5.4)\\
\mathcal{D}\left({}^{\lfloor I}f \middle|\middle| f \right) &\equiv \int {}^{\lfloor I}f(d_p(\mathring{t})) \ln\left(\frac{{}^{\lfloor I}f(d_0p(\mathring{t}))}{f(d_p(\mathring{t}))} \right) dd_p(\mathring{t}).
\end{aligned}
$$

Thus, for our design purposes, the outer description of the o-system $f(\cdot)$ and the optimized ideal pdf ${}^{\lfloor I}f(\cdot)$ have to be specified on $d_p^*(\mathring{t})$ only.

Proof. Using the basic rules for pdfs — the chain rule and normalization; see Proposition 2.4 — we can directly verify the following identities

$$
\begin{aligned}
\mathcal{D}\left(f \middle|\middle| {}^{\lfloor I}f \right) &\equiv \int f(d_p(\mathring{t}), d_{o+}(\mathring{t})) \ln\left(\frac{f(d_p(\mathring{t}), d_{o+}(\mathring{t}))}{{}^{\lfloor I}f(d_p(\mathring{t}), d_{o+}(\mathring{t}))} \right) d(d_p(\mathring{t}), d_{o+}(\mathring{t}))\\
&\underbrace{=}_{\text{chain rule}} \int f(d_{o+}(\mathring{t})|d_p(\mathring{t})) f(d_p(\mathring{t}))\\
&\times \quad \ln\left(\frac{f(d_{o+}(\mathring{t})|d_p(\mathring{t})) f(d_p(\mathring{t}))}{{}^{\lfloor I}f(d_{o+}(\mathring{t})|d_p(\mathring{t}))\, {}^{\lfloor I}f(d_p(\mathring{t}))} \right) d(d_p(\mathring{t}), d_{o+}(\mathring{t}))\\
&\underbrace{=}_{(5.3)} \int f(d_{o+}(\mathring{t})|d_p(\mathring{t})) f(d_p(\mathring{t})) \ln\left(\frac{f(d_p(\mathring{t}))}{{}^{\lfloor I}f(d_p(\mathring{t}))} \right) d(d_p(\mathring{t}), d_{o+}(\mathring{t}))\\
&\underbrace{=}_{\text{normalization, Proposition 2.4}} \int f(d_p(\mathring{t})) \ln\left(\frac{f(d_p(\mathring{t}))}{{}^{\lfloor I}f(d_p(\mathring{t}))} \right) dd_p(\mathring{t}).
\end{aligned}
$$

The equality for $\mathcal{D}\left({}^{\lfloor I}f||f \right)$ can be proved in the same way. □

Consequently, the accepted condition (5.3) allows us to leave the surplus data space of the operator $d_{o+}^*(\mathring{t})$ completely out of our consideration. Thus, under (5.3), the design results obtained for empty and nonempty $d_{o+}^*(\mathring{t})$ are the same. Notation is, however, simpler if $d_{o+}^*(\mathring{t}) = \emptyset$. In this case, the definition (5.1) implies

$$
d_o^*(\mathring{t}) = d_{op}^*(\mathring{t}) \;\Rightarrow d_o^*(\mathring{t}) \subset d_p^*(\mathring{t}). \qquad (5.5)
$$

From now on, we adopt this formal simplification that the surplus data space of the operator $d_{o+}^*(\mathring{t})$ is empty and thus the data space of the operator $d_o^*(\mathring{t})$ is a subset of the data space of the advisory system $d_p^*(\mathring{t})$.

5.1.4 Construction of a true user's ideal pdf

The considered design of an optimal advisory system assumes that the aims of management and advising can be expressed by the user's ideal pdf describing the desired behavior of the p-data $d_p(\mathring{t})$.

The user expresses its aims in terms of quantitative indicators, called *quality markers*, that can be evaluated using the o-data only. The quality markers $m(\mathring{t}) \in m^*(\mathring{t})$ qualify behavior of the o-system through a known function

$$d_o^*(\mathring{t}) \rightarrow m^*(\mathring{t}) \equiv \text{(partially) ordered space.} \tag{5.6}$$

Desired properties of quality markers, expressing management aims, are assumed to be "translated" into the *true user's ideal pdf* ${}^{\lfloor U}f(d_o(\mathring{t}))$ that characterizes desired distribution of o-data available to the operator. The following "translation" methods are at our disposal.

- A simple, typically normal, pdf is chosen. Set point for $d_{o;t}$ is defined as its mean. Covariance matrix is chosen so that the significant probability is allocated to the multivariate interval covering desirable ranges of d_o-entries.
- A simple, typically normal, model is estimated on historical records $d_o(\mathring{t}_o)$ and its mean is replaced by the desired set-point. The covariance is shrunk so that a desirable improvement is enforced.

Remark(s) 5.2

1. *A Monte Carlo-based translation, as elaborated in connection with DESIGNER project [126], can be also used.*
2. *Within this book, the true user's ideal pdf is taken as unimodal pdf. Recently, it was found that the adopted fully probabilistic design can be practically applied even when the user's ideal pdf is a finite mixture [129, 130]. This extends applicability of the fully probabilistic design on decision-making problems with multiple criteria!*

Problem 5.1 (Advising on internal quantities) *Generally, the user cares about inner, directly unobservable states, too. The operator has to guess them using observed data and also optimize behavior of these estimates. Finally, the observed-data dependent loss function is optimized. For simplicity, the advising is formulated directly as the data-driven design, cf. Agreement 2.8. The relevant theory should, however, be extended explicitly to the case when the operator deals with internal, considered but directly unobservable states, too.*

5.1.5 Extension of a true user's ideal pdf to the surplus p-data

The adopted condition (5.5) implies that an implementation of conceptual Algorithm 5.1 requires specification of a user's ideal pdf ${}^{\lfloor U}f(\cdot)$ on the data space of the advisory system $d_p^*(\mathring{t})$. The true user's ideal pdf is, however, defined "naturally" on the data space of the operator $d_o^*(\mathring{t}) \equiv d_{op}^*(\mathring{t})$; see (5.5) and Section 5.1.4. At the same time, the fact that the advisory system deals with a wider data set than the operator should be exploited for reaching a better quality of advices. Thus, it is necessary to extend the user's ideal pdf from $d_o^*(\mathring{t})$ to $d_p^*(\mathring{t})$.

By definition, the operator is not aware and consequently interested in the surplus data of the advisory system $d_{p+}(\mathring{t})$. Thus, the operator has to leave this surplus data to the fate determined by the managed system and the influence of the advisory system. It leads us to the choice

$$
\begin{aligned}
{}^{\lfloor U}f(d_p(\mathring{t})) &\underbrace{=}_{\text{chain rule}} {}^{\lfloor U}f(d_{p;\mathring{t}}|d_p(\mathring{t}-1))\, {}^{\lfloor U}f(d_p(\mathring{t}-1)) \underbrace{=}_{(5.1),(5.5)} \\
&= {}^{\lfloor U}f(d_{o;\mathring{t}}|d_{p+;\mathring{t}}, d_p(\mathring{t}-1))\, {}^{\lfloor U}f(d_{p+;\mathring{t}}|d_p(\mathring{t}-1))\, {}^{\lfloor U}f(d_p(\mathring{t}-1)) \\
&\equiv {}^{\lfloor U}f(d_{o;\mathring{t}}|d_{p+;\mathring{t}}, d_p(\mathring{t}-1))\, {}^{\lfloor I}f(d_{p+;\mathring{t}}|d_p(\mathring{t}-1))\, {}^{\lfloor U}f(d_p(\mathring{t}-1)) \\
&\equiv {}^{\lfloor U}f(d_{o;\mathring{t}}|d_o(\mathring{t}-1))\, {}^{\lfloor I}f(d_{p+;\mathring{t}}|d_p(\mathring{t}-1))\, {}^{\lfloor U}f(d_p(\mathring{t}-1)) \\
&= \prod_{t\in t^*} {}^{\lfloor U}f(d_{o;t}|d_o(t-1))\, {}^{\lfloor I}f(d_{p+;t}|d_p(t-1)).
\end{aligned} \tag{5.7}
$$

The first equivalence in the third row means that $d^*_{p+;\mathring{t}}$ is left to its fate: the operator leaves the p-system to identify this part of the user's ideal pdf with the option made by the p-system. The second equivalence in the third row says that the user's ideal pdf on $d^*_{o;\mathring{t}}$ is constructed irrespective of the inaccessible values in $d^*_{p+}(\mathring{t})$. The final identity in (5.7) is adopted as a basic assumption.

Remark(s) 5.3

1. *This subsection completes construction of the user's ideal pdf on the union of data spaces needed in the first step of Algorithm 5.1. Step 2 is discussed in Section 5.2 and elaborated in Chapters 6, 8, and 10. Steps 3 and 4 are covered by Sections 5.4 and 5.5 and elaborated in Chapters 7, 9, and 11.*
2. *Operator may be aware of but uninterested in some entries of $d_{o;t}$. Then, they can be formally thought as entries in the surplus data space of an operator.*
3. *The designer may use the entries of p-data as tuning knobs in design. This option is indeed exploited when we try to create a user-friendly advisory system that controls the burden on the operator. For instance, the frequency of changes in advising is kept low.*

5.2 Learning conditions

The design of the advisory system relies on a good mixture model of the o-system. This sections discusses necessary modelling and learning conditions we suppose to be met.

Before progressing in a formal way, it is worth stressing the following facts.

The need for any advisory system arises when there are good and bad modes in managing the considered m-system.

The chance to design a good advisory system exists if both good and bad modes are reflected in its data space d^*_p. It implies that the pdf $f(d_p(\mathring{t}))$ describing $d^*_p(\mathring{t})$ completely is expected to have <u>multiple modes</u>.

The possibility to design an efficient advisory system arises if the available design experience contains well-pronounced information about all significant operation modes that may occur while the operator handles the managed system. In other words, the data providing this experience have to be rich and informative.

The top position of the operator in the control hierarchy implies (see Chapter 1) that the experience of the advisory system is to be predominantly data-based.

The rate of the operator's actions is often (much) slower than the sampling rate of the sensors and controllers within the m-system. Consequently, all considered data can be and should be grouped or reduced to a sequence generated with the rate of the operator's actions. We suppose it done.

The experience of the p-system also includes surplus data of the advisory system $d^*_{p+}(\mathring{t}) \equiv d^*_p(\mathring{t}) \setminus d^*_o(\mathring{t})$. They are directly at the disposal of the advisory system but not to the operator. These data may be invaluable in discovering various operating modes. They have to be considered in the design of any efficient p-system. Consequently, we have to deal with all data $d_p(\mathring{t})$ and thus we can mostly drop the subscript p further on and adopt the identity

$$d(\mathring{t}) \equiv d_p(\mathring{t}). \tag{5.8}$$

The general theory (see Chapter 2) and the above discussion imply that a *multiple-mode pdf* $f(d(\mathring{t})|d(0))$ is the description required for a design of the considered advisory system. The condition $d(0) \equiv d_p(0)$ stands for the experience available to the p-system before its use. Further on, $d(0)$ is fixed and formally included into $d(\mathring{t})$, Then, the notation can be simplified by "hiding" $d(0)$.

The pdf $f(d(\mathring{t}))$ can be factorized according to the chain rule $f(d(\mathring{t})) = \prod_{t \in t^*} f(d_t|d(t-1))$. The assumed low action rate of the operator implies that the faster dynamics of the system is diminished in the grouped or reduced data. Thus, slow transitions among quasi-steady-state working points are modelled. This can be well described by a *Markov model of a finite order.* Thus, we can assume that the unguided o-system is described by the joint pdf

$$f(d(\mathring{t})) = \prod_{t \in t^*} f(d_t|\phi_{t-1}), \text{where}$$

$\phi_t \in \phi^*,\ t \in t^*$, are known finite-dimensional vectors called *observable states.*

The intended *online use* of the advisory system implies that we have to deal with models containing ϕ_t that can be evaluated in a recursive manner, i.e., $\phi_t = \mathit{\Phi}(\phi_{t-1}, d_t)$ with a known function $\mathit{\Phi}$.

The advisory system can be successful only if the relationships learned from experience are valid almost permanently. Thus, we have to assume that the outer model $\{f(d_t|\phi_{t-1})\}_{t \in t^*}$, Agreement 2.7, of the unguided o-system built on the p-data $d_t \equiv d_{p;t}$ is time invariant.

It is known that the needed multiple-mode probabilistic models can (almost) always be approximated by a *finite mixture* of unimodal models [49], called *components*. We estimate this model using — at least approximately — the standard Bayesian methodology.

For reference purposes, we summarize learning conditions adopted in the construction of the advisory system.

Requirement 5.2 (Learning conditions for the design)

1. *The data space* $d^*(\mathring{t})$ *of the advisory system (the p-system) has a nonempty intersection with the data space* $d_o^*(\mathring{t})$ *of the o-system.*
2. *The sampling rate is harmonized with the operating rate. The number* $\mathring{d}$ *of data items sent to the p-system at each time moment is fixed.*
3. *No attempt is made to influence quantities lying in the data space of the o-system and being out of the data space of the p-system.*
4. *The pdf on records* $d(\mathring{t})$*, available to the p-system, is modelled by the pdf* $f(d(\mathring{t})|\Theta) \equiv \prod_{t \in t^*} f(d_t|\phi_{t-1}, \Theta)$*, where the pdf* $f(d_t|\phi_{t-1}, \Theta)$*, determined by the measurable state* ϕ_{t-1} *and by the parameter* Θ*, is a time invariant finite mixture; see (5.9) below.*
5. *The approximate Bayesian learning described in Section 6.5 provides the pdf* $f(d(\mathring{t})) \equiv f(d(\mathring{t})|d(0)) = \prod_{t \in t^*} f(d_t|d(t-1))$ *needed for the design of the p-system.*

The finite mixture approximating distribution of data $d(\mathring{t})$ with multiple modes is assumed to have the form

$$f(d(\mathring{t})|\Theta) \equiv \prod_{t \in t^*} f(d_t|\phi_{t-1}, \Theta) \text{ with } \textit{finite mixtures} \text{ as parameterized models}$$

$$f(d_t|\phi_{t-1}, \Theta) \equiv \sum_{c \in c^*} \alpha_c f(d_t|\phi_{c;t-1}, \Theta_c, c),\ c^* = \{1, \ldots, \mathring{c}\},\ \mathring{c} < \infty, \tag{5.9}$$

$f(d_t|\phi_{c;t-1}, \Theta_c, c)$is called *component* given by parameters Θ_c and the state

$\phi_{c;t} = \Phi_c(\phi_{c;t-1}, d_t)$, i.e., the state $\phi_{c;t-1}$ can be recursively updated data d_t,

$\alpha_c \equiv$ the probabilistic *component weight*

$\Theta \equiv$ mixture parameter formed by component parameters and weights in

$$\Theta^* \equiv \left\{ \{\Theta_c \in \Theta_c^*\}_{c \in c^*},\ \alpha \equiv [\alpha_1, \ldots, \alpha_{\mathring{c}}] \in \alpha^* \equiv \left\{ \alpha_c \geq 0,\ \sum_{c \in c^*} \alpha_c = 1 \right\} \right\}.$$

The mixture parameter $\Theta = [\alpha_c, \Theta_c]_{c \in c^*}$ may also include the number of components $\mathring{c}$.

The entries of d_t can be permuted in each component and some permutations may lead to parameterizations with fewer parameters. It makes us include into the model description the permutations

$$d \to d_c \ \text{ with } \ d_{ic} = d_{j_{ic}}, \tag{5.10}$$

where j_{ic} is ith entry of the permuted indexes $[1, \ldots, \mathring{d}]$.

The assignment (5.10) is applied component-wise and together with the chain rule, Proposition 2.4, give

$$f(d_t|\phi_{c;t-1},\Theta_c,c) = \prod_{i\in i^*} f(d_{ic;t}|d_{(i+1)\cdots\mathring{d}c;t},\phi_{c;t-1},\Theta_{ic},c)$$
$$\equiv \prod_{i\in i^*} f(d_{ic;t}|\psi_{ic;t},\Theta_{ic},c). \quad (5.11)$$

The additional subscript i of the parameter Θ_{ic} indicates that only some entries of Θ_c may occur in the ith pdf predicting the ith scalar entry of $d_{ic;t}$. Similarly, the introduced *regression vector* $\psi_{ic;t}$ is generally a subvector of the vector

$$[d_{(i+1)c;t},\ldots,d_{\mathring{d}c;t},\phi'_{c;t-1}]'. \quad (5.12)$$

Now we can fix nomenclature related to the mixture.

Agreement 5.4 (Nomenclature related to mixtures)

Pdfs: The pdf $f(d_t|\phi_{c;t-1},\Theta_c,c)$ *in (5.9) is called the* parameterized component *of a mixture and* α_c *is the* weight of the cth parameterized component.

The pdf $f(d_{ic;t}|\psi_{ic;t},\Theta_{ic},c)$ *in (5.11) is called the* parameterized factor.

A parameterized factor occurring in several components is the common parameterized factor.

Factors are called the adjacent factors *if* $d_{(i+1)c;t}$ *is in the regression vector of the factor predicting* $d_{ic;t}$ *or vice versa.*

The predictive pdf $f(d_t|d(t-1),c)$ *is called the* component.

The predictive pdf $f(d_{ic;t}|d_{(i+1)\cdots\mathring{d}c;t},d(t-1),c)$ *is called the* factor.

The pdf $f(\Theta_{ic}|d(t))$ *is called the* factor estimate.

The pdf $f(\Theta_c|d(t))$ *is called the* component estimate.

The pdf $f(\alpha|d(t))$ *is called the* component-weight estimate.

The pdf $f(\Theta|d(t))$ *is called the* mixture estimate.

The estimate is called the prior estimate *if* $t=0$.

The estimates are called posterior estimates *if* $t\in t^*$. *Often, the posterior estimate is called the estimate only.*

Data: The vector d_t *containing data measured at time* t *is called the* data record.

The predicted scalar quantity $d_{ic;t}$ *is called the* output of the factor.

The entry $d_{i;t}$, $i=1,\ldots,\mathring{d}$, *of the data record* d_t *is also called the* i*th* data channel *or simply the* i*th* channel. *These terms are used in the implementation context.*

Data entries that never play the role of the output of a factor are called nonmodelled data.

The vector $\phi_{c;t-1}$, *that can be updated recursively using the newest data* d_t, *is the observable* state of the parameterized component.

The parameterized factor is determined by the regression vector $\psi_{ic;t}$ *formed by a subselection of entries from the vector* $[d_{(i+1)\cdots\mathring{d}c;t},\phi'_{c;t-1}]'$ *(5.12).*

The coupling $\Psi_{ic;t}\equiv[d_{ic;t},\psi'_{ic;t}]'$ *is called the* data vector *of the factor.*

The data vector Ψ is in the phase form *if it consists of a selection of entries from the data record d_t and its several delayed values $d_t, \ldots, d_{t-\partial}$, $\partial \in \partial^* \equiv \{0, 1, \ldots, \mathring{\partial}\}$, $\mathring{\partial} < \infty$. The corresponding state vector ϕ_{t-1} is also said to be in the phase form.*

Structures: Model structure is defined in a hierarchical way starting from the simplest elements, i.e., factors.

The list of regressors, *meaning the entries of the regression vector of a parameterized factor, is called the* structure of the parameterized factor.

The structure of the parameterized factor in the phase form *is the list of pairs (j, ∂_j) stating that $d_{j;t-\partial_j}$, $\partial_j \in \partial^*$, belongs to the data vector Ψ_t.*

The structure of the parameterized component *is an ordered list of factors creating it. The order characterizes the chosen permutation (5.10).*

Structure of the parameterized mixture *is the list of parameterized components creating it.*

Remark(s) 5.4

1. *The introduced factors, predicting individual entries of the modelled data,*
   - *provide flexibility of the parametric description,*
   - *allow us to jointly describe continuous and discrete valued quantities,*
   - *permit us to respect dependencies reflected in several components,*
   - *open a way for use of different models for different entries of d_t.*
2. *The adopted dynamic mixture model is not sufficiently general. The component weights should also depend on the state vector. The choice is driven by our inability to estimate this "natural" and more realistic model. This important aspect is discussed more deeply in Section 5.3. The restrictive assumption is partially relaxed in Chapter 13.*
3. *The modelled p-data may contain a part that is not in the data space of the o-system. Some of them might just bring complementary information and their evolution need not be modelled. They are used only as entries in the state vectors ϕ_t. In this way, the introduced nonmodelled data may arise. It is worth stressing that their use is limited more or less to one-step-ahead predictions. Otherwise, such quantities have to be modelled as their predictions are needed.*
4. *Redundancy and contradictions in specification of structures have to be checked when being defined.*
5. *Let us assume that*

$$d_t \sim \mathcal{N}_{d_t}\left(\begin{bmatrix} \theta \\ 0 \end{bmatrix}, \begin{bmatrix} r_1 & r_{12} \\ r_{12} & r_2 \end{bmatrix} \right),$$

   where the symbol $\sim$ *expresses that the two-dimensional vector d_t is distributed according to the normal pdf $\mathcal{N}_d(\mu, R)$ with the expectation μ and covariance matrix R. The scalar parameter $\theta \neq 0$ determines μ.*

   It is straightforward to show that both possible factorized parameterizations differ in the number of parameters whenever $r_{12} \neq 0$. This provides *the counterexample to the conjecture that the number of parameters to be estimated does not depend on the order of factors. Thus, the notation of the component structure as an* ordered list of factors *is meaningful.*

6. *Formally, the measured data, the mixture model and a suitably specified prior pdf allow us to get the needed model of the unguided o-system $f(d(\mathring{t}))$ through the Bayesian estimation and prediction. Practically, the estimation of mixture models on tens of thousands of records d_t with tens of entries is computationally very intensive. Thus, we are forced to use an approximations developed in detail in Section 6.5.*

5.3 Mixtures as approximate models and predictors

The considered mixture has dynamic components but constant weights. They express that the cth component of the o-system is active for $100 \times \alpha_c\%$ of the operating time. The adopted model (5.9) allows mutually independent as well as past-state-independent changes of active components at any time moment. For dynamic mixtures, this is an "unnatural" choice. The restriction to constant component weights has a pragmatic motivation. The assumption of constant α and a proposal as to how to weaken it are discussed here. A more systematic attempt to avoid this restriction is given in Section 13.2.

Let us consider that the parameterized model is a projection of a more complex model labelled by Θ and by an additional parameter $M \in M^* \equiv \cup_{c\in c^*} M_c^*$ where the finite collection of sets $\{M_c^*\}_{c\in c^*}$ covers M^*, i.e. $M_c^* \cap M_{\tilde{c}}^* = \emptyset$ for $c \neq \tilde{c}$. Then,

$$\begin{aligned} f(d_t|d(t-1),\Theta) &= \int f(d_t, M|d(t-1),\Theta)\, dM \\ &= \int f(d_t|d(t-1),\Theta,M) f(M|d(t-1),\Theta)\, dM \\ &= \sum_{c\in c^*} \int_{M_c^*} f(d_t|d(t-1),\Theta,M) f(M|d(t-1),\Theta)\, dM. \quad (5.13) \end{aligned}$$

Let us assume that for each $c \in c^*$ there is an $M_c \in M_c^*$ for which the following approximation

$$f(d_t|d(t-1),\Theta,M) \approx f(d_t|d(t-1),\Theta,M_c) \equiv f(d_t|d(t-1),\Theta_c,c)$$

is good for all $M \in M_c^*$ and for all possible data sequences $d(t),\ t \in t^*$. Using this assumption, we arrive at the finite mixture with data-dependent weights

$$f(d_t|d(t-1),\Theta) \approx \sum_{c\in c^*} f(d_t|d(t-1),\Theta_c,c)\tilde{\alpha}_c(d(t-1),\Theta), \qquad (5.14)$$

where

$$\begin{aligned}\tilde{\alpha}_c(d(t-1),\Theta) &= \int_{M_c^*} f(M|d(t-1),\Theta)\, dM \\ &= \text{Probability}(M \in M_c^*|d(t-1),\Theta) \equiv f(M_c^*|d(t-1),\Theta).\end{aligned}$$

The approximating mixture on the right-hand side of (5.14) has to be pdf. The individual components are pdfs; thus the right-hand side of (5.14) becomes pdf iff we assume $\sum_{c\in c^*} \tilde{\alpha}_c(d(t-1),\Theta) = 1$ for all $d(t-1)$. In order to make the consequences of this condition explicit, we write $\tilde{\alpha}_c(d(t-1),\Theta)$ as a normalized product of constant probabilistic weights α_c and of a nonnegative parameterized functions of data $\beta_c(d(t-1),\Theta)$

$$\tilde{\alpha}_c(d(t-1),\Theta) = \frac{\alpha_c\beta_c(d(t-1),\Theta)}{\sum_{\tilde{c}\in c^*} \alpha_{\tilde{c}}\beta_{\tilde{c}}(d(t-1),\Theta)}.$$

The mixtures with constant component weights are obtained if the functions $\{\beta_c(\cdot)\}_{c\in c^*}$ are constant. If they really depend on data, we get the mixture

$$f(d_t|d(t-1),\Theta) = \sum_{c\in c^*} \alpha_c \frac{f(d_t|d(t-1),\Theta_c,c)\beta_c(d(t-1),\Theta)}{\sum_{\tilde{c}\in c^*} \alpha_{\tilde{c}}\beta_{\tilde{c}}(d(t-1),\Theta)}. \tag{5.15}$$

For nontrivial $\beta_c(\cdot)$, the components of the mixture (5.15) do not belong to the exponential family so that their efficient estimation on large data sets is extremely difficult. Consequently, no efficient algorithm for estimation of the overall mixture is known except a novel attempt in Section 13.2.

We show, however, that our restricted model (5.9) with constant weights can be interpreted as a limit of the model (5.14).

Proposition 5.2 (Constant weights approximation (5.14))
Weights of the approximating pdf (5.14) converge almost surely to constant values. Thus, the adopted model with constant weights can be viewed as an asymptotic version of the "correct" approximating pdf (5.14).

Proof. For a fixed c and Θ, it holds that

$$\begin{aligned}\mathcal{E}[\tilde{\alpha}_c(d(t),\Theta)|d(t-1),\Theta] &\equiv \int f(M_c^*|d(t),\Theta) f(d_t|d(t-1),\Theta)\, dd_t \\ &\underbrace{=}_{\text{chain rule}} \int \frac{f(M_c^*,d_t|d(t-1),\Theta)}{f(d_t|d(t-1),\Theta)} f(d_t|d(t-1),\Theta)\, dd_t \\ &\underbrace{=}_{\text{canceling}} \int f(M_c^*,d_t|d(t-1),\Theta)\, dd_t \\ &\underbrace{=}_{\text{marginalization}} f(M_c^*|d(t-1),\Theta) \equiv \tilde{\alpha}_c(d(t-1),\Theta).\end{aligned}$$

Thus, $\{\tilde{\alpha}_c(d(t),\Theta), d(t)\}_{t\in t^*}$ is a martingale, which is moreover nonnegative and bounded by 1 as $\tilde{\alpha}_c(d(t),\Theta)$ is a probability. Thus, the martingale convergence theorem [81] applies and $\tilde{\alpha}_c(d(t),\Theta)$ converges almost surely to a constant probability. □

The mixture serves mainly as the predictor of the future behavior of the o-system. The assumed invariance of weights may deteriorate its quality substantially. It can be seen on a simple mixture with a well-separated pair of components of similar symmetric shapes and equal weights. For such a mixture, the expected value, which is taken as a good point estimate of the future values d_t, sits in the improbable area between them. The following proposition shows how the problem can be resolved in a generic way that fits into the considered context.

Proposition 5.3 (Mixtures on grouped data) *Let us decompose data sequence $d(\mathring{t})$ into adjacent nonoverlapping groups of a length $n > 1$. Thus, we consider the following probabilistic description of the whole sequence (with a negligible exception of initial and terminal groups of the length n)*

$$f(d(\mathring{t})|\Theta) = \prod_{\tau=1}^{\frac{\mathring{t}}{n}} f(d_{[(\tau-1)n+1]\cdots\tau n}|d((\tau-1)n),\Theta). \tag{5.16}$$

Let the individual parameterized pdfs have the mixture form with constant component weights

$$f(d_{[(\tau-1)n+1]\cdots\tau n}|d((\tau-1)n),\Theta) = \sum_{c\in c^*} \alpha_c f(d_{[(\tau-1)n+1]\cdots\tau n}|d((\tau-1)n),\Theta_c,c). \tag{5.17}$$

Then, the predictor of $d_{\tau n}$ based on $d(\tau n-1)$ (notice the braces!) has data-dependent weights. Specifically,

$$f(d_{\tau n}|d(\tau n-1),\Theta) = \sum_{c\in c^*} \tilde{\alpha}_c(d(\tau n-1),\Theta) f(d_{\tau n}|d(\tau n-1),\Theta_c,c)$$

$$\tilde{\alpha}_c(d(\tau n-1),\Theta) = \frac{\alpha_c f(d_{[(\tau-1)n+1]\cdots(\tau n-1)}|d((\tau-1)n),\Theta_c,c)}{\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c}} f(d_{[(\tau-1)n+1]\cdots(\tau n-1)}|d((\tau-1)n),\Theta_{\tilde{c}},\tilde{c})}. \tag{5.18}$$

Proof. The derived formula is directly implied by the chain rule for pdfs. □

Remark(s) 5.5
Note that we predict the data at the end of the grouping interval. It is sufficient for the assumed low rate of operator interventions for which the predictions serve. The other data records in the group can be predicted similarly, but the very first data record in the group is still predicted with constant weights.

Problem 5.2 (Mixtures at factor level) *The mixture modelling of grouped data indicates that mixtures can be used at various levels of decomposition* *of the pdf $f(d(\mathring{t}))$. For instance, modelling of factors by mixtures could bring additional freedom. For instance, non-normal factors can be approximated or modes living at the factor level can be modelled.*

Problem 5.3 (Shifted and repeatedly used predictors) *Poorer predictions of "earlier" entries in the group can be suppressed by dealing with n mutually shifted predictors or by using the single one working on shifted data. Experiments indicate that these techniques are worth elaborating.*

Problem 5.4 (Rational approximations) *The discussed way of getting data-dependent weights seems to be acceptable. In spite of this, a theoretical and algorithmic solution admitting rational forms instead of finite mixtures would be highly desirable. An attempt presented in Section 13.2 is based on the technique given in [131]. Alternatively, the approximate estimation developed in Section 6.5 could be extended by applying it both to the numerator and the denominator of the "rational" approximation.*

5.4 Design of advisory systems

Here, elements related to the optimization part of the basic design Algorithm 5.1 are presented.

5.4.1 Types of advisory systems

Under the adopted notations and assumptions, we are able to specify a formal description of advisory systems and to distinguish their basic types.

Agreement 5.5 (Fixed and adaptive advisory systems) *The advisory system is a system that constructs the ideal pdf ${}^{\lfloor I}f(d(\mathring{t})|\mathcal{P})$ and presents its projections to the operator. The following types of advisory systems are distinguished according to the exploited experience $\mathcal{P}$.*

The fixed advisory system *constructs the ideal pdf in offline mode using the experience $\mathcal{P} \equiv d(0)$, cf. Chapter 2, obtained before the use of the advisory system in its online mode. Thus, its construction is formally described by the mapping*

$$ {}^{\lfloor U}f(d(\mathring{t})),\, f(\Theta|d(0)) \to \left\{ {}^{\lfloor I}f(d(\mathring{t})|d(0)) \right\}. $$

The pdf ${}^{\lfloor U}f(d(\mathring{t}))$ is the user's ideal pdf obtained by extending the true user's ideal pdf, Section 5.1.4, in the way described in Section 5.1.5. The involved parameter estimate $f(\Theta|d(0))$ is based on the experience collected before online usage of the advisory system. Thus, the fixed advisory system exploits the unguided model $f(d(\mathring{t})|d(0))$ that results from the estimation based on the data produced without the use of the p-system, cf. Agreement 5.3.

The adaptive advisory system *extends its experience during its online use. Thus, its construction is formally described by the sequence of mappings*

$$\left\{ {}^{\lfloor U}f(d_{t+1},\cdots,d_{\mathring{t}}|d(t)),\, f(\Theta|d(t)) \to \left\{ {}^{\lfloor I}f(d_{t+1},\cdots,d_{\mathring{t}}|d(t))\right\}\right\}_{t\in\{0\}\cup t^*}.$$

Thus, the adaptive advisory system exploits the guided model $f(d_t,\cdots,d_{\mathring{t}}|d(t))$ that results from the estimation based also on the data produced with the use of the p-system, cf. Agreement 5.3.

The use of the advisory system in the online mode consists of a sequential presentation of the optimized ideal pdf ${}^{\lfloor I}f(\cdot)$. It is evaluated at measured and (or) contemplated arguments $d_o(t)$ as well as at the measured values $d_{p+}(t),\ t\in t^*$.

As anticipated by conceptual Algorithm 5.1, only some optional parts of the estimated model are optimized. Basic variants are discussed now.

Generally, the p-system may not be aware which quantities in $d_o^*(\mathring{t})$ are operator actions. For instance, the operator can change both pressure and temperature of a managed gas system. The advisory system measures changes of both of them but may have no information on the command button pressed by the operator. Note that such a situation is more frequent in medical or societal applications.

This incomplete knowledge is another important difference that makes the design of the advisory system a very specific decision-making problem. The following agreement singles out the case when this information lack is complete.

Agreement 5.6 (Academic advisory system) *The advisory system designed without knowing which entries of $d_{o;t}$ belong to the action space of the operator $a_o^*(\mathring{t})$ is called the* academic advisory system. *The corresponding design is called the* academic design.

Agreement 5.7 (Industrial and simultaneous advisory systems) *The advisory system designed with knowledge of a nonempty part of the action space of the operator, say $u_o^*(\mathring{t}) \subset a_o^*(\mathring{t}),\ u_o^*(\mathring{t}) \neq \emptyset$, is called the* industrial advisory system.

The o-actions $u_{o;t}$ are called the recognizable actions*; see Fig. 5.1.*

The design of recognizable actions is called industrial design. *The joint academic and industrial design is called* simultaneous design.

Remark(s) 5.6

Obviously, the industrial advisory system is to be optimized through the simultaneous design whenever possible. Sometimes, however, the weights have physical meaning and cannot be influenced by operator's strategy.

5.4.2 Advices as actions of the p-system

During development of the advisory system, basic types of its actions emerged. Here we classify them (a wider discussion follows).

Agreement 5.8 (Nomenclature of actions of the p-system) *The actions available to p-systems*

$$a_{p;t} \equiv (c_t, u_{o;t}, z_t, s_t) \text{ are interpreted as follows.} \tag{5.19}$$

Recommended pointers $\{c_t\}_{t\in t^*}$ *are pointers to the components that are recommended to be kept active at respective time moments. Recommended pointers are* academic advices. *Academic advices cannot be directly communicated to the operator and have to be reflected in the ideal pdf* ${}^{\lfloor I}f(d_{o;t}|d(t-1))$ *offered to him.*

Recommended recognizable actions $\{u_{o;t}\}_{t\in t^*}$ *guide the operator in selecting recognizable actions. These advices result from the industrial or simultaneous designs. The recommended recognizable actions can be interpreted as ordinary inputs of the o-system with the operator serving as an imperfect actuator. Ideally, the recommended recognizable actions should be directly fed into the o-system. This insures that the identical notation is being used for the* <u>recommended</u> *recognizable actions and the recognizable actions.*

Priority actions $\{z_t\}_{t\in t^*}$ *select entries of* $\{d_t\}_{t\in t^*}$ *whose ideal behavior is shown to the operator. These advices result from optimized* assigning priorities. *The priority action* z_t *is* $\mathring{z}$ *vector* ($\mathring{z} \le \mathring{d}_o$) *of differing indexes* $z_{i;t} \in \{1,\ldots,\mathring{d}_o\}$, $i \in \{1,\ldots,\mathring{z}\}$. *The operator gets the marginal pdfs* $\left\{ {}^{\lfloor I}f(d_{z_t;t}|d(t-1)\right\}_{t\in t^*}$ *of the ideal pdf resulting from the previous design.*

Signaling actions $\{s_t \in s^* \equiv \{0,1\}\}_{t\in t^*}$ *stimulate the operator to take some measures when behavior of the o-system significantly differs from the desired one. These advices result from optimized* signaling. *The operator gets probabilistic information whether an intervention is needed or not. It is coded, for instance, as traffic lights.*

5.4.3 Unguided and guided models for respective designs

We assume that the learning part of the advisory system provides a good model $f(d_t|d(t-1))$ of the o-system uninfluenced by the p-system, i.e., the model of the unguided o-system.

A systematic design of the p-system, as described in Algorithm 5.1, requires a model relating its advices to responses of the o-system, i.e., the model of the guided o-system.

The needed but *speculative class of models* is proposed gradually in subsequent sections. The term "speculative" underlines that the model is based on hypothesis that the operator fully cooperates and is able to drive the o-system to the desired state. Possible adverse consequences of this speculation can be suppressed by using an adaptive advisory system with models having advices as an explicit part of experience. Initially, and often permanently, we have to rely on such speculative models as the basis of the systematic academic, industrial and simultaneous design, respectively. Speculative guided models are also used for assigning priorities and signalling.

The consistent transition from the unguided model to the guided one has the following common structure.

The considered data split

$$d_t \equiv d_{p;t} = (\mathit{\Delta}_{p;t}, a_{p;t}) \equiv (\text{p-innovations}, \text{p-actions}) \equiv (\text{p-innovations}, \text{advices}).$$

The corresponding factorization of the unguided model by the chain rule reads

$$f(d_t|d(t-1)) = f(\mathit{\Delta}_{p;t}|a_{p;t}, d(t-1)) f(a_{p;t}|d(t-1)).$$

Assuming (speculating) that the recommended actions and actions realized by the o-system coincide, the first factor on the right-hand side describes the reaction of the o-system on advices. It is given by its physical nature and cannot be changed. The second factor describes the rule of generating $a_{p;t}$. Exactly this rule should be changed by the optimized advising. This gives the general form of the optimized guided model

$$\begin{aligned} {}^{\lfloor I}f(d_t|d(t-1)) &= f(\mathit{\Delta}_{p;t}|a_{p;t}, d(t-1))\, {}^{\lfloor I}f(a_{p;t}|d(t-1)) \\ &= \frac{f(d_t|d(t-1))}{\int f(d_t|d(t-1))\, d\mathit{\Delta}_{p;t}}\, {}^{\lfloor I}f(a_{p;t}|d(t-1)). \end{aligned} \tag{5.20}$$

As discussed in Section 5.4.1, respective designs differ in the available advices, and consequently they lead to different guided models.

Remark(s) 5.7

Note that conditioning used in (5.20) is potential source of computational difficulties. We can say beforehand that for the academic and simultaneous designs the obtained guided models are relatively simple, unlike for the industrial design and assigning priorities. The complexity of signalling related model is somewhere in between these two cases.

5.4.4 Academic design

We have at disposal the multiple-mode model of the unguided o-system $f(d(\mathring{t}))$. As discussed above; see Agreement 5.5, the advisory system maps $f(d(\mathring{t}))$ on an ideal pdf ${}^{\lfloor I}f(d(\mathring{t}))$, whose projections are presented to the operator, preferably in a graphic form.

The joint pdf $f(d(\mathring{t}))$ describes the probability distribution of achievable modes within the data space of the advisory system. Thus, the reachable ideal pdf should be created from these modes. The selection of modes leading to a higher management quality should be advised. The academic design selects the recommended mode through the recommended pointer $c_t \in c^*$ to a particular component (mode) by defining the ideal pdf

$${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t})) \equiv \prod_{t \in t^*} {}^{\lfloor I}f(d_t, c_t|d(t-1)) \equiv \prod_{t \in t^*} f(d_t|d(t-1), c_t)\, {}^{\lfloor I}f(c_t|d(t-1)).$$

The pdfs $f(d_t|d(t-1), c_t),\ c_t \in c^*$ are estimated components and the optional probabilities ${}^{\lfloor I}f(c_t|d(t-1))$ describe the randomized causal strategy

$$\{d^*(t-1) \to c_t \in c^*\}_{t\in t^*} \quad \text{to be designed.} \tag{5.21}$$

Remark(s) 5.8

1. *Usage of the notation* ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ *is a bit inconsistent as* $c(\mathring{t})$ *is a part of* $d(\mathring{t})$. *It helps, however, to focus attention on the discussed advices. Further on, both variants are used. Context should prevent possible misunderstandings.*
2. *The optimized ideal pdf* ${}^{\lfloor I}(d_t|d(t-1))$ *is projected on low-dimensional pdfs on subsets of* $d^*_{o;t}$. *They describe entries of o-data and are also called advisory mixtures. This agreement is used for all basic designs.*

The recommended pointer $c_t \in d^*_{p+;t}$, thus, a projection on d^*_o has to be presented to the operator. The corresponding marginal predictive pdf has the form

$${}^{\lfloor I}f(d_{o;t}|d(t-1)) = \sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|d(t-1))f(d_{o;t}|d(t-1), \Theta_{c_t}, c_t). \tag{5.22}$$

The advising strategy $\left\{ {}^{\lfloor I}f(c_t|d(t-1))\right\}_{t\in t^*}$ is optimized in the sense of the fully probabilistic design; see Section 2.4.2. For that, we have to specify a user *ideal pdf for the interconnection of the o- and p-systems* ${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t})) \equiv {}^{\lfloor I}f(d_t|d(t-1))\, {}^{\lfloor U}f(c_t|d(t-1))$. The constructed ideal pdf ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor I}f(d_t|d(t-1))\, {}^{\lfloor U}f(c_t|d(t-1))$ should be as close as possible to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t}))$.

The recommended pointers c_t belong to $d^*_{p+;t}$ so that the general extension of the user's ideal pdf could be used; see Section 5.1.5. Practical reasons make us to take this part of the user's ideal pdf as tuning knob of the design. For instance, it is reasonable to inhibit fast changes of values of $c(\mathring{t})$ in order to get relatively stable advices given to the operator. This can be achieved by selecting such a user's ideal pdf $\prod_{t\in t^*} {}^{\lfloor U}f(c_t|d(t-1))$ that assigns high probabilities to sequences $c(\mathring{t})$ with small differences $c_t - c_{t-1}$. Also, an offline analysis may discourage operating at some $\underline{\text{dangerous modes}}$ of $f(d(\mathring{t}))$ completely. For that, it is sufficient to restrict support of ${}^{\lfloor U}f(c_t|d(t-1))$ on pointers to the nondangerous components. Such considerations determine the discussed part of the user's ideal pdf

$$\left\{ {}^{\lfloor U}f(c_t|d(t-1))\right\}_{t\in t^*}.$$

With it, the complete user's ideal pdf gets the form

$${}^{\lfloor U}f(d(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor I}f(d_{p+;t}|d(t-1))\, {}^{\lfloor U}f(d_{o;t}|d_o(t-1))\, {}^{\lfloor U}f(c_t|d(t-1)). \tag{5.23}$$

It follows from the discussed extension of the true user's ideal pdf ${}^{\lfloor U}f(d_o(\mathring{t}))$ to ${}^{\lfloor U}f(d(\mathring{t}))$ (see Section 5.1.5) from the chain rule and the fact, that the user can deal with quantities which he is aware of.

This selection of the extended user's ideal pdf completes the formulation of the academic design that makes the general fully probabilistic design formally applicable; see Proposition 2.11. The practical evaluation requires approximations that are discussed in Chapter 7.

As mentioned above, some components might be handled as "dangerous". We take a component as dangerous if a permanent operation on it leads to an unacceptable behavior of the o-system.

Agreement 5.9 (Dangerous component) *Let $f(\Psi|c)$ be the* steady-state pdf *of the data vector Ψ assigned to the permanent activity of the cth component of the mixture*

$$f(\Psi|c) = \int f(\Psi|\tilde{\Psi}, c) f(\tilde{\Psi}|c)\, d\tilde{\Psi}. \tag{5.24}$$

Here, $f(\Psi|\tilde{\Psi}, c) \equiv f(\Psi_t = \Psi|\Psi_{t-1} = \tilde{\Psi}, c)$ is a formal, state-space version of the cth component that describes the evolution of the data vector Ψ_t.

Let us consider average marker *of the form*

$$\frac{1}{\mathring{t}} \sum_{t \in t^*} m(\Psi_t) \tag{5.25}$$

given by a partial quality marker $m(\Psi_t)$. Then, the component c is called dangerous if the probability of a given set $\bar{m}^*$ of non-acceptable values of $m(\Psi)$

$$\int \chi_{\bar{m}^*}(m(\Psi)) f(\Psi|c)\, d\Psi \tag{5.26}$$

is too high. The symbol $\chi_{x^*}(\cdot)$ means indicator of the set x^*.

5.4.5 Industrial design

The industrial design optimizes recommended recognizable actions $u_{o;t}$. Ideally, these actions are directly fed into the o-system and their consequences are predicted by the model of the unguided o-system. They are similar to ordinary inputs of the o-system with the operator serving as an imperfect actuator.

The constructed randomized strategy is described by the pdfs $\left\{ {}^{\lfloor I}f(u_{o;t}|d(t-1)) \right\}_{t \in t^*}$. These pdfs replace $\{f(u_{o;t}|d(t-1))\}_{t \in t^*}$ forming a part of the estimated unguided model $f(d_t|d(t-1))$. Thus, the ideal pdf generated by this design has the form

$$\begin{aligned} {}^{\lfloor I}f(d_t|d(t-1)) &= f(\Delta_t|u_{o;t}, d(t-1))\, {}^{\lfloor I}f(u_{o;t}|d(t-1)) = \qquad (5.27)\\ = {}^{\lfloor I}f(u_{o;t}|d(t-1)) & \frac{\sum_{c_t \in c^*} \alpha_{c_t} f(\Delta_t|u_{o;t}, d(t-1), c_t) f(u_{o;t}|d(t-1), c_t)}{\sum_{c_t \in c^*} \alpha_{c_t} f(u_{o;t}|d(t-1), c_t)}. \end{aligned}$$

The strategy determining the optimal ${}^{\lfloor I}f(u_{o;t}|d(t-1))$ is obtained through the fully probabilistic design, Proposition 2.11. The needed user's ideal pdf

$$ {}^{\lfloor U}f(d(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor U}f(\varDelta_t|u_{o;t}, d(t-1))\, {}^{\lfloor U}f(u_{o;t}|d(t-1)), $$

that includes the target for the recognizable actions, is constructed exactly as described in Section 5.1.5.

Remark(s) 5.9

1. *Note that $u_{o;t}$ belongs to $d^*_{o;t}$, thus it is the only advice that can be* <u>*directly*</u> *presented to the operator.*
2. *The estimated strategy, generating the recognizable actions of the unguided o-system and described by the marginal pdfs $f(u_{o;t}|d(t-1), c_t)$ of individual components, influences the resulting ideal pdf; see (5.27). This effect is specific for nontrivial mixtures, in which the strategies used at various components do not cancel in the inspected conditional pdf (5.27).*
3. *Properties of the components creating the ideal pdf are influenced by the advising strategy $\{ {}^{\lfloor I}f(u_{o;t}|d(t-1))\}_{t\in t^*}$ influencing the recognizable actions $u_o(\mathring{t})$ applied. It may, for instance, convert the dangerous components in nondangerous ones and vice versa.*

5.4.6 Simultaneous academic and industrial design

The simultaneous design optimizes both recommended pointers to components and recommended recognizable actions. It should lead to a better advising strategy than a sequential use of academic and industrial designs. The simultaneous design is surprisingly simpler than the industrial one; see Chapter 7.

The ideal pdf ${}^{\lfloor I}f(d_t|d(t-1))$, generated by the fully probabilistic design, Proposition 2.11, minimizes the KL divergence to the user's ideal pdf

$$ {}^{\lfloor U}f(d(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor U}f(\varDelta_{o;t}|u_{o;t}, d_o(t-1))\, {}^{\lfloor I}f(\varDelta_{p+;t}|u_{o;t}, d(t-1)) $$
$$ \times\, {}^{\lfloor U}f(u_{o;t}|d(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1)), $$

which includes the user's ideal pdf ${}^{\lfloor U}f(u_{o;t}|d_o(t-1))$ for the recognizable actions $u_{o;t}$ as well as the target probability ${}^{\lfloor U}f(c_t|u_{o;t}, d(t-1))$ for the recommended pointers. The latter element is discussed in connection with the academic design, Section 5.4.4. In the simultaneous design, it may depend on $u_{o;t}$, too.

The result is similar to (5.27) with the component weights replaced by the designed probabilities ${}^{\lfloor I}f(c_t|u_{o;t}, d(t-1))$ of recommended actions. Specifically, it holds that

$$
\begin{aligned}
{}^{\lfloor I}f(d_t|d(t-1)) &= {}^{\lfloor I}f(\Delta_t|u_{o;t},d(t-1))\,{}^{\lfloor I}f(u_{o;t}|d(t-1)) \equiv {}^{\lfloor I}f(u_{o;t}|d(t-1)) \\
&\times \frac{\sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|u_{o;t},d(t-1))f(\Delta_t|u_{o;t},d(t-1),c_t)f(u_{o;t}|d(t-1),c_t)}{\sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|u_{o;t},d(t-1))f(u_{o;t}|d(t-1),c_t)}.
\end{aligned}
\tag{5.28}
$$

Remark(s) 5.10
It is worth repeating that the elements in the formula (5.28) with the superscript ${}^{\lfloor I}$ *are optimized. The elements without the superscript* ${}^{\lfloor I}$ *are those obtained through estimation of the unguided o-system. They reflect those parts of operating practice that are not expected to be changed by the advising.*

5.5 Interactions with the operator

The discussed advisory systems can be seen as specific versions of a high-level control system. The ideal pdf ${}^{\lfloor I}f(d(\mathring{t}))$ resulting from the designs discussed above has to be, however, perceived by a human being. This implies a nonstandard task, namely, the optimization of the *presentation of advices.* Essentially, low-dimensional projections of the high-dimensional ideal pdf have to be shown to the operator. This calls for generating additional actions of the p-system caring about interaction of the p-system with the operator. Specifically, it has to be taken into account that

- A few selected quantities can only be shown to the operator including those which should be changed the most urgently in order to minimize risk of malfunctioning or maximize benefit. In other words, presentation priorities have to be dynamically assigned to the o-data.
- The information load on the operator has to be controlled by demanding changes of the o-actions only when needed.
  In other words, signaling that controls dynamically operator's attention has to be designed.

These problems are addressed in Sections 5.5.1 and 5.5.2 under typical advising scenarios.

5.5.1 Assigning priorities

Presentation of quantities worth the operator interest is simple when priorities of critical quantities to be shown are fixed by technological prescriptions.

The situation is also simple if a full question and answer mode of the dialog is adopted. In this case, the operator can ask the p-system:

> What happens to a quantity $d_{i;t}$ if I assign a value $\bar{d}_{j;t}$ to the quantity $d_{j;t}$?

In this case, the optimized (guided) predictive pdf ${}^{\lfloor I}f(d_{i;t}|\bar{d}_{j;t},d(t-1))$ is simply shown. This pdf obtained via marginalization and conditioning from the optimized guided pdf ${}^{\lfloor I}f(d_t|d(t-1))$.

A similar simple case arises when a few recognizable actions $u_{o;t}$ are to be recommended only. Then, the marginal pdfs ${}^{\lfloor I}f(u_{io;t}|d(t-1))$, $i = 1, \ldots, \mathring{u}_o$, of the optimized pdf ${}^{\lfloor I}f(u_{o;t}|d(t-1))$ are evaluated and shown to the operator.

The situation becomes more complex when there is a need to show only marginal pdfs of a few critical quantities contained in the extensive full data record d_t. It leads to the following specific decision problem.

The p-system is given task to generate a $\mathring{z}$-vector of priority actions z_t with entries $z_{i;t} \in \{1, \ldots, \mathring{d}_o\}$, $i \in i^* \equiv \{1, \ldots, \mathring{z}\}$, $\mathring{z} < \mathring{d}_o$. The value of $z_{i;t} = j$ means that recommendations on jth entry of $d_{o;t}$ should be shown to the operator.

Note that the number of shown quantities $\mathring{z}$ has to be small, say 5, in order to respect limited perceiving abilities of human beings.

The ideal pdf ${}^{\lfloor I}f(d(\mathring{t}))$ resulting from academic, industrial or simultaneous designs as well as the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$ are assumed to be fixed and available when designing the presentation strategy $\left\{ {}^{\lfloor I}f(z_t|d(t-1))\right\}_{t \in t^*}$.

Similarly as in the academic design, the target pdf for the adopted fully probabilistic design ${}^{\lfloor U}f(d(\mathring{t}))$ is extended by the factor $\prod_{t \in t^*} {}^{\lfloor U}f(z_t|d(t-1))$. This tuning knob allows us to respect technological preferences and (or) to restrict rate of changes of the quantities selected for the presentation to the operator.

Using the redundant notation $d(\mathring{t}), z(\mathring{t})$ instead of $d(\mathring{t})$ (cf. Remark 5.8) the optimized model, describing the influence of presentation actions, gets the form

$$ {}^{\lfloor I}f(d(\mathring{t}), z(\mathring{t})) = \prod_{t \in t^*} {}^{\lfloor I}f(d_t|z_t, d(t-1))\, {}^{\lfloor I}f(z_t|d(t-1)), \tag{5.29} $$

where the pfs $\left\{ {}^{\lfloor I}f(z_t|d(t-1))\right\}_{t \in t^*}$ describe the randomized *presentation strategy* to be designed.

Let us assume again that the operator cooperates fully when given the ideal pdf for entries of d_t with indexes z_t. Then, the operator is expected to act so that the behavior of the o-system has the distribution with the marginal pdf ${}^{\lfloor I}f(d_{z_t;t}|d(t-1))$, while entries not shown follow the model of the unguided o-system. Here, the pdf ${}^{\lfloor I}f(d_t|d(t-1))$ is the ideal pdf offered by the advisory system that results from the previous optimization.

Thus, the compromise ${}^{\lfloor I}f(d_t|z_t, d(t-1))$ between

- the unguided model of the o-system $f(d_t|d(t-1))$, acting without the p-system, and
- the optimized guided model ${}^{\lfloor I}f(d_t|d(t-1))$, acting fully according to the p-system that presents $d_{z_t;t}$

looks as follows (5.20):

$$ {}^{\lfloor I}f(d_t|z_t, d(t-1)) = f(d_t|d(t-1)) \frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))}. \tag{5.30} $$

In this model, the marginal pdf $f(d_{z_t;t}|d(t-1))$ of the presented quantities $d_{z_t;t}$ — computed from the estimated mixture $f(d_t|d(t-1))$ — is replaced by the corresponding marginal pdf ${}^{\lfloor I}f(d_{z_t;t}|d(t-1))$ gained from the previously designed ideal pdf ${}^{\lfloor I}f(d_t|d(t-1))$, which is generally also mixture. Thus, the model (5.30) is rather complex ratios of mixtures. Consequently, the fully probabilistic design must be approximated in order to get a feasible presentation strategy; cf. Chapter 7.

Moreover, the number of variants to be compared $\binom{\mathring{d}_o}{\mathring{z}}$ during optimization is mostly very large. It motivates us to select $\mathring{z} = 1$ and to use ${}^{\lfloor I}f(z_t = i|d(t-1))$, $i \in \{1, \ldots, \mathring{d}_o\}$ as a degree of the presentation priority assigned to the ith entry of d_t.

Problem 5.5 (Alternative design of presentation priorities) *This part of the design is, a bit surprisingly, the hardest one. It calls for an alternative formulation. For instance, it would be possible to perform design predecessors with alternative fixed choices of presented quantities and then to compare the predicted quality of the guided closed loop behavior. This formulation is worth considering.*

5.5.2 Stimulating the operator

Operator may actively call the p-system for advices. Typically, however, his attention has to be attracted when the state of the managed system requires it. The problem how to attract the operator's attention is addressed here. In modelling of signalling influence, we proceed similarly as in previous sections.

The p-system is given the task to generate actions $\{s_t\}_{t\in t^*}$, $s_t \in \{0,1\}$. The value $s_t = 0$ means that system is in a good state and no extra operator activity is needed. The value of $s_t = 1$ urgently demands operator actions. We search for an admissible *signaling strategy* described by the causal rules

$$\{d^*(t-1) \to s_t^* \equiv \{0,1\}\}_{t\in t^*}\,.$$

The user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$ as well as the ideal pdf ${}^{\lfloor I}f(d(\mathring{t}))$, resulting from academic or industrial or simultaneous design, are assumed to be fixed here. The user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$ is extended to signaling actions $s(\mathring{t})$ by the pf ${}^{\lfloor U}f(s_t|d(t-1))$ reflecting the desired damping of the stimulation.

Stimulation, when respected by the operator,

- results into the guided behavior of the o-system, i.e., $f(d_t|s_t = 1, d(t-1)) \equiv {}^{\lfloor I}f(d_t|d(t-1))$ if $s_t = 1$,
- leaves the o-system unguided, i.e., $f(d_t|s_t = 0, d(t-1)) \equiv f(d_t|d(t-1))$ if $s_t = 0$

Thus, the resulting model of the optimized guided o-system has the form

$${}^{\lfloor I}f(d(\mathring{t}), s(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor I}f(d_t|s_t, d(t-1))\, {}^{\lfloor I}f(s_t|d(t-1)),$$

where the probabilities $\left\{ {}^{\lfloor I}f(s_t|d(t-1)) \right\}_{t\in t^*}$ describe the constructed *signaling strategy.* The influence of the signalling action is described by the model

$${}^{\lfloor I}f(d_t|s_t, d(t-1)) = \delta_{s_t,0} f(d_t|d(t-1)) + (1-\delta_{s_t,0})\, {}^{\lfloor I}f(d_t|d(t-1)).$$

As $s_t \in d^*_{p+;t}$, it cannot be directly provided to the operator. Instead,

$$\hat{s}_t = \mathcal{E}[\delta_{s_t,0}|d(t-1)] = {}^{\lfloor I}f(s_t = 0|d(t-1))$$

is shown, usually, as a colored traffic light. Here, *Kronecker symbol* is used

$$\delta_{s,\tilde{s}} = \begin{cases} 1 \text{ if } & s = \tilde{s} \\ 0 \text{ otherwise} \end{cases}. \tag{5.31}$$

5.6 Design summary

For reference purposes, a review of the proposed models, expressing the expected influence of particular advises, is given, Subsection 5.6.1. The overall design scenario in Subsection 5.6.2 anticipates the decision subtasks to be solved on the way towards the constructed advisory system.

5.6.1 Influence of advices on the o-system

During the discussion of particular designs, the following overall model of the interconnecting of the o-system and the p-system has arisen. Recall, the superscript ${}^{\lfloor I}$ indicates that the corresponding object results from the optimization of advices $a_{p;t} = (c_t, u_{o;t}, z_t, s_t)$. The overall model has the form

$$f(d_{o;t}|a_{p;t}, d(t-1)) \equiv \delta_{s_t,0} f(d_{o;t}|d(t-1)) + (1-\delta_{s_t,0})\, {}^{\lfloor I}f(d_{o;t}|d(t-1))$$

s_t is 0 if operator does not use recommended actions, otherwise s_t is 1.

He reacts on signaling chosen according to ${}^{\lfloor I}f(s_t|d(t-1))$.

The first predictor above is obtained in learning of the unguided system

$$f(d_{o;t}|d(t-1)) \equiv \sum_{c\in c^*} \alpha_c f(d_{o;t}|d(t-1), c) \text{ and} \tag{5.32}$$

$${}^{\lfloor I}f(d_{o;t}|d(t-1)) \equiv f(d_{o;t}|d(t-1)) \frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))}, \text{ where}$$

z_t is the presentation action chosen according to ${}^{\lfloor I}f(z_t|d(t-1))$

$${}^{\lfloor I}f(d_{o;t}|d(t-1)) \equiv {}^{\lfloor I}f(u_{o;t}|d(t-1)) \times$$
$$\times \frac{\sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|u_{o;t}, d(t-1)) f(\varDelta_{o;t}|u_{o;t}, d(t-1), c_t) f(u_o|d(t-1), c_t)}{\sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|u_{o;t}, d(t-1)) f(u_o|d(t-1), c_t)}.$$

The probabilities $\left\{ {}^{\lfloor I}f(s_t|d(t-1)) \right\}_{t\in t^*}$, $s_t \in s^* \equiv \{0,1\}$, describe signaling strategy. They result from the corresponding, fully probabilistic, signaling design. This design is the last in the sequence of respective designs.

The probabilities $\left\{ {}^{\lfloor I}f(z_t|d(t-1)) \right\}_{t \in t^*}$,

$$z_t \in z^* \equiv \left\{ [z_1, \ldots, z_{\mathring{z}}],\ z_i \in \{1, \ldots, \mathring{d}_o\} \right\}$$

describe presentation strategy. They result from the presentation design made after finishing some of the designs listed below.

The pdfs $\left\{ {}^{\lfloor I}f(u_{o;t}|d(t-1)) \right\}_{t \in t^*}$, $u_{o;t} \in u_o^*$, describe strategy generating recommended recognizable actions. They result from industrial or simultaneous design. The industrial design relies on availability of the component weights $f(c_t|u_{o;t}, d(t-1))$. They are obtained either from learning, then $f(c_t|u_{o;t}, d(t-1)) = \alpha_c$, or from the previous academic design, then $f(c_t|u_{o;t}, d(t-1)) = {}^{\lfloor I}f(c_t|u_{o;t}, d(t-1))$. The simultaneous design — when adopted — is the first in the sequence of designs.

The probabilities $\left\{ {}^{\lfloor I}f(c_t|d(t-1)) \right\}_{t \in t^*}$, $c_t \in c^* \equiv \{1, \ldots, \mathring{c}\}$, describe strategy generating academic advices. They result from the academic or simultaneous designs that start the overall sequence of designs.

Remark(s) 5.11

Recall that the adopted models relating advices to the behavior of the guided o-system are of a speculative nature. They serve as the necessary departing point, but the model should be corrected through the use of the p-system: we should estimate model $f(d_t|a_{p;t}, d(t-1), \Theta)$ describing explicitly dependence of data d_t on the adopted advices $a_{p;t}$. For it, the use of the adaptive version of the p-system becomes highly desirable.

5.6.2 Overall scenario and design subtasks

The solution of the overall design problem includes a number of particular steps anticipated in the algorithm below. The algorithm serves us as a design guide. Its steps are discussed in Chapters 6 and 7 at the general level and specialized to specific mixture elements in subsequent Chapters. For fixed advisory system, all steps, except steps 16c and 17, are made in offline mode. For adaptive advisory system, steps 7, 14, 15 have to be run in online mode, too. Note that the recursively performed estimation step 7 is computationally cheap and redesigns 14, 15 can be performed with much slower rate without a significant harm.

Algorithm 5.2 (Design of the advisory system)

1. *Collect the learning data that have to reflect all modes of operating.*
2. *Select quality markers and express their desirable values as the true user's ideal pdf ${}^{\lfloor U}f(d_o(\mathring{t}))$.*
3. *Collect prior physical information on the managed system, operating and measuring conditions.*
4. *Preprocess the data available for learning by*

   a) *grouping data records according to the operator's perceiving and acting rates,*
   b) *reducing dimensionality using expert knowledge that should exclude surely irrelevant record entries,*
   c) *removing outliers,*
   d) *replacing missing data,*
   e) *suppressing high-frequency noise.*

5. *Select the largest acceptable structure of the estimated mixture.*
6. *Select a prior pdf initializing estimation of the mixture describing the o-system.*
7. *Estimate the mixture describing the o-system; if need be, go back to step 6 or even go to step 4.*
8. *Estimate structure of the mixture in its full hierarchy from factors to the whole mixture and, if need be, go back to step 6.*
9. *Reduce dimensionality of the problem by removing data that do not influence even indirectly quantities in* $^{\lfloor U}f(d_o(\mathring{t}))$ *and, if need be, go back to Steps 4 or 6.*
10. *Use physical prior information for individual factors and, if need be, go back to Step 6.*
11. *Validate quality of the obtained model using expert opinion as well as independent testing data. And, if the result is unsatisfactory, repeat whole learning procedure, possibly from Step 3.*
12. *Analyze individual components and qualify recommendable and dangerous ones.*
13. *Select basic advising scenario (academic, industrial or simultaneous).*
14. *Design the advising strategy.*
15. *Design presentation and signaling strategies.*
16. *Validate advising strategy by*
    a) *comparing real actions of the operator with those generated by advising system (without closing the advising loop),*
    b) *judging proximity of good modes in data with the behavior stimulated by a good operator,*
    c) *using advising system at the full scale.*
17. *Use the p-system in the fixed or adaptive mode by feeding the currently measured data into the advising mixture as well as into presentation and signaling strategies. Show the low-dimensional projection of the advising mixture to the operator. The shown quantities are selected by the presentation strategy and the call for o-actions is driven by the signaling strategy.*

6

Solution and principles of its approximation: learning part

Chapter 5 specified all elements needed for the design of an advisory system according to the theory recalled in Chapter 2. This formal solution of the design of the advisory system helps us undoubtedly to clarify the structure of evaluations that should be made. Their practical usefulness is, however, restricted by the "curse of dimensionality". Thus, this conceptual solution has to be complemented by approximate but feasible evaluations. A discussion of their principles for the *learning part of the advisory system* is presented here. The offline mode is predominantly addressed and data are mostly historical.

Specifically, the treated advisory system is based on the parameterized model of the o-system. Its behavior observed by the p-system, grouped according to the rate of operator actions, is described by the mixture (5.9)

$$f(d_t|d(t-1),\Theta) = f(d_t|\phi_{1\cdots\mathring{c};t-1},\Theta) = \sum_{c\in c^*} \alpha_c f(d_t|\phi_{c;t-1},\Theta_c,c). \qquad (6.1)$$

Each parameterized component $f(d_t|\phi_{c;t-1},\Theta_c,c)$ is decomposed into the product of parameterized factors (see Agreement 5.4)

$$f(d_t|\phi_{c;t-1},\Theta_c,c) = \prod_{i\in i^*} \underbrace{f(d_{ic;t}|\psi_{ic;t},\Theta_{ic},c)}_{i\text{th factor}},\ i^* \equiv \left\{1,\ldots,\mathring{d}\right\}, \qquad (6.2)$$

where the regression vectors $\psi_{ic;t}$ are made of $[d'_{(i+1)\cdots\mathring{d}c;t}, \phi'_{c;t-1}]'$. The factors $f(d_{ic;t}|\psi_{ic;t},\Theta_{ic},c)$ are parameterized by individual parameters Θ_{ic}. Collection of these parameters, together with the probabilistic weights of components $\alpha \in \alpha^*$ (5.9), form the multivariate parameter Θ of the mixture.

Parameters, structures of factors and components, as well as the structure of the mixture (see Agreement 5.4) have to be estimated. The estimation is inspected, here.

Section 6.1 prepares common tools used throughout this chapter. First, the considered class of Bayesian estimates is specified. Then, predictors serving for comparison of alternative estimates are presented in Sections 6.1.1 and

6.1.2. The alternatives are generated through versions of branch-and-bound techniques; see Section 6.1.3.

The formal Bayesian estimation is described by Proposition 2.14 with experience formed by the p-data. These data result from preprocessing the raw data observed by the p-system; see discussion in Section 6.2.

Application of the Bayesian paradigm requires specification of a prior pdf $f(\Theta)$. Use of the prior knowledge to this purpose is treated in Section 6.3. Even with such knowledge available, proper specification of the prior pdf is nontrivial. An extensive discussion of promising ways of its construction is described in Section 6.4.

For a chosen $f(\Theta)$, the evaluations of the likelihood function $\mathcal{L}(\Theta, \mathcal{P}_{a^*_{t+1}})$ (2.45) and its integral $\mathcal{I}(\mathcal{P}_{a^*_{t+1}})$ (2.46) represent the main computational burden. The difficulty of the mixture estimation stems from the fact that the likelihood function is a product of sums of pdfs depending on data $d(t)$ and on the unknown parameter Θ. Thus, its formal analytical expression contains a huge number of terms that cannot be handled exactly. For this reason, an approximate treatment is necessary. It is discussed in Section 6.5.

The important structure estimation task is outlined in Section 6.6. Section 6.7 covers model validation. It forms a natural bridge to Chapter 7, which discusses design of the advising and presentation strategies, cf. Section 5.4.

6.1 Common tools

Here, tools used throughout this chapter are prepared. Their description requires specification of learning conditions.

Agreement 6.1 (Considered forms of pdfs on Θ^*) *The prior pdf $f(\Theta) \equiv f(\Theta|d(0))$ and posterior pdf $f(\Theta|d(t))$ defining generalized Bayesian estimates of the mixture (Agreement 5.4) are considered in the common form*

$$f(\Theta|d(t)) = Di_\alpha(\kappa_t) \prod_{i\in i^*, c\in c^*} f(\Theta_{ic}|d(t)), \ t \in \{0\} \cup t^*, \tag{6.3}$$

$$Di_\alpha(\kappa_t) \equiv \mathcal{B}^{-1}(\kappa_t) \prod_{c\in c^*} \alpha_c^{\kappa_{c;t}-1} \chi_{\alpha^*}(\alpha) \equiv \textit{ Dirichlet pdf on}$$

$$\alpha^* \equiv \left\{ \alpha_c, \ \sum_{c\in c^*} \alpha_c = 1 \right\}$$

$$\mathcal{B}(\kappa_t) \equiv \frac{\prod_{c\in c^*} \Gamma(\kappa_{c;t})}{\Gamma\left(\sum_{c\in c^*} \kappa_{c;t}\right)} \equiv \textit{ multivariate beta function, where} \tag{6.4}$$

$$\kappa_t \equiv [\kappa_{1;t}, \ldots, \kappa_{\mathring{c};t}]' \in \kappa^* \equiv \{\kappa_t : \kappa_{c;t} > 0\}.$$

Verbally, parameters Θ_{ic}, $i \in i^* \equiv \{1, \ldots, \mathring{d}\}$, $c \in c^*$, of individual parameterized factors are mutually conditionally independent and independent of *the component weights α. The component weights have Dirichlet distribution $Di_\alpha(\kappa)$ with its support on the* probabilistic simplex α^*.

The Dirichlet distribution $Di_\alpha(\kappa_t)$ is analyzed in Chapter 10 in detail. Here, we only need to know that the expectation assigned to $Di_\alpha(\kappa_t)$ is

$$\mathcal{E}\left[\alpha_c|d(t)\right] = \mathcal{E}[\alpha_c|\kappa_t] = \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c};t}} \equiv \hat{\alpha}_{c;t}. \tag{6.5}$$

6.1.1 Prediction and model selection

The constructed p-system predicts consequences of the observed behavior and operator actions on the future behavior of the o-system. Thus, its performance depends heavily on the quality of the used predictive pdf. Under Agreement 6.1, the value of the predictive pdf of a component c at a possible data vector $\Psi_{t+1} = [d'_{t+1}, \phi'_t]'$, Agreement 5.4, conditioned on measured data $d(t)$ is

$$\begin{aligned} f(d_{t+1}|d(t), c) &\equiv \mathcal{E}\left[\left.\prod_{i\in i^*} f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c)\right| d(t), c\right] \\ &= \prod_{i\in i^*}\int f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c) f(\Theta_{ic}|d(t))\, d\Theta_{ic} \\ &= \prod_{i\in i^*} f\left(d_{ic;t+1}|\psi_{ic;t+1}, d(t), c\right). \end{aligned} \tag{6.6}$$

The overall predictive pdf of a mixture is

$$\begin{aligned} f(d_{t+1}|d(t)) &\equiv \mathcal{E}\left[\left.\sum_{c\in c^*}\alpha_c \prod_{i\in i^*} f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c)\right| d(t)\right] \\ = \sum_{c\in c^*}&\frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c};t}}\prod_{i\in i^*}\int f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c) f(\Theta_{ic}|d(t))\, d\Theta_{ic} \\ \underbrace{=}_{(6.5)} \sum_{c\in c^*}&\hat{\alpha}_{c;t}\prod_{i\in i^*} f(d_{ic;t+1}|\psi_{ic;t+1}, d(t), c). \end{aligned} \tag{6.7}$$

Similarly, we get a predictor for the fixed advisory system. For reference purposes, we summarize these predictors in the following proposition.

Proposition 6.1 (Mixture-based one-step-ahead predictor) *Under Agreement 6.1, the estimation within the adaptive advisory system provides the value of the one-step-ahead predictor at a possible data vector $\Psi_{t+1} = [d'_{t+1}, \phi'_t]'$ (Agreement 5.4) in the form*

$$f(d_{t+1}|d(t)) = \sum_{c\in c^*}\hat{\alpha}_{c;t}\prod_{i\in i^*} f(d_{ic;t+1}|\psi_{ic;t+1}, d(t), c). \tag{6.8}$$

The adaptive predictor *of a factor output* $d_{ic;t+1}$

$$f(d_{ic;t+1}|\psi_{ic;t+1}, d(t), c) \equiv \int f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c) f(\Theta_{ic}|d(t))\, d\Theta_{ic} \quad (6.9)$$

eliminates the unknown parameters of the parameterized factor by its integration weighted by the posterior pdf $f(\Theta_{ic}|d(t))$, *by the posterior factor estimate, Agreement 5.4.*

The component weights are replaced by the expectation (6.5) determined by the statistic κ_t.

In the fixed advisory system, measured data do not enter the condition of the distribution on parameters, i.e., $f(\Theta|d(t)) \equiv f(\Theta|d(0)) \equiv f(\Theta)$. *Thus, the one-step-ahead predictor becomes*

$$f(d_{t+1}|d(t)) \equiv \sum_{c \in c^*} \hat{\alpha}_{c;0} \prod_{i \in i^*} f(d_{ic;t+1}|\psi_{ic;t+1}, c). \quad (6.10)$$

The fixed predictor *of a factor output* $d_{ic;t+1}$

$$f(d_{ic;t+1}|\psi_{ic;t+1}, c) \equiv \int f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c) f(\Theta_{ic})\, d\Theta_{ic} \quad (6.11)$$

eliminates the unknown parameters of the parameterized factor by its integration weighted by the prior pdf $f(\Theta_{ic})$, *by the prior factor estimate, Agreement 5.4 .*

The component weights are replaced by the expectation (6.5) determined by the prior statistic κ_0.

6.1.2 Likelihood on variants

During the learning phase of the p-system construction, the predictors (6.8) or (6.10) serve for selecting the best variant among those differing in initialization, forgetting, structure, etc. Let $d(\mathring{t})$ be learning data and let $v \in v^*$ label a finite collection of such, a priori equally probable, variants. Then, the Bayes rule provides their posterior probabilities

$$f(v|d(\mathring{t})) \propto f(d(\mathring{t})|v).$$

The variants v with high values of the predictive pdf $f(d(\mathring{t})|v)$ evaluated at measured data $d(\mathring{t})$ are obviously preferable. For conciseness, we call the repeatedly evaluated values $f(d(\mathring{t})|v)$ v-*likelihood*s.

For the adaptive advisory system, the chain rule for pdfs implies that the v-likelihood is obtained as the product of one-step-ahead predictors with d_{t+1} equal to the measured data item. This is not true for the fixed advisory system as it holds that

$$f(d(\mathring{t})|v) = \int \prod_{t\in t^*} f(d_t|d(t-1),\Theta,v)\, f(\Theta|v)\, d\Theta$$
$$\neq \prod_{t\in t^*} \int f(d_t|d(t-1),\Theta,v)\, f(\Theta|v)\, d\Theta.$$

The product of fixed one-step-ahead predictors is just an approximation of the v-likelihood corresponding to the fixed advisory system. This approximation is reasonable if the fixed prior estimate $f(\Theta|v)$ of the mixture parameters, Agreement 5.4, has almost one-point support. Otherwise, the adopted approximation may be misleading.

The branch-and-bound algorithm (Section 6.1.3) searches for the highest v-likelihood among a finite set of variants. They are, however, evaluated approximately only. Thus, the maximum value can be blurred by the "approximation noise". Thus, a tool is needed distinguishing whether compared v-likelihood functions are practically equal or not. The following simple test of a hypothesis may serve for this purpose.

Proposition 6.2 (Test on equality of log-likelihoods) *Let us consider a pair of random sequences $l_1(\mathring{t})$, $l_2(\mathring{t})$ of real scalars $l_{1;t}, l_{2;t}$. Let us formulate the following hypotheses*

$$\begin{aligned} &H_0 : l(\mathring{t}) \equiv \big(l_1(\mathring{t}) - l_2(\mathring{t})\big) \sim \mathcal{N}(0, rI_{\mathring{t}}), \quad \textit{with an unknown variance } r > 0,\\ &H_1 : l(\mathring{t}) \equiv \big(l_1(\mathring{t}) - l_2(\mathring{t})\big) \sim \mathcal{N}(m\mathbf{1}_{\mathring{t}}, rI_{\mathring{t}}), \\ &\qquad\qquad \textit{with an unknown mean } m \in (-\infty,\infty) \textit{ and } r > 0, \end{aligned} \tag{6.12}$$

*where $I_{\mathring{t}}$ is $(\mathring{t},\mathring{t})$-*unit matrix *and $\mathbf{1}_{\mathring{t}}$ is $\mathring{t}$-vector consisting of units.*

Let both hypotheses be a priori equally probable and conjugate prior Gauss–inverse–Wishart pdfs $GiW_r(\varepsilon,\varepsilon)$, $GiW_{m,r}(\varepsilon I_2,\varepsilon)$, $\varepsilon > 0$ are chosen; see Chapter 8. Then, for $\varepsilon \to 0$,

$$f(H_1|l(\mathring{t})) = \left[1 + \left(1 - \frac{\bar{l}^2}{\overline{l^2}}\right)^{0.5\mathring{t}}\right]^{-1}, \quad \bar{l} \equiv \frac{1}{\mathring{t}}\sum_{t\in t^*} l_t, \quad \overline{l^2} \equiv \frac{1}{\mathring{t}}\sum_{t\in t^*} l_t^2. \tag{6.13}$$

Thus, for a given probability $\beta \in (0,1), \beta \approx 1$, we take both variants as equivalent iff

$$1 - \left(\frac{\beta}{1-\beta}\right)^{2/\mathring{t}} \geq \frac{\bar{l}^2}{\overline{l^2}}. \tag{6.14}$$

Proof. This is a special case of Bayesian prediction and hypothesis testing related to normal pdfs that are discussed in detail in Chapter 8. The general sufficient statistics are just given a specific form. □

Problem 6.1 (Alternative tests of equality) *Experiments indicate excessive sharpness of the test. Other ways have to be inspected. For instance, assuming asymptotic normality of the sequence* $t^{-0.5}\sum_{\tau=1}^{t} l_\tau$*, the test on the zero mean of the limiting distribution applies. Alternatively, expected value in* H_1 *may be assumed as a* $\mathring{t}$*-vector* m*, or recently proposed general stopping rules [128] can be adopted.*

6.1.3 Branch-and-bound techniques

In subsequent sections, we search for the highest v-likelihood in a relatively complex setting. The search itself can be seen as a version of *branch-and-bound techniques*. This methodology is widely used in optimization, e.g., [132, 133]. It generates sets of alternatives, evaluates values of the optimized functional (function) and then bounds the set of the inspected alternatives.

More formally, let us search for maximum of the functional (function) $F : X^* \rightarrow (-\infty, \infty)$ over X^*. Then, the considered generic algorithm looks as follows.

Algorithm 6.1 (Generic branch-and-bound algorithm)

Initial mode

- *Select the upper bound* $\mathring{n}$ *on the number* n *of iterations.*
- *Set the iteration counter* $n = 1$ *and the initial guess of the maximum value* $\bar{F} = -\infty$.
- *Select the initial set of alternatives* $X_n^* \equiv \{X_{\iota n} \in X^*,\ \iota = 1, \ldots, \mathring{\iota}_n\},\ \ \mathring{\iota}_n < \infty$.
- *Select optional parameters of branching and bounding mappings.*

Iterative mode

1. *Evaluate the values* $F_n^* \equiv \{F(X_{\iota n}),\ X_{\iota n} \in X_n^*\}$ *of* F *on the set of alternatives* X_n^*.
2. *Find* $\bar{F}_n \equiv \max F_n^*$.
3. *Take a maximizing argument* $\bar{X}_n$ *of* F *on* X_n^* *as an approximation of the maximizing argument searched for. Set* $\bar{F}_n = F(\bar{X}_n)$.
4. *Take* $\bar{X}_{n-1}$ *as the maximizing argument of* F *on* X^* *and stop if* $\bar{F}_n \leq \bar{F}$. *Otherwise set* $\bar{F} = \bar{F}_n$ *and go to the next step.*
5. *Branch the set of alternatives* X_n^* *by a* branching mapping $\mathcal{A}$

$$\mathcal{A} : (X_n^*, F_n^*) \rightarrow X_{n+1|n}^* \equiv \{X_{\iota(n+1|n)} \in X^*,\ \iota = 1, \ldots, \mathring{\iota}_{n+1|n}\},\ \mathring{\iota}_{n+1|n} \geq \mathring{\iota}_n. \tag{6.15}$$

6. *Evaluate the values of* F *on the set* $X_{n+1|n}^*$ *of alternatives*

$$F_{n+1|n}^* \equiv \left\{F(X_{\iota(n+1|n)}),\ X_{\iota(n+1|n)} \in X_{n+1|n}^*\right\}.$$

7. *Bound the set of alternatives* $X^*_{n+1|n}$ *by a* bounding mapping $\mathcal{U}$

$$\begin{aligned}\mathcal{U}:\ (X^*_{n+1|n}, F^*_{n+1|n}) &\to X^*_{n+1} \\ &\equiv \{X_{\iota(n+1)} \in X^*,\ \iota = 1,\ldots,\mathring{\iota}_{n+1}\},\ \mathring{\iota}_{n+1} \le \mathring{\iota}_{n+1|n}.\end{aligned} \tag{6.16}$$

8. *Increase the counter* $n = n + 1$. *Take* $\bar{X}_n$ *as the maximizing argument of* F *and stop if* $n > \mathring{n}$. *Otherwise go to the beginning of* Iterative mode.

Proposition 6.3 (Properties of branch-and-bound algorithm) *Let us select* $\bar{X}_n \in \operatorname{Arg\,max} F^*_n$ *and* $\bar{X}_{n+1|n} \in \operatorname{Arg\,max} F^*_{n+1|n}$. *Let* $\bar{X}_n$ *stay in* $X^*_{n+1|n}$ *and* $\bar{X}_{n+1|n}$ *in* X^*_{n+1}. *Then, Algorithm 6.1 may stop at the local maximum only in Step 4. The argument found during the premature break at Step 8 cannot be worse then the initial guess.*

Proof. The statement is implied directly by construction of Algorithm 6.1 and by the assumption that the best alternative is either preserved or even improved during each complete step of the algorithm. □

Remark(s) 6.1

1. *The algorithm may not stop at Step 4.*
2. *Attaining the global maximum is guaranteed when the maximized* F *has a single mode and the algorithm stops at Step 4, i.e., when it avoids the enforced break in Step 8.*
3. *The specific algorithm is obtained by selecting*
   - *the initial set of alternatives* X^*_1,
   - *the branching* $\mathcal{A}$ *and bounding* $\mathcal{U}$ *mappings.*

   Efficiency of this algorithm depends on the adapted choice of these tuning knobs. Section 6.4 specifies the functional F *we maximize and justifies options* X^*_1, $\mathcal{A}, \mathcal{U}$ *we consider as promising.*
4. *Selection of the bounding mapping* $\mathcal{U}$ *is straightforward if we want to respect conditions of Proposition 6.3. Entries* $X^*_{n+1|n}$ *with low values of* F *are simply omitted. Thus, the choice of the number of the preserved arguments* $\mathring{\iota}_{n+1}$ *is the only real option made in connection with bounding mapping* $\mathcal{U}$. *In our application, this number is restricted from above by computational complexity. It is also restricted from below as* X^*_{n+1} *has to provide sufficient number of points for branching. Thus, the choice of the initial set* X^*_1 *and of the problem-adapted branching mapping* $\mathcal{A}$ *decide on the resulting efficiency.*

Problem 6.2 (Critical review of branching mapping) *Choice of the branching mapping is the hardest problem addressed. The options made below could be and should be complemented by "stealing" the most promising ideas occurring in various fields of optimization, for instance, in genetic algorithms [133].*

6.2 Data preprocessing

The advisory system relies on its ability to model significant relationships observable on the o-system; see Agreement 5.1. An attempt to reflect all relationships is usually hopeless as it leads to a hardly manageable model. For this reason, it is necessary to separate rare events and suppress superfluous details from modelling. Data have to be preprocessed. The preprocessing tasks can be classified as follows.

- Data transformation covers:
  - *data scaling* that suppresses numerical problems and simplifies the choice of optional learning and design parameters; for instance, it allows us to standardize prior pdfs;
  - *reduction of dimensionality* that decreases the computational and numerical load by taking into account usual measurement redundancy;
  - *tailoring of the time-scale* that harmonizes the modelling and operator-actions rates.
- Outlier removal excludes rarely occurring data that differ significantly from the behavior of the rest of data. Note that the mixture estimation separates frequently occurring outliers automatically as it assigns them specific components. The outlier preprocessing can often be decomposed into:
  - *outlier detection* when the outlying data items are marked as missing ones;
  - *data interpolation* when missing data are substituted by their estimates.
- Filtering removes parts of data unrelated to the underlying dynamic relationships of interest. The removed parts reflect deficiencies of the measurement process. The noise is suppressed by a variety of signal processing techniques [118]. Typically, differences in frequency content are exploited.

The final aim of the data processing has to be kept in mind while applying it. Any data preprocessing adds a dynamic module into the closed loop to be handled by the advisory system. This simple statement implies that some classical signal-processing techniques, which add a substantial dynamic delay, have to be avoided. For instance, outlier detection is often based on use of nonlinear median or linear moving average filters [134, 135, 136]. In the given context, they can be used only with a narrow windowing. In that respect, the detection methods exploiting dynamically changing boundaries for outliers detection may help [137].

The mixture estimation adds the following strict constraint:

Preprocessing of individual signals must not mix data from various components!

Techniques solving the discussed tasks partially overlap. The Table 6.1 guides in this respect.

Table 6.1. Preprocessing tasks and methods

Tasks	**Methods**	**Sec.**
Data transformation		6.2.1
Data scaling Reduction of dimensionality Tailoring of time scale	Physical range and sample moments Principal component analysis Local filtering	
Outlier removal		6.2.2
Detection Interpolation	Check of physical boundaries, model-based tests, Mixtures of normal and outlying data, model-based interpolation, filter-based interpolation, wavelet-based filtering, local filtering	
Filtering		6.2.3
Removal of well separable noise Removal of poorly separable noise and grouping Noise removal by noncausal filtering	Classical smoothing Local filtering Wavelet-based filtering	

6.2.1 Data transformation

Data scaling

For numerical reasons, it is necessary to scale the processed data to similar numerical ranges. Data scaling and shifting should be done on the basis of the usual ranges of considered quantities. The widespread scaling based on sample moments may be sensitive to outlying data. Use of extreme values for scaling is even less robust.

It is reasonable to deal with scaled data throughout learning and design of the advisory system. Rescaling to the user units is needed just in presentation of the design results to the operator during advising; see Section 9.1.4.

Dimensionality reduction

Principal component analysis (PCA) is one of the most important techniques for dimensionality reduction of strongly correlated multidimensional data [17]. The standard PCA assumes implicitly or explicitly unimodal multivariate normal distribution. The data clustering techniques assume that data are generated by a mixture of (normal) distributions. So, a direct use of PCA is slightly illogical. It would be better to fit a mixture in the original data space and then apply PCA to individual normal components. We need, however, to reduce the dimensionality before clustering in order to improve it. A sort of data presegmenting provides a compromise between these contradictory needs. Sometimes, it is possible in the learning phase when the processed data can be segmented into unimodal classes. PCA then can be applied separately to each of them; cf. Section 6.7.1.

Alternatively, mixtures with components having low-rank expectations could be and should be considered. The idea of functional approximation [36, 38] that approximates Bayesian estimation by minimizing of the KL divergence over a suitable set of approximating pdfs seems to be proper way for doing this task.

Problem 6.3 (PCA on components) *Feasibility and real influence of the discussed ways of PCA application on components should be studied in a quantitative way.*

Tailoring of the time scale

The discrete-time models are built for the sampling rate given by the highest rate feasible by the human being. Data are mostly collected with higher rates. Thus, a representative has to be searched for a group of measured data. This can be done with help of a *local filter* [138, 139]. Its use allows us, moreover, to suppress high-frequency noise and also partially suppress outliers. The skipping of data, often recommended, falls into this category, but it is obvious that it loses precious information.

6.2.2 Outlier removal

Outlier removal consists of interconnected detection of outliers and interpolation of data they hide.

Outlier detection based on physical boundaries

Mostly, the inspected data d_t have a specific physical meaning with well defined boundaries $\underline{d}$, $\overline{d}$ on their usual values, $\underline{d} \le d_t \le \overline{d}$.

The censored data that do not fall in this hypercube can be immediately interpolated or marked as missing for a postponed interpolation. Often, the ranges on signal-rate changes are also available. Then, a similar hypercube on temporal changes of data can be defined and exploited in the same way.

Model-based detection

Comparison of the current signal value with a value predicted by a suitable model is the efficient method for detection of an unexpected value. The detection discussed in the previous paragraph is essentially of this type with physically justified expected characteristics of the inspected signal.

Comparison of the output of a fixed filter with the value in question rests on the assumption that the normal signal passes the filter without distortion, on the assumption that the filter models its course properly. Of course, the model can be built explicitly and even recursively estimated.

The simplest, most widely used technique checks the inequality $\sigma_{i;t}h \leq |d_{i;t} - \mu_{i;t}|$. In it, $\mu_{i;t}$ and $\sigma_{i;t}$ are the sample mean or median and the standard deviation of measurements $d_{i;t-\overset{\circ}{k}}, \ldots, d_{i;t+\overset{\circ}{k}}$, respectively. The size of the window is $2\overset{\circ}{k}+1$ and h is an optional threshold. The outlier is detected if the inequality is satisfied. This is the noncausal version of the detector. The median and standard deviation are calculated from $d_{i;t-\overset{\circ}{k}}, \ldots, d_{i;t-1}$ in its causal counterpart.

The two plots in Fig. 6.1 show the outlier removal for artificial three-dimensional data. The discussed outlier removal technique was used. The data $d_{i;t}$ marked as outliers are interpolated by $\mu_{i;t}$.

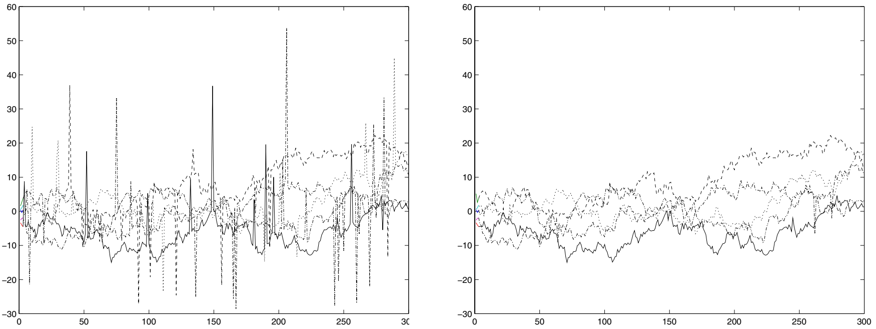


Fig. 6.1. Outlier removal based on checking deviations from a sample mean computed on a moving window. Shown are raw and processed data.

Mixtures of normal and outlying data

Rare outliers can be processed at the level of an individual signal. Frequently occurring outliers fed into a mixture estimation create naturally extra component(s). The overall mixture estimation finds the component to which the current signal value belongs. Thus, we can assume that the processed signal comes from a single "global" component and to model it by a local simple, typically two-component mixture. This model is recursively estimated using quasi-Bayes estimation with forgetting; Section 6.5. Again, the estimation detects the component to which the inspected value belongs. It detects essentially whether the value is an outlier or not and uses it for correction of the corresponding parameter estimates. Data interpolation is then straightforward. The outlier is replaced by the value predicted by the component corresponding to "normal" course of the treated signal. Thus, the model-based interpolation is used.

Success of the specific application of this idea depends heavily on a good decision of whether the signal belongs to a single global component. A careful selection of forgetting factors used for estimation of the local mixture is even more important. The component describing the "normal" course has to be highly adapted in order to use effectively a local model that modifies the normal signal as little as possible. The component describing the outliers should be adapted very little; otherwise there is danger that the separation of both modes will be lost.

Model-based interpolation

This method replaces a missing data item by the point prediction made by a simple model fitted to the outlier-free data. The replacement of the invalid data item by the last measured value is its practically attractive variant. It adds small filtering dynamics. The assumption that the basic signal varies slowly represents the model behind this variant.

Generally, model-based methods are strongly application-domain dependent. For example, the paper [140] describes interesting, model-based technique for data interpolation of a degraded audio signal. Another method of the model-based interpolation, as well as the dynamic-boundaries test, are described in [137].

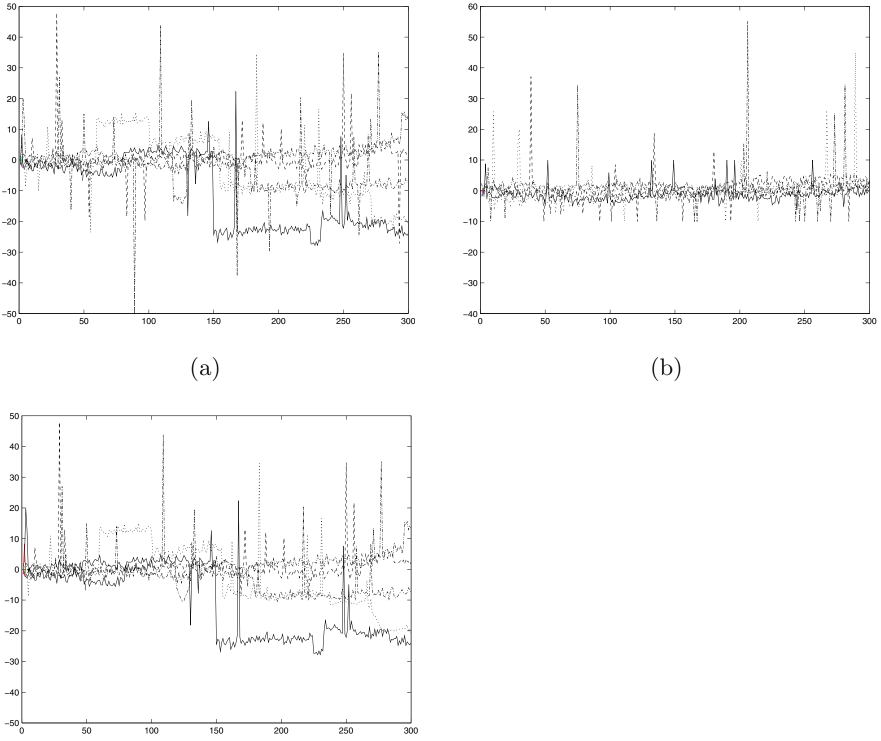
Filter-based interpolation

This type of interpolation generates approximation of an actual data item by classical smoothing filters.

The three plots in Fig. 6.2 illustrate the interpolation of missing data. The artificial data with outliers are processed. The raw data are in the first plot. The data out of the given physical boundaries are marked as missing. The lower bound $\underline{d} = -10$ is considered only. In the second plot, the data marked as missing are set to be equal to the value of this boundary. At the third plot, the data are interpolated by a filter-based smoothing.

6.2.3 Filtering

High-frequency measurement noise is a typical disturbing element that should be removed. There are a lot of techniques available to this purpose. They are excellent if the frequency content of the removed noise differs substantially from that of the recovered signal. Otherwise, they add often substantial dynamic delay that is harmful in the closed loop.



(a)

(b)

(c)

Fig. 6.2. Interpolation of missing data. Shown are (a) raw data, (b) data cut at a given lower bound, and (c) smoothed data.

Local filters overcome this problem. They are applicable whenever grouping of data is desirable. In offline mode even noncausal filtering is acceptable. Then, the wavelet filters are a powerful tool for removing selected, typically high, frequencies.

Classical filters for smoothing

The classical smoothing filters remove the high-frequency noise. The mean and median filters are justified whenever we can assume that the system behaves for a long time period according to a single component. Forgetting or moving-window versions can be employed. The mean filter calculated on a moving window is defined $x_{i;t} = \frac{1}{\mathring{k}} \sum_{k=1}^{\mathring{k}} d_{i;t-k+1}$, where $x_{i;t}$ is the filter output, and $\mathring{k}$ is the window size. The *median filter* on the moving window is defined $x_{i;t} = \mathsf{median}(d_{i;t}, d_{i;t-1}, \ldots, d_{i;t-\mathring{k}+1})$, where the function $\mathsf{median}(\cdot)$ evaluates the median of the argument. The *mean filter* with forgetting is generated recursively $x_{i;t} = \lambda_i x_{i;t-1} + (1-\lambda_i) d_{i;t}$ and it is determined by the forgetting factor $\lambda_i \in (0,1)$. The initial filtered value is set $x_{i;1} = d_{i;1}$. The filtered value is the weighted mean of all previous values of $d_{i;t-k+1}$

$$x_{i;t} = \sum_{k=1}^{t-1} \rho_{ik;t} d_{i;t-k+1}.$$

The weights are $\rho_{ik;t} = (1-\lambda_i)\lambda_i^{k-1}$, except for $\rho_{i1;t} = \lambda_i^{t-1}$. Notice that $\sum_{k=1}^{t} \rho_{ik;t} = 1$.

Analogy to the mean filter with forgetting leads to the median filter with forgetting defined by $x_{i;t} = \mathsf{wmedian}(d_{i;t}, d_{i;t-1}, \ldots, d_{i;t-\mathring{k}+1}; \rho_{i1;t}, \ldots, \rho_{i\mathring{k};t})$, where $\mathsf{wmedian}$ denotes the median of data weighted by $\rho_{i1;t}, \ldots, \rho_{i\mathring{k};t}$, [141].

Local filters

Local filters are based on the assumption that the underlying signal does not change too quickly and on the possibility to group several data [138]. Then, the signal can be modelled by a simple function, say by a straight line, within the timespan determined by a single period of operator actions. The redundant data are used for estimating this function whose single value is offered as the grouped and filtered value. Models of the corrupting noise (light-tailed Gaussian, mixture of a pair of Gaussian pdfs differing in noise variance, heavy-tailed Cauchy, one-sided exponential, etc.), fed into the Bayesian estimation and prediction, provide directly the required estimate. The piecewise application of the filter guarantees that a small dynamic delay is introduced. It is at most a single period at the operator rate. Irregularities in the original sampling can be suppressed by choosing evenly distributed time moments at which the filtered samples are evaluated.

Problem 6.4 (Outlier cancelling filters) *Local regression with heavy-tailed noise seems to be the proper direction inspected. Obsolete filters of this type [142] should be renewed using advanced techniques for online nonlinear filtering; see Section 3.4.*

Wavelet filtering

Wavelet transformation converts original data into different levels of resolution with respect to their spectrum. By manipulating the gained spectra, high frequencies can be suppressed without removing "edges" from the original signal. This noncausal filter can be used in the learning phase only.

Fig. 6.3 shows an example, of wavelet de-noising applied to artificial data containing significant edges (steps). Data are processed for each channel $d_i(\mathring{t})$ separately. The figure shows a single channel processed. The Haar wavelet is employed. The original signal is analyzed into wavelet coefficients, Fig. 6.4,

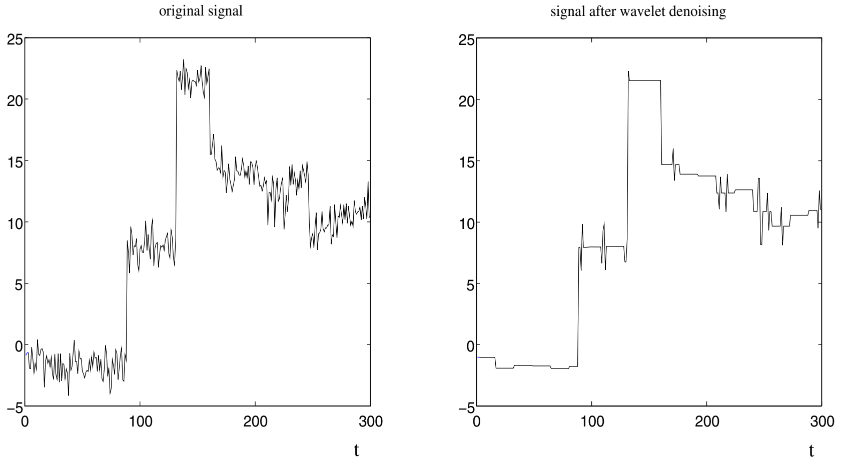


Fig. 6.3. Wavelet de-noising of a piecewise smooth signal. Raw and de-noised data are shown.

by the discrete wavelet transformation [143]. H_0 denotes the low-pass filter operator and H_1 the high-pass filter operator and the symbol ↓2 denotes down sampling at the rate of 2:1. Thereafter, the smallest coefficients are suppressed

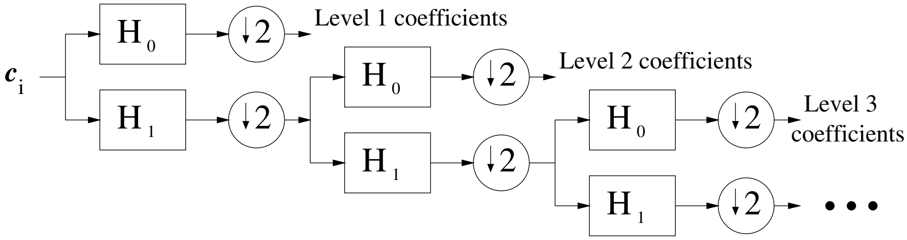


Fig. 6.4. Multilevel Haar-wavelet decomposition of the signal by the filter bank. The notation ↓ 2 denotes down-sampling and H_0, H_1 mark filter elements.

— using an appropriate threshold — and the signal is reconstructed by a synthesis process, which is the inverse to that shown in Fig. 6.4.

Problem 6.5 (Comparison of quality of filters) *Mostly, preprocessing reduces effects observable on reality. It makes conceptually hard to compare quality of competitive filters. Models obtained on differently filtered data surely cannot be compared on them.*

Probably, the performance on predictions of raw data conditioned on filtered data should be compared. It is reasonable when preprocessing is "mild", but it is not clear whether such a comparison is fair; for instance, when removing outliers. Obviously, a solution of this problem is of a direct practical interest.

Remark(s) 6.2
Recently a novel filtering technique based on a mixture whose components differ just in filtering used has been proposed [144]. It seems to be an efficient and promising way to merge a bank of filters while decreasing sensitivity to the choice of filter parameters.

6.2.4 Filters generating factors in an exponential family

The algorithms presented in this text served for creating an efficient software basis [66] that can be simply tailored to a specific application. The original software version was predominantly oriented towards *normal ARX factor*s (auto-regressive models with external inputs) and Markov-chain factors. They both belong to the dynamic exponential family, Section 3.2, Agreement 3.1, that provides the practically feasible factors. A rich supply of other feasible factors is obtained from them by data transformations $d \rightarrow \tilde{d}$ and filtering, possibly applied also to regressors forming the regression vector ψ. This Section discusses and classifies possible ways of generating them in order to allow their systematic implementation.

The ith parameterized factor, predicting the factor output $d_{i;t}$ and belonging to the dynamic exponential family is described by the p(d)f

$$\begin{aligned} f(d_{i;t}|d_{(i+1)\cdots \mathring{d};t}, d(t-1), \Theta) &\equiv f(d_{i;t}|\psi_{i;t}, \Theta) \\ &= A(\Theta)\exp\left[\langle B(\Psi_{i;t}), C(\Theta)\rangle + D(\Psi_{i;t})\right]. \end{aligned}$$

Two ingredients determine it.

1. The filter $F(\cdot)$
$$[\Psi_{i;t}, \Omega_t] = F\left(d_{i\cdots \mathring{d};t}, \Psi_{i;t-1}, \Omega_{t-1}\right) \quad (6.17)$$
updating recursively the filtered data vector
$$\Psi'_{i;t} \equiv [\tilde{d}_{i;t}, \psi'_{i;t}] \equiv [\textit{filtered factor output}, \text{filtered regression vector}].$$
The symbol Ω_t denotes a finite-dimensional *filter state.*
2. The elements in the functional form of the exponential family,
$$\begin{aligned} \langle\cdot,\cdot\rangle &\equiv \text{a functional, linear in the first argument,} \quad (6.18)\\ A(\Theta) &\equiv \text{ a nonnegative scalar function defined on } \Theta^*,\\ B(\Psi),\ C(\Theta) &\equiv \text{compatible array functions defined on } \Psi^*,\ \Theta^*, \text{ and}\\ D(\Psi) &\equiv \text{scalar function on } \Psi^*. \end{aligned}$$

The intended use of the filtered data vectors in prediction and design of actions imposes constraints on the admissible filters.

Requirement 6.1 (Allowed filters) *The filter (6.17)*

1. *determines the filtered regression vector ψ_t independently of the value $d_{i;t}$;*

2. *allows us to reconstruct uniquely the original* $d_{i;t}$ *when* $\tilde{d}_{i;t}$ *and* $\psi_{i;t}$ *are given;*
3. *allows us to reconstruct uniquely the original actions among* $d_{i+1;t}, \ldots, d_{\mathring{d};t}$ *when* $\psi_{i;t}$ *is given.*

A limited supply of the practically available filters (6.17) and elements (6.18) permits us to label them by the *data vector type* $\equiv {}^{\dagger}\mathit{\Psi} \in {}^{\dagger}\mathit{\Psi}^{*} \equiv \{1, \ldots, {}^{\dagger}\mathring{\mathit{\Psi}}\}$ and by the *functional form type* $\equiv {}^{\dagger}F \in {}^{\dagger}F^{*} \equiv \{1, \ldots, {}^{\dagger}\mathring{F}\}$.

The following allowed filters are simply available.

Allowed filtering of data vectors

- State ϕ_t in the phase form; see Agreement 5.4.
  The time-invariant filter state $\mathit{\Omega}$ is the list determining the structure of the data vector constructed from the state in the phase form.
- State ϕ_t in the phase form transformed by a fixed affine mapping.
  Scaling factor, additive term and structure of the data vectors form the time-invariant filter state $\mathit{\Omega}$.
- Generalized ARX models [69] with a transformed factor output obtained from a one-to-one, unknown-parameter-free transformation of the factor output.
  The filter state is in the phase form and the Jacobian of the discussed transformation is independent of unknown parameters.
  The log-normal factors, Section 8.1.6, with the state in the phase form (see Agreement 5.4) serves as an important example of this type. Generally, this simple class increases the modelling power surprisingly.
- *ARMAX factors*, an extension of ARX factor by a known moving average (MA) model of the noise, is converted to an ARX model by time-varying prewhitening [145].
  The filter state $\mathit{\Omega}_t$ contains the time-invariant description of the noise correlations, i.e., both structure and parameters. It also contains the time-varying filter resulting from it; see [145].
  The factor is equivalent to a special state-space model that leads to a special Kalman filtering [146], which can be combined with the parameter estimation without an approximation.
- A mixture of ARMAX factors of the above type with a common ARX part [144].
  The filter state $\mathit{\Omega}_t$ contains several different $\mathit{\Omega}_{j;t}$, each of the type described in the previous item.
- A mixture of ARX factors with the common ARX part and components differing in filters applied to the original state in the phase form.
  The filter state $\mathit{\Omega}_t$ contains information on the structure and time-invariant parameters of the underlying filters.

Available functional forms within the exponential family

- Normal factors, Chapter 8.

- Markov-chain factors, Chapter 10.
- MT normal factors, Chapter 11.
- *Static factors* in the exponential family.

Both these lists can surely be extended. For instance, specific incremental versions of the state in the phase form may be found as a useful specific instance. Nonlinear/non-Gaussian factors embedded in the exponential family in the way described in Section 3.4 serve as another example.

6.2.5 Statistics for the exponential family

The Bayesian estimation in the exponential family is described by Proposition 3.2. We use conjugate prior pdfs. It has the form

$$f(\Theta) \equiv f(\Theta|d(0)) \propto A(\Theta)^{\nu_0} \exp\left[\langle V_0, C(\Theta)\rangle\right].$$

The estimation is combined with stabilized forgetting, Algorithm 3.1. It is specified by the forgetting factor $\lambda \in [0,1]$ and alternative pdf ${}^{\lfloor A}f(\Theta|d(t))$. We select it as a conjugate pdf, too

$${}^{\lfloor A}f(\Theta|d(t)) \equiv {}^{\lfloor A}f(\Theta) \propto A(\Theta)^{\,{}^{\lfloor A}\nu_t} \exp\left[\left\langle {}^{\lfloor A}V_t, C(\Theta)\right\rangle\right].$$

In the discussed context, it is mostly chosen as a time invariant pdf. So that, without substantial restriction of generality, we can assume that the formal structure of its updating coincides with that of the factor statistics in question.

With the options made, the posterior pdf then has conjugate form and its statistics ν_t, V_t can be updated recursively. Approximate estimations, Section 6.5, modify these recursions, giving a weight $w_t \in [0,1]$ to increments of these statistics. The compound recursions, also describing evolution of the data vector, have the following form. The first part reflects the data updating; cf. Proposition 2.13:

$$\begin{aligned}
&\text{Filtering: } [\Psi_t, \Omega_t] = F(d_t, \Psi_{t-1}, \Omega_{t-1}),\ \Psi_0,\ \Omega_0 \text{ given} && (6.19)\\
&\text{Weighted data updating: } \tilde{V}_t = V_{t-1} + w_t B(\Psi_t), \quad \tilde{\nu}_t = \nu_{t-1} + w_t\\
&\quad V_0,\ \nu_0 \text{ are chosen as a description of the prior pdf,}\\
&\text{Filtering of the alternative: } \left[{}^{\lfloor A}\Psi_t,\ {}^{\lfloor A}\Omega_t\right] = {}^{\lfloor A}F(d_t,\ {}^{\lfloor A}\Psi_{t-1},\ {}^{\lfloor A}\Omega_{t-1})\\
&\quad {}^{\lfloor A}\Psi_0,\ {}^{\lfloor A}\Omega_0 \text{ given}\\
&\text{Alternative data updating: } {}^{\lfloor A}V_t = {}^{\lfloor A}V_{t-1} + {}^{\lfloor A}w_t\, {}^{\lfloor A}B\left({}^{\lfloor A}\Psi_t\right)\\
&\quad {}^{\lfloor A}\nu_t = {}^{\lfloor A}\nu_{t-1} + {}^{\lfloor A}w_t,\ {}^{\lfloor A}V_0,\ {}^{\lfloor A}\nu_0 \text{ characterize a prior alternative.}
\end{aligned}$$

The time updating closes the recursion by the stabilized forgetting formula, Algorithm 3.1,

$$V_t = \lambda \tilde{V}_t + (1-\lambda)\, {}^{\lfloor A}V_t, \quad \nu_t = \lambda \tilde{\nu}_t + (1-\lambda)\, {}^{\lfloor A}\nu_t. \qquad (6.20)$$

The externally supplied weights w_t, ${}^{\lfloor A}w_t$, the functions $B(\cdot)$, ${}^{\lfloor A}B(\cdot)$ and updating of data vectors Ψ_t, ${}^{\lfloor A}\Psi_t$ may differ. Both values $B(\cdot)$, ${}^{\lfloor A}B(\cdot)$ have to be compatible with the commonly considered array function $C(\Theta)$. This can be checked when preparing the time updating.

Thus, the complete description of respective factors can be composed of the above elements. Let us make it formal.

Agreement 6.2 (Description of factors in the exponential family) *The factor in the exponential family is described by*

$$\left({}^{\dagger}\Psi,\ {}^{\dagger}F, \Omega_\tau, \Psi_\tau, V_\tau, \nu_\tau\right) \equiv (\textit{filter type, functional type, filter state,}$$
$$\textit{data vector, } V\textit{-statistics,degrees of freedom}) .$$

The initial values, $\tau = 0$, are stored for application of the iterative learning, Section 6.4.1, in addition to the current values with $\tau = t$.

The used filter has to meet Requirement 6.1.

6.2.6 Prediction in EF with statistics gained by filtering

The term $D(\Psi)$ occurring in the definition of the exponential family does not influence estimation. Also, the estimation need not refer to the original data and can completely use the filtered versions of data vectors. For prediction, however, both these aspects are significant.

Proposition 3.2 implies that the prediction of the ith filtered factor output has the form (3.8)

$$f(\tilde{d}_{i;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1)) = \frac{\mathcal{I}(V_{i;t-1} + B(\Psi_{i;t}), \nu_{i;t-1} + 1)}{\mathcal{I}(V_{i;t-1}, \nu_{i;t-1})} \exp[D(\Psi_{i;t})], \text{ with}$$
$$\mathcal{I}(V, \nu) \equiv \int A(\Theta)^{\nu} \exp\left[\langle V, C(\Theta)\rangle\right] d\Theta. \qquad (6.21)$$

Proposition 2.5 on transformation of random quantities gives

$$f(d_{i;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1)) = J(\Psi_{i;t})\frac{\mathcal{I}(V_{i;t-1} + B(\Psi_{i;t}), \nu_{i;t-1} + 1)}{\mathcal{I}(V_{i;t-1}, \nu_{i;t-1})} \exp[D(\Psi_{i;t})]$$
$$J(\Psi_{i;t}) \equiv \frac{\partial F(d_{i\cdots\mathring{d};t}, \Psi_{i;t-1}, \Omega_{i;t-1})}{\partial d_{i;t}}. \qquad (6.22)$$

This prediction serves for presenting predictions to the operator as well as for computing v-likelihood values (see Section 6.1.2) used for comparing of various estimation variants. These v-likelihood values can be used even for comparison of different filters applied to the same signal. This hints how to solve Problem 6.5.

6.3 Use of prior knowledge at the factor level

The design of the advisory system relies mostly on vast amount of informative data, but a strong prior knowledge may improve the quality of the resulting model. This opportunity should not be missed and its inspection forms the core of this section.

We deal with prior knowledge at the factor level: a prior pdf for each factor is considered as an entity on its own. This fits well the product form of the considered pdfs; see Agreement 6.1.

Computational constraints force us to use predefined functional forms of prior estimates of factors. Mostly, they are taken as conjugate prior pdfs (3.13). In this case, values of statistics determining them serve for expressing the prior knowledge.

Initially, the available knowledge pieces are translated into a common basis of *fictitious data*, i.e., the data observed or potentially observable by the advisory system. Fictitious data are split into internally consistent data blocks. Each of them is processed by the Bayes rule with *forgetting*; see Section 3.1. Then, the estimation results are merged. The merging task is nontrivial due to the unknown, possibly not fully consistent, relationships of the processed data blocks. The particular *knowledge sources* are below labelled by $K_k,\ k \in k^* \equiv \{1, \ldots, \mathring{k}\}$.

6.3.1 Internally consistent fictitious data blocks

Some sources of knowledge provide knowledge directly in the form of data blocks. These blocks include obsolete data, data from identification experiments differing from usual working conditions — like measurement of step responses — and data gained from a realistic simulation. Section 6.3.2 discusses other sources of knowledge that have to be "translated" into data blocks. The common expression as data blocks allows us to deal with all of them in a unified way.

We sort the knowledge sources into groups whose members provide *internally consistent data blocks.*

Agreement 6.3 (Internally consistent data blocks) *The data block D_K reflecting the knowledge piece K is called internally consistent iff $f(\Theta|K)$ can be expressed as the result of Bayesian estimation, possibly with appropriately chosen forgetting; see Section 3.1. The estimation inserts data D_K into the considered parameterized model and starts from a flat* pre-prior pdf *$\bar{f}(\Theta)$.*

The adequate forgetting should always be used in order to counteract undermodelling. This is of an extreme importance for simulators that may provide a huge number of data and make overfitting of the posterior pdfs possible.

Algorithm 6.2 (Processing of internally consistent data blocks)

Initial mode

- *Select the structure of the considered factor.*
- *Specify, usually very flat, pre-prior pdf* $\bar{f}(\Theta)$.
- *Select the set of alternative forgetting factors* $\lambda \in \lambda^* \equiv \{\lambda_1, \ldots, \lambda_{\mathring{\lambda}}\}$, $\lambda_i \in (0, 1]$. *Typically,*
$$\lambda_i = 1 - 10^{-i}, \; i = 1, 2, 3, 4, 5, 6. \tag{6.23}$$

The choice is implied by the knowledge that the effective memory is $(1 - \lambda_i)^{-1}$ *and by the wish to cover a wide time span with a few grid points.*

Processing mode

1. *Perform Bayesian estimation and prediction starting from* $\bar{f}(\Theta)$ *and using the data block* D_K *for all* $\lambda \in \lambda^*$, *i.e., obtain* $f(\Theta|K, \lambda_i) \equiv f(\Theta|D_K, \lambda_i)$ *and v-likelihood* $f(d(\mathring{t})|K, \lambda_i)$, $i \in \{1, \ldots, \mathring{\lambda}\}$ *evaluated on real data* $d(\mathring{t})$.
2. *Choose* $f(\Theta|K) \equiv f(\Theta|K, \bar{\lambda})$, *where* $\bar{\lambda}$ *is a point estimate of the forgetting factor within the set* λ^*. *The value* $\bar{\lambda}$ *is mostly taken as the maximizing argument of the obtained v-likelihood* $f(d(\mathring{t})|K, \lambda_i)$.

Remark(s) 6.3

1. *The construction is applicable if the data* D_K *allow us to evaluate the value of the parameterized model* $f(d|\psi, \Theta)$ *at least for a single data vector* $\Psi = [d, \psi']'$.
2. *Selection of the best forgetting factor can be completely avoided if we take the pairs* (K, λ_i), $i \in i^*$, *as different knowledge items and merge them as described in Section 6.3.3. The weighting used there distinguishes differences in quality implied by differences in forgetting.*

6.3.2 Translation of input-output characteristics into data

Often, guessed or designed input-output characteristics of the modelled system are available. Static gain, a point on frequency response, and time-constants serve as examples. The majority, if not all, have an experimental basis. They describe the expected response of the system to a given stimulus. Such knowledge item K gives characteristics of the predictor

$$f(d|\psi) \equiv f(d|\psi, K) \equiv \int f(d|\psi, \Theta) f(\Theta|K) \, d\Theta \tag{6.24}$$

for a scalar d (a factor is dealt with) and a fixed regression vector ψ. Mostly, some moments of the predictive pdf can be guessed. Formally, we assume the knowledge piece in the form

$$h(\psi) = \int H(d, \psi) f(d|\psi) \, dd, \tag{6.25}$$

where $h(\psi)$ and $H(\Psi) \equiv H(d,\psi)$ are known vector functions. Typically, we know

$$\hat{d} = \int d f(d|\psi)\, dd, \qquad r_d = \int (d-\hat{d})^2 f(d|\psi)\, dd. \tag{6.26}$$

The case (6.26) corresponds to the knowledge $h(\psi) = \left[\hat{d}, r_d\right]$, $H(d,\psi) \equiv H(\Psi) = \left[d, (d-\hat{d})^2\right]$.

If no pdf $f(\Theta|K)$ fulfilling (6.25) exists then this information source cannot be internally consistent (see Agreement 6.3), and has to be split into internally consistent parts. However, as a rule, the knowledge item (6.24), (6.25) does not determine the constructed pdf $f(\Theta|K)$ completely. It is reasonable to construct a pdf $f(\Theta|K)$ that expresses just the considered information item. Thus, it makes sense to choose such a $f(\Theta|K)$ that is the nearest one to the flat pre-prior pdf $\bar{f}(\Theta)$. The prior estimate $f(\Theta|K)$ is searched for among all pdfs

$$f(\Theta|K) \in f_K^* \equiv \{\text{pdfs fulfilling (6.24), (6.25)}\}. \tag{6.27}$$

It motivates our option of $f(\Theta|K)$ respecting (6.24) and (6.25) as

$$f(\Theta|K) \in \text{Arg}\min_{f\in f_K^*} \mathcal{D}(f||\bar{f}). \tag{6.28}$$

Proposition 6.4 (Knowledge of input-output characteristics) *The minimizing argument $f(\Theta|K)$ of (6.28) respecting the constraint (6.25) has the form*

$$f(\Theta|K) = \frac{\bar{f}(\Theta)\exp[\mu' g(\psi,\Theta)]}{\int \bar{f}(\Theta)\exp[\mu' g(\psi,\Theta)]\, d\Theta}, \quad \textit{where} \tag{6.29}$$

$$g(\psi,\Theta) \equiv \int H(d,\psi) f(d|\psi,\Theta)\, dd \quad \textit{and } \mu \textit{ solves} \tag{6.30}$$

$$h(\psi) = \frac{\int g(\psi,\Theta)\exp[\mu' g(\psi,\Theta)\bar{f}(\Theta)]\, d\Theta}{\int \exp[\mu' g(\psi,\Theta)\bar{f}(\Theta)]\, d\Theta}. \tag{6.31}$$

Proof. Inserting (6.24), determined by an arbitrary $f(\Theta) \equiv f(\Theta|K) \in f_K^*$, into the constraint (6.25), we get

$$h(\psi) = \int H(d,\psi)\left[\int f(d|\psi,\Theta) f(\Theta)\, d\Theta\right] dd = \int g(\psi,\Theta) f(\Theta)\, d\Theta. \tag{6.32}$$

The second equality is implied by the Fubini theorem on multiple integrations and by the definition (6.30). The equality (6.32), normalization $\int f(\Theta)\, d\Theta = 1$ and non-negativity $f(\Theta) \geq 0$ determine the convex set where the optimal $f(\Theta|K)$ is searched for. Thus, we optimize the convex KL divergence on the convex set. The relevant Lagrangian, given by the vector multiplier μ, is

$$\int f(\Theta)\left[\ln\left(\frac{f(\Theta)}{\bar{f}(\Theta)}\right) - \mu' g(\psi,\Theta)\right] d\Theta$$
$$= \int f(\Theta)\left[\ln\left(\frac{f(\Theta)}{\bar{f}(\Theta)\exp[\mu' g(\psi,\Theta)]}\right)\right] d\Theta$$
$$= \int f(\Theta)\left[\ln\left(\frac{f(\Theta)}{\frac{\bar{f}(\Theta)\exp[\mu' g(\psi,\Theta)]}{\int \bar{f}(\tilde{\Theta})\exp[\mu' g(\psi,\tilde{\Theta})]\,d\tilde{\Theta}}}\right)\right] d\Theta$$
$$- \ln\left(\int \bar{f}(\Theta)\exp[\mu' g(\psi,\Theta)]\,d\Theta\right).$$

Properties of the KL divergence imply that the first term is minimized by the pdf (6.29). It is the overall minimizer as the second term is not influenced by this choice at all. The multiplier μ has to be chosen so that the constraints (6.25) are met. □

Remark(s) 6.4

1. *Solution of the algebraic equation (6.31) is often a difficult task.*
2. *The obtained form differs generally from the conjugate prior pdf. For this reason, a specific optimization is needed when the conjugate prior pdf (3.13) is required.*
3. *The majority of elaborated solutions are related to a single-input u single-output y linear system with the state in the phase form*

$$\psi_t = [u_t, \ldots, u_{t-\partial_u}, y_{t-1}, \ldots, y_{t-\partial_y}, 1].$$

Knowledge of the static gain g is the simplest example of expressing the knowledge pieces in terms of (6.26). The following correspondence holds: $\mathcal{E}[g] \approx \hat{g} = \hat{d}$, $cov[g] \approx r_g = r_d$ and

$$\psi' \equiv [\underbrace{1,\ldots,1}_{(\partial_u+1)\,times}, \underbrace{\hat{g},\ldots,\hat{g}}_{\partial_y\,times}, 1].$$

Dominant time constant, knowledge of a point on frequency response, measured step response, and smoothness of the step response are other examples elaborated for this model [115, 147, 148].

6.3.3 Merging of knowledge pieces

A combination of results in Sections 6.3.1 6.3.2, gives a collection of prior estimates $f(\Theta|K_k)$, $k \in k^* \equiv \{1,\ldots,\mathring{k}\}$ obtained by the processing of individual, internally consistent, data blocks and of individual knowledge items. We have to use them for constructing a single estimate $\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k}))$ of the posterior pdf $f(\Theta|d(\mathring{t}), K(\mathring{k}))$ that reflects all of them as well as the measured data $d(\mathring{t})$. This posterior pdf should be close to the unknown posterior pdf $f(\Theta|d(\mathring{t}), K(\mathring{k}))$ that arises from the application of the Bayes rule starting from an unknown "objective" combination $f(\Theta|K(\mathring{k}))$. For the construction of the estimate $\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k}))$, we

- can use available real data $d(\mathring{t})$ and pdfs $f(\Theta|K_k),\ k \in k^*$: they form our experience $\mathcal{P}$ for the addressed estimation task,
- do not know mutual relationships of the prior estimates $f(\Theta|K_k),\ k \in k^*$,
- are aware that the quality of individual pdfs $f(\Theta|K_k),\ k \in k^*$ may vary substantially but in an unknown manner.

We face the problem of estimating the unknown pdf $f(\Theta|d(\mathring{t}), K(\mathring{k})) \in f^*$. The estimation of this infinite-dimensional object is hard. We solve it by adopting the following approximation.

Agreement 6.4 (Approximation of the estimated pdf) *Let $f_k^* \subset f^*$ contain those posterior pdfs $f(\Theta|d(\mathring{t}), K(\mathring{k}))$ for which $f(\Theta|d(\mathring{t}), K_k)$ is the best approximating pdf among $\left\{f(\Theta|d(\mathring{t}), K_k)\right\}_{k\in k^*}$.*

On f_k^*, we approximate the unknown objective pdf $f \equiv f(\Theta|d(\mathring{t}), K(\mathring{k}))$, combining prior knowledge $K(\mathring{k})$ and real data, by the pdf $f_k \equiv f(\Theta|d(\mathring{t}), K_k)$, i.e.,

$$f \equiv f(\Theta|d(\mathring{t}), K(\mathring{k})) \approx f(\Theta|d(\mathring{t}), K_k) \equiv f_k,\ \forall f(\Theta|d(\mathring{t}), K(\mathring{k})) \in f_k^*. \quad (6.33)$$

Proposition 6.5 (Merging of knowledge pieces) *Let the approximation (6.33) be adopted, the natural conditions of decision making (2.36) met and the sets f_j^* be a priori equally probable. Then, the merger $\hat{f}$ minimizing the expected KL divergence $\mathcal{E}\left[\mathcal{D}\left(\hat{f}||f\right)\right]$, exploiting the experience $\mathcal{P}$ consisting of measured data $d(\mathring{t})$ and prior pdfs corresponding to individual internally consistent knowledge pieces $\{f(\Theta|K_k)\}_{k\in k^*}$ has the form*

$$\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k})) \propto \mathcal{L}(\Theta, d(\mathring{t})) \prod_{k\in k^*} [f(\Theta|K_k)]^{\beta_{k|d(\mathring{t})}}. \quad (6.34)$$

In (6.34), $\mathcal{L}(\Theta, d(\mathring{t}))$ is the likelihood corresponding to the considered parameterized factor and measured data $d(\mathring{t})$. The posterior probabilities $\beta_{k|d(\mathring{t})}$ of sets f_k^* are

$$\beta_{k|d(\mathring{t})} \propto \int \mathcal{L}(\Theta, d(\mathring{t})) f(\Theta|K_k)\, d\Theta. \quad (6.35)$$

Thus, the probabilities $\beta_{k|d(\mathring{t})}$ are equal to the normalized v-likelihood corresponding to respective choices $f(\Theta|K_k)$ of the prior pdfs.

Proof. Under (6.33), the minimization of the approximate expected KL divergence can be performed as follows.

$$
\begin{aligned}
&\mathrm{Arg}\min_{\hat{f}} \mathcal{E}\left[\mathcal{D}\left(\hat{f}\middle|\middle| f\right)\right] = \mathrm{Arg}\min_{\hat{f}} \mathcal{E}\left\{\sum_{k\in k^*} \chi_{f_k^*}(f)\mathcal{E}\left[\mathcal{D}\left(\hat{f}\middle|\middle| f\right)\middle|\mathcal{P}\right]\right\} \\
&\underbrace{\approx}_{(6.33)} \mathrm{Arg}\min_{\hat{f}} \mathcal{E}\left\{\sum_{k\in k^*} \mathcal{E}\left[\chi_{f_k^*}(f)\,\mathcal{D}\left(\hat{f}\middle|\middle| f_k\right)\middle|\mathcal{P}\right]\right\} = \\
&\underbrace{=}_{\text{Proposition 2.7}} \mathrm{Arg}\min_{\hat{f}} \sum_{k\in k^*} \mathcal{E}\left[\chi_{f_k^*}(f)\,\mathcal{D}\left(\hat{f}\middle|\middle| f_k\right)\middle|\mathcal{P}\right] \underbrace{=}_{\substack{\beta_{k|d(\mathring{t})} \equiv \mathcal{E}[\chi_{f_k^*}(f)|\mathcal{P}] \\ \text{definition of } \mathcal{D}}} \\
&= \mathrm{Arg}\min_{\hat{f}} \int \hat{f}(\Theta|d(\mathring{t}),K(\mathring{k}))\ln\left[\frac{\hat{f}(\Theta|d(\mathring{t}),K(\mathring{k}))}{\prod_{k\in k^*}[f(\Theta|d(\mathring{t}),K_k)]^{\beta_{k|d(\mathring{t})}}}\right] d\Theta.
\end{aligned}
\tag{6.36}
$$

Normalization of the denominator in the last fraction to a pdf and the use of the second property of the KL divergence, Proposition 2.10, imply the form (6.34) of the approximating pdf. To prove remaining statements, it is sufficient to realize that the pdfs combined into the geometric mean have the common factor equal to the likelihood function $\mathcal{L}(\Theta, d(\mathring{t}))$. The weight $\beta_{k|d(\mathring{t})}$ is the posterior probability that $f_k \equiv f(\Theta|K_k)$ is the closest pdf to $f(\Theta|K(\mathring{k}))$, i.e., $f(\Theta|K(\mathring{k})) \in f_k^*$. Values of these posterior probabilities are simply obtained by the Bayes rule applied to hypotheses $f(\Theta|K(\mathring{k})) \in f_k^*,\ k \in k^*$, starting from the assumed uniform prior on them. It proves the formula (6.35). □

Algorithm 6.3 (Merging knowledge pieces)

1. *Apply Algorithm 6.2 to get* $f(\Theta|K_k),\ k \in k^*$ *on internally consistent data blocks.*
2. *Quantify individual knowledge pieces as described by Proposition 6.4.*
3. *Evaluate the likelihood function* $\mathcal{L}(\Theta, d(\mathring{t}))$ *corresponding to the considered parameterized model.*
4. *Evaluate the v-likelihood*

$$f(d(\mathring{t})|K_k) = \int \mathcal{L}(\Theta, d(\mathring{t}))f(\Theta|K_k)\,d\Theta \quad \textit{with prior pdfs} \ \ f(\Theta|K_k),\ k \in k^*.$$

5. *Evaluate weights*

$$\beta_{k|d(\mathring{t})} = \frac{f(d(\mathring{t})|K_k)}{\sum_{\tilde{k}\in k^*} f(d(\mathring{t})|K_{\tilde{k}})},\ \ k \in k^*.$$

6. *Determine the merger as the posterior pdf to be used*

$$\hat{f}(\Theta|d(\mathring{t}),K(\mathring{k})) \propto \mathcal{L}(\Theta, d(\mathring{t}))\prod_{k\in k^*}[f(\Theta|K_k)]^{\beta_{k|d(\mathring{t})}}.$$

Note that the used "prior pdf" $\prod_{k\in k^*}[f(\Theta|K_k)]^{\beta_{k|d(\mathring{t})}}$ is seen in the last step of the merging algorithm.

6.4 Construction of the prior estimate

It is known [49] that mixtures cannot be identified uniquely. This means that the estimation results depend on the prior parameter estimates of the individual factors and on the estimates of component weights as well as on the overall structure used. The need to use an approximate estimation, Section 6.5, makes the situation even harder. Thus, the *choice of the prior pdf* significantly determines the quality of the results.

All clustering and mixture estimation techniques face a similar *initialization problem.* There is a lot of algorithms for approaching it, for instance, [149]. We have, however, found no good solution suitable for estimation of mixtures with high-dimensional *dynamic components* and a large estimation horizon $\mathring{t}$.

Our probably novel approach is gradually presented here. It applies tailored branch-and-bound Algorithm 6.1 for inspection of alternative prior pdfs. Subsection 6.4.1 interprets this approach as a feasible approximation of Bayesian estimation of unknown prior pdf and demonstrates that the corresponding v-likelihood is the relevant functional to be maximized. Omission of variants with smaller v-likelihoods is justified as common bounding mapping in Section 6.4.2. The major problem of selecting promising alternatives, i.e., the design of branching mappings, forms the body of this important section. The variants exploit guesses of posterior pdfs obtained from previous guesses of prior pdfs. The posterior pdfs are too sharp to be combined directly into new variants of prior pdfs. Thus, they have to be flattened in some way. Flattening mappings are inspected in Subsection 6.4.3. Then, a group of branching mappings is studied that preserves the structure of inspected mixture variants, Sections 6.4.4, 6.4.5, 6.4.6, 6.4.7. They differ in complexity and applicability. They serve as building blocks of the most important technique, proposed in Section 6.4.8, which inspects prior pdfs corresponding to differing structures of the estimated mixture.

A significant part of the complexity of the initialization problem stems from the dynamic nature of the estimated mixture. Restriction to static mixtures makes the problem simpler. The relationships of dynamic and static mixtures are inspected in concluding Section 6.4.9, which sheds some light on the applicability of static clustering techniques to dynamic data records.

6.4.1 Iterative construction of the prior pdf

The Bayes rule specifies the posterior estimates $f(\Theta|d(\mathring{t})) \propto f(d(\mathring{t})|\Theta)f(\Theta)$ in a noniterative fashion. At the same time, iterative data-based constructions of $f(\Theta)$ are practically and successfully used. They face the following common and often overlooked danger.

Let the pdf $f_n(\Theta|d(\mathring{t}))$ be the pdf obtained through the nth formal repetition of the Bayes rule

$$f_n(\Theta|d(\mathring{t})) \propto f(d(\mathring{t})|\Theta) f_{n-1}(\Theta|d(\mathring{t}))$$

with $f_0(\Theta|d(\mathring{t})) \equiv f(\Theta)$. Then, $f_n(\Theta|d(\mathring{t})) \propto [f(d(\mathring{t})|\Theta)]^n f(\Theta)$. For $f(\Theta) > 0$ on Θ^*, the pdf $f_n(\Theta|d(\mathring{t}))$ concentrates on Θ maximizing globally the likelihood function $f(d(\mathring{t})|\Theta)$. This seems to be a good property. The convergence to a meaningful point is, however, likely to be spoiled when the involved likelihood $f(d(\mathring{t})|\Theta)$ is evaluated approximately only, as must be done for mixtures. Moreover, the pdf obtained in this way completely loses information on precision of such an estimate.

The following re-interpretation and modification of the iterative evaluations resolves the outlined problem.

Iterations are used when we are uncertain about the adequate prior pdf. Moreover, the likelihood function is evaluated approximately only and the quality of the approximation depends on the prior pdf used. Thus, the *likelihood function is uncertain*, too. According to the adopted Bayesian philosophy, any uncertainty should be treated as randomness. Thus, the joint pdf $f(d(\mathring{t}), \Theta)$ of data $d(\mathring{t})$ and parameters Θ becomes an infinite-dimensional unknown (hyper)parameter. The mixture estimation of interest specifies the set $f^*(d(\mathring{t}), \Theta)$ of the possible joint pdfs. The considered estimation task is then characterized by the following agreement.

Agreement 6.5 (Estimation of the joint pdf of data and parameters) *An estimate $\hat{f}(d(\mathring{t}), \Theta)$ of the joint pdf of data and parameters $f(d(\mathring{t}), \Theta)$ is searched for within the set*

$$f^*(\cdot,\cdot) \equiv \left\{ f(d(\mathring{t}), \Theta) \equiv \mathcal{L}(\Theta, d(\mathring{t})) f(\Theta) \right\} \quad \textit{with the likelihood}$$

$$\mathcal{L}(\Theta, d(\mathring{t})) = \prod_{t \in t^*} f(d_t|d(t-1), \Theta) \ \textit{given by the parameterized model}$$

$$f(d_t|d(t-1), \Theta) = \sum_{c \in c^*} \alpha_c f(d_t|d(t-1), \Theta_c, c). \tag{6.37}$$

Its components are products of factors

$$f(d_t|d(t-1), \Theta_c, c) = \prod_{i \in i^*} f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)$$

and the prior mixture estimate $f(\Theta)$ is a product of (conjugate) prior pdfs

$$f(\Theta) = Di_\alpha(\kappa_0) \prod_{i \in i^*} f(\Theta_{ic}); \ \textit{see Agreement 6.1.}$$

The constructed estimator $d^*(\mathring{t}) \to f^*$ is required to specify a point estimate $\hat{f}(\cdot,\cdot) \in f^*(\cdot,\cdot)$ of the joint pdf $f(\cdot,\cdot)$ of data and parameters. The evaluations have to stop after a finite number of evaluation steps.

The application of the general decision-making theory needs a specification of the expected loss to be optimized. It calls for specification of probabilistic measure on the set $f^*(\cdot,\cdot)$ of pdfs. It is difficult both technically and practically. In order to avoid this difficulty, we interpret the maximum likelihood estimate as a point Bayesian estimate corresponding to a special prior distribution. This interpretation is correct under some technical conditions, e.g., [150]. We assume that these conditions are applicable in our case. Then, we have to describe the corresponding (hyper)likelihood function ${}^{\lfloor h}\mathcal{L}(f(\cdot,\cdot), d(\mathring{t}))$ only. It means that we have to specify the functional dependence of the distribution of the observed data $d(\mathring{t})$ on the estimated infinite-dimensional (hyper)parameter $f(\cdot,\cdot) \equiv f(d(\mathring{t}), \Theta)$. This likelihood equals to the marginal pdf of data $d(\mathring{t})$ computed for each unknown hyperparameter formed by the unknown joint pdf $f(\cdot,\cdot)$ of data and parameters. Thus,

$${}^{\lfloor h}\mathcal{L}(f(\cdot,\cdot), d(\mathring{t})) \equiv f\left(d(\mathring{t})|f(\cdot,\cdot)\right) = \int f(d(\mathring{t}), \Theta)\, d\Theta. \tag{6.38}$$

The last equality is implied by the marginalization rule, Proposition 2.4.

These considerations lead us to the choice of the estimate $\hat{f}(d(\mathring{t}), \Theta)$ of the joint pdf $f(d(\mathring{t}), \Theta)$ of data and parameters as

$$\hat{f}(d(\mathring{t}), \Theta) \in \mathrm{Arg} \max_{\tilde{f} \in f^*(\cdot,\cdot)} \int \tilde{f}(d(\mathring{t}), \Theta)\, d\Theta. \tag{6.39}$$

For any $\tilde{f} \in f^*(\cdot,\cdot)$ (6.37), we are able to evaluate (approximately) the maximized functional (6.39). We face an “ordinary” optimization problem over the infinite-dimensional domain (6.37). We tackle it by the branch-and-bound techniques prepared in Section 6.1.3. Moreover, we use the fact that for a chosen guess of the prior pdf $\hat{f}(\Theta)$ the approximation of the likelihood $\mathcal{L}$ in (6.37) is uniquely determined by the algorithm adopted for the approximate estimation; see Section 6.5.

Note that forgetting can be used during the search. The final estimates have to be done without forgetting as we need a time-invariant description of the o-system. We summarize the above thoughts.

Agreement 6.6 (Search for the joint pdf of data and parameters) *We search for a variant of $\hat{f}(\Theta)$ for which the joint pdf $\hat{f}(d(\mathring{t}), \Theta)$ in $f^*(\cdot,\cdot)$ (6.37) evaluated by an approximate algorithm, Section 6.5, with the unit forgetting exhibits the highest v-likelihood $\hat{f}(d(\mathring{t}))$. The dimensionality of the problem directs us towards the maximization by the branch-and-bound techniques; Section 6.1.3.*

6.4.2 Common bounding mapping

The bounding mapping $\mathcal{U}$ (6.16) adopted further on is simple and universal. The arguments $X_{\iota(n+1|n)} \equiv f_{\iota(n+1|n)}(\cdot,\cdot)$ in (6.37) with smallest values of v-likelihoods

$$F(X_{\iota(n+1|n)}) \equiv f\left(d(\mathring{t})|f_{\iota(n+1|n)}(\cdot,\cdot)\right) = \int f_{\iota(n+1|n)}(\cdot,\cdot)\, d\Theta \equiv f_{\iota(n+1|n)}(d(\mathring{t}))$$

are omitted when forming the new set of candidates $X_{n+1}^* \subset f^*(\cdot,\cdot)$; see (6.37).

It is necessary to stress that for application it is useful to take into account the following points.

- Values $F(X_{\iota(n+1|n)}) \equiv f_{\iota(n+1|n)}(d(\mathring{t}))$ for omitted guesses of prior pdfs have to be known and compared with the best value of the v-likelihood $\bar{F}_{n+1|n} \equiv \bar{f}_{n+1|n}$ found up to this moment. This trivial observation can be and was overlooked in a complex optimization setting.
- The number of preserved or omitted arguments is the only optional parameter of this mapping. It is mostly dictated by the adopted branching mapping that usually needs several arguments in $X_n^* \subset f^*(\cdot,\cdot)$ for generating new promising candidates to form new candidates $X_{n+1|n}^* \subset f^*(\cdot,\cdot)$; see (6.37).

6.4.3 Flattening mapping

Our discussion on the choice of the prior estimate started with an inspection of the iterative application of the Bayes rule. Its basic idea — that posterior pdf $f(\Theta|d(\mathring{t}))$ probably gives a better clue about Θ than the prior pdf $f(\Theta)$ — is elaborated here.

The posterior estimate of parameters cannot be directly used as a new guess of the prior pdf as it is much more concentrated than a reasonable prior pdf. It is dangerous in the discussed context, as the original prior has been assumed to be unreliable, and thus the posterior pdf may be concentrated at a false subset of Θ^*. In order to avoid this danger, we have to apply a *flattening mapping* $\mathcal{G}$

$$\mathcal{G} : f(\Theta|d(\mathring{t})) \rightarrow \hat{f}(\Theta) \tag{6.40}$$

to get a new guess of $\hat{f}(\Theta)$ of the prior pdf $f(\Theta)$. It has to be designed so that a compromise between the following contradictory requirements is reached.

- $\hat{f}(\Theta)$ should resemble $f(\Theta|d(t))$,
- $\hat{f}(\Theta)$ should be flat enough.

According to the adopted development methodology, we design $\mathcal{G}$ by solving a suitable decision-making task.

For the fixed $d(\mathring{t})$, let us denote the posterior pdf $\tilde{f} \equiv \tilde{f}(\Theta) \equiv f(\Theta|d(\mathring{t}))$. Moreover, we select a prototype of a flat pdf $\bar{f} \equiv \bar{f}(\Theta)$, called also *pre-prior pdf*. A uniform pdf (even an improper one) suits often to this purpose.

The pdf $\hat{f} \equiv f(\Theta)$ generated by the constructed flattening mapping $\mathcal{G}$ (6.40) is found as a minimizing argument of the functional

$$\mathcal{D}\left(\hat{f}||\tilde{f}\right) + q\mathcal{D}\left(\hat{f}||\bar{f}\right) \text{ specified by the optional } q > 0. \tag{6.41}$$

The KL divergence (2.25) $\mathcal{D}\left(\hat{f}||\tilde{f}\right)$ reflects the first requirement and $\mathcal{D}\left(\hat{f}||\bar{f}\right)$ the second one. The positive weight q is the design parameter that controls the compromise we are seeking for. It can be interpreted as the ratio of probabilities assigned to hypotheses that the true pdf equals $\hat{f}$ and $\bar{f}$, respectively.

Proposition 6.6 (Optimal flattening mapping) *Let $\tilde{f}$ and $\bar{f}$ be a given pair of pdfs defined on Θ^*. Then, the pdf $\hat{f} \in f^* \equiv \{$pdfs defined on $\Theta^*\}$ minimizing the functional (6.41) has the form*

$$\hat{f} \propto \tilde{f}^{\Lambda} \bar{f}^{1-\Lambda} \quad \textit{with} \quad \Lambda = 1/(1+q) \in (0,1). \tag{6.42}$$

Proof. It holds $\hat{f} \in \mathrm{Arg\,min}_{f\in f^*} \mathcal{D}\left(f||\tilde{f}\right) + q\mathcal{D}\left(f||\bar{f}\right) \equiv$

$$\underbrace{\equiv}_{\text{definition of}\,\mathcal{D}} \quad \mathrm{Arg}\min_{f\in f^*} \int f \left[\ln\left(\frac{f}{\tilde{f}}\right) + q\ln\left(\frac{f}{\bar{f}}\right)\right] d\Theta$$

$$= \mathrm{Arg}\min_{f\in f^*} (1+q) \int f \ln\left(\frac{f}{\tilde{f}^{1/(1+q)}\,\bar{f}^{q/(1+q)}}\, d\Theta\right) \underbrace{\equiv}_{(6.42)} \hat{f}.$$

The last identity is implied by independence of the normalizing shift

$$\ln\left(\int \tilde{f}^{1/(1+q)}\,\bar{f}^{q/(1+q)}\, d\Theta\right)$$

of the optional pdf f, by the positivity of $1+q$, the definition of Λ (6.42), the definition of $\hat{f}$ (6.42) and Proposition 2.10. □

The pdf $\hat{f}$ found in Proposition 6.6 coincides formally with a geometric mean of a pair of pdfs; see Proposition 3.1. The power Λ has, however, another interpretation. It is controlled by the optional weight $q > 0 \Leftrightarrow \Lambda \in (0,1)$ balancing the Bayes rule and the flattening.

The adopted Agreement 6.1 implies that the estimate of mixture parameters Θ is a product of the Dirichlet pdf describing component weights α and the product of pdfs describing parameters Θ_{ic} of individual factors. The KL divergence of a pair of such products is simply sum of KL divergences between the corresponding marginal pdfs creating the product. Consequently, Proposition 6.6 can be applied to each of them with its specific weight and thus its specific Λ.

Flattening was motivated by an iterative application of the Bayes rule. The combination of the Bayes rule and flattening is described by the operator

$$\mathcal{A}_{\Lambda}:\ f(\Theta) \rightarrow \frac{\left[\mathcal{L}(\Theta, d(\mathring{t}))f(\Theta)\right]^{\Lambda} [\bar{f}(\Theta)]^{1-\Lambda}}{\int \left[\mathcal{L}(\Theta, d(\mathring{t}))f(\Theta)\right]^{\Lambda} [\bar{f}(\Theta)]^{1-\Lambda}\, d\Theta}. \tag{6.43}$$

Its use requires careful specification of the *flattening rate* Λ.

Choice of flattening rate in branching

Flattening is often applied as a part of branching. Then, it is effectively used just once and the asymptotic analysis in a subsequent section gives no hint about how to choose the flattening parameter Λ. We discuss an adequate choice here.

For the Dirichlet pdf $Di_{\alpha}(\kappa)$ (6.4) estimating component weights, the sum $\sum_{c\in c^*}\kappa_c$ is the effective number of observed and fictitious data; see Chapter 10. Thus, the prior and posterior values of this sum differ by the effective number of processed data. This difference should be suppressed if the flattened pdf serves as the prior pdf. The flattened version should be based on the same number of (fictitious) data as the prior pdf. This gives the following recommendation.

Proposition 6.7 (Flattening of the Dirichlet pdf in branching) *Let the pre-prior pdf be Dirichlet $Di_{\alpha}(\bar{\kappa})$, the prior one be $Di_{\alpha}(\kappa_0)$ and the posterior one $Di_{\alpha}(\kappa_{\mathring{t}})$. Then, the new prior $Di_{\alpha}(\hat{\kappa}_0)$, obtained by flattening with*

$$\hat{\kappa}_0 = \Lambda\kappa_{\mathring{t}} + (1-\Lambda)\bar{\kappa}, \quad \textit{where} \;\; \Lambda = \frac{\sum_{c\in c^*}(\kappa_{c;0}-\bar{\kappa}_c)}{\sum_{c\in c^*}(\kappa_{c;\mathring{t}}-\bar{\kappa}_c)}, \tag{6.44}$$

guarantees equality

$$\sum_{c\in c^*}\hat{\kappa}_{c;0} = \sum_{c\in c^*}\kappa_{c;0} \equiv {}^{\lfloor\kappa}K_0. \tag{6.45}$$

Proof. Omitted. □

Remark(s) 6.5

1. *Experiments indicate that $\bar{\kappa}\approx 0$ and $\sum_{c\in c^*}\kappa_{c;0}\approx 0.1\mathring{t}$ are suitable options. Then, $\Lambda\approx 0.1$.*
2. *A positive flattening rate Λ is obtained iff (see (6.45))*

$$^{\lfloor\kappa}K_0 > {}^{\lfloor\kappa}\bar{K} \equiv \sum_{c\in c^*}\bar{\kappa}_c. \tag{6.46}$$

   For the discussed noniterative flattening, it is sufficient and reasonable to take the lowest value $\bar{\kappa}=0$.
3. *Meaningful $\Lambda<1$ is obtained if $^{\lfloor\kappa}K_{\mathring{t}} \equiv \sum_{c\in c^*}\kappa_{c;\mathring{t}} > {}^{\lfloor\kappa}K_0$. It is naturally fulfilled when estimating without forgetting. This condition has to be checked whenever exponential forgetting is employed. For the constant forgetting factor with a flat alternative pdf, this condition is equivalent to the requirement, cf. (6.45), (6.46),*

$$\textit{forgetting factor} > 1 - \frac{1}{^{\lfloor\kappa}K_0 - {}^{\lfloor\kappa}\bar{K}}. \tag{6.47}$$

Let us discuss now the choice of the flattening rate related to parameters Θ_{ic}; cf. Agreement 6.1.

Mostly, we assume that all pdfs in $\prod_{i\in i^*}\prod_{c\in c^*} f(\Theta_{ic})$ are conjugate pdfs of the parameterized factors belonging to the exponential family; Section 3.2. In this family, the effective number of exploited, possibly fictitious, records is stored in counters ν_{ic} for each in $f(\Theta_{ic})$. All considered approximate estimations (see Section 6.5) add a fraction of the data mass to individual factors. Even without forgetting, we can guarantee at most that $\sum_{ic}\nu_{ic;\mathring{t}} = \mathring{d}\mathring{t} + \sum_{ic}\nu_{ic;0}$. Thus, even for single flattening used in branching, we cannot flatten factor-wise. We have to use a common flattening rate to all of them. Similarly as in the case of component weights, the common flattening rate is chosen so that $\sum_{ic}\nu_{ic;\mathring{t}}$ decreases to $\sum_{ic}\nu_{ic;0}$. This implies the following proposition.

Proposition 6.8 (Flattening of $\prod_{ic} f(\Theta_{ic}|d(\mathring{t}))$ in branching) *Let us consider factors in the exponential family and the product forms of*

$$\textit{conjugate pre-prior pdf} \quad \prod_{i\in i^*, c\in c^*} A^{\bar{\nu}_{ic}}(\Theta_{ic})\exp\left[\left\langle \bar{V}_{ic}, C(\Theta)\right\rangle\right]\chi_{\Theta^*_{ic}}(\Theta_{ic})$$

$$\textit{conjugate prior pdf} \quad \prod_{i\in i^*, c\in c^*} A^{\nu_{ic;0}}(\Theta_{ic})\exp\left[\left\langle V_{ic;0}, C(\Theta)\right\rangle\right]\chi_{\Theta^*_{ic}}(\Theta_{ic})$$

$$\textit{posterior pdf} \quad \prod_{i\in i^*, c\in c^*} A^{\nu_{ic;\mathring{t}}}(\Theta_{ic})\exp\left[\left\langle V_{ic;\mathring{t}}, C(\Theta)\right\rangle\right]\chi_{\Theta^*_{ic}}(\Theta_{ic}). \tag{6.48}$$

Then, the prior pdf obtained by flattening

$$\prod_{i\in i^*, c\in c^*} A^{\hat{\nu}_{ic;0}}(\Theta)\exp\left[\left\langle \hat{V}_{ic;0}, C(\Theta)\right\rangle\right]\chi_{\Theta^*_{ic}}(\Theta_{ic}) \quad \textit{with}$$

$$\hat{V}_{ic;0} = \Lambda V_{ic;\mathring{t}} + (1-\Lambda)\bar{V}_{ic}, \quad \hat{\nu}_{ic;0} = \Lambda\nu_{ic;\mathring{t}} + (1-\Lambda)\bar{\nu}_{ic} \tag{6.49}$$

$$\forall i \in i^* \equiv \{1,\ldots,\mathring{d}\},\ c\in c^*,\ \textit{and}\ \Lambda = \frac{\sum_{i\in i^*}\sum_{c\in c^*}(\nu_{ic;0}-\bar{\nu}_{ic})}{\sum_{i\in i^*}\sum_{c\in c^*}(\nu_{ic;\mathring{t}}-\bar{\nu}_{ic})}$$

guarantees that

$$\sum_{i\in i^*}\sum_{c\in c^*}\hat{\nu}_{ic;0} = \sum_{i\in i^*}\sum_{c\in c^*}\nu_{ic;0} \equiv {}^{\lfloor\nu}K_0. \tag{6.50}$$

Proof. Omitted. □

Remark(s) 6.6

1. *A positive flattening rate Λ is obtained iff*

$$^{\lfloor\nu}\kappa_0 > \sum_{i\in i^*}\sum_{c\in c^*}\bar{\nu}_{ic} \equiv {}^{\lfloor\nu}\bar{K}. \tag{6.51}$$

It is guaranteed when we select $\bar{\nu}_c = 0,\ c\in c^*$.

2. *Meaningful $\Lambda < 1$ is obtained if ${}^{\lfloor\nu}K_{\mathring{t}} \equiv \sum_{i\in i^*}\sum_{c\in c^*}\nu_{ic;\mathring{t}} > {}^{\lfloor\nu}K_0$. It is naturally fulfilled when the estimation is performed without forgetting. This condition has to be checked whenever the exponential forgetting is employed. For the constant forgetting factor with a flat alternative pdf, this condition is equivalent to the requirement (cf. (6.50), (6.51))*

$$\textit{forgetting factor} > 1 - \frac{1}{{}^{\lfloor\nu}K_0 - {}^{\lfloor\nu}\bar{K}}. \tag{6.52}$$

3. *Use of stabilized forgetting avoids the problem discussed in the previous item. It can also be used for the Dirichlet part of the treated prior and posterior pdfs. For the respective factors in the exponential family, it is sufficient to select a positive degree $\nu_{ic;0}$ of the factor $A(\Theta_{ic})$ in the exponential family cf. (3.6) serving as the alternative pdf. Then, the inspected sums do not fall and no special actions are needed during flattening.*

Choice of flattening rate in iterative learning

In order to give justified recommendations of the flattening rates in iterations, we inspect influence of flattening rate Λ on results of the repetitive use of $\mathcal{A}_\Lambda$. The analysis is made in the ideal case when the likelihood $\mathcal{L}(\Theta, d(\mathring{t})) \equiv f(d(\mathring{t})|\Theta)$ is evaluated exactly in each iteration.

Proposition 6.9 (Asymptotic of the repetitive use of $\mathcal{A}_\Lambda$) *Let $\Lambda(n) \in [0, 1-\varepsilon]^n$, $\varepsilon \in (0,1)$, be a sequence of weights determining operators $\mathcal{A}_n \equiv \mathcal{A}_{\Lambda_n}$ (6.43) with a common likelihood $\mathcal{L}(\Theta, d(\mathring{t})) \equiv f(d(\mathring{t})|\Theta)$. Let the common support of the prior pdf $f(\Theta)$ and the (flat) pre-prior pdf $\bar{f}(\Theta)$ contain the support of $\mathcal{L}(\Theta, d(\mathring{t}))$. Then, the pdfs $f_n \equiv \mathcal{A}_n\mathcal{A}_{n-1}\cdots\mathcal{A}_1[f]$ have the following possible stationary value*

$$f_\infty \equiv \overline{\lim}_{n\to\infty} f_n \propto \mathcal{L}^{\,{}^{\lfloor\Lambda}A}\,\bar{f}^{\,{}^{\lfloor\Lambda}B}, \quad \textit{where the exponents are} \tag{6.53}$$

$$\begin{aligned}
{}^{\lfloor\Lambda}A &\equiv \overline{\lim}_{n\to\infty}\, {}^{\lfloor\Lambda}A_n, \quad {}^{\lfloor\Lambda}A_n \equiv \sum_{i=1}^{n}\prod_{j=i}^{n}\Lambda_j \equiv \Lambda_n\left(1 + {}^{\lfloor\Lambda}A_{n-1}\right)\\
&{}^{\lfloor\Lambda}A_0 = 0
\end{aligned} \tag{6.54}$$

$$\begin{aligned}
{}^{\lfloor\Lambda}B &\equiv \overline{\lim}_{n\to\infty}\, {}^{\lfloor\Lambda}B_n, \quad {}^{\lfloor\Lambda}B_n \equiv \sum_{i=1}^{n}(1-\Lambda_i)\prod_{j=i+1}^{n}\Lambda_j \equiv (1-\Lambda_n) + \Lambda_n\, {}^{\lfloor\Lambda}B_{n-1}\\
&{}^{\lfloor\Lambda}B_0 = 0.
\end{aligned} \tag{6.55}$$

The asymptotic pdf f_∞ equals to $\mathcal{L}^{\Lambda/(1-\Lambda)}\bar{f}$ for a constant Λ. For $\Lambda = 0.5$, the pdf f_∞ coincides with that resulting from the Bayes rule initiated by the pre-prior pdf $\bar{f}$.

The weight $\Lambda = 0.5$ is the desirable asymptotic value of varying Λ_n. With $\Lambda = 0.5$, we get ${}^{\lfloor\Lambda}A = {}^{\lfloor\Lambda}B = 1$, leading to f_∞ coinciding with the posterior pdf for the prior pdf equal to $\bar{f}$.

Proof. The repetitive application of $\mathcal{A}_n$ gives $f_n = \mathcal{L}^{\lfloor\Lambda A_n} f^{\prod_{\tilde{n}=1}^{n} \Lambda_{\tilde{n}}} \bar{f}^{\lfloor\Lambda B_n}$. The nonnegative sequence $\left\{ {}^{\lfloor\Lambda}A_n \right\}_{n\geq 1}$ is given by the formula (6.54) and is bounded from above by the sum of the geometric sequence with the first term and quotient equal to $1-\varepsilon$. Thus, the $\overline{\lim}_{n\to\infty} {}^{\lfloor\Lambda}A_n$ exists and it is smaller than $1/\varepsilon - 1$.

The power of the prior pdf f converges quickly to zero as all factors $\Lambda_n \leq 1-\varepsilon$. The power ${}^{\lfloor\Lambda}B_n$ of the pre-prior pdf $\bar{f}$ evolves according to (6.55) and it is bounded by unity. Thus, its $\overline{\lim}_{n\to\infty}$ is smaller than one. For a constant Λ, ${}^{\lfloor\Lambda}A_n$ becomes the sum of a geometric sequence with value converging to $\Lambda/(1-\Lambda)$ and ${}^{\lfloor\Lambda}B_n$ to unity. □

Remark(s) 6.7

1. *Uniform separation of Λ_n from unity by $\varepsilon > 0$ corresponds to a uniformly positive penalty $q_n \geq q > 0$ (6.41). This desirable constraint prevents repetitive application of the plain Bayes rule and avoids the drawbacks discussed in Section 6.4.1.*
2. *The chosen "informative" prior pdf does not influence the asymptotic solution. Its preservation, however, may be of extreme importance at the beginning of iterations as it may keep the approximately evaluated likelihood close to the correct one.*

A change of the pre-prior pdf used in the flattening during iterative learning offers a remedy of the problem mentioned in the above remark. The simplest possibility is described by the following algorithm.

Algorithm 6.4 (Iterative learning with the preserved prior pdf)
Initial mode

- *Select the upper bound $\mathring{n}$ on the number n of iterations.*
- *Set the iteration counter $n = 1$ and select the initial flattening rate $\Lambda_1 \in (0,1)$.*
- *Select a flat pre-prior pdf, typically, improper one $\bar{f} \propto 1$.*
- *Specify a proper prior pdf $f_0 = f$.*
- *Compute a new guess of the prior pdf f_1 using the approximate estimation and flattening (cf. (6.43))*

$$f_1 \propto \left[\mathcal{L}_0 f_0\right]^{\Lambda_0} \bar{f}^{1-\Lambda_1}, \quad \textit{with } \mathcal{L}_0 \approx \mathcal{L}(\Theta, d(\mathring{t})).$$

- *Re-define the pre-prior pdf to f_0.*

Iterative mode

1. *Stop if $n \geq \mathring{n}$.*
2. *Increase the iteration counter $n = n+1$ and select the flattening rate Λ_n.*
3. *Compute a new guess of the prior pdf f_n using approximate estimation and flattening with f_0 serving as the pre-prior pdf*

$$f_n \propto \left[\mathcal{L}_{n-1} f_{n-1}\right]^{\Lambda_n} f_0^{1-\Lambda_n}.$$

4. *Go to the beginning of* Iterative mode.

Proposition 6.10 (Asymptotic of Algorithm 6.4) *Let $\Lambda(n) \in [0, 1-\varepsilon]^n$, $\varepsilon \in (0,1)$, be a sequence of weights determining operators $\mathcal{A}_n \equiv \mathcal{A}_{\Lambda_n}$ (6.43) with a common likelihood $\mathcal{L}(\Theta, d(\mathring{t})) \equiv f(d(\mathring{t})|\Theta)$. Let the common support of the prior pdf $f_0(\Theta) \equiv f(\Theta)$ and the (flat) pre-prior pdf $\bar{f}(\Theta)$ contain the support of $\mathcal{L}(\Theta, d(\mathring{t}))$. Then, the pdfs f_n generated by Algorithm 6.4 have the following possible stationary value*

$$\begin{aligned}
&f_\infty \equiv \overline{\lim}_{n\to\infty} f_n \propto \mathcal{L}^{\,{}^{\lfloor\Lambda}A} f_0^{\,{}^{\lfloor\Lambda}B}, \quad \textit{where the exponents are} && (6.56)\\
&{}^{\lfloor\Lambda}A \equiv \overline{\lim}_{n\to\infty}\, {}^{\lfloor\Lambda}A_n = \overline{\lim}_{n\to\infty} \Lambda_n(1 + {}^{\lfloor\Lambda}A_{n-1}),\ \underline{{}^{\lfloor\Lambda}A_1 = \Lambda_1}\\
&{}^{\lfloor\Lambda}B \equiv \overline{\lim}_{n\to\infty}\, {}^{\lfloor\Lambda}B_n = {}^{\lfloor\Lambda}B_n \equiv (1-\Lambda_n) + \Lambda_n\, {}^{\lfloor\Lambda}B_{n-1},\ \underline{{}^{\lfloor\Lambda}B_1 = \Lambda_1}.
\end{aligned}$$

The asymptotic pdf becomes $f_\infty = \mathcal{L}^{\Lambda/(1-\Lambda)} f_0$, for a constant $\Lambda = \Lambda_n,\ n > 1$. For $\Lambda = 0.5$, the pdf f_∞ coincides with that resulting from the Bayes rule initiated by the prior pdf f_0.

The value $\Lambda = 0.5$ is the desirable limit of varying Λ_n. For it, we get ${}^{\lfloor\Lambda}A = {}^{\lfloor\Lambda}B = 1$, leading to f_∞ coinciding with the posterior pdf determined by the prior pdf f_0.

Proof. Omitted, as it essentially copies the proof of Proposition 6.9. □

In the realistic iterative mode when the likelihood is evaluated approximately, it is reasonable to use a varying flattening rate Λ_n. We would like to have ${}^{\lfloor\Lambda}A_{\mathring{n}} = 1$ after the specified number of iterations $\mathring{n} > 1$. For $\mathring{n} = 1$, the values of Λ are recommended in Propositions 6.7 and 6.8. We take them respectively as initial conditions ${}^{\lfloor\Lambda}A_1 \equiv \Lambda_1$ for the Dirichlet (6.4) and exponential factors (6.48). If we specify, moreover, a function describing the desired increase of ${}^{\lfloor\Lambda}A_n$ to ${}^{\lfloor\Lambda}A_{\mathring{n}} = 1$ we get the rule as to how to compute respective values of Λ_n. We do that for a simple linear function.

Proposition 6.11 (Flattening rate for linearly growing confidence) *Let us repeat flattening for $\mathring{n} > 1$ steps and select*

$$\begin{aligned}
{}^{\lfloor\Lambda}A_{1D} &\equiv \Lambda_{1D} = \frac{\sum_{c\in c^*}(\kappa_{c;0} - \bar{\kappa}_c)}{\sum_{c\in c^*}(\kappa_{c;\mathring{t}} - \bar{\kappa}_c)} \quad \textit{for the Dirichlet factor in (6.3)}\\
{}^{\lfloor\Lambda}A_1 &\equiv \Lambda_1 = \frac{\sum_{i\in i^*, c\in c^*}(\nu_{ic;0} - \bar{\nu}_{ic})}{\sum_{i\in i^*, c\in c^*}(\nu_{ic;\mathring{t}} - \bar{\nu}_{ic})} \quad \textit{for the factors (6.48) in (6.3).}
\end{aligned}$$

Let us require ${}^{\lfloor\Lambda}A_{\mathring{n}} = 1$ and ${}^{\lfloor\Lambda}A_{\mathring{n}D} = 1$ after a linear growth. Then, for $n > 1$

$$\begin{aligned}
{}^{\lfloor\Lambda}A_{nD} &= {}^{\lfloor\Lambda}A_{(n-1)D} + k_D,\ k_D \equiv \frac{1 - {}^{\lfloor\Lambda}A_{1D}}{\mathring{n} - 1},\quad \Lambda_{nD} = \frac{{}^{\lfloor\Lambda}A_{nD}}{1 + {}^{\lfloor\Lambda}A_{(n-1)D}}\\
{}^{\lfloor\Lambda}A_n &= {}^{\lfloor\Lambda}A_{n-1} + k,\ k \equiv \frac{1 - {}^{\lfloor\Lambda}A_1}{\mathring{n} - 1},\quad \Lambda_n = \frac{{}^{\lfloor\Lambda}A_n}{1 + {}^{\lfloor\Lambda}A_{n-1}}. \qquad (6.57)
\end{aligned}$$

Proof. The required linear growth and the fixed initial condition implies ${}^{\lfloor\Lambda}A_n = {}^{\lfloor\Lambda}A_1 + k(n-1)$. The requirement ${}^{\lfloor\Lambda}A_{\mathring{n}} = 1$ gives $k = \frac{1-{}^{\lfloor\Lambda}A_1}{\mathring{n}-1}$ and (6.54) implies

$$\Lambda_n = \frac{{}^{\lfloor\Lambda}A_n}{1 + {}^{\lfloor\Lambda}A_{n-1}} \equiv \frac{{}^{\lfloor\Lambda}A_1 + k(n-1)}{1 + {}^{\lfloor\Lambda}A_1 + k(n-2)}.$$

Formulas are formally the same for the Dirichlet part that differs in initial conditions.

□

Asymptotic analysis of the flattening is made for the common exactly evaluated likelihood $\mathcal{L}(\Theta, d(\mathring{t}))$. Flattening is, however, justified by a violation of this assumption. Thus, it makes sense to model this violation and consequently to modify variations of the flattening rate. In this respect, we assume that in the nth flattening iteration

$$\mathcal{L}_n(\Theta, d(\mathring{t})) = \left[\mathcal{L}(\Theta, d(\mathring{t}))\right]^{\beta_n}, \tag{6.58}$$

where $\beta_n \in (0,1]$ is a known scalar. This model is motivated by experiments with estimation employing forgetting; see Section 6.4.7. With forgetting $\sum_{c\in c^*} \nu_{c;\mathring{t}} < \sum_{c\in c^*} \nu_{c;0} + \mathring{d}\mathring{t}$ (cf. Remarks 6.6) and thus $\mathcal{L}_n(\Theta, d(\mathring{t}))$ cannot be equal to the exact $\mathcal{L}(\Theta, d(\mathring{t}))$. The likelihood is simply more flat than that assumed up to now. The ratio

$$\beta_n = \frac{\sum_{c\in c^*} \nu_{nc;\mathring{t}}}{\sum_{c\in c^*} \nu_{nc;0} + \sum_{t=1}^{\mathring{t}} \lambda^t}, \text{ where } \lambda \text{ is the used forgetting factor,} \tag{6.59}$$

indicates the degree of flattening in the nth iterative step. Let us repeat the asymptotic analysis of Algorithm 6.4 under the assumption (6.58).

Proposition 6.12 (Asymptotic of Algorithm 6.4 under (6.58)) *Let $\Lambda(n) \in [0, 1-\varepsilon]^n$, $\varepsilon \in (0,1)$, be a sequence of weights determining operators $\mathcal{A}_n \equiv \mathcal{A}_{\Lambda_n}$ (6.43) with the varying likelihood $\mathcal{L}_n(\Theta, d(\mathring{t})) = [\mathcal{L}(\Theta, d(\mathring{t}))]^{\beta_n} \equiv [f(d(\mathring{t})|\Theta)]^{\beta_n}$. Let the common support of the prior pdf $f(\Theta) = f_0(\Theta)$ and of the (flat) pre-prior pdf $\bar{f}(\Theta)$ contain the support of $\mathcal{L}(\Theta, d(\mathring{t}))$. Then, the pdfs f_n generated by Algorithm 6.4 have the following possible stationary value*

$$\begin{aligned} & f_\infty = \overline{\lim}_{n\to\infty} f_n \propto \mathcal{L}^{\,{}^{\lfloor\Lambda}A} f_0^{\,{}^{\lfloor\Lambda}B}, \quad \textit{where} \\ & {}^{\lfloor\Lambda}A = \overline{\lim}_{n\to\infty}\, {}^{\lfloor\Lambda}A_n = \overline{\lim}_{n\to\infty} \Lambda_n(\beta_n + {}^{\lfloor\Lambda}A_{n-1}), \quad {}^{\lfloor\Lambda}A_1 = \Lambda_1, \quad \textit{and} \\ & {}^{\lfloor\Lambda}B = \overline{\lim}_{n\to\infty}\, {}^{\lfloor\Lambda}B_n = {}^{\lfloor\Lambda}B_n \equiv (1-\Lambda_n) + \Lambda_n\, {}^{\lfloor\Lambda}B_{n-1}, \quad {}^{\lfloor\Lambda}B_1 = \Lambda_1. \end{aligned} \tag{6.60}$$

The desirable limit is $A = B = 1$ when f_∞ coincides with the posterior pdf given by the prior pdf f_0.

Proof. Let us assume that, for $n > 1$, $f_{n-1} = \mathcal{L}^{\,{}^{\lfloor\Lambda}A_{n-1}} f_0^{\,{}^{\lfloor\Lambda}B_{n-1}} \bar{f}^{\,{}^{\lfloor\Lambda}C_{n-1}}$. The estimation and flattening steps, together with the last assumption and with (6.58), imply

$$f_n = [\mathcal{L}_n f_{n-1}]^{\Lambda_n} f_0^{1-\Lambda_n} = \left[\mathcal{L}^{\beta_n + {}^{\lfloor\Lambda}A_{n-1}} f_0^{{}^{\lfloor\Lambda}B_{n-1}} \bar{f}^{{}^{\lfloor\Lambda}C_{n-1}}\right]^{\Lambda_n} f_0^{1-\Lambda_n} \Rightarrow$$

$$\begin{aligned}
&{}^{\lfloor\Lambda}A_n = \Lambda_n(\beta_n + {}^{\lfloor\Lambda}A_{n-1}),\ \ {}^{\lfloor\Lambda}B_n - 1 = \Lambda_n\left({}^{\lfloor\Lambda}B_{n-1} - 1\right),\\
&{}^{\lfloor\Lambda}C_n = \Lambda_n\, {}^{\lfloor\Lambda}C_{n-1}.
\end{aligned}$$

Thus, ${}^{\lfloor\Lambda}B_n \to 1$ and ${}^{\lfloor\Lambda}C_n \to 0$ whenever $\prod_{n=1}^{\mathring{n}} \Lambda_n \to 0$. For it, it is sufficient to have $\Lambda_n \in [0, 1-\varepsilon],\ 1 > \varepsilon > 0$. Then, also the nonstationary difference equation for ${}^{\lfloor\Lambda}A_n$ is stable and the requirement ${}^{\lfloor\Lambda}A_n \to 1$ is the only serious constraint on the sequence Λ_n. □

The discussed modification leads to the following modification of Proposition 6.11.

Proposition 6.13 (Almost linearly growing confidence in $\mathcal{L}_n$ (6.58)) *Let us assume that β_n, β_{nD} are given and repeat flattening for $\mathring{n} > 1$ steps. Let us select*

$${}^{\lfloor\Lambda}A_{1D} \equiv \beta_{1D}\Lambda_{1D},\ \ \Lambda_{1D} = \frac{\sum_{c\in c^*}(\kappa_{c;0} - \bar{\kappa}_c)}{\sum_{c\in c^*}(\kappa_{c;\mathring{t}} - \bar{\kappa}_c)}$$

for flattening of Dirichlet factor (6.4) in (6.3) and

$${}^{\lfloor\Lambda}A_1 \equiv \beta_1\Lambda_1,\ \ \Lambda_1 = \frac{\sum_{i\in i^*, c\in c^*}(\nu_{ic;0} - \bar{\nu}_{ic})}{\sum_{i\in i^*, c\in c^*}(\nu_{ic;\mathring{t}} - \bar{\nu}_{ic})}$$

for the factors (6.48) in (6.3). Let us require ${}^{\lfloor\Lambda}A_{\mathring{n}} = 1$ after a linear growth under (6.58). Then, for $n > 1$,

$$\begin{aligned}
{}^{\lfloor\Lambda}A_n &= {}^{\lfloor\Lambda}A_{n-1} + k_n,\ \ k_n \equiv \min\left(\beta_n, \frac{1 - {}^{\lfloor\Lambda}A_{n-1}}{\mathring{n} - n + 1}\right) && (6.61)\\
\Lambda_n &= \frac{{}^{\lfloor\Lambda}A_n}{\beta_n + {}^{\lfloor\Lambda}A_{n-1}}\\
{}^{\lfloor\Lambda}A_{nD} &= {}^{\lfloor\Lambda}A_{(n-1)D} + k_{nD},\ \ k_{nD} \equiv \min\left(\beta_{nD}, \frac{1 - {}^{\lfloor\Lambda}A_{(n-1)D}}{\mathring{n} - n + 1}\right)\\
\Lambda_{nD} &= \frac{{}^{\lfloor\Lambda}A_{nD}}{\beta_n + {}^{\lfloor\Lambda}A_{(n-1)D}}
\end{aligned}$$

It may happen that either ${}^{\lfloor\Lambda}A_{\mathring{n}} < 1$ or ${}^{\lfloor\Lambda}A_{\mathring{n}D} < 1$. Then, $\mathring{n}$ must be increased by 1*. The necessary number of such extensions for reaching the value* $\max({}^{\lfloor\Lambda}A_{\mathring{n}}, {}^{\lfloor\Lambda}A_{\mathring{n}D})$ *is always finite.*

Proof. It is sufficient to inspect the exponential-family factor case as the Dirichlet case is identical.

Let us consider a generic step n with ${}^{\lfloor\Lambda}A_{n-1}$ and β_n given. We would like to have ${}^{\lfloor\Lambda}A_{\mathring{n}} = 1$ after a linear growth. With ${}^{\lfloor\Lambda}A_n = {}^{\lfloor\Lambda}A_{n-1} + k_n$, we require $1 = {}^{\lfloor\Lambda}A_{n-1} + \sum_{\tilde{n}=n}^{\mathring{n}} k_{\tilde{n}}$. Hoping for constant $k_{\tilde{n}}$ for $\tilde{n} \geq n$, we get

$$k_n = \frac{1 - {}^{\lfloor \Lambda}A_{n-1}}{\mathring{n} - n + 1} \quad \text{and} \quad \Lambda_n = \frac{k_n + {}^{\lfloor \Lambda}A_{n-1}}{\beta_n + {}^{\lfloor \Lambda}A_{n-1}}.$$

This is a legal step iff $\Lambda_n \leq 1 \Leftrightarrow k_n \leq \beta_n$. When this condition is met, we apply the proposed Λ_n. We use $\Lambda_n = 1$ otherwise, which corresponds to $k_n = \beta_n$. This will be slower than the intended linear growth but such a Λ_n will increase β_{n+1} at most to 1. Thus, it can be expected that the linear increase will be possible in the next iteration step.

It may well happen that, for the planned $\mathring{n}$, ${}^{\lfloor \Lambda}A_{\mathring{n}}$ does not reach unity. Then, it has to be increased. The need for this extension will end up in a finite number of attempts as the increase of A_n in each step is positive and the target value is finite. □

The proposition specifies a general flattening algorithm applicable both in iterative learning and in branching. It is presented for factors in the exponential family $f(d_i|\psi_{ic}, \Theta_{ic}, ic) = A(\Theta_{ic}) \exp[\langle B(\Psi_{ic}), C(\Theta_{ic})\rangle]$ and estimates $f(\Theta_{ic}|d(\mathring{t})) \propto A^{\nu_{ic;\mathring{t}}}(\Theta_{ic}) \exp\left[\left\langle V_{ic;\mathring{t}}, C(\Theta_{ic})\right\rangle\right]$ corresponding to the conjugate prior pdfs $f(\Theta_{ic}) \propto A^{\nu_{ic;0}}(\Theta_{ic}) \exp[\langle V_{ic;0}, C(\Theta_{ic})\rangle]$ and the flat conjugate alternative $\bar{f}(\Theta_{ic}) \propto A^{\bar{\nu}_{ic}}(\Theta_{ic}) \exp\left[\left\langle \bar{V}_{ic}, C(\Theta_{ic})\right\rangle\right]$.

Dirichlet pdfs $Di_\alpha(\bar{\kappa})$, $Di_\alpha(\kappa_0)$ are used as flat pre-prior pdf and prior pdf, respectively. The adopted approximate estimation preserves this form so that the posterior pdf is also $Di_\alpha(\kappa_{\mathring{t}})$.

Algorithm 6.5 (Universal flattening with an almost linear rate)

Initial mode

- *Select upper bound $\mathring{n}$ on the number n of iterations and set $n = 1$.*
- *Select the statistics $\bar{\nu}, \bar{V}, \bar{\kappa}$ determining the pre-prior pdf $\bar{f}(\Theta)$.*

Iterative mode

1. *Perform approximate mixture estimation; see Section 6.5.*
2. *Evaluate*

$$T = \begin{cases} \mathring{t} & \textit{if no forgetting is used} \\ \sum_{t=1}^{\mathring{t}} \lambda^t \approx \frac{1}{1-\lambda} & \textit{if forgetting with factor } \lambda < 1 \textit{ is used} \end{cases}$$

$$\beta_n = \frac{\sum_{i \in i^*, c \in c^*} \nu_{ic;\mathring{t}}}{\mathring{d}T + \sum_{i \in i^*, c \in c^*} \nu_{ic;0}} \quad \textit{for the "factor" part of the posterior pdf}$$

$$\beta_{nD} = \frac{\sum_{c \in c^*} \kappa_{c;\mathring{t}}}{T + \sum_{c \in c^*} \kappa_{c;0}} \quad \textit{for the Dirichlet part of the posterior pdf.}$$

3. *Compute flattening rates Λ_n, Λ_{nD} and the powers of the likelihood ${}^{\lfloor \Lambda}A_n$, ${}^{\lfloor \Lambda}A_{nD}$ for the factor as well as for Dirichlet parts.*

$$\textit{For } n = 1, \quad \Lambda_1 \equiv \frac{\sum_{i \in i^*, c \in c^*} (\nu_{ic;0} - \bar{\nu}_{ic})}{\sum_{i \in i^*, c \in c^*} (\nu_{ic;\mathring{t}} - \bar{\nu}_{ic})} \textit{ and set } {}^{\lfloor \Lambda}A_1 \equiv \beta_1 \Lambda_1$$

$$\Lambda_{1D} \equiv \frac{\sum_{c\in c^*}\left(\kappa_{c;0} - \bar{\kappa}_c\right)}{\sum_{c\in c^*}\left(\kappa_{c;\mathring{t}} - \bar{\kappa}_c\right)} \quad \textit{and set } {}^{\lfloor\Lambda}A_{1D} \equiv \beta_{1D}\Lambda_{1D}.$$

$$\textit{For } n > 1, \quad k_n \equiv \min\left(\beta_n, \frac{1 - {}^{\lfloor\Lambda}A_{n-1}}{\mathring{n} - n + 1}\right), \ {}^{\lfloor\Lambda}A_n = {}^{\lfloor\Lambda}A_{n-1} + k_n$$

$$\Lambda_n = \frac{{}^{\lfloor\Lambda}A_n}{\beta_n + {}^{\lfloor\Lambda}A_{n-1}}$$

$$k_{nD} \equiv \min\left(\beta_{nD}, \frac{1 - {}^{\lfloor\Lambda}A_{(n-1)D}}{\mathring{n} - n + 1}\right)$$

$${}^{\lfloor\Lambda}A_{nD} = {}^{\lfloor\Lambda}A_{(n-1)D} + k_{nD}$$

$$\Lambda_{nD} = \frac{{}^{\lfloor\Lambda}A_{nD}}{\beta_{nD} + {}^{\lfloor\Lambda}A_{(n-1)D}}.$$

4. *Apply flattening as follows,* $i = 1, \ldots, \mathring{d}, \ c \in c^*$.

$$\begin{aligned}
\textit{For } n = 1, \ & V_{ic(n+1);0} = \Lambda_n V_{icn;\mathring{t}} + (1 - \Lambda_n)\bar{V}_{ic} \\
& \nu_{ic(n+1);0} = \Lambda_n \nu_{icn;\mathring{t}} + (1 - \Lambda_n)\bar{\nu}_{ic} \\
& \kappa_{c(n+1);0} = \Lambda_{nD}\kappa_{cn;\mathring{t}} + (1 - \Lambda_{nD})\bar{\kappa}_c \\
\textit{For } n > 1, \ & V_{ic(n+1);0} = \Lambda_n V_{icn;\mathring{t}} + (1 - \Lambda_n) V_{ic;0} \\
& \nu_{ic(n+1);0} = \Lambda_n \nu_{icn;\mathring{t}} + (1 - \Lambda_n)\nu_{ic;0} \\
& \kappa_{c(n+1);0} = \Lambda_{nD}\kappa_{cn;\mathring{t}} + (1 - \Lambda_{nD})\kappa_{c;0}.
\end{aligned}$$

5. *Stop if* $\mathring{n} = 1$ *and take the result as the newly generated prior pdf.*
6. *Increase* $n = n + 1$.
7. *Stop if* $n > \mathring{n} > 1$ *and* $\max\left({}^{\lfloor\Lambda}A_{\mathring{n}}, {}^{\lfloor\Lambda}A_{\mathring{n}D}\right) = 1$ *and take the result as the newly generated prior pdf.*
8. *Go to the beginning of* Iterative mode *if* $n \leq \mathring{n}$.
9. *Increase* $\mathring{n} = \mathring{n} + 1$ *and go to the beginning of* Iterative mode *if* $n > \mathring{n}$ *and* $\max\left({}^{\lfloor\Lambda}A_{\mathring{n}}, {}^{\lfloor\Lambda}A_{\mathring{n}D}\right) < 1$.

Remark(s) 6.8

1. *It is possible to take the result of the first estimation that started from a flat pre-prior pdf and was followed by flattening as the prior pdf if there is no justified alternative option.*
2. *The powers* ${}^{\lfloor\Lambda}A_{\mathring{n}}$, ${}^{\lfloor\Lambda}A_{\mathring{n}D}$ *have to be simultaneously checked for reaching unity. Generally, they reach unity after different numbers of iterations. We have selected conservative stopping leading to a flatter final pdf. It corresponds to the experience that over-confidence is more dangerous than under-confidence.*
3. *Use of this flattening should allow us to use forgetting extensively as the presented algorithm avoids the check on values of the forgetting factor discussed in Remarks 6.6.*

Problem 6.6 (Improvements of flattening strategies) *Experiments indicate superiority of forgetting with linear growth. Seemingly more realistic "almost linear growth" provides poorer results. It indicates that the understanding of the problem is still limited and there is a space for further improvements. Ideally, a sort of feedback during iterations could and should be used.*

6.4.4 Geometric mean as branching mapping

This subsection starts description of several branching mappings. They tailor the general branch-and-bound technique (see, Section 6.1.3) to the initialization problem summarized in Agreement 6.6.

The inspected estimates of the prior pdf $\hat{f}(\Theta)$ and the adopted approximate estimator, Section 6.5, determine the joint pdf of data and mixture parameters $X \equiv \hat{f}(\cdot,\cdot) \equiv \hat{f}(d(\mathring{t}),\Theta)$. This pdf is the argument of the maximized functional (6.38)

$$F(X) \equiv {}^{\lfloor h}\mathcal{L}\left(\hat{f}(\cdot,\cdot), d(\mathring{t})\right) \equiv \int \hat{f}(d(\mathring{t}),\Theta)\, d\Theta = \hat{f}(d(\mathring{t})).$$

While searching for maximizing argument we always preserve the best argument found until the considered step; cf. Proposition 6.3. Thus, for a description of the branching mapping, it is necessary to describe only how novel alternative guesses $\{\hat{f}(\Theta)\}$ of the best prior pdf are generated.

Here, we inspect the case when the set of candidates contains just a pair of pdfs $\hat{f}_1(\Theta)$, $\hat{f}_2(\Theta)$ and a new singleton is generated. These guesses of the prior pdf define the approximate posterior pdfs through an approximate estimation. We want to find the best point estimate of the posterior pdf. It is unknown because of the lack of complete knowledge about the adequate prior pdf.

Formally, we are facing exactly the same situation as discussed in Section 6.3.3. Similarly to Agreement 6.4, we introduce the unknown probability β_1 that the unknown prior pdf $f(\Theta)$ belongs to the set f_1^* containing pdfs that are closer to the guess $\hat{f}_1(\Theta)$ than to the guess $\hat{f}_2(\Theta)$. Adopting the approximation $f(\Theta) \approx \hat{f}_1(\Theta)$ on f_1^* and $f(\Theta) \approx \hat{f}_2(\Theta)$ on the complement of f_1^*, we get the (approximate) minimizing argument $\hat{f}(\Theta|d(\mathring{t}))$ of the expected KL divergence $\mathcal{E}\left[\mathcal{D}\left(f||\hat{f}\right)\right]$. It is a hot candidate for the value of the branching operator. This motivates the definition of the *geometric branching mapping*

$$\begin{aligned} \mathcal{A} \; : \; & \left\{\hat{f}_1(d(\mathring{t})), \hat{f}_2(d(\mathring{t})), \hat{f}_1(\Theta|d(\mathring{t})),\, \hat{f}_2(\Theta|d(\mathring{t}))\right\} \\ & \to \hat{f}(\Theta|d(\mathring{t})) \propto [\hat{f}_1\left(\Theta|d(\mathring{t})\right)]^{\lambda}[\hat{f}_2\left(\Theta|d(\mathring{t})\right)]^{1-\lambda} \\ \lambda \equiv \beta_{1|d(\mathring{t})} & = \frac{\hat{f}_1(d(\mathring{t}))}{\hat{f}_1(d(\mathring{t})) + \hat{f}_2(d(\mathring{t}))}. \end{aligned} \tag{6.62}$$

The used values of v-likelihood $\hat{f}_i(d(\mathring{t})) \equiv \int \hat{f}_i(d(\mathring{t})|\Theta)\hat{f}_i(\Theta)\, d\Theta,\ i = 1, 2$, are obtained as by-products of the quasi-Bayes algorithm (see formula (6.7) in

Section 6.1.1) or have to be approximated for a fixed predictor as discussed in Section 6.1.2.

Remark(s) 6.9

1. *We may use the geometric branching (almost) universally in conjunction with various initial guess of alternatives and various additional branching mappings $\mathcal{A}$. In this way, we get various versions of the general branch-and-bound Algorithm 6.1.*
2. *It is straightforward to specify a geometric branching that combines more than two variants into a single pdf. Simply,*

$$\hat{f}(d(\mathring{t}),\Theta) \propto \prod_{i=1}^{\mathring{i}} [\hat{f}_i(d(\mathring{t}),\Theta)]^{\lambda_i}, \quad \lambda_i \propto \int \hat{f}_i(d(\mathring{t}),\Theta)\, d\Theta, \ i = 1, 2, \ldots, \mathring{i}.$$

   It does not seem that it makes sense to deal with this extension in the discussed context.
3. *The structure of the estimated mixture does not change during iterations. Thus, it has to be sufficiently reflected even in the prior pdf $f(\Theta)$. Techniques allowing changes of the structure are described in Section 6.4.8.*

Problem 6.7 (Likelihood assigned to the geometric mean) *It is not certain that the found geometric mean gives a better results. Generally, it must be checked by repeating the estimation after flattening. The change in the corresponding v-likelihood can be predicted, however, without repeating the estimation. It is implied by the Bayes rule (2.8)*

$$\hat{f}_i(\Theta|d(\mathring{t})) = \frac{\mathcal{L}(\Theta, d(\mathring{t}))\hat{f}_i(\Theta)}{\int \mathcal{L}(\Theta, d(\mathring{t}))\hat{f}_i(\Theta)\, d\Theta} = \frac{\mathcal{L}(\Theta, d(\mathring{t})) f_i(\Theta)}{\hat{f}_i(d(\mathring{t}))}, \ i = 1, 2, \underbrace{\Rightarrow}_{(6.62)}$$

$$\hat{f}_3(\Theta|d(\mathring{t})) = \frac{[\hat{f}_1(\Theta|d(\mathring{t}))]^{\lambda}[\hat{f}_2(\Theta|d(\mathring{t}))]^{1-\lambda}}{\int [\hat{f}_1(\Theta|d(\mathring{t}))]^{\lambda}[\hat{f}_2(\Theta|d(\mathring{t}))]^{1-\lambda}\, d\Theta} \underbrace{\Rightarrow}_{(2.8)}$$

$$\hat{f}_3(\Theta|d(\mathring{t})) = \frac{\mathcal{L}(\Theta, d(\mathring{t}))\hat{f}_1^{\lambda}(\Theta)\hat{f}_2^{1-\lambda}(\Theta)}{\hat{f}_3(d(\mathring{t}))\int \hat{f}_1^{\lambda}(\Theta)\hat{f}_2^{1-\lambda}(\Theta)\, d\Theta} \Rightarrow$$

$$\hat{f}_3(d(\mathring{t})) = [\hat{f}_1(d(\mathring{t}))]^{\lambda}[\hat{f}_2(d(\mathring{t}))]^{1-\lambda} \frac{\int \left[\hat{f}_1(\Theta|d(\mathring{t}))\right]^{\lambda} \left[\hat{f}_2(\Theta|d(\mathring{t}))\right]^{1-\lambda} d\Theta}{\int \left[\hat{f}_1(\Theta)\right]^{\lambda} \left[\hat{f}_2(\Theta)\right]^{1-\lambda} d\Theta}.$$

The above formula is an approximate one as the approximately evaluated likelihood function depends on the chosen prior pdf. The derived relationship has not been tested sufficiently. Its analogy is, however, widely and successfully exploited for merging components and factors; see Section 6.6.4. Thus, it is worth trying it as it could substantially enhance the potential behind the geometric-mean mapping.

6.4.5 Random branching of statistics

The deterministic optimization algorithms are prone to find just a local optimum. It is known that the global optimization calls for a random search [132]. Computational complexity prevents us to use the safe, completely random search. Thus, we should use a combination of deterministic and random searches. Here, we touch on possible random steps.

According to Agreement 6.1, all considered pdfs on unknown parameters of the mixture have the form (6.3). In learning, we have to consider the cases when the whole data history $d(\mathring{t})$ reduces to a finite-dimensional (almost) sufficient statistics, say $\{\mathcal{V}_{ic}\}$. It means that estimates are searched in the form $\hat{f}(\Theta_{ic}|d(\mathring{t})) = \hat{f}(\Theta_{ic}|\mathcal{V}_{ic})$, $i \in \{1,\ldots,\mathring{d}\}$, $c \in c^*$. The functional form of the right-hand sides as well as the set $\mathcal{V}_{ic}^*$ of possible values of $\mathcal{V}_{ic}$ are known.

Knowing this, we can generate random samples $\tilde{\mathcal{V}}_{ic}$ in $\mathcal{V}_{ic}^*$. It gives a new guess $\tilde{f}(\Theta_{ic}|d(\mathring{t})) \equiv \hat{f}(\Theta_{ic}|\tilde{\mathcal{V}}_{ic})$. By flattening (see Proposition 6.6), we get a new guess of the prior pdf $\hat{f}_{new}(\Theta_{ic})$.

Remark(s) 6.10

1. *The random generating should be applied to selected factors only. It decreases the computational load and, moreover, preserves those estimates that are satisfactory. Typically, the candidate factors are those that provide predictors similar to predictors constructed from the prior pdf only.*
2. *The random sample $\tilde{\mathcal{V}}_{ic}$ may or may not be related to the given $\mathcal{V}_{ic}$. The latter option allows us a less restricted search and increases our chance of reaching the best possible result. Computationally it may be prohibitive.*
3. *Samples $\tilde{\mathcal{V}}_{ic}$ around $\mathcal{V}_{ic}$ should be searched for, if worries of the previous item apply. The Bayes rule implies that*
$$f(\mathcal{V}_{ic}) = f(d(\mathring{t})) = \frac{f(d(\mathring{t})|\Theta_{icj})f(\Theta_{icj})}{f(\Theta_{icj}|d(\mathring{t}))}$$
for any $\{\Theta_{icj}\}_{j\in j^*} \in \Theta^*$. Thus, we know the functional form of $f(\mathcal{V}_{ic})$ and we are able to evaluate this pdf for any values $\mathcal{V}_{ic}$. Thus, we can get a clue this pdf looks like and use it potentially for generating highly probable samples of the statistics.

6.4.6 Prior-posterior branching

The posterior pdf flattened back to the space of potential prior pdfs provides an alternative to the original prior pdf. Thus, $\mathcal{A}_\Lambda$ (6.43) can be used as a promising branching mapping. Flattening combined with geometric branching (6.62) we call *prior-posterior branching.* It gives the following iterative learning algorithm.

Algorithm 6.6 (Prior-posterior branching)

Initial mode

- *Select an upper bound* $\mathring{n}$ *on the number* n *of iterations and set* $n = 1$.
- *Select a sufficiently rich structure of the mixture; see Agreement 5.4.*
- *Select the pre-prior* $\bar{f}(\Theta) =$ *a flat prior pdf used for flattening.*
- *Select the prior pdf* $\hat{f}_{1n}(\Theta) \equiv f(\Theta)$ *(not necessarily equal to* $\bar{f}(\Theta)$*).*
- *Compute the approximate posterior pdf* $\tilde{f}_{1n}(\Theta|d(\mathring{t})) \propto f(d(\mathring{t})|\Theta)\hat{f}_{1n}(\Theta)$ *using an approximate Bayesian estimation; see Section 6.5.*
- *Evaluate the v-likelihood* l_{1n} *resulting from the use of* $\hat{f}_{1n}(\Theta)$.
- *Apply the flattening operation to* $\tilde{f}_{1n}(\Theta|d(\mathring{t}))$ *according to Propositions 6.7 (on the Dirichlet marginal) and 6.8 (on pdfs describing factors). Denote the resulting pdf* $\hat{f}_{2n}(\Theta)$.
- *Compute the approximate posterior pdf* $\tilde{f}_{2n}(\Theta|d(\mathring{t})) \propto f(d(\mathring{t})|\Theta)\hat{f}_{2n}(\Theta)$ *using an approximate Bayesian estimation; see Section 6.5.*
- *Evaluate the v-likelihood* l_{2n} *resulting from the use of* $\hat{f}_{2n}(\Theta)$.
- *Set* $\bar{l}_n = \max(l_{1n}, l_{2n})$.

Iterative mode

1. *Apply geometric branching to the pair* $\tilde{f}_{1n}(\Theta|d(\mathring{t}))$, $\tilde{f}_{2n}(\Theta|d(\mathring{t}))$ *with the v-likelihood* l_{1n}, l_{2n}, *respectively. For* $\lambda \equiv \frac{l_{1n}}{l_{1n}+l_{2n}}$, *it gives*
$$\tilde{f}_{3n}(\Theta|d(\mathring{t})) \propto \left[\tilde{f}_{1n}(\Theta|d(\mathring{t}))\right]^{\lambda} \left[\tilde{f}_{2n}(\Theta|d(\mathring{t}))\right]^{1-\lambda}.$$
2. *Apply the flattening operation on* $\tilde{f}_{3n}(\Theta|d(\mathring{t}))$ *according to Propositions 6.7 (on the Dirichlet marginal) and 6.8 (on pdfs describing factors). Denote the resulting pdf* $\hat{f}_{3n}(\Theta)$.
3. *Compute the approximate posterior pdf* $\tilde{f}_{3n}(\Theta|d(\mathring{t}))$ *resulting from the use of* $\hat{f}_{3n}(\Theta)$ *as the prior pdf. Evaluate the v-likelihood* l_{3n} *related to the use of* $\hat{f}_{3n}(\Theta)$ *as the prior pdf.*
4. *Choose among* $\tilde{f}_{in}(\Theta|d(\mathring{t}))$, $i \in \{1, 2, 3\}$ *the pair with the highest v-likelihood* l_{in}. *Call them* $\tilde{f}_{1(n+1)}(\Theta|d(\mathring{t}))$, $\tilde{f}_{2(n+1)}(\Theta|d(\mathring{t}))$ *with the v-likelihood* $l_{1(n+1)}$, $l_{2(n+1)}$.
5. *Go to the beginning of* Iterative mode *with* $n = n + 1$ *if*
$$\bar{l}_{n+1} \equiv \max(l_{1(n+1)}, l_{2(n+1)}) > \bar{l}_n,$$
or if $\bar{l}_{n+1}, \bar{l}_n$ *are the same according to Proposition 6.2, and if* $n < \mathring{n}$.
6. *Stop and select among* $\hat{f}_{in}(\Theta)$, $i = 1, 2$, *the pdf leading to the higher value of* l_{in} *as the prior pdf constructed.*

Remark(s) 6.11

1. *No improvement can be expected when the values of* λ *stabilize around 0.5. This observation may serve as an additional stopping rule.*

2. *The algorithm is applicable out of the considered context of the mixture estimation.*
3. *The structure of the estimated mixture does not change during iterations. Thus, it has to be sufficiently reflected even in the prior pdf $f(\Theta)$. Techniques allowing changes of the structure are described in Section 6.4.8.*

Problem 6.8 (Controlled prior-posterior branching) *Prediction of the v-likelihood after making geometric mean (see Problem 6.7) can be used for controlling the branching algorithm. The comparison of the predicted and evaluated v-likelihood values indicates how much the likelihood varies due to the approximate estimation used.*

6.4.7 Branching by forgetting

Each iteration of the Bayes rule described in the previous section is performed over all data available $d(\mathring{t})$. It might be a waste of precious computational resources. Moreover, it cannot be used in online mode. The technique described here helps us to avoid these drawbacks. Essentially, the considered alternatives are generated through parallel quasi-Bayes estimations, Algorithm 6.13, differing in forgetting factors used.

For a given $\hat{f}(\Theta|d(t))$, two approximate estimations run in parallel, one with the unit forgetting and the other one with a forgetting $\lambda << 1$. Their predictive abilities are evaluated through their v-likelihood. Their geometric mean coincides with one of the compared pdfs if the absolute difference of their v-log-likelihood values is high enough. At this moment, the poorer variant can be reset to the better alternative. This description is formalized in the following algorithm.

Algorithm 6.7 (Online branching with forgetting)

Initial mode

- *Select a sufficiently rich structure of the mixture.*
- *Select a constant $\rho \approx$ 3-5 defining the significant difference of v-log-likelihood values.*
- *Set the record counter $t = 0$ and select the prior pdf $\hat{f}(\Theta|d(t)) \equiv f(\Theta)$.*
- *Choose a fixed, relatively small forgetting factor $\lambda < 1$.*
- *Select a fixed alternative pdf used in the stabilized forgetting; see Proposition 3.1. The alternative is either flat pre-prior pdf or the chosen prior pdf.*

Data processing mode

1. *Set $\hat{f}_1(\Theta|d(t)) = \hat{f}_\lambda(\Theta|d(t)) = \hat{f}(\Theta|d(t))$.*
2. *Initialize v-log-likelihood values assigned to considered alternatives $l_{1;t} = 0$, $l_{\lambda;t} = 0$.*
3. *Collect new data d_{t+1}.*

4. *Update* $\hat{f}_1(\Theta|d(t))$ *to* $\hat{f}_1(\Theta|d(t+1))$ *using approximate estimation, Section 6.5, with the forgetting factor 1.*
5. *Recompute the v-log-likelihood* $l_{1;t}$ *to* $l_{1;t+1}$ *by adding the logarithm of the mixture-based prediction* $\ln(f(d(t+1)|d(t)))$ *obtained for the "prior" pdf* $\hat{f}_1(\Theta|d(t))$.
6. *Update* $\hat{f}_\lambda(\Theta|d(t))$ *to* $\hat{f}_\lambda(\Theta|d(t+1))$ *using approximate estimation with the forgetting factor* λ *and the chosen alternative pdf.*
7. *Recompute the v-log-likelihood* $l_{\lambda;t}$ *to* $l_{\lambda;t+1}$ *by adding logarithm of the mixture-based prediction* $\ln(f(d(t+1)|d(t)))$ *obtained for the "prior" pdf* $\hat{f}_\lambda(\Theta|d(t))$.
8. *Go to Step 3 with* $t = t+1$ *if* $|l_{1;t+1} - l_{\lambda;t+1}| < \rho$.
9. *Set* $\hat{f}(\Theta|d(t+1)) = \hat{f}_1(\Theta|d(t+1))$ *if* $l_{1;t+1} > l_{\lambda;t}$*, otherwise assign* $\hat{f}(\Theta|d(t+1)) = \hat{f}_\lambda(\Theta|d(t+1))$.
10. *Set* $t = t+1$ *and go to the beginning of* Data processing mode *if* $t \leq \mathring{t}$.
11. *Take* $\hat{f}(\Theta|d(\mathring{t}))$ *as the final estimate of the posterior pdf.*

Remark(s) 6.12

1. *The decision on the appropriate model, based on the difference of the involved v-likelihood, is more sensitive to the choice of* ρ *than expected. An insufficient reliability of the posterior pdfs seems to be the real reason for this behavior. This hints at the appropriate remedy: decide on the model quality after checking that the probability of the branch with no forgetting stabilizes in the time course.*
2. *Speeding up of the learning is the main expectation connected with this algorithm. The model with no forgetting is expected to be the winner in long run. Thus, the estimation with forgetting can be switched off when the estimation without forgetting is better majority of the time. The rule for switching off the estimation with forgetting might look like as follows.* Switch off the estimation with forgetting if the estimation without it is better for a sufficiently large portion of time.
3. *There is a danger of a numerical breakdown when the update with forgetting wins for a longer time. This case indicates either a very poor start or a significantly wrong model structure. The use of a proper alternative in forgetting removes this danger.*
4. *The estimation with no forgetting should always be included as we search for a time invariant model of the o-system.*
5. *The technique can be directly combined with prior-posterior branching.*
6. *The choice of the fixed forgetting factor* λ *is based on experience. At present, the option* $\lambda \approx 0.6$ *seems to be satisfactory.*
7. *The necessary condition needed for standard flattening* $\sum_{i\in i^*,c\in c^*} \nu_{ic;\mathring{t}} > \sum_{i\in i^*,c\in c^*} \nu_{ic;0}$ *may be violated when forgetting* $\lambda < 1$ *wins too often and noninformative alternative pdf is used. Use of the stabilized forgetting with a simplified alternative which preserves the degrees of freedom is the recommended way to avoid the problem.*

Problem 6.9 (Branching by forgetting at factor level) *The algorithm is presented at the mixture level: the whole mixture is a winner if its v-likelihood is sufficiently higher than its competitor. The strategy can be and should be extended up to the factor level. This extension looks quite promising and worth trying.*

6.4.8 Branching by factor splitting

The design of the branching mapping (6.15) presented in this section is unique as it suits the cases when there is no clue on mixture structure, i.e., structure of factors, presence of common factors and the number of components.

The basic idea of the branching by factor splitting is simple. Let us assume that the mixture structure, Agreement 5.4, and the estimate of the prior pdf $\hat{f}_n(\Theta)$ are selected in an nth iteration. The approximate estimation, Section 6.5, gives the posterior pdf $\tilde{f}_n(\Theta|d(\mathring{t})) \propto f(d(\mathring{t})|\Theta)\hat{f}_n(\Theta)$.

During approximate estimation, Section 6.5, we can check whether a particular factor covers two modes of the pdf predicting $d_{ic;t}$. We split each factor that covers two modes into a new pair that preserves some properties of the original factor but has a chance to fit the processed data better.

Structure of factors is estimated and possibly modified during splitting. We get a new, substantially larger, number of components that are re-estimated after flattening. Then, similar components are merged or spurious ones cancelled. Consequently a simpler mixture is obtained. The whole process is repeated while the v-likelihood increases.

For reference purposes, we list steps that have to be elaborated in detail.

Algorithm 6.8 (Key steps of branching by splitting)

1. *Selection of the very initial mixture to be split.*
2. *Selection of factors for splitting.*
3. *Splitting of factors.*
4. *Estimation of the factor structure; see Section 6.6.1.*
5. *Construction of a new mixture.*
6. *Merging of components; see Section 6.6.4.*
7. *Cancelling of components; see Section 6.6.4.*

The steps without reference are discussed in this section. The structure estimation has a wider use and its parts are discussed in the referred subsections of Section 6.6.

Selection of the initial mixture: Step 1 in Algorithm 6.8

The very initial mixture, corresponding to the selection of the initial set of the alternatives in branch-and-bound Algorithm 6.1, has to be selected first. It should

- give a chance to find the objective model,
- exclude part of the data space that contains (almost) no data.

The first condition requires selection of sufficiently rich structures of factors involved. For reference purposes, we formulate it formally.

Requirement 6.2 (Over-parameterized factors) *The regression vectors* $\psi_{ic;t}$ *of parameterized factors*

$$f(d_{i;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1), \Theta_{ic}, c) \equiv f(d_{i;t}|\psi_{ic;t}, \Theta_{ic}, c)$$

include regression vectors ${}^{\lfloor o}\psi_{i;t}$ *of the objective pdf*

$${}^{\lfloor o}f(d_{i;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1)) \equiv {}^{\lfloor o}f(d_{i;t}|\,{}^{\lfloor o}\psi_{i;t}), \text{ cf. Section 2.3.1.}$$

This requirement is relatively simply fulfilled due to the slow dynamics of the inspected o-system; cf. Section 5.2. It must, however, be respected when reducing the model structure; see Section 6.6.

To meet the second requirement on the very initial mixture, we used the single-component mixture. It safely excludes the space out of the union of support of the objective components. It cannot exclude emptiness of the inter-component space. This often causes bad behavior during the search. Typically, the attempts to improve the one-component mixture fail and the initialization stops prematurely. Initial splitting performed without the check of the v-likelihood brings practical remedy of the problem. The depth $\mathring{m}$ of this initial splitting is usually restricted by the problem complexity so that its choice is not difficult. This uncontrolled splitting relies on our ability to merge similar components; see Section 6.6.4.

Algorithm 6.9 (Selection of the very initial mixture)

Initial mode

- *Select quantities in* d_t *to be modelled.*
- *Select a structure of the richest regression vector.*
- *Select an initial-splitting depth* $\mathring{m} > 1$ *and set the depth counter* $m = 0$.
- *Select a (flat) prior pdf* $f(\Theta)$ *on the single component mixture.*
- *Select an initial value of the statistics* κ_0 *describing the Dirichlet pdf (10.2). It serves for flattening according to Proposition 6.7.*

Evaluation mode

1. *Increase* $m = m + 1$.
2. *Perform (approximate) estimation of the mixture; see Section 6.5.*
3. *Split factors as described in subsequent sections.*
4. *Apply flattening operation, Propositions 6.7, 6.8, so that a new initial mixture estimate is obtained.*
5. *Go to the beginning of* Evaluation mode *if* $m < \mathring{m}$. *Otherwise take the resulting mixture as the very initial one for further evaluations.*

Remark(s) 6.13
Other choices, like random starts, are possible and should be considered in specific cases. A random start, however, does not guarantee that we shall not stay in a weakly populated part of the data space. Moreover, available generators of this type care mostly about initial positions only. They do not provide a complete description of the prior pdf.

Hierarchical selection of split factors and their split: Steps 2 and 3 in Algorithm 6.8

The *hierarchical splitting* described here solves both referred steps. It relies on our ability to estimate three factors instead of a single one.

During the approximate estimation, Section 6.5, the ith factor within the cth component is assigned a weight $w_{ic;t} \in [0,1]$ that expresses the expectation that the data item $d_{ic;t}$ is generated by this factor. Then, the modified parameterized factor

$$f_w(d_{ic;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1)) \equiv f_w(d_{ic;t}|\psi_{ic;t},\Theta_{ic}) \propto \left[f(d_{ic;t}|\psi_{ic;t},\Theta_{ic})\right]^{w_{ic;t}}$$

is used in the "ordinary" Bayes rule for data updating of the posterior pdf describing the estimate of factor parameters. Thus, for estimation purposes, we can inspect this factor without taking into account others. It allows us to simplify notation and temporarily drop the subscripts w, c.

We want to check whether the discussed factor explains data that should be modelled by more factors while assuming that Requirement 6.2 is met. We formulate hypothesis H_0 that the objective pdf has two components

$$ {}^{\lfloor o}f(d_t|\psi_t) = \beta f(d_t|\psi_t,\Theta_1) + (1-\beta) f(d_t|\psi_t,\Theta_2), \ \beta \in (0,1), \tag{6.63}$$

where Θ_1, Θ_2 have the same structure as the parameter Θ in the split factor.

The alternative hypothesis H_1 assumes that ${}^{\lfloor o}f(d_t|\psi_t) = f(d_t|\psi_t, {}^{\lfloor o}\Theta)$, i.e., denies the need for splitting.

If we can afford it, we estimate parameters of the mixture (6.63) together with the factor in question in order to decide on the need to split. With $f(H_0) = f(H_1)$, modelling no prejudice, the Bayes rule gives

$$\text{Probability(split is needed}|d(\mathring{t})) \equiv f(H_0|d(\mathring{t})) = \frac{f(d(\mathring{t})|H_0)}{f(d(\mathring{t})|H_0) + f(d(\mathring{t})|H_1)}. \tag{6.64}$$

The v-likelihood values $f(d(\mathring{t})|H_0)$, $f(d(\mathring{t})|H_1)$ are obtained as a byproduct of the approximate estimation applied to the simple mixture (6.63), Section 6.5.

The factor is split if the probability (6.64) is high enough. The estimated factors of the mixture (6.63) are natural candidates for its replacement.

For reference purposes, let us write down the corresponding algorithm.

Algorithm 6.10 (Hierarchical split of factors)

Initial mode

- *Construct the very initial estimate* $f(\Theta)$ *of the mixture, Algorithm 6.9.*
- *Select significance level* $\bar{P} \in (0,1)$, $\bar{P} \approx 1$ *controlling the need to split.*
- *Assign to each factor* ic, $i = 1, \ldots, \mathring{d}$, $c \in c^*$, *a prior pdf* $f(\beta_{ic}, \Theta_{1ic}, \Theta_{2ic})$ *on parameters of the mixture (6.63). Indices ic stress the "membership" of this mixture to the factor ic.*
- *Initialize v-likelihoods* $l_{ic;0|H}$ *corresponding to respective hypotheses* $H \in \{H_0, H_1\}$ *and factors* $i \in i^*$, $c \in c^*$.

Sequential mode, *running for* $t = 1, 2, \ldots,$

1. *Acquire data* d_t *and complement the data vectors* $\Psi_{ic;t} = [d_{ic;t}, \psi'_{ic;t}]'$, $i = 1, \ldots, \mathring{d}$, $c \in c^*$.
2. *Perform a single step of an approximate estimation (see Section 6.5), i.e., update* $f(\Theta|d(t-1)) \to f(\Theta|d(t))$. *The estimation generates weights* $w_{ic;t}$ *expressing the degree with which the data* d_t *are assigned to the parameterized factor ic. Update at the same time the values of the v-likelihood* $l_{ic;t|H_1}$.
3. *Weight the current data by respective* $w_{ic;t}$. *In the exponential family (3.6), it replaces unit power of* $A(\Theta_{ic})$ *by* $w_{ic;t}$ *and* $B(\Psi_{ic;t})$ *by* $w_{ic;t}B(\Psi_{ic;t}) \equiv B(\Psi_{wic;t})$. *Update estimates of parameters of the models (6.63)*
$$f(\beta_{ic}, \Theta_{1ic}, \Theta_{2ic}|d(t-1)) \to f(\beta_{ic}, \Theta_{1ic}, \Theta_{2ic}|d(t)), \quad \textit{using } \Psi_{wic;t}.$$
Update also values of the v-likelihood $l_{ic;t|H_0}$ *using the predictors*
$$f(d_{wic;t}|\psi_{wic;t}, d(t-1)), \ \textit{with } \Psi_{wic;t} \equiv [d_{wic;t}, \psi'_{wic;t}]' \textit{ and}$$
constructed from the estimated two-component mixture (6.63).
4. *Go to the beginning of* Sequential mode *while* $t \leq \mathring{t}$.

Selection mode *running for* $i = 1, \ldots, \mathring{d}$, $c \in c^*$

Split $f(\Theta_{ic}|d(\mathring{t})) \to \left(f(\Theta_{1ic}|d(\mathring{t})), f(\Theta_{2ic}|d(\mathring{t}))\right)$ *if*
$$\frac{l_{ic;\mathring{t}|H_0}}{l_{ic;\mathring{t}|H_0} + l_{ic;\mathring{t}|H_1}} \geq \bar{P}.$$

Heuristic choice of split factors: an alternative for Step 2 in Algorithm 6.8

The hierarchically-based splitting depends heavily on our ability to face the computational demands caused by the need to estimate three factors instead of each original factor. In this and subsequent paragraph, less demanding ways are discussed.

Simplification is reached by leaving out the full estimation of subfactors hidden in the factor to be split. It is done by

- a heuristic step, for instance, by inspecting only the most flat factors or factors in the most populated component,
- a tailored hypothesis testing,
- exploiting a reduced version of the hierarchical split; see Chapter 8.

Other variants were also considered. Generally, experience is rather mixed but the last item seems to be the most preferable.

Optimization-based factor splitting: an alternative of Step 3 in Algorithm 6.8

Simplified splitting of a selected factor is discussed here. The first variant takes new factors as flattened or sharpened versions of the factor in question.

Proposition 6.14 (Flattened / sharpened factors) *Let pdfs $f, \bar{f}$ have a common support Θ^*. Let the pdf $\hat{f}$ minimize the KL divergence $\mathcal{D}\left(\hat{f}||f\right)$ and have the prespecified KL divergence*

$$\mathcal{D}\left(\hat{f}||\bar{f}\right) = \omega\mathcal{D}\left(f||\bar{f}\right),\ \omega \neq 1. \ \text{ Then, it has the form } \hat{f} \propto f^{\lambda}\bar{f}^{1-\lambda}. \tag{6.65}$$

The scalar λ is chosen so that the constraint in (6.65) is met.

Proof. We minimize the convex functional and the minimizing argument of the Lagrangian is searched for. It coincides, however, with the functional (6.41), where q is now the Lagrangian multiplier. It gives a flattened version if $q > 0 \Leftrightarrow \lambda < 1$ and a sharpened version if $q < 0 \Leftrightarrow \lambda > 1$. □

The obtained factor has a limited use since it preserves the original-factor mode. Moreover, the choice of the optional value ω is generally unsolved.

More promising is the version that enforces explicitly a shift of the original-factor mode. The mode is characterized through the expected value of a function $g(\Theta)$ depending on the estimated parameter Θ.

Proposition 6.15 (Shifted factor) *Let pdfs $\hat{f}, f$ have a common support Θ^* and let $g(\Theta)$ be an array-valued function defined on Θ^*. Let the pdf $\hat{f}$ minimize the KL divergence $\mathcal{D}\left(\hat{f}||f\right)$ and have the prescribed norm of the shift in expectation of the function $g(\Theta)$*

$$\omega = \left\langle \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta, \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta \right\rangle,\ \omega > 0.$$

$$\text{Then, it has the form } \hat{f}(\Theta) \propto f(\Theta)\exp[s\left\langle \mu, g(\Theta)\right\rangle],\ s \in \{-1, 1\}, \tag{6.66}$$

determined by the constant array μ. The array μ is compatible with the scalar product $\langle\cdot,\cdot\rangle$ and the solution of the considered task is in the set of functions (6.66) "indexed" by μ. The array μ has to fulfill the equality

$$\mu = \rho \int g(\Theta) \left(\frac{f(\Theta)\exp[-\langle \mu, g(\Theta)\rangle]}{\int f(\tilde{\Theta})\exp[-\left\langle \mu, g(\tilde{\Theta})\right\rangle]\, d\tilde{\Theta}} - f(\Theta) \right) d\Theta \tag{6.67}$$

with the scalar ρ determined by the identity $\langle \mu, \mu \rangle = \rho^2 \omega$.

Proof. The Lagrangian of this optimization task is

$$\begin{aligned} \mathcal{J} \equiv \mathcal{D}\left(\hat{f}||f\right) + \nu \int \hat{f}(\Theta)\, d\Theta \\ +\rho \left\langle \int g(\Theta)\left(\hat{f}(\Theta) - f(\Theta)\right) d\Theta, \int g(\Theta)\left(\hat{f}(\Theta) - f(\Theta)\right) d\Theta \right\rangle. \end{aligned}$$

It is determined by scalars ν, ρ that have to be chosen so that constraints, normalization and the restriction on the shift norm, are met. Let us denote $\mu \equiv \rho \int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta$. Then, the Lagrangian can be written in the form

$$\begin{aligned} &\mathcal{J} = constant \\ &+ \int \hat{f}(\Theta) \ln \left(\frac{\hat{f}(\Theta)}{\frac{f(\Theta)\exp[-\langle \mu, g(\Theta)\rangle]}{\int f(\Theta)\exp[-\langle \mu, g(\Theta)\rangle]\, d\Theta}} \right) d\Theta - \ln\left[\int f(\Theta)\exp[-\langle \mu, g(\Theta)\rangle]\, d\Theta \right]. \end{aligned}$$

Basic properties of the KL divergence imply the claimed form of the minimizer. The found form of the minimizer and definition of μ imply

$$\mu = \rho \int g(\Theta) \left(\frac{f(\Theta)\exp[-\langle \mu, g(\Theta)\rangle]}{\int f(\tilde{\Theta})\exp[-\left\langle \mu, g(\tilde{\Theta})\right\rangle]\, d\tilde{\Theta}} - f(\Theta) \right) d\Theta.$$

The restriction on the norm of the shift determines the constraint on the length of the vector μ.

The freedom in the choice of s is implied by the even nature of the constraint in (6.66). □

Remark(s) 6.14

1. *The proposed shifting may fail if the over-parameterized factor is split. The shift tends to directions, where the split factor is less concentrated. In the over-parameterized case, such a shift brings nothing positive. Thus, it is necessary to perform* structure estimation *on the factor selected for splitting and get rid of superfluous parameters.*
2. *Constraints in Propositions 6.14, 6.15 can be used simultaneously, giving the candidate pdfs in the predictable form*

$$\hat{f}(\Theta) \propto f^{\lambda}(\Theta)\bar{f}^{1-\lambda}(\Theta)\exp\left[\langle \mu, g(\Theta)\rangle\right] \tag{6.68}$$

parameterized by a scalar λ and an array μ. The optimum can be searched for them only.

3. *A proper choice of the function $g(\Theta)$ helps us to stay within a desirable class of pdfs. Typically, we try to stay within the computationally advantageous class of conjugate pdfs, (3.13).*
4. *A linear term in $\int g(\Theta)(\hat{f}(\Theta) - f(\Theta))\, d\Theta$ can be included in the constraint (6.66) without changing the form of the results. It adds a freedom that may be used for obtaining a suitable form of $\hat{f}$.*
5. *The constraint $\hat{f} \geq 0$ was not explicitly considered while optimizing. The solution meets it automatically.*

The initialization results based on Proposition 6.15 were found to be too sensitive with respect to a specific choice of the optional parameters $\langle \cdot, \cdot \rangle$, $g(\Theta)$, ω. For this reason, an alternative formulation was tried. The following proposition reflects a step made in this direction. It gets rid of the choice of the product $\langle \cdot, \cdot \rangle$ and supports intuitive reasons for selecting $g(\Theta) = \Theta$.

Proposition 6.16 (Alternative shifted factor) *Let us consider pdfs $\hat{f}, f$ with a common support Θ^* and finite gradients on Θ^*. Moreover, let a function $g : \Theta^* \to \Theta^*$ and scalars $\beta, b \in (0,1)$ be chosen. We search for such a pdf $\hat{f}$ that*

- *minimizes the KL divergence $\mathcal{D}\left(\hat{f}||f\right)$,*
- *has the prescribed shift in expectation of $g(\Theta)$*

$$f\left(\int g(\Theta)\hat{f}(\Theta)\, d\Theta\right) = \beta f\left(\int g(\Theta) f(\Theta)\, d\Theta\right), \ \beta \in (0,1), \tag{6.69}$$

- *has the prescribed ratio of "peaks"*

$$\hat{f}\left(\int g(\Theta)\hat{f}(\Theta)\, d\Theta\right) = b^{-1} f\left(\int g(\Theta) f(\Theta)\, d\Theta\right), \ b \in (0,1). \tag{6.70}$$

Then, it has the form

$$\hat{f}(\Theta) \propto f(\Theta) \exp[\mu' g(\Theta)] \tag{6.71}$$

determined by a vector μ chosen so that the constraints (6.69), (6.70) are met.

Proof. The Lagrangian of this optimization task is

$$\mathcal{J} \equiv \mathcal{D}(\hat{f}||f) + \nu \int \hat{f}(\Theta)\, d\Theta + \rho f\left(\int g(\Theta)\hat{f}(\Theta)\, d\Theta\right) + \lambda \hat{f}\left(\int g(\Theta)\hat{f}(\Theta)\, d\Theta\right).$$

It is determined by scalars ν, ρ and λ that have to be chosen so that constraints — normalization, (6.69) and (6.70) — are met. Its variation $\delta\mathcal{J}$, which should be equal to zero for the optimal $\hat{f}$, reads

$$\delta\mathcal{J} = \int \hat{\delta}\left[\ln(\hat{f}/f) + \nu + 1 + g'(\Theta)\mu\right] d\Theta, \text{ with } \mu = \rho\frac{\partial f(\Theta)}{\partial\Theta} + \lambda\frac{\partial \hat{f}(\Theta)}{\partial\Theta}$$

$$\text{evaluated at the point } \int g(\Theta)\hat{f}(\Theta)\,d\Theta.$$

This implies the claimed results. □

Remark(s) 6.15

1. *The requirement (6.69) has the following meaning. The positions of the new factor $\hat{f}(\Theta)$, characterized by $\int g(\Theta)\hat{f}(\Theta)\,d\Theta$ should be still sufficiently probable when judged from the viewpoint of the split pdf $f(\Theta)$. It should, however, differ from the original position of $f(\Theta)$ characterized by $\int g(\Theta)f(\Theta)\,d\Theta$. The selection $\beta < 1$ guarantees it.*
2. *As a rule, there are many pdfs $\hat{f}$ that minimize the divergence $\mathcal{D}\left(\hat{f}||f\right)$ under the constraint (6.69). We can use this freedom for selecting a more concentrated pdf than the split one so that the corresponding parameter uncertainty projected into the data space is smaller. The requirement (6.70) expresses this wish. The requirement fixes ratio of "peaks" of the involved pdfs. The pdf values at expectations determining respective positions are taken as the compared peaks. The selection $b < 1$ guarantees that the new pdf has a higher peak and thus is narrower than the split factor f.*
3. *Intuitively, for each pair $(\beta, b) \in (0,1)\times(0,1)$ there are pdfs fulfilling both conditions (6.69), (6.70). Nonemptiness of this set has to be, of course, checked. Those pdfs $\hat{f}$ that minimize the distance to the split pdf f while meeting these conditions are taken as a result of splitting. Additional conditions can be formulated if the formulated task has more solutions.*
4. *The specific choices β, b have to be studied in order to find a compromise between the contradictory requirements. We want to*
   - *make the shift as large as possible,*
   - *stay within the domain of attraction of the factor split,*
   - *make the resulting pdf reasonably narrow.*

Intuition and experiments indicate that the choice $g(\Theta) = \Theta$ is reasonable. The results were, however, found sensitive to the optional parameters β, b. This stimulated attempts to search for a more objective choice of constraints. Proposition 2.15 provides a guide. According to it, the posterior pdf $f(\Theta|d(t))$ concentrates on minimizing arguments of entropy rate (2.48). For motivating thoughts, we assume that it converges to its expected value and that we deal with the parameterized model from the exponential family, Section 3.2. Thus, the posterior pdf $f(\Theta|d(t))$, conjugated to this family, concentrates on

$$\lim_{\mathring{t}\to\infty} \text{supp}\left[f(\Theta|d(\mathring{t}))\right] = \text{Arg}\min_{\bar{\Theta}\in\Theta^*}\left\{-\ln(A(\bar{\Theta})) - {}^{\lfloor o}\mathcal{E}\left[\langle B([d,\psi']'), C(\bar{\Theta})\rangle\right]\right\}$$

$$\equiv \text{Arg}\min_{\bar{\Theta}\in\Theta^*} \mathcal{H}_\infty\left({}^{\lfloor o}f||\bar{\Theta}\right). \quad (6.72)$$

Here, $\bar{\Theta}$ is a nonrandom argument of the posterior pdf $f(\Theta = \bar{\Theta}|d(t))$ and ${}^{\lfloor o}\mathcal{E}[\cdot]$ is the expectation assigned to the objective pdf ${}^{\lfloor o}f(d,\psi) \equiv {}^{\lfloor o}f(\Psi)$.

The weighting used in connection with the approximate estimation, Section 6.5, guarantees that we deal with data vectors $\Psi = [d', \psi']'$ in a relatively narrow set. Thus, we can assume that the mixture character of the objective pdf is reflected in the conditional pdf (6.63) only. It allows us to assume

$$ {}^{\lfloor o}f(d,\psi) = \left[\beta\, {}^{\lfloor o}f(d|\psi,\Theta_1) + (1-\beta)\, {}^{\lfloor o}f(d|\psi,\Theta_2)\right] {}^{\lfloor o}f(\psi), \tag{6.73} $$

where the pdf ${}^{\lfloor o}f(\psi)$ brings no information on $\beta, \Theta_1, \Theta_2$. In other words, meeting of a sort of natural conditions of decision making (2.36) is expected. Then, the identity (6.72) relates points in $\mathrm{supp}\left[\, f(\Theta|d(\mathring{t}))\right]$, to the parameters β, Θ_1, Θ_2 and specifies a constraint on the split factors. This observation is elaborated in Chapter 8. Due to it, the split into a pair of pdfs $f(\Theta_1|d(\mathring{t}))$, $f(\Theta_2|d(\mathring{t}))$ is made without a full estimation of the mixture (6.63). It decreases significantly the computational burden connected with the hierarchical splitting.

Construction of a new mixture: Step 5 in Algorithm 6.8

Splitting of factors provides a potentially huge number of components made of them. A prospective cth component is formed from new factors ic with parameters described by pdfs $\hat{f}_{j_i}(\Theta_{ic}|d(\mathring{t}))$, $j_i \in \{1,2\}$. A deductive selection of a reasonably small subset of such components seems to be impossible. For this reason, a nondeductive rule has to be chosen. At present, we let each component split at most into a pair of components by selecting respective factors in a predefined way. Random choice seems to be a better alternative and tests confirm it.

Remark(s) 6.16
The discussed construction of a new mixture belongs to a wide class of problems in which simple elements can be combined in a more complex object in many ways. A systematic solution for a class of such integer-programming tasks is proposed in [151]. This methodology was proposed in connection with the so-called shadow cancelling problem; see Section 12.3.2. It is worth elaborating for the problem of this paragraph, too.

6.4.9 Techniques applicable to static mixtures

The majority of available techniques are applicable to static mixtures that have constant regression vectors. The following section demonstrates that these techniques are approximately applicable to dynamic mixtures. Then, promising techniques suitable for static mixtures are outlined.

Static clustering of dynamic systems

Mixture estimation constructs models with multiple modes. It is perceived as a branch of *cluster analysis* [149]. Cluster algorithms are often used for an approximate dynamic clustering, e.g., [54, 152, 153, 154]. They simply neglect mutual dependence of data vectors $\Psi(\mathring{t})$ and treat them as independent records. Thus, it is reasonable to analyze properties of this approximation. We rely on asymptotic analysis of the type presented in [82] and in Proposition 2.15. For presentation simplicity, we assume a restricted but important special case of discrete-valued data and the state in the phase form, with an over-estimated order ∂. Thus, our analysis is made under the following restrictions.

Requirement 6.3 (On static treatment of dynamic mixtures) *We assume that*

- *data d_t have discrete values,*
- *data vectors Ψ_t are modelled by a static mixture*

$$f(\Psi_t|\Psi(t-1),\Theta) \equiv f(\Psi_t|\Theta); \tag{6.74}$$

- *the chosen prior pdf $f(\Theta)$ is positive on Θ^*,*
- *data vectors are in the phase form $\Psi_t = [d'_t,\ldots,d'_{t-\partial},1]' \equiv [d'_t,\phi'_{t-1}]'$ with a finite order ∂,*
- *the state of the objective pdf of data ${}^{\llcorner o}f(d(\mathring{t}))$, cf. Chapter 2, is uniquely determined by the model state ϕ_{t-1}, $t \in t^*$,*

$${}^{\llcorner o}f(d_t|d(t-1)) = {}^{\llcorner o}f(d_t|\phi_{t-1}) \Rightarrow {}^{\llcorner o}f(d(\mathring{t})) = \prod_{t\in t^*} {}^{\llcorner o}f(d_t|\phi_{t-1}). \tag{6.75}$$

Proposition 6.17 (On independently clustered data vectors) *Let Requirement 6.3 be met. If a unique ${}^{\llcorner o}\Theta \in \Theta^* \cap \mathrm{supp}\,[\,f(\Theta)]$ exists such that*

$$f(\Psi_t|\,{}^{\llcorner o}\Theta) = {}^{\llcorner o}f(\Psi_t),\ \forall \Psi_t \in \Psi^*_t, \tag{6.76}$$

then the formal posterior pdf $\prod_{t\in t^*} f(\Psi_t|\Theta)f(\Theta)$ concentrates on ${}^{\llcorner o}\Theta$. Otherwise, the posterior pdf $\prod_{t\in t^*} f(\Psi_t|\Theta)f(\Theta)$ concentrates on minimizing arguments of the (relative) entropy rate

$$\mathrm{Arg}\min_{\Theta\in\Theta^*} \overline{\lim}_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t}\ln\left(\frac{{}^{\llcorner o}f(\Psi_\tau)}{f(\Psi_\tau|\Theta)}\right).$$

The statement holds in spite of the neglected dependence of data vectors $\Psi(\mathring{t})$.

Proof. The <u>formal</u> approximate posterior pdf can be expressed as follows.

$$f(\Theta|\Psi(t)) \propto f(\Theta)\exp\left\{-t\frac{1}{t}\sum_{\tau=1}^{t}\ln\left(\frac{{}^{\llcorner o}f(d_\tau|\phi_{\tau-1})}{f(\Psi_\tau|\Theta)}\right)\right\}. \tag{6.77}$$

Let ${}^{\lfloor o}\mathcal{E}$ denote the objective expectation corresponding to the objective pdf ${}^{\lfloor o}f(d(\mathring{t})$. Then, for any fixed $\Theta \in \Theta^*$ and it holds

$$
\begin{aligned}
&{}^{\lfloor o}\mathcal{E}\left[\prod_{\tau=1}^{t} \frac{f(\Psi_\tau|\Theta)}{{}^{\lfloor o}f(d_\tau|\phi_{\tau-1})}\right] = \int \left[\prod_{\tau=1}^{t} \frac{f(\Psi_\tau|\Theta)}{{}^{\lfloor o}f(d_\tau|\phi_{\tau-1})} \, {}^{\lfloor o}f(d_\tau|d(\tau-1))\right] dd(t)\\
&\underbrace{=}_{(6.75)} \int \left[\prod_{\tau=1}^{t} \frac{f(\Psi_\tau|\Theta)}{{}^{\lfloor o}f(d_\tau|\phi_{\tau-1})} \, {}^{\lfloor o}f(d_\tau|\phi_{\tau-1})\right] dd(t)\\
&\underbrace{=}_{\text{canceling}} \int \prod_{\tau=1}^{t} f(\Psi_\tau|\Theta)\, dd(t) \underbrace{=}_{\text{chain rule}} \int \prod_{\tau=1}^{t} f(d_\tau|\phi_{\tau-1},\Theta) f(\phi_{\tau-1}|\Theta)\, dd(t)\\
&\underbrace{\leq}_{\text{discrete } \phi_{t-1}} \int \prod_{\tau=1}^{t} f(d_\tau|\phi_{\tau-1},\Theta)\, dd(t) \underbrace{=}_{\text{normalisation}} 1. \qquad (6.78)
\end{aligned}
$$

In summary, for the discrete-valued ϕ_{t-1}, it holds that

$$
0 < {}^{\lfloor o}\mathcal{E}\left[\prod_{\tau}^{t} \frac{f(\Psi_\tau|\Theta)}{{}^{\lfloor o}f(d_\tau|\phi_{\tau-1})}\right] \leq 1. \qquad (6.79)
$$

Similar manipulations with the conditional version of ${}^{\lfloor o}\mathcal{E}$ pdf reveals that

$$
\zeta_t \equiv \prod_{\tau}^{t} \frac{f(\Psi_\tau|\Theta)}{{}^{\lfloor o}f(d_\tau|\phi_{\tau-1})}
$$

is nonnegative martingale with respect to $d(t)$. Thus, [81], it converges almost surely to a nonnegative quantity smaller than one. Consequently, the inversion ζ_t^{-1} converges almost surely either to infinity or to a finite value greater than one. Consequently,

$$
\overline{\lim}_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \ln\left(\frac{{}^{\lfloor o}f(d_\tau|\phi_{t-1})}{f(\Psi_\tau|\Theta)}\right) \geq 0
$$

almost surely for $t \to \infty$. The form of (6.77), notice the factor $-t$, implies that the formal posterior pdf concentrates on

$$
\begin{aligned}
&\text{Arg} \min_{\Theta\in\Theta^*} \overline{\lim}_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \ln\left(\frac{{}^{\lfloor o}f(d_\tau|\phi_{\tau-1})}{f(\Psi_\tau|\Theta)}\right)\\
&= \text{Arg} \min_{\Theta\in\Theta^*} \overline{\lim}_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \ln\left(\frac{{}^{\lfloor o}f(\Psi_\tau)}{f(\Psi_\tau|\Theta)}\right) = \text{Arg} \min_{\Theta\in\Theta^*} \overline{\lim}_{t\to\infty} \mathcal{H}_t({}^{\lfloor o}f(\cdot)||\Theta).
\end{aligned}
$$

The first equality is implied by the chain rule ${}^{\lfloor o}f(\Psi_\tau) = {}^{\lfloor o}f(d_\tau|\phi_{\tau-1})\, {}^{\lfloor o}f(\phi_{\tau-1})$ and by the fact that the added term $\frac{1}{t}\sum_{\tau=1}^{t} \ln({}^{\lfloor o}f(\phi_{\tau-1}))$ does not influence minimization. The last equality is just the definition of the entropy rate.

If the unique ${}^{\llcorner o}\Theta \in \Theta^*$ exists such that ${}^{\llcorner o}f(\Psi_\tau) = f\left(\Psi_\tau | {}^{\llcorner o}\Theta\right)$, then it is the unique minimizing argument. □

Remark(s) 6.17

1. *The result is not surprising: a dynamic model introduces a specific dependence structure among data records $\{d_t\}$. The use of data vectors Ψ_t instead of d_t offers a similar but less structured freedom. Hence, the asymptotic coincidence is possible whenever $\Psi_t,\ t \in t^*$, "overlap" the dependence structure of the correct dynamic model.*
2. *The previous statement can be illustrated nicely on the special case of normal components.*
3. *The performed analysis indicates that the clustering and estimation algorithms designed for static mixtures are applicable in the dynamic case, at least asymptotically.*
4. *The parameterized mixture can be interpreted as a superposition of two random generators. First, a pointer $c_t \in c^*$ is selected with probability α_c. Then, data are randomly generated from this component. In the considered mixtures, the generated pointers $\{c_t\}_{t\in t^*}$ are mutually independent. This independence is illogical in the dynamic context. modelling of pointers by Markov chains would be more adequate. With such an option, the estimation becomes much harder. The pointers, however, can be approximately treated as pairs c_t, c_{t-1} with dependence on c_{t-1}, c_{t-2} neglected. The discussed result makes this extension of the mixture model worth considering.*

Problem 6.10 (Weakening of the conditions in Proposition 6.17) *The conditions of Proposition 6.17 can be almost surely made weaker. It would be worthwhile to make inspection in this respect as there is a chance for a better understanding of the result. Also, estimates of the convergence rate could be and should be searched for.*

Promising initializations of static mixtures

Here, potential directions in solution of the initialization problem restricted to static mixtures are outlined without a detailed elaboration.

Splitting of data space

The idea of this technique is similar to that of branching by factor splitting. All operations are, however, made in the data space. The technique is applicable to static mixtures and can be extended to dynamic mixtures when data vectors Ψ are treated as independent data records.

Gradual extension of low-dimensional mixtures

The approximate Bayesian mixture estimation for low-dimensional data records is a computationally feasible task. Even the initial locations and widths can be spaced equally within the region, where data and thus their static mean values occur. Then, the approximate Bayesian estimation is expected to provide good results within a reasonable computational time. This speculation has motivated the first solution we have ever used. The solution is described in terms of data records $\{d_t\}$, but it is directly applicable to the data vectors $\{\Psi_t\}$.

Algorithm 6.11 (Extension of low-variate static mixtures)

Initial mode

- *Select the upper bound $\mathring{c}_i$ on the number c_i of components on each scalar axis $i \in \{1, \ldots, \mathring{d}\}$.*
- *Specify prior pdfs $f(\Theta_i)$ related to single-dimensional mixtures assigned to each entry of d_i.*
- *Estimate the individual mixtures $f(\Theta_i | d_i(\mathring{t}))$, $i = 1, \ldots, \mathring{d}$, by some approximate technique; see Section 6.5.*
- *Select the most significant components (coinciding with factors); see Section 6.6.*
- *Apply branching versions of the flattening operation; Propositions 6.7, 6.8. This gives good single-variate prior pdfs $f(\Theta_i)$, $i = 1, \ldots, \mathring{d}$.*
- *Set the dimension of the data space to be extended $\mathring{\underline{d}} = 1$.*
- *Denote $f(\Theta_{1\cdots\mathring{\underline{d}}}) = f(\Theta_1)$.*

Extension mode

1. *Set $\mathring{\underline{d}} = \mathring{\underline{d}} + 1$.*
2. *Assume the mixture defined on data $d_{1\cdots\mathring{\underline{d}}}$ in the product form*

$$f\left(d_{1\cdots\mathring{\underline{d}}}\middle|\Theta_{1\cdots\mathring{\underline{d}}}\right) \equiv f\left(d_{1\cdots(\mathring{\underline{d}}-1)}\middle|\Theta_{1\cdots(\mathring{\underline{d}}-1)}\right) f\left(d_{\mathring{\underline{d}}}\middle|\Theta_{\mathring{\underline{d}}}\right).$$

3. *Define the prior pdf on its parameters as $f\left(\Theta_{1\cdots\mathring{\underline{d}}}\right) = f\left(\Theta_{1\cdots(\mathring{\underline{d}}-1)})f(\Theta_{\mathring{\underline{d}}}\right)$.*
4. *Perform an approximate parameter estimation using data $d_{1\cdots\mathring{\underline{d}}}(\mathring{t})$ to get $f\left(\Theta_{1\cdots\mathring{\underline{d}}}\middle| d_{1\cdots\mathring{\underline{d}}}(\mathring{t})\right)$.*
5. *Select the most significant components; see Section 6.6.*
6. *Apply branching versions of the flattening operation, Propositions 6.7, 6.8. This gives a good higher-variate prior pdf $f\left(\Theta_{1\cdots\mathring{\underline{d}}}\right)$.*
7. *Go to the beginning of* Extension mode *if $\mathring{\underline{d}} < \mathring{d}$. Otherwise stop and take the resulting prior pdf as the prior on the whole parameter space.*

Remark(s) 6.18
The experimental results obtained are reasonable. They exhibit a dependence on the order in which the extension is performed. This may be compensated by starting the procedure from the technologically most important quantities or by a random permutation of the ordering used.

A direct extension of single variate components

The following variant of the extension provides another possibility to solve the initialization problem.

Again, we perform estimation of single variate mixtures for all entries d_i (or Ψ_i). Each entry d_i (Ψ_i) is individually described by a mixture

$$f(d_i|\Theta_i) = \sum_{c_i \in c_i^*} \alpha_{c_i} f(d_i|\Theta_{ic_i}, c_i), \; i = 1, \ldots, \mathring{d}. \tag{6.80}$$

Its components are identical with factors. This provides a set of factors for building multivariate components. We take the single-variate components as independent factors forming multivariate components. We have insufficient reasons for other choice at this stage of the construction; see also Proposition 12.4. The potential multivariate components have structures determined by the one-to-one composition rule π

$$\begin{aligned} \pi :\; & c^* \to [c_1^*, \ldots, c_{\mathring{d}}^*] \Leftrightarrow & (6.81)\\ \pi :\; & \text{the multivariate component } c \text{ is the product of factors that coincide} \\ & \text{with components } \pi(c)_1 \in c_1^*, \ldots, \pi(c)_{\mathring{d}} \in c_{\mathring{d}}^* \text{ in (6.80).} \end{aligned}$$

The number of such mappings π is very large. Thus, we cannot inspect them fully and suitable candidates have to be guessed from the available single-variate results. The selection of promising composition rules is called the *shadow cancelling problem.* Essentially, we want to avoid the combinations that contain little data and appear due to the nonuniqueness of reconstructing multivariate objects from their low-dimensional "shadows". A solution is proposed in Chapter 12.

Clustering of pointers

The troubles with the discussed shadow cancelling problem led to the idea of recording weights of individual single-variate components generated in an approximate estimation, Section 6.5, and clustering these pointers. This clustering is of a similar complexity as the original one. We expect, however, a better separation of the "weight" clusters. Moreover, a coupling with MT algorithm [67], Chapter 12, suits it ideally as the static clustering is adequate and the box width determining it can simply be chosen: the probabilistic weights are known to be in the interval $[0, 1]$.

Combination of solutions

Feasible solutions are mostly nondeductive and thus susceptible to failures or to stopping at local optimum. Thus, a combination of techniques is the must. Generally,

- a random element in the adopted search increases the chance of reaching the global optimum,
- the quality of the results can hardly be a monotonous function of the problem parameters; thus, both extensions of low-dimensional solutions as well as reductions of high-dimensional ones have to be used,
- the v-likelihood, Section 6.1.2, seems to be the only viable universal tool for comparing quality of the alternative results.

Problem 6.11 (Extension of the model set) *The exponential family contains a wider set of pdfs than the* <u>*dynamic*</u> *exponential family. Thus, the static treatment of dynamic components opens the possibility to extend the set of feasible models. This possibility should be widely exploited.*

Problem 6.12 (Elaborate the above ideas and bring in new ones) *This section outlines an incomplete but wide range of possible directions to be followed. They should be complemented and elaborated. Even conclusions that some of the sketched ideas are crazy would be invaluable.*

6.5 Approximate parameter estimation

The estimation described by Proposition 2.14 is quite formal for the mixture model (6.1) since the resulting posterior pdf consists of a product of sums of functions of unknown parameters. The number of terms that should be stored and handled blows up exponentially. An approximation is therefore required.

The existing algorithms do not support estimation of dynamic mixtures. This has motivated our search for an adequate algorithm. A slight extension of a known algorithm [49], labelled the *quasi-Bayes estimation* (*QB*), was proposed in [155] and its properties illustrated on simulation examples. It is discussed in Section 6.5.1.

The insight gained into the approximate learning has allowed us to extend the classical *expectation-maximization algorithm (EM)* to the dynamic mixtures; see Section 6.5.2. It serves us as an auxiliary initialization tool and as a golden standard suitable for comparisons. Note that Markov Chain Monte Carlo (MCMC) techniques are also worth inspecting but the known applications [50] still deal with much lower dimensions than needed for advising.

The quasi-Bayes estimation has good properties, Bayesian motivation and interpretation as well as predictable and tractable computational complexity. It is of a recursive nature so that it may serve in the adaptive advisory system. These reasons justify our preference for it. At the same time, its construction implies that its results depend on the order of the data processing. This drawback was found significant in some applications dealing with static mixtures. This made us design a hybrid of the quasi-Bayes and EM algorithms. This algorithm, called *batch quasi-Bayes estimation* (*BQB*), is described in Section 6.5.3. It preserves the essence of the Bayesian paradigm needed in other tasks such as in structure estimation; see Section 6.6. At the same time, its results are independent of the processing order.

The *quasi-EM algorithm* is another hybrid proposed. It can be viewed as a rather simple quasi-Bayes $\underline{\text{point}}$ estimation.

6.5.1 Quasi-Bayes estimation

For the design and description of the quasi-Bayes estimation, we introduce discrete random pointers $c(\mathring{t}) = (c_1, \ldots, c_{\mathring{t}})$, $c_t \in c^*$, to particular components. They are assumed to be mutually independent with time-invariant probabilities $f(c_t = c|d(t-1), \Theta) = \alpha_c$. With these pointers, the parameterized model (6.1) can be interpreted as the marginal pdf $f(d_t|d(t-1), \Theta)$ of the joint pdf (cf. Agreement 5.1)

$$f(d_t, c_t|d(t-1), \Theta) = \prod_{c \in c^*} \left[\alpha_c f(d_t|\phi_{c;t-1}, \Theta_c, c)\right]^{\delta_{c,c_t}} \tag{6.82}$$

$$\underbrace{=}_{(6.2)} \prod_{c \in c^*} \left[\alpha_c \prod_{i \in i^*} f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)\right]^{\delta_{c,c_t}},$$

where δ is the *Kronecker symbol* defined by the formula

$$\delta_{c,\tilde{c}} \equiv \begin{cases} 1 \text{ if } & c = \tilde{c} \\ 0 \text{ otherwise} \end{cases}.$$

Indeed, the marginal pdf $f(d_t|d(t-1), \Theta)$ of the pdf $f(d_t, c_t|d(t-1), \Theta)$, which is found by summing (6.82) over $c_t \in c^*$, has the form (6.1), (6.2).

Let the approximate pdf $f(\Theta|d(t-1))$ be of the product form, cf. Agreement 6.1,

$$f(\Theta|d(t-1)) = Di_\alpha(\kappa_{t-1}) \prod_{c \in c^*} \prod_{i \in i^*} f(\Theta_{ic}|d(t-1)) \propto$$

$$\propto \prod_{c \in c^*} \alpha_c^{\kappa_{c;t-1}-1} \prod_{i \in i^*} f(\Theta_{ic}|d(t-1)). \tag{6.83}$$

The pdfs $f(\Theta_{ic}|d(t-1))$ describe parameters of factors and $Di_\alpha(\kappa_{t-1})$ is the pdf describing component weights. It is determined by the vector κ_{t-1} with positive entries $\kappa_{c;t-1} > 0$. They are known values of functions of $d(t-1)$. The proportionality in (6.83) is implied by the definition of the *Dirichlet pdf*

$$Di_\alpha(\kappa) \equiv \frac{\prod_{c \in c^*} \alpha_c^{\kappa_c - 1}}{\mathcal{B}(\kappa)}, \ \mathcal{B}(\kappa) \equiv \frac{\prod_{c \in c^*} \Gamma(\kappa_c)}{\Gamma\left(\sum_{c \in c^*} \kappa_c\right)}.$$

Its properties are discussed in detail in Chapter 10.

The assumption (6.83), the formula (6.82) and the Bayes rule (2.8) give

$$f(\Theta, c_t|d(t)) \propto \prod_{c\in c^*} \alpha_c^{\kappa_{c;t-1}+\delta_{c,c_t}-1} \prod_{i\in i^*} [f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)]^{\delta_{c,c_t}} f(\Theta_{ic}|d(t-1)). \tag{6.84}$$

In order to obtain an approximation of the desired pdf $f(\Theta|d(t))$, we have to eliminate c_t from (6.84). The correct marginalization with respect to c_t destroys the feasible product form (6.83). It is preserved if δ_{c,c_t} is replaced by its point estimate. We approximate δ_{c,c_t} by its conditional expectation

$$\delta_{c,c_t} \approx \mathcal{E}[\delta_{c,c_t}|d(t)] = f(c_t = c|d(t)). \tag{6.85}$$

By integrating (6.84) over the parameters Θ, cf. Proposition 10.1, we get this probability in the form

$$\begin{aligned} w_{c;t} \equiv f(c_t = c|d(t)) &= \frac{\hat{\alpha}_{c;t-1} \prod_{i\in i^*} f(d_{ic;t}|\psi_{ic;t}, d(t-1), c)}{\sum_{\tilde{c}\in c^*} \hat{\alpha}_{\tilde{c};t-1} \prod_{i\in i^*} f(d_{i\tilde{c};t}|\psi_{i\tilde{c};t}, d(t-1), \tilde{c})} \\ &= \frac{\kappa_{c;t-1} \prod_{i\in i^*} f(d_{ic;t}|\psi_{ic;t}, d(t-1), c)}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c};t-1} \prod_{i\in i^*} f(d_{i\tilde{c};t}|\psi_{i\tilde{c};t}, d(t-1), \tilde{c})}. \end{aligned} \tag{6.86}$$

$$\text{Here,} \quad f(d_{ic;t}|\psi_{ic;t}, d(t-1), c) \equiv \int f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) f(\Theta_{ic}|d(t-1))\, d\Theta_{ic}$$

$$\hat{\alpha}_{c;t-1} = \frac{\kappa_{c;t-1}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c};t-1}} \tag{6.87}$$

are the Bayesian prediction (2.38) for a single factor identified by indexes ic. The values $\hat{\alpha}_{c;t-1}$ are the conditional expectations of α_c; cf. (6.5) and Chapter 10. The formula (6.86) can be interpreted as the Bayes rule applied to the discrete unknown random variable $c_t \in c^*$ with the prior probability $\hat{\alpha}_{c;t-1} \propto \kappa_{c;t-1}$.

By inserting the approximation (6.85), (6.86) into (6.84), the approximately updated posterior pdf has the same product form as in (6.83) with

$$\kappa_{c;t} = \kappa_{c;t-1} + w_{c;t} \tag{6.88}$$

$$f(\Theta_{ic}|d(t)) \propto [f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)]^{w_{c;t}} f(\Theta_{ic}|d(t-1)). \tag{6.89}$$

This step completes the design of the estimation algorithm.

Algorithm 6.12 (Quasi-Bayes estimation without common factors)

Initial (offline) mode

- *Select the complete structure of the mixture.*
- *Select the prior pdfs $f(\Theta_{ic})$ of the individual factors, ideally, in the conjugate form with respect to the parameterized factors $f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)$.*
- *Select initial values $\kappa_{c;0} > 0$ of statistics describing prior pdf α. Intuitively motivated values about 0.1$\mathring{t}/\mathring{c}$ were found reasonable.*

- *Compute the initial estimates of the component weights* $\hat{\alpha}_{c;0} = \frac{\kappa_{c;0}}{\sum_{\tilde{c} \in c^*} \kappa_{\tilde{c};0}}$.

Sequential (online) mode, *running for* $t = 1, 2, \ldots,$

1. *Acquire the data record* d_t.
2. *Compute the values of the predictive pdfs*

$$f(d_{ic;t}|\psi_{ic;t}, d(t-1), c) = \int f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) f(\Theta_{ic}|d(t-1))\, d\Theta_{ic}$$
$$= \frac{\mathcal{I}(d(t)|ic)}{\mathcal{I}(d(t-1)|ic)},$$

see (2.47), for each individual factor, $i \in i^* = \{1, \ldots, \mathring{d}\}$, *in all components,* $c \in c^*$, *and the measured data record* d_t. *The adaptive, one-stage-ahead predictor (see Section 6.1.2) is thus used.*

3. *Compute values of the predictive pdfs*

$$f(d_t|d(t-1), c) = \prod_{i \in i^*} f(d_{ic;t}|\psi_{ic;t}, d(t-1), c)$$

for the measured data d_t *and each individual component* $c \in c^*$.

4. *Compute the probabilistic weights* $w_{c;t}$ *using the formula (6.86).*
5. *Update the scalars* $\kappa_{c;t} = \kappa_{c;t-1} + w_{c;t}$ *according to (6.88).*
6. *Update the Bayesian parameter estimates* $f(\Theta_{ic}|d(t-1))$ *of individual factors according to the weighted Bayes rule(6.89)*

$$f(\Theta_{ic}|d(t)) \propto [f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)]^{w_{c;t}} f(\Theta_{ic}|d(t-1)).$$

7. *Evaluate the point estimates of the mixing weights*

$$\mathcal{E}[\alpha_c|d(t)] = \frac{\kappa_{c;t}}{\sum_{\tilde{c} \in c^*} \kappa_{\tilde{c};t}} \equiv \hat{\alpha}_{c;t}$$

and, if need be, characteristics of $f(\Theta_{ic}|d(t))$ *describing parameters* Θ_{ic}.

8. *Go to the beginning of* Sequential mode *while data are available.*

Remark(s) 6.19

1. *The algorithm is applicable whenever the Bayesian estimation and prediction for each factor can be simply calculated. This singles out parameterized models admitting finite-dimensional statistics as the candidates among which factors should be chosen. This class consists essentially of the* exponential family *augmented by the uniform distribution with unknown boundaries, Section 3.2. They have conjugate (self-reproducing) prior pdfs for which only finite-dimensional statistics need to be stored and updated. Normal regression models, Chapter 8, or Markov chains, Chapter 10, are prominent examples of such dynamic factors. Other cases are more or less restricted to static factors.*

2. *It is worth noticing that the predictions of individual factors without data weighting is performed for evaluation of the weights $w_{c;t+1}$. The parameter estimation algorithm performs almost the same algebraic operations but with weighted data. This simple observation is exploited in implementation of the algorithm. Moreover, it is widely used in judging of estimation quality; see Section 6.1.2.*

Problem 6.13 (Alternative approximation of unknown δ_{c,c_t})
The proposed solution can be improved by searching for a specific product-form-preserving approximation $\hat{f}(\Theta|d(t+1))$ (containing no c_{t+1}), which is the nearest one to (6.84) in terms of the KL divergence [37]. This task is feasible. This type of solution is elaborated in [131] and indeed provides a better solution. The quasi-Bayes estimation is there also shown to be reasonable approximation of this solution. The problem of accumulating of approximation errors, cf. Section 3.4, remains to be open.

The described quasi-Bayes algorithm 6.12 does not take into account the possibility that some factors are common to various components. The extension to this useful case is, however, straightforward.

Algorithm 6.13 (Quasi-Bayes estimation with common factors)
Initial (offline) mode

- *Select the complete structure of the mixture.*
- *Select prior pdfs $f(\Theta_{ic})$ of individual factors, ideally, in the conjugate form (3.13) with respect to the parameterized factors $f(d_{ic;t+1}|\psi_{ic;t+1},\Theta_{ic},c)$.*
- *Select initial values $\kappa_{c;0} > 0$, say, about $0.1\mathring{t}/\mathring{c}$, describing prior pdf of the component weights α.*
- *Compute the initial point estimates of the component weights $\hat{\alpha}_{c;0} = \frac{\kappa_{c;0}}{\sum_{\tilde{c}c^*}\kappa_{\tilde{c};0}}$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots$,*

1. *Acquire the data record d_t.*
2. *Compute values of the predictive pdfs*
$$f(d_{ic;t}|\psi_{ic;t},d(t-1),c) = \int f(d_{ic;t}|\psi_{ic;t},\Theta_{ic})f(\Theta_{ic}|d(t-1))\,d\Theta_{ic}$$
$$= \frac{\mathcal{I}(d(t)|ic)}{\mathcal{I}(d(t-1)|ic)},$$
see (2.47), for $\underline{\text{different}}$ factors, $i \in i^* = \{1,\ldots,\mathring{d}\}$, in all components, $c \in c^*$, and the measured data record d_t.
The adaptive, one-stage-ahead predictor (see Section 6.1.2) is used.
3. *Compute the values of predictive pdfs*
$$f(d_t|d(t-1),c) = \prod_{i\in i^*} f(d_{ic;t}|\psi_{ic;t},d(t-1),c)$$
for the measured data d_t and each individual component $c \in c^*$.

4. *Compute the probabilistic weights* $w_{c;t}$ *using the formula (6.86).*
5. *Update the scalars* $\kappa_{c;t} = \kappa_{c;t-1} + w_{c;t}$ *according to (6.88).*
6. *Update the Bayesian parameter estimates of* *different* *factors*

$$f(\Theta_{ic}|d(t)) \propto [f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)]^{w_{ic;t}} f(\Theta_{ic}|d(t-1)), \quad \textit{where} \qquad (6.90)$$

$$w_{ic;t} = \sum_{\tilde{c} \in c_i^*} w_{\tilde{c};t}.$$

The set c_i^* *includes pointers* c *to components that contain* i*th factor. Handling of different factors only and this step distinguish the presented algorithm from Algorithm 6.12.*

7. *Evaluate, the point estimate of the mixing weights*

$$\mathcal{E}[\alpha_c|d(t)] = \frac{\kappa_{c;t}}{\sum_{\tilde{c} \in c^*} \kappa_{\tilde{c};t}} \equiv \hat{\alpha}_{c;t}$$

and, if need be, characteristics of $f(\Theta_{ic}|d(t))$ *describing parameters* Θ_{ic}.

8. *Go to the beginning of* Sequential mode *while data are available.*

6.5.2 EM estimation

The expectation-maximization (EM) algorithm is a classical and successful way of the mixture estimation [49]. The EM algorithm provides a local maximum of the posterior pdf (2.44) over Θ^*.

Its convergence is known to be monotonic but slow. The generic multimodality of the posterior pdf calls for a repetitive search using various initial conditions. AutoClass [42] is a typical and successful implementation of the EM algorithm. It combats the multimodality by using random initial conditions. It offers a high chance of finding the global maximum but it is computationally expensive. Essence of the EM algorithm is described here.

We are given the joint pdf of the observed data $d(\mathring{t})$ and nonmeasured pointers $c(\mathring{t})$ to individual components conditioned on unknown parameters $\Theta \in \Theta^*$. For fixed measured data $d(\mathring{t})$ and chosen structure of parameterized models, they specify the logarithmic likelihood function

$$\mathcal{L}(\Theta, d(\mathring{t}), c(\mathring{t})) \equiv \ln(f(d(\mathring{t}), c(\mathring{t})|\Theta)) = \sum_{t \in t^*} \sum_{c \in c^*} \delta_{c,c_t} \ln\left(\alpha_c f(d_t|\phi_{t-1}, c, \Theta)\right). \qquad (6.91)$$

Similarly as in the quasi-Bayes estimation, a reasonable approximation of δ_{c,c_t} is searched for. Unlike for the quasi-Bayes estimation, a point estimate $\hat{\Theta}$ of Θ is searched for and batch processing is supported.

Algorithm 6.14 (EM algorithm for mixture estimation)

Initial mode

- *Select the complete structure of the mixture.*
- *Select the upper bound* $\mathring{n}$ *on the number* n *of iterations and set* $n = 0$.
- *Select the initial point estimate* $\hat{\Theta}_n$ *of* Θ.

Iterative mode

1. *Approximate in batch all δ_{c,c_t}, $t \in t^*$. Use all measured data $d(\mathring{t})$ when taking*

$$\delta_{c,c_t} \approx \mathcal{E}\left[\delta_{c,c_t}|d(\mathring{t}), \hat{\Theta}_n\right] \equiv f\left(c_t = c|d(\mathring{t}), \hat{\Theta}_n\right), \ t \in t^*. \tag{6.92}$$

   This is called the E(xpectation) step.
2. *Insert the approximations (6.92) into the log-likelihood function (6.91).*
3. *Find the new point estimate $\hat{\Theta}_{n+1}$ of Θ as the maximizing argument of this approximate likelihood function. This is called the M(aximization) step.*
4. *Evaluate the value of the approximate log-likelihood attained by the maximizing argument.*
5. *Stop if the value of the approximate log-likelihood has not increased or $n \geq \mathring{n}$. Otherwise, set $n = n + 1$ and go to the beginning of* Iterative mode.

To make the EM algorithm applicable we need to evaluate the estimate of the Kronecker symbol.

Proposition 6.18 (Batch estimate of pointers to components) *Let*

$$f(d_t, c_t|d(t-1), c(t-1), \Theta) \equiv f(d_t|\phi_{c_t;t-1}, c_t, \Theta_{c_t})\alpha_{c_t}, \ t \in t^*,$$

and $\hat{\Theta} \equiv (\hat{\alpha}_1, \ldots, \hat{\alpha}_{\mathring{c}}, \hat{\Theta}_1, \ldots, \hat{\Theta}_{\mathring{c}}) \in \Theta^*$ be a fixed instance of

$$\Theta \equiv (\alpha_1, \ldots, \alpha_{\mathring{c}}, \Theta_1, \ldots, \Theta_{\mathring{c}}).$$

Then, the probability of the random pointer $c_t \in c^*$, conditioned on all data $d(\mathring{t})$ and $\hat{\Theta}$, is given by the formula

$$w_{c_t;t} \equiv f(c_t|d(\mathring{t}), \hat{\Theta}) = \frac{\hat{\alpha}_{c_t} f(d_t|\phi_{c_t;t-1}, c_t, \hat{\Theta})}{\sum_{c \in c^*} \hat{\alpha}_c f(d_t|\phi_{c;t-1}, c, \hat{\Theta})} \propto \hat{\alpha}_{c_t} f(d_t|\phi_{c_t;t-1}, c_t, \hat{\Theta}). \tag{6.93}$$

Proof. The the chain rule and assumed dependencies imply

$$f(d(\mathring{t}), c(\mathring{t})|\hat{\Theta}) = \prod_{t \in t^*} f(d_t|\phi_{c_t;t-1}, c_t, \hat{\Theta}_c)\hat{\alpha}_{c_t}.$$

The pdf $f(d(\mathring{t}), c_t|\hat{\Theta})$ is obtained by marginalization

$$\begin{aligned} f(d(\mathring{t}), c_t|\hat{\Theta}) &= \sum_{\{c_\tau \in c^*\}_{\tau \in t^* \setminus t}} f(d(\mathring{t}), c(\mathring{t})|\hat{\Theta}) \\ &= f\left(d_t|\phi_{c_t;t-1}, c_t, \hat{\Theta}_{c_t}\right) \hat{\alpha}_{c_t} \times \text{term\_independent\_of\_}c_t. \end{aligned}$$

The desired conditional pdf is proportional to the above pdf with the proportion factor implied by normalization. In it, the complex term independent of c_t cancels and we get the claimed form. □

Remark(s) 6.20

1. *The guaranteed monotonic increase of the likelihood [49] to a local maximum is the key advantage of EM algorithm and justifies the stopping rule used.*
2. *As a rule, EM algorithm is described and used for static components. Here, the needed dynamic version is described.*
3. *The adopted interpretation of the EM algorithm differs from a standard interpretation. It is close to the description of the quasi-Bayes algorithm. It helped us to create a novel batch quasi-Bayes algorithm 6.16.*

EM algorithm 6.14 is described at the component level. Its applicability is extended when it is used at factor level, possibly with common factors.

Algorithm 6.15 (EM mixture estimation with common factors)
Initial mode

- *Select the complete structure of the mixture.*
- *Select the upper bound* $\mathring{n}$ *on the number* n *of iterations and set* $n = 0$.
- *Select initial point estimates* $\hat{\Theta}_{icn}$ *of parameters* Θ_{ic} *characterizing* i*th factor within* c*th component.*
- *Select point estimates* $\hat{\alpha}_{cn}$ *of the components weights* α_c*, typically, as the uniform pf.*

Iterative mode

1. *Set the* $\mathring{c}$*-vector of auxiliary statistics* $\kappa_{n;0} = 0$ *and use the current point estimates* $\hat{\Theta}_n$ *in the following evaluations.*
2. *Set the approximate log-likelihood functions* $\mathcal{L}_n(\Theta_{ic}, d(0))$ *to be accumulated to zero. They correspond to log-likelihoods of individual factors within individual components.*

   Sequential mode, *running for* $t = 1, 2, \ldots,$

   a) *Construct the data vectors* $\Psi_{ic;t}$, $i \in \{1, \ldots, \mathring{d}\}$, $c \in c^*$.
   b) *Compute the values of the predictive pdfs* $f(d_{ic;t}|\psi_{ic;t}, \hat{\Theta}_{icn}, c)$ *for different factors,* $i \in i^* = \{1, \ldots, \mathring{d}\}$, *in all components,* $c \in c^*$, *using the parameter estimates* $\hat{\Theta}_{icn}$ *that are constant during the time cycle within the* n*th iteration. Thus, the approximate, fixed one-stage-ahead predictor with certainty-equivalence approximation is used.*
   c) *Compute the values of predictive pdfs*

   $$f(d_t|\phi_{c;t-1}, \hat{\Theta}_{cn}, c) \equiv \prod_{i \in i^*} f(d_{ic;t}|\psi_{ic;t}, \hat{\Theta}_{icn}, c)$$

   for each individual component $c \in c^*$.
   d) *Compute the weights* $w_{cn;t}$ *approximating the Kronecker symbol* δ_{c,c_t}

   $$w_{cn;t} = \frac{\hat{\alpha}_{cn} f(d_t|\phi_{c;t-1}, \hat{\Theta}_{cn}, c)}{\sum_{\tilde{c} \in c^*} \hat{\alpha}_{\tilde{c}n} f(d_t|\phi_{\tilde{c};t-1}, \hat{\Theta}_{\tilde{c}n}, \tilde{c})}.$$

e) Update the statistics $\kappa_{cn;t} = \kappa_{cn;t-1} + w_{cn;t}$.

f) Update the log-likelihood functions describing different factors

$$\mathcal{L}_n(\Theta_{ic}, d(t)) = w_{icn;t} \ln \left[f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c) \right] + \mathcal{L}_n(\Theta_{ic}, d(t-1)), \quad \textit{with}$$
$$w_{icn;t} = \sum_{\tilde{c} \in c_i^*} w_{\tilde{c}n;t}. \qquad (6.94)$$

The set c_i^* *includes pointers to components that contain the ith factor.*

g) Go to the beginning of Sequential mode *if* $t \leq \mathring{t}$. *Otherwise continue.*

3. *Find the new point estimates* $\hat{\Theta}_{n+1}$ *of* Θ *as the maximizing arguments of the approximate log-likelihood*

$$\sum_{c \in c^*} \left[\kappa_{cn;\mathring{t}} \ln(\alpha_c) + \sum_{i \in i^*} \mathcal{L}_n(\Theta_{ic}, d(\mathring{t})) \right].$$

4. *Evaluate the value of the approximate log-likelihood attained at the maximizing argument.*
5. *Stop if the value of the approximate log-likelihood has not increased or* $n \geq \mathring{n}$. *Otherwise, set* $n = n + 1$ *and go to the beginning of* Iterative mode.

Remark(s) 6.21

1. *The algorithm can be used for searching maximum a posteriori probability estimate (MAP) by selecting nontrivial initial likelihood function* $\mathcal{L}(\Theta, d(0))$ *and positive* $\kappa_{n;0}$.
2. *Obviously, the EM algorithm is applicable iff the considered log-likelihood function can be represented on the computer. It essentially means that we have to deal with factors belonging to the exponential family; see Section 3.2.*
3. *The algorithm is intentionally written in the style of the quasi-Bayes algorithm; Algorithm 6.13. It indicates their similarity, which enriches both of them. For instance, the maximizing argument can be searched and immediately used within the time cycle. It leads to the adaptive* quasi-EM algorithm. *It can be used in online learning and its convergence can be faster. The loss of the monotonic convergence and of the processing-order independence is the price paid for it.*

6.5.3 Batch quasi-Bayes estimation

The construction of the quasi-Bayes estimation implies that its results depend on the order in which data are processed. Experiments have shown that this dependence may be significant. The EM algorithm avoids this drawback but its orientation to the point estimation prevents us to embed it into the advantageous Bayesian framework. For instance, the construction of the prior pdf by splitting (see Section 6.4.8) cannot be exploited. The presented interpretation of the EM algorithm allows us to create a novel batch quasi-Bayes estimation algorithm that is processing-order independent. Essentially, it estimates pointers to components within nth iteration by using approximation of pdfs of the mixture parameters. This pdf is fixed within the nth stage of the iterative search.

Algorithm 6.16 (BQB mixture estimation with common factors)

Initial mode

- *Select the complete structure of the mixture.*
- *Select the upper bound* $\mathring{n}$ *on the number* n *of iterations set* $n = 0$.
- *Set maximum of the v-log-likelihood* $\bar{l} = -\infty$.
- *Select the (flat) pre-prior pdfs* $\bar{f}(\Theta_{ic})$ *related to the individual factors*
$$f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c),$$
ideally, in the conjugate form.
- *Select the pre-prior values* $\bar{\kappa}_c > 0$ *used in flattening.*
- *Select the initial guess of the prior pdfs* $f_n(\Theta_{ic})$ *related to the individual factors* $f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c)$, *ideally, in the conjugate form.*
- *Select initial values* $\kappa_{cn} > 0$ *determining the prior Dirichlet pdf on the component weights.*

Iterative mode

1. *Make copies* $f(\Theta_{ic}|d(0)) = f_n(\Theta_{ic})$ *and* $\kappa_{c;0} = \kappa_{cn}$.
2. *Set the value of the v-log-likelihood* $l_{n;0} = 0$.
3. *Compute the point estimates of the component weights* $\hat{\alpha}_{cn} = \frac{\kappa_{cn}}{\sum_{\tilde{c} \in c^*} \kappa_{\tilde{c}n}}$.

   Sequential mode, *running for* $t = 1, 2, \ldots,$

   a) *Construct the data vectors* $\Psi_{ic;t}$, $i \in i^* \equiv \{1, \ldots, \mathring{d}\}$, $c \in c^*$.

   b) *Compute the values of the predictive pdfs*
   $$f_n(d_{ic;t}|\psi_{ic;t}, c) = \int f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c) f_n(\Theta_{ic}) \, d\Theta_{ic}$$
   for each individual factor $i \in i^*$ *in all components* $c \in c^*$. *The prior pdf* $f_n(\Theta)$ *is constant during the time cycle. The fixed one-step-ahead predictor (see Section 6.1.2) is thus used.*

   c) *Compute the values of the predictive pdfs*
   $$f_n(d_t|\psi_{c;t}, c) \equiv \prod_{i \in i^*} f_n(d_{ic;t}|\psi_{ic;t}, c)$$
   for each individual component $c \in c^*$.

   d) *Update v-log-likelihood* $l_{n;t} = l_{n;t-1} + \ln\left(\sum_{c \in c^*} \hat{\alpha}_{cn} f_n(d_t|\psi_{c;t}, c)\right)$.

   e) *Compute the weights* $w_{cn;t}$ *approximating the Kronecker symbol* δ_{c,c_t}
   $$w_{cn;t} = \frac{\hat{\alpha}_{cn} f_n(d_t|\psi_{c;t}, c)}{\sum_{\tilde{c} \in c^*} \hat{\alpha}_{\tilde{c}n} f_n(d_t|\psi_{\tilde{c};t}, \tilde{c})}.$$

f) Update copies of the statistic $\kappa_{c;t} = \kappa_{c;t-1} + w_{cn;t}$ *determining Dirichlet pdf.*

g) Update copies of the prior pdfs

$$f(\Theta_{ic}|d(t)) \propto [f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)]^{w_{icn;t}} f(\Theta_{ic}|d(t-1)), \quad (6.95)$$
$$w_{icn;t} = \sum_{\tilde{c} \in c_i^*} w_{\tilde{c}n;t}.$$

The set c_i^* *includes pointers to components that contain the* i*th factor.*

h) Go to the beginning of Sequential mode *if* $t \leq \mathring{t}$. *Otherwise continue.*

4. *Stop if* $l_{n;\mathring{t}} < \bar{l}$ *or the compared valued are taken as equal (see Proposition 6.2) or* $n \geq \mathring{n}$. *Otherwise set* $\bar{l} = l_{n;\mathring{t}}$, *increase the iteration counter* $n = n+1$ *and apply the flattening operation to*

$$Di_{\alpha}(\kappa_{c;\mathring{t}}) \prod_{c \in c^*} \prod_{i \in i^*} f(\Theta_{ic}|d(\mathring{t})).$$

It provides the new estimate of the prior pdf (6.3) given by $f_n(\Theta_{ic})$, κ_{cn}. *The flattening rate corresponding to the iterative Bayesian learning, cf. Proposition 6.9, has to be used.*

5. *Go to the beginning of* Iterative mode.

Remark(s) 6.22

1. *The BQB algorithm uses Bayesian predictors for estimating the Kronecker symbol* δ_{c,c_t}. *They, among other, respect uncertainty of the current estimates of unknown parameters. Predictions become too cautious if this uncertainty is too high. It may break down the algorithm completely. Knowing it, the remedy is simple. Essentially, predictions used in the EM algorithm that ignore these uncertainties have to be used in several initial iterative steps of the algorithm.*
2. *The improvement anticipated in Problem 6.13 can surely be applied to BQB estimation, too.*

6.6 Structure estimation

The structure of the mixture is determined by the number and composition of components. Structure of individual factors is included in it; see Agreement 5.4. We can list the set of potential candidates for such a selection and label promising subselections by a "structural" pointer $S \in S^*$. As its best value is unknown, it becomes a part of the unknown mixture parameter Θ. Its estimation essentially reduces to evaluation of the corresponding marginal probability $f(S|d(\mathring{t}))$. For it, the predictive pdfs $f(d_t|d(t-1), S)$, are computed in parallel for the competing structures $S \in S^*$ and used for computing the posterior pf

$$f(S|d(\mathring{t})) \propto \prod_{t \in t^*} f(d_t|d(t-1), S)\, f(S) = \frac{\mathcal{I}(d(\mathring{t})|S)}{\mathcal{I}(d(0)|S)} f(S). \tag{6.96}$$

Recall that $\mathcal{I}(d(\mathring{t})|S)$ is the normalization constant (2.46) with $\mathcal{P}_{a^*_{\mathring{t}+1}} = d(\mathring{t})$, computed within the Sth structure. The pf $f(S)$ is the optional prior probability assigned $S \in S^*$.

The complexity of the structure estimation increases with the number of compared structures. This number blows up exponentially with the dimensionality of data space, lengths of regression vectors and the number of components. Thus, we can compare only a small subselection of competitive structures. The following feasible iterative procedure is used.

Algorithm 6.17 (Conceptual structure estimation)

1. *The mixture is estimated with the richest acceptable structure, both with respect to the number of components and complexity of the involved factors.*
2. *Structures of the individual factors are estimated; see Section 6.6.1.*
3. *Quantities having no (even indirect) influence on quality markers are excluded from the considered structures; cf. Section 5.1.4.*
4. *Structures of the individual components are estimated; see Section 6.6.3.*
5. *Similar components are merged; see Section 6.6.4.*
6. *Superfluous components are cancelled; see Section 6.6.4.*
7. *The resulting mixture is flattened and the evaluation sequence is iterated until convergence.*

Problem 6.14 (Feasibility and reliability of structure estimation) *Structure estimation is known to be a hard task even in the unimodal case. The known problems are surely enhanced in the mixture estimation. Sufficient amount of data seems to be the only advantage of the addressed problem. On the other hand, it increases the computational load. Both experimental design [117] and specific instances of Algorithm 6.17 decide on the feasibility and reliability of structure estimation. Adequate and practically feasible experimental design is not available for the considered estimation of mixtures. Also, the proposed way of structure estimation can be and should be improved.*

6.6.1 Estimation of factor structure

In this case, structures of unimodal models with single outputs have to be estimated. This is a classical, well-supported problem. For normal factors, efficient numerical procedures exist [93, 95]. For other factors (Markov, MT; see Chapters 10 and 12), the adequate procedures have to be prepared but the basic ideas can be “stolen” from the normal case.

6.6.2 Structure estimation in factor splitting

Structure estimation before splitting of factors, Section 6.4.8, was found indispensable when splitting is made in the direction of largest uncertainty of the posterior pdf $f(\Theta|d(\mathring{t}))$. This happens in the case of optimization-based splitting.

Originally, superfluous parameters were excluded when recognized. This action is, however, dangerous when the estimates are still too far from the correct positions. Requirement 6.2 on sufficient richness of the factor structure is then easily violated. The parameters labelled as superfluous for an intermediate mixture estimate can be significant in the final best estimate. The following combination of the structure estimation with splitting provides the remedy.

Let us apply structure estimation. It splits $\Theta = (\Theta_a, \Theta_b) \equiv \underline{\text{temporarily}}$ (superfluous, significant) parameters. Using some technique of Section 6.4.8, the marginal pdf $f(\Theta_b|d(\mathring{t}))$ is split into a pair $\hat{f}_j(\Theta_b|d(\mathring{t})),\ j = 1, 2$. The part, that is temporarily taken as superfluous, remains unchanged. It defines the pair of pdfs on the original parameter space

$$\hat{f}_j(\Theta|d(\mathring{t})) = f(\Theta_a|\Theta_b, d(\mathring{t}))\hat{f}_j(\Theta_b|d(\mathring{t})),\ j = 1, 2. \tag{6.97}$$

These pdfs are used in further iterations so that the possibility to move temporarily superfluous parameter entries among significant ones and vice versa is preserved.

6.6.3 Estimation of component structure

A component is a product of factors (see Agreement 5.4)

$$\begin{aligned} f(d_t|\phi_{t-1}, \Theta_c, c) &= \prod_{i \in i^*} f(d_{ic;t}|d_{1c;t}, \ldots, d_{(i-1)c;t}, \phi_{t-1}, \Theta_{ic}) \\ &\equiv \prod_{i \in i^*} f(d_{ic;t}|, \psi_{ic;t}, \Theta_{ic}). \end{aligned} \tag{6.98}$$

The permutation π_c (5.10) of d_t to $d_{c;t}$ influences the overall description of the parameterized component. Thus, it makes sense to search for the best ordering, to search for the structure of the component. It can be done by the following conceptual algorithm.

Algorithm 6.18 (MAP estimate of a component structure) *Specify the set of compared structures of the component, and for all its entries execute the following actions.*

- *Estimate structures of the involved factors according to Section 6.6.1.*
- *Evaluate the likelihood value assigned to the inspected component structure and the best structures of factors.*

Finally, compute the posterior probability of the component structures within the considered set and select the MAP estimate.

The algorithm is parameterized by the set of compared component structures. This set is always somehow restricted. The following constraints are common.

- Discrete-valued entries have to be placed at the end of d-entries in order to get Markov-chain factors.
- Structures in which "privileged quantities" defining quality markers, Section 5.1.4, depend on the smallest number of quantities are preferred.

Problem 6.15 (Algorithmic estimation of component structure) *In spite of the discussed restrictions, the complete set is again too large to be inspected fully. Thus, a similar strategy to the factor-structure estimation has to be adopted. A full local search around the best current guess of the structure is to be performed while an improvement is observable. An algorithmic solution of this problem is, however, missing.*

6.6.4 Merging and cancelling of components

The initialization by the factor splitting, Section 6.4.8, is almost always employed for determining the overall number of components. During splitting, the number of components increases rather quickly, leading to mixtures with too many components. It calls for algorithms reducing the *number of components*. They are discussed here.

We deal with mixtures with different number of components. In order to distinguish them, we attach the left superscript that stresses the number of components involved.

Two types of reductions of number of components are possible. The *merging of components* deals with mixtures

$$ {}^{\lfloor\mathring{c}}f(d_t|d(t-1),\Theta) = \sum_{c=1}^{\mathring{c}} \alpha_c f(d_t|d(t-1),\Theta_c,c) $$

that contain almost identical components. We merge them and estimate the mixture with less components.

The *cancelling of components* deals with mixtures containing negligible components. We drop them and estimate the mixture with less components.

For a proper treatment of both cases, subsequent paragraphs inspect

- how to modify prior and posterior pdfs on parameters after reducing the number of components,
- how to find the component to be modified,
- when to stop reductions.

How to modify pdfs on parameters?

The mixture is invariant with respect to permutations of its components. Thus, we can inspect merging of the components $\mathring{c}-1, \mathring{c}$ and cancelling of the component $\mathring{c}$.

The verbal description of the component merging implies that

$$\begin{aligned}
&{}^{\lfloor \mathring{c}-1}f(\alpha_1,\ldots,\alpha_{\mathring{c}-1},\Theta_1,\ldots,\Theta_{\mathring{c}-1})\\
&\quad\propto {}^{\lfloor \mathring{c}}f(\alpha_1,\ldots,\alpha_{\mathring{c}},\Theta_1,\ldots,\Theta_{\mathring{c}}|\alpha_{\mathring{c}-1}=\alpha_{\mathring{c}},\Theta_{\mathring{c}-1}=\Theta_{\mathring{c}})\\
&\quad\underbrace{\propto}_{(6.3)} \prod_{c=1}^{\mathring{c}-2}\left[\alpha^{\,{}^{\lfloor \mathring{c}}\kappa_c-1}\;{}^{\lfloor \mathring{c}}f(\Theta_c)\right]\alpha_{\mathring{c}-1}^{\,{}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1}+\,{}^{\lfloor \mathring{c}}\kappa_{\mathring{c}}-2}\;{}^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}-1})\;{}^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}}=\Theta_{\mathring{c}-1}).
\end{aligned}\tag{6.99}$$

It gives the merged component in the form

$${}^{\lfloor \mathring{c}-1}f(\Theta_{\mathring{c}-1}) \propto {}^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}-1})\;{}^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}}=\Theta_{\mathring{c}-1}) \tag{6.100}$$

and its weight is characterized by the statistic

$${}^{\lfloor \mathring{c}-1}\kappa_{\mathring{c}-1} = {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}} - 1$$

while pdfs describing the remaining components and weights are unchanged.

The verbal description of the component cancelling implies that

$$\begin{aligned}
&{}^{\lfloor \mathring{c}-1}f(\alpha_1,\ldots,\alpha_{\mathring{c}-1},\Theta_1,\ldots,\Theta_{\mathring{c}-1})\\
&\quad= {}^{\lfloor \mathring{c}}f(\alpha_1,\ldots,\alpha_{\mathring{c}},\Theta_1,\ldots,\Theta_{\mathring{c}-1},\Theta_{\mathring{c}}|\alpha_{\mathring{c}}=0,\Theta_{\mathring{c}}\text{ has no influence})\\
&\quad\underbrace{\propto}_{(6.3)} \prod_{c=1}^{\mathring{c}-1}\alpha_c^{\,{}^{\lfloor \mathring{c}}\kappa_c-1}\;{}^{\lfloor \mathring{c}}f(\Theta_c).
\end{aligned}\tag{6.101}$$

Thus, the spurious component is completely neglected and statistics of the remaining pdfs are unchanged.

Remark(s) 6.23

1. *The considered prior and posterior pdfs have the common form (6.3) so that the inspected question is answered both for prior and posterior pdfs.*
2. *The merging leads to a proper pdf only if* ${}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}} - 1 > 0$. *This condition may be critical for prior pdfs as the component is ready for cancelling if this condition is violated for the posterior pdf.*
3. *The merging (6.99) increases expected weights of components with indexes* $c = 1,\ldots,\mathring{c}-2$ *as*

$${}^{\lfloor \mathring{c}-1}\mathcal{E}[\alpha_c] = \frac{{}^{\lfloor \mathring{c}}\kappa_c}{\sum_{\tilde{c}=1}^{\mathring{c}}\,{}^{\lfloor \mathring{c}}\kappa_{\tilde{c}} - 1} \geq \frac{{}^{\lfloor \mathring{c}}\kappa_c}{\sum_{\tilde{c}=1}^{\mathring{c}}\,{}^{\lfloor \mathring{c}}\kappa_{\tilde{c}}} \equiv {}^{\lfloor \mathring{c}}\mathcal{E}[\alpha_c].$$

Also, a straightforward algebra reveals that

$${}^{\lfloor \mathring{c}-1}\mathcal{E}[\alpha_{\mathring{c}-1}] \geq {}^{\lfloor \mathring{c}}\mathcal{E}[\alpha_{\mathring{c}-1}] + {}^{\lfloor \mathring{c}}\mathcal{E}[\alpha_{\mathring{c}}].$$

Thus, values of likelihoods assigned to respective components decide whether the merging improves the overall mixture likelihood.

4. *The cancelling (6.101) implies*

$${}^{\lfloor \mathring{c}-1}\mathcal{E}[\alpha_c] \equiv \frac{{}^{\lfloor \mathring{c}}\kappa_c}{\sum_{\tilde{c}=1}^{\mathring{c}-1} {}^{\lfloor \mathring{c}}\kappa_{\tilde{c}}} \geq \frac{{}^{\lfloor \mathring{c}}\kappa_c}{\sum_{\tilde{c}=1}^{\mathring{c}} {}^{\lfloor \mathring{c}}\kappa_{\tilde{c}}} \equiv {}^{\lfloor \mathring{c}}\mathcal{E}[\alpha_c],\ c = 1,\ldots,\mathring{c}-1.$$

   It indicates that there is a chance for an increase of the v-likelihood by omitting a nonnegative term in the parameterized mixture.

5. *The v-likelihood need not be recomputed after component merging and cancelling if the error of the approximate estimation can be neglected. This fact, demonstrated below, significantly decreases the computational burden.*

What are candidates for merging of components

Generally, we select the merging candidate by evaluating a distance between the posterior pdf ${}^{\lfloor \mathring{c}}f(\Theta|d(\mathring{t}))$, describing the original mixture, and the pdf ${}^{\lfloor \mathring{c}-1}f(\Theta|d(\mathring{t}))$ obtained by merging. We have to extend the merged pdf to the original parameter space. Then, we can use the KL divergence as the evaluated distance.

Let us start with the extension of the merger.

The prior and posterior pdfs assigned to the original and merged pdf have the form (6.3). The merging does not change initial component and their estimates, i.e.,

$${}^{\lfloor \mathring{c}-1}f(\Theta_c|d(\mathring{t})) = {}^{\lfloor \mathring{c}}f(\Theta_c|d(\mathring{t})),\ {}^{\lfloor \mathring{c}-1}\kappa_{c;\mathring{t}} = {}^{\lfloor \mathring{c}}\kappa_{c;\mathring{t}},\ c = 1,\ldots,\mathring{c}-2.$$

The same identity holds for prior pdfs.

The posterior pdf corresponding to the component obtained by merging (6.100) has the form

$$\begin{aligned} {}^{\lfloor \mathring{c}-1}f(\Theta_{\mathring{c}-1}|d(\mathring{t})) &\propto {}^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}-1}|d(\mathring{t}))\, {}^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}} = \Theta_{\mathring{c}-1}|d(\mathring{t})) \\ {}^{\lfloor \mathring{c}-1}\kappa_{\mathring{c}-1;\mathring{t}} &= {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}} - 1. \end{aligned}$$

The pdf ${}^{\lfloor \mathring{c}-1}f(\Theta|d(\mathring{t}))$ acts on the reduced space of parameters. For judging the distance of ${}^{\lfloor \mathring{c}-1}f$ and ${}^{\lfloor \mathring{c}}f$, we have to stay within the original parameter space corresponding to $\mathring{c}$ components. For achieving it, we extend the merger on the original space by using identical estimates $\tilde{f}(\Theta_{\mathring{c}-1}|d(\mathring{t}))$, $\tilde{f}(\Theta_{\mathring{c}}|d(\mathring{t}))$ of the last two components. Both these pdfs are determined by the identical statistics that are selected so that the merger (6.100) is obtained after their merging. Taking into account the form (6.3), we can formalize this requirement at the factor level and set

$$\begin{aligned} \tilde{f}(\Theta_{i(\mathring{c}-1)} = \Theta_i|d(\mathring{t})) &\equiv \tilde{f}(\Theta_{i\mathring{c}} = \Theta_i|d(\mathring{t})) \\ &\propto \sqrt{{}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)} = \Theta_i|d(\mathring{t}))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_i|d(\mathring{t}))} \\ \tilde{\kappa}_{\mathring{t}} &\equiv 0.5\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}}\right). \end{aligned} \tag{6.102}$$

This completes extension of the prospective merger to the original space. The posterior pdf ${}^{\lfloor\mathring{c}}\tilde{f}(\Theta|d(\mathring{t}))$ has the form (6.3) with factors describing components $c = 1,\ldots,\mathring{c}-2$ identical with those describing original mixture. The last two components have the posterior pdfs assigned to them in the form (6.102).

With this, we are ready to search for merging candidates. The components $\mathring{c}-1$, $\mathring{c}$ are taken candidate for merging if the following KL divergence is small

$$\begin{aligned}
\mathcal{D}_{(\mathring{c}-1)\mathring{c}} &\equiv \mathcal{D}\left({}^{\lfloor\mathring{c}}f(\cdot|d(\mathring{t}))\middle|\middle|\, {}^{\lfloor\mathring{c}}\tilde{f}(\cdot|d(\mathring{t}))\right) \\
&= \sum_{i=1}^{\mathring{d}} \Big[\mathcal{D}\left({}^{\lfloor\mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\middle|\middle|\, \tilde{f}(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\right) \\
&\quad + \mathcal{D}\left({}^{\lfloor\mathring{c}}f(\Theta_{i\mathring{c}}|d(\mathring{t}))\middle|\middle|\, \tilde{f}(\Theta_{i\mathring{c}}|d(\mathring{t}))\right)\Big] \\
&\quad + \mathcal{D}\left(Di_{\alpha_{\mathring{c}-1},\alpha_{\mathring{c}}}\left({}^{\lfloor\mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}},\, {}^{\lfloor\mathring{c}}\kappa_{\mathring{c};\mathring{t}}\right)\middle|\middle|\, Di_{\alpha_{\mathring{c}-1},\alpha_{\mathring{c}}}\left(\tilde{\kappa}_{\mathring{t}}, \tilde{\kappa}_{\mathring{t}}\right)\right).
\end{aligned} \tag{6.103}$$

This allows us to order candidates for merging. We take also into account that

- a difference in the structure of components prevents their merging as they are infinitely far,
- the distance matrix (6.103) is symmetric,
- the numerical construction of the trial merger is necessary for finding components whose merging makes the smallest change of the posterior pdf.

Algorithm 6.19 (Selection and merging of a pair of components)
Initial mode

- *Perform estimation of the mixture with a sufficient number of components* $\mathring{c}$ *so that the pdfs* ${}^{\lfloor\mathring{c}}f(\Theta_{ic}|d(t))$ *and statistics* ${}^{\lfloor\mathring{c}}\kappa_t$, $t \in \{0, \mathring{t}\}$, $c \in c^*$, $i = 1,\ldots,\mathring{d}$, *are at disposal.*
- *Set the minimum value of the KL divergence* $\underline{\mathcal{D}} = +\infty$ *and initialize the indexes* $(\underline{c}, \underline{\tilde{c}}) = (0,0)$ *of the component pair to be merged.*

Evaluation mode

1. *Select the pair whose merging leads to the smallest KL divergence between the original mixture and mixture with the merged components as follows.*

```
For $c = 1,\ldots,\mathring{c}-1$
  For $\tilde{c} = c+1,\ldots,\mathring{c}$
        Set the indicator of the common structure $cs = 0$.
    For $i = 1,\ldots,\mathring{d}$
        Set $cs = -1$ and break the cycle over $i$ if the structures of
        $\Theta_{ic}$ and $\Theta_{i\tilde{c}}$ differ.
```

```
              Create the pdf $\tilde{f}$ (6.102) needed in the distance measuring
              with $\mathring{c}-1 \leftrightarrow c,\ \mathring{c} \leftrightarrow \tilde{c}$ ($\leftrightarrow$ means mutual correspondence)
        end   of the cycle over $i$
            Do if $cs = 0$
              Complete definition of the pdf (6.102) used in
              the KL divergence and set $\tilde{\kappa}_{\mathring{t}} = \frac{1}{2}\left(\kappa_{c;\mathring{t}} + \kappa_{\tilde{c};\mathring{t}}\right)$.
             Evaluate the KL divergence $\tilde{\mathcal{D}} \equiv \mathcal{D}_{(\mathring{c}-1)\mathring{c}}$ according to (6.103)
              with $\mathring{c}-1 \leftrightarrow c,\ \mathring{c} \leftrightarrow \tilde{c}$
              If $\tilde{\mathcal{D}} < \underline{\mathcal{D}}$
                 Set $\underline{\mathcal{D}} = \tilde{\mathcal{D}},\ (\underline{c}, \underline{\tilde{c}}) = (c, \tilde{c})$ and store the pdf $\tilde{f},\ \tilde{\kappa}$.
              end of the test on the value $\tilde{\mathcal{D}}$
           end of the condition $cs = 0$
      end   of the cycle over $\tilde{c}$
    end   of the cycle over $c$
```

2. *Stop and announce that no merging is possible if* $(\underline{c}, \underline{\tilde{c}}) = (0, 0)$. *It happens if no pair of components has the identical structure of components.*
3. *Rename components so that* $(\mathring{c}-1, \mathring{c}) = (\underline{c}, \underline{\tilde{c}})$.
4. *Finish the merging of the best components found and apply also the same operation to the prior pdf, i.e.,*

$$
\begin{aligned}
{}^{\lfloor \mathring{c}-1}\kappa_{\mathring{c}-1;\mathring{t}} &\equiv 2\tilde{\kappa}_{\mathring{t}} - 1 \equiv {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}} - 1 \\
{}^{\lfloor \mathring{c}-1}\kappa_{\mathring{c}-1;0} &\equiv {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;0} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c};0} - 1 \text{ and}
\end{aligned}
$$

For $i = 1, \ldots, \mathring{d}$

$$
\begin{aligned}
&{}^{\lfloor \mathring{c}-1}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t})) \\
&\equiv \frac{{}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_{i(\mathring{c}-1)}|d(\mathring{t}))}{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, d\Theta_{i(\mathring{c}-1)}} \\
&\propto \left[\tilde{f}(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\right]^2 \qquad (6.104) \\
&{}^{\lfloor \mathring{c}-1}f(\Theta_{i(\mathring{c}-1)}|d(0)) \equiv \\
&\equiv \frac{{}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(0))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_{i(\mathring{c}-1)}|d(0))}{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(0))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_{i(\mathring{c}-1)}|d(0))\, d\Theta_{i(\mathring{c}-1)}}.
\end{aligned}
$$

end of the cycle over i.

5. *Leave the remaining components and statistics unchanged by the considered merging. It completes the definition of the prior and posterior pdf on the mixture with* $\mathring{c}-1$ *components.*

The necessary condition for obtaining the proper pdfs $\kappa_{\mathring{c}-1;t}+\kappa_{\mathring{t};t}>1,\ t\in\{0,\mathring{t}\}$ should be checked. Such components should not be merged but they are hot candidates for cancelling.

The complexity of the underlying evaluations is clearer to see on a practical version of this algorithm written for parameterized factors in the exponential family

$$f(d_{ic;t}|\psi_{ic;t},\Theta_{ic}) = A(\Theta_{ic})\exp\left[\langle B(\Psi_{ic;t}), C(\Theta_{ic})\rangle + D(\Psi_{ic;t})\right],$$

with scalar functions $A(\Theta_{ic}) \geq 0$, $D(\Psi_{ic;t})$, $\Psi_{ic;t} = [d_{ic;t}, \psi'_{ic;t}]'$, and $\langle\cdot,\cdot\rangle$ being a scalar product of array-valued functions $B(\cdot), C(\cdot)$; see Section 3.2. Its conjugate pdfs have the form, $t\in\{0,\mathring{t}\}$,

$$\begin{aligned}{}^{\lfloor\mathring{c}}f\left(\Theta\middle|\,{}^{\lfloor\mathring{c}}V_t,\,{}^{\lfloor\mathring{c}}\nu_t,\,{}^{\lfloor\mathring{c}}\kappa_t\right) &= \mathcal{B}^{-1}\left({}^{\lfloor\mathring{c}}\kappa_t\right)\prod_{c=1}^{\mathring{c}}\alpha_c^{\,{}^{\lfloor\mathring{c}}\kappa_{c;t}-1}\\ &\times\prod_{i=1}^{\mathring{d}}\frac{A(\Theta_{ic})^{\,{}^{\lfloor\mathring{c}}\nu_{ic;t}}\exp\left[\left\langle\,{}^{\lfloor\mathring{c}}V_{ic;t}, C_{ic}(\Theta_{ic})\right\rangle\right]}{\underbrace{\int A(\Theta_{ic})^{\,{}^{\lfloor\mathring{c}}\nu_{ic;t}}\exp\left[\left\langle\,{}^{\lfloor\mathring{c}}V_{ic;t}, C_{ic}(\Theta_{ic})\right\rangle\right]d\Theta_{ic}}_{\mathcal{I}\left(\,{}^{\lfloor\mathring{c}}V_{ic;t},\,{}^{\lfloor\mathring{c}}\nu_{ic;t}\right)}}.\end{aligned} \tag{6.105}$$

Algorithm 6.20 (Merging of a component pair in EF)

Initial mode

- *Perform estimation of a mixture with a sufficient number of components* $\mathring{c}$ *so that statistics describing them* ${}^{\lfloor\mathring{c}}\kappa_t,\ {}^{\lfloor\mathring{c}}\nu_{ic;t},\ {}^{\lfloor\mathring{c}}V_{ic;t},\ t\in\{0,\mathring{t}\},\ c\in c^*,\ i=1,\ldots,\mathring{d}$, *are at disposal.*
- *Set the minimum value of the KL divergence* $\underline{\mathcal{D}}=+\infty$ *and initialize the indexes* $(\underline{c},\underline{\tilde{c}})=(0,0)$ *of the component pair to be merged.*

Evaluation mode

1. *Select the pair whose merging leads to the smallest KL divergence between the original mixture and mixture with the merged components as follows.*

```
For  $c = 1,\ldots,\mathring{c}-1$
  For  $\tilde{c} = c+1,\ldots,\mathring{c}$
            Set the indicator of the common structure  $cs = 0$.
    For  $i = 1,\ldots,\mathring{d}$
            Set $cs = -1$ and break the cycle over $i$ if the structures of
            $\Theta_{ic}$ and $\Theta_{i\tilde{c}}$ differ.
            Create the statistics
            $\tilde{\nu}_{i;\mathring{t}} = 0.5(\nu_{ic;\mathring{t}} + \nu_{i\tilde{c};\mathring{t}}),\quad \tilde{V}_{i;\mathring{t}} = 0.5(V_{ic;\mathring{t}} + V_{i\tilde{c};\mathring{t}})$.
    end  of the cycle over $i$
```

```
            Do if $cs = 0$
                Complete creating the statistics for the KL divergence
                 measuring and set $\tilde{\kappa}_{\mathring{t}} = 0.5(\kappa_{c;\mathring{t}} + \kappa_{\tilde{c};\mathring{t}})$.
                Evaluate the common part of the KL divergence
                $\tilde{\mathcal{D}} = \mathcal{D}\left(Di_{\alpha_c,\alpha_{\tilde{c}}}(\kappa_{c;\mathring{t}}, \kappa_{\tilde{c};\mathring{t}})||Di_{\alpha_c,\alpha_{\tilde{c}}}(\tilde{\kappa}_{\mathring{t}}, \tilde{\kappa}_{\mathring{t}})\right)$.
                Complete the evaluation of the KL divergence
        For   $i = 1, \ldots, \mathring{d}$
                $\tilde{\mathcal{D}} = \tilde{\mathcal{D}} + \mathcal{D}\ \left(f(\Theta_{ic}|V_{ic;\mathring{t}}, \nu_{ic;\mathring{t}})\right|\left| f(\Theta_{ic}|\tilde{V}_{i;\mathring{t}}, \tilde{\nu}_{i;\mathring{t}})\right)$
                $+\mathcal{D}\ \left(f(\Theta_{i\tilde{c}}|V_{i\tilde{c};\mathring{t}}, \nu_{i\tilde{c};\mathring{t}})\right|\left| f(\Theta_{ic}|\tilde{V}_{i;\mathring{t}}, \tilde{\nu}_{i;\mathring{t}})\right)$.
        end   of the cycle over i
               If $\tilde{\mathcal{D}} < \underline{\mathcal{D}}$
                   Set $\underline{\mathcal{D}} = \tilde{\mathcal{D}}$, $(\underline{c}, \underline{\tilde{c}}) = (c, \tilde{c})$ and store values of
                    the trial statistics $\tilde{\kappa}_{\mathring{t}}$, $\tilde{\nu}_{i;\mathring{t}}$, $\tilde{V}_{i;\mathring{t}}$, $\forall i \in i^*$.
               end of the test on the value $\tilde{\mathcal{D}}$.
            end of the condition $cs = 0$
    end   of the cycle over $\tilde{c}$
end   of the cycle over c
```

2. *Stop and announce that no merging is possible if* $(\underline{c}, \underline{\tilde{c}}) = (0, 0)$.
3. *Rename components so that* $(\mathring{c} - 1, \mathring{c}) = (\underline{c}, \underline{\tilde{c}})$.
4. *Finish the merging of the best components found and also apply the same operation to the prior pdf*

$$
\begin{aligned}
&{}^{\lfloor \mathring{c}-1}\kappa_{\mathring{c}-1;\mathring{t}} \equiv 2\tilde{\kappa}_{\mathring{t}} - 1 \equiv {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}} - 1 \qquad (6.106)\\
&{}^{\lfloor \mathring{c}-1}\kappa_{\mathring{c}-1;0} \equiv {}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;0} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c};0} - 1 \quad \textit{and}
\end{aligned}
$$

```
    For     $i = 1, \ldots, \mathring{d}$
                $^{\lfloor \mathring{c}-1}\nu_{i(\mathring{c}-1);\mathring{t}} \equiv 2\tilde{\nu}_{i;\mathring{t}} \equiv {}^{\lfloor \mathring{c}}\nu_{i(\mathring{c}-1);\mathring{t}} + {}^{\lfloor \mathring{c}}\nu_{i\mathring{c};\mathring{t}}$
                $^{\lfloor \mathring{c}-1}V_{i(\mathring{c}-1);\mathring{t}} \equiv 2\tilde{V}_{i;\mathring{t}} \equiv {}^{\lfloor \mathring{c}}V_{i(\mathring{c}-1);\mathring{t}} + {}^{\lfloor \mathring{c}}V_{i\mathring{c};\mathring{t}}$
                $^{\lfloor \mathring{c}-1}\nu_{i(\mathring{c}-1);0} \equiv {}^{\lfloor \mathring{c}}\nu_{i(\mathring{c}-1);0} + {}^{\lfloor \mathring{c}}\nu_{i\mathring{c};0}$
                $^{\lfloor \mathring{c}-1}V_{i(\mathring{c}-1);0} \equiv {}^{\lfloor \mathring{c}}V_{i(\mathring{c}-1);0} + {}^{\lfloor \mathring{c}}V_{i\mathring{c};0}$.        (6.107)
    end     of the cycle over i
```

5. *Leave the remaining statistics unchanged by the considered merging. It completes the definition of the conjugate prior and posterior pdfs on the mixture with* $\mathring{c} - 1$ *components.*

Remark(s) 6.24
Merging can be performed at the factor level. It is done in an independent paragraph. Factor merging can reveal common factors that reflect a common physical nature of some modelled relationships.

Merging of a group of components

The previous discussion describes how to merge a component pair. The resulting estimate has to be flattened and the estimation repeated. It provides both corrected merger and the v-likelihood that indicates success or failure of this merging. This way is extremely costly for the considered large data sets. Thus, it makes sense to merge several pairs of components before re-estimation. It calls for a rule determining how many components should be merged. In other words, the estimation results obtained for $\mathring{c}$ components have to be used for prediction of the value of the v-likelihood after group merging. This problem is addressed here.

The merging provides both ${}^{\lfloor \mathring{c}-1}f(\Theta|d(\mathring{t}))$ and ${}^{\lfloor \mathring{c}-1}f(\Theta)$. We also know ${}^{\lfloor \mathring{c}}f(\Theta|d(\mathring{t}))$ and ${}^{\lfloor \mathring{c}}f(\Theta)$. Parameter estimates of both mixtures, labelled by $\tilde{c} \in \{\mathring{c}-1, \mathring{c}\}$, obey $\underline{\text{approximately}}$ the Bayes rule

$$\overbrace{{}^{\lfloor \tilde{c}}f(\Theta|d(\mathring{t}))}^{\text{known}} = \frac{{}^{\lfloor \tilde{c}}f(d(\mathring{t})|\Theta)\ \overbrace{{}^{\lfloor \tilde{c}}f(\Theta)}^{\text{known}}}{{}^{\lfloor \tilde{c}}f(d(\mathring{t}))}. \tag{6.108}$$

For the mixture with $\mathring{c}$ terms, we even know the v-likelihood ${}^{\lfloor \mathring{c}}f(d(\mathring{t}))$. Let us consider the prior and posterior pdfs describing parameters of both mixtures on the subset ${}^{\lfloor \mathring{c}}\Theta^*$ determined by the merging conditions

$$\alpha_{\mathring{c}-1} = \alpha_{\mathring{c}},\ \Theta_{i(\mathring{c}-1)} = \Theta_{i\mathring{c}},\ i = 1, \ldots, \mathring{d}. \tag{6.109}$$

On this subset ${}^{\lfloor \mathring{c}}f(d(\mathring{t})|\Theta) = {}^{\lfloor \mathring{c}-1}f(d(\mathring{t})|\Theta)$ and this factor cancels in the ratio of expressions (6.108)

$$\begin{aligned}
&\frac{{}^{\lfloor \mathring{c}-1}f(\Theta|d(\mathring{t}))}{{}^{\lfloor \mathring{c}}f(\Theta|d(\mathring{t}))} \qquad (6.110)\\
&= \frac{\mathcal{B}\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{t}}\right)}{\mathcal{B}\left({}^{\lfloor \mathring{c}-1}\kappa_{\mathring{t}}\right)} \prod_{i=1}^{\mathring{d}} \frac{{}^{\lfloor \mathring{c}-1}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))}{{}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\ {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_{i(\mathring{c}-1)}|d(\mathring{t}))}\\
&\underbrace{=}_{(6.104),(6.4),(6.109)} \frac{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}}\right)\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}}\right)\Gamma\left(\left(\sum_{c=1}^{\mathring{c}}{}^{\lfloor \mathring{c}}\kappa_{c;\mathring{t}}\right)-1\right)}{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}} + {}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}} - 1\right)\Gamma\left(\sum_{c=1}^{\mathring{c}}{}^{\lfloor \mathring{c}}\kappa_{c;\mathring{t}}\right)}\\
&\times \prod_{i=1}^{\mathring{d}} \frac{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, d\Theta_{i(\mathring{c}-1)} \int {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}}|d(\mathring{t}))\, d\Theta_{i\mathring{c}}}{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\ {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, d\Theta_{i(\mathring{c}-1)}}
\end{aligned}$$

$$\underbrace{=}_{(6.108),(6.104),(6.109)} \frac{{}^{\lfloor \mathring{c}}f(d(\mathring{t}))}{{}^{\lfloor \mathring{c}-1}f(d(\mathring{t}))} \frac{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;0}\right)\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c};0}\right)\Gamma\left(\left(\sum_{c=1}^{\mathring{c}} {}^{\lfloor \mathring{c}}\kappa_{c;0}\right)-1\right)}{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;0}+{}^{\lfloor \mathring{c}}\kappa_{\mathring{c};0}-1\right)\Gamma\left(\sum_{c=1}^{\mathring{c}} {}^{\lfloor \mathring{c}}\kappa_{c;0}\right)}$$

$$\times \prod_{i=1}^{\mathring{d}} \frac{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(0))\, d\Theta_{i(\mathring{c}-1)} \int {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}}|d(0))\, d\Theta_{i\mathring{c}}}{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(0))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}} = \Theta_{i(\mathring{c}-1)}|d(0))\, d\Theta_{i(\mathring{c}-1)}}.$$

The last equality in (6.110) gives the identity that allows us to evaluate the v-likelihood ${}^{\lfloor \mathring{c}-1}f(d(\mathring{t}))$ without repeating the whole estimation. Using the formula $\Gamma(x+1) = x\Gamma(x)$, [156], we get

$$ {}^{\lfloor \mathring{c}-1}f(d(\mathring{t})) = {}^{\lfloor \mathring{c}}f(d(\mathring{t})) \frac{\dfrac{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;0}\right)\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c};0}\right)\left(\left(\sum_{c=1}^{\mathring{c}} {}^{\lfloor \mathring{c}}\kappa_{c;0}\right)-1\right)}{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;0}+{}^{\lfloor \mathring{c}}\kappa_{\mathring{c};0}-1\right)}}{\dfrac{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}}\right)\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}}\right)\left(\left(\sum_{c=1}^{\mathring{c}} {}^{\lfloor \mathring{c}}\kappa_{c;\mathring{t}}\right)-1\right)}{\Gamma\left({}^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1;\mathring{t}}+{}^{\lfloor \mathring{c}}\kappa_{\mathring{c};\mathring{t}}-1\right)}} \tag{6.111}$$

$$\times \frac{\prod_{i=1}^{\mathring{d}} \dfrac{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(0))\, d\Theta_{i(\mathring{c}-1)} \int {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}}|d(0))\, d\Theta_{i\mathring{c}}}{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(0))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}}=\Theta_{i(\mathring{c}-1)}|d(0))\, d\Theta_{i(\mathring{c}-1)}}}{\prod_{i=1}^{\mathring{d}} \dfrac{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, d\Theta_{i(\mathring{c}-1)} \int {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}}|d(\mathring{t}))\, d\Theta_{i\mathring{c}}}{\int {}^{\lfloor \mathring{c}}f(\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, {}^{\lfloor \mathring{c}}f(\Theta_{i\mathring{c}}=\Theta_{i(\mathring{c}-1)}|d(\mathring{t}))\, d\Theta_{i(\mathring{c}-1)}}}.$$

Thus, we can judge whether trial merging of a component pair increases the v-likelihood. This gives us a chance to merge a group components without costly re-estimation.

We write the resulting algorithm for the practically significant case of components in the exponential family. For this model, the formula (6.111) gets a more specific form. It can be written in terms of the statistics of the original mixture so that the superscript $\mathring{c}$ can be dropped.

$$ {}^{\lfloor \mathring{c}-1}f(d(\mathring{t})) =$$

$$= {}^{\lfloor \mathring{c}}f(d(\mathring{t})) \frac{\Gamma\left(\kappa_{\mathring{c}-1;\mathring{t}}+\kappa_{\mathring{c};\mathring{t}}-1\right)\Gamma\left(\kappa_{\mathring{c}-1;0}\right)\Gamma\left(\kappa_{\mathring{c};0}\right)\left(\left(\sum_{c=1}^{\mathring{c}}\kappa_{c;0}\right)-1\right)}{\Gamma\left(\kappa_{\mathring{c}-1;0}+\kappa_{\mathring{c};0}-1\right)\Gamma\left(\kappa_{\mathring{c}-1;\mathring{t}}\right)\Gamma\left(\kappa_{\mathring{c};\mathring{t}}\right)\left(\left(\sum_{c=1}^{\mathring{c}}\kappa_{c;\mathring{t}}\right)-1\right)}$$

$$\times \prod_{i=1}^{\mathring{d}} \frac{\mathcal{I}(V_{i(\mathring{c}-1);\mathring{t}}+V_{i\mathring{c};\mathring{t}}, \nu_{i(\mathring{c}-1);\mathring{t}}+\nu_{i\mathring{c};\mathring{t}})}{\mathcal{I}(V_{i(\mathring{c}-1);0}+V_{i\mathring{c};0}, \nu_{i(\mathring{c}-1);0}+\nu_{i\mathring{c};0})} \frac{\mathcal{I}(V_{i(\mathring{c}-1);0}, \nu_{i(\mathring{c}-1);0})\mathcal{I}(V_{i\mathring{c};0}, \nu_{i\mathring{c};0})}{\mathcal{I}(V_{i(\mathring{c}-1);\mathring{t}}, \nu_{i(\mathring{c}-1);\mathring{t}})\mathcal{I}(V_{i\mathring{c};\mathring{t}}, \nu_{i\mathring{c};\mathring{t}})}.$$

The last formula allows us to modify Algorithm 6.20 so that promising candidates are gradually merged up to the moment when the v-likelihood is predicted to drop.

Algorithm 6.21 (Merging of a group of components in EF)

Initial mode

- *Perform estimation of a mixture with components in the exponential family and conjugate prior pdfs. Let us do that for an over-estimated number of components* $\mathring{c} \geq 2$. *Thus, the statistics* ${}^{\lfloor \mathring{c}}\kappa_t$, ${}^{\lfloor \mathring{c}}\nu_{ic;t}$, ${}^{\lfloor \mathring{c}}V_{ic;t}$, $t \in \{0, \mathring{t}\}$, $c = 1, \ldots, \mathring{c}$, $i = 1, \ldots, \mathring{d}$, *are at our disposal.*

- *Set pointers $c = 1, \tilde{c} = 2$ to trial components to be merged.*

Evaluation mode

```
        Set the indicator of the common structure $cs = 0$.
For   $i = 1, \ldots, \mathring{d}$
        Set $cs = -1$ and break the cycle over $i$ if the structures of
        $\Theta_{ic}$ and $\Theta_{i\tilde{c}}$ differ.
end   of the cycle over $i$
    Do if $cs = 0$
        Evaluate the common part of the trial merger
        $\tilde{\kappa}_{\mathring{t}} = \kappa_{c;\mathring{t}} + \kappa_{\tilde{c};\mathring{t}} - 1,\ \tilde{\kappa}_0 = \kappa_{c;0} + \kappa_{\tilde{c};0} - 1.$
        Evaluate and store the factor-related parts of the trial merger
For   $i = 1, \ldots, \mathring{d}$
        $\tilde{\nu}_{i;\mathring{t}} = \nu_{ic;\mathring{t}} + \nu_{i\tilde{c};\mathring{t}}, \quad \tilde{V}_{i;\mathring{t}} = V_{ic;\mathring{t}} + V_{i\tilde{c};\mathring{t}}.$
end   of the cycle over $i$
        Evaluate the change $\tilde{l}$ of log-v-likelihood expected after
        the definite merging
  $\tilde{l} = \left\{ -\ln\left(\Gamma(\kappa_{c;\mathring{t}})\right) - \ln\left(\Gamma(\kappa_{\tilde{c};\mathring{t}})\right) + \ln\left(\Gamma(\tilde{\kappa}_{\mathring{t}})\right) - \ln\left(\left(\sum_{c=1}^{\mathring{c}-1} \kappa_{c;\mathring{t}}\right) - 1\right)\right\}$
  $- \left\{ -\ln\left(\Gamma(\kappa_{c;0})\right) - \ln\left(\Gamma(\kappa_{\tilde{c};0})\right) + \ln\left(\Gamma(\tilde{\kappa}_0)\right) - \ln\left(\left(\sum_{c=1}^{\mathring{c}} \kappa_{c;0}\right) - 1\right)\right\}.$
For   $i = 1, \ldots, \mathring{d}$
        (factor parts)
        $\tilde{l} = \tilde{l} + \left\{\ln(\mathcal{I}(\tilde{V}_{i;\mathring{t}}, \tilde{\nu}_{i;\mathring{t}})) - \ln(\mathcal{I}(V_{ic;\mathring{t}}, \nu_{ic;\mathring{t}})) - \ln(\mathcal{I}(V_{i\tilde{c};\mathring{t}}, \nu_{i\tilde{c};\mathring{t}}))\right\}$
              $- \left\{\ln(\mathcal{I}(\tilde{V}_{i;0}, \tilde{\nu}_{i;0})) - \ln(\mathcal{I}(V_{ic;0}, \nu_{ic;0})) - \ln(\mathcal{I}(V_{i\tilde{c};0}, \nu_{i\tilde{c};0}))\right\}.$
end   of the cycle over $i$
    end of the condition $cs = 0$
      If $\tilde{l} \leq 0$ or $cs < 0$
        Set $\tilde{c} = \tilde{c} + 1$.
        Go to the beginning of Evaluation mode if $\tilde{c} \leq \mathring{c}$. Otherwise continue.
        Set $c = c + 1$ and $\tilde{c} = c + 1$.
        Go to the beginning of Evaluation mode if $c < \mathring{c}$. Otherwise stop.
      else replace statistics related to the component $c$ by
        $\tilde{\kappa}_{\mathring{t}},\ \tilde{\kappa}_0,\ \left\{\tilde{V}_{i;\mathring{t}},\ \tilde{V}_{i;0},\ \tilde{\nu}_{i;\mathring{t}},\ \tilde{\nu}_{i;0}\right\}_{i=1}^{\mathring{d}}.$
        Swap the components $\mathring{c}$ and $\tilde{c}$.
        Decrease $\mathring{c} = \mathring{c} - 1$, i.e., omit the component $\tilde{c}$.
        Set $\tilde{c} = c + 1$ if $\tilde{c} > \mathring{c}$.
      end of the test on improvement of v-likelihood and of $cs < 0$
        Stop if $\mathring{c} = 1$. Otherwise go to the beginning of Evaluation mode.
```

Factor-based merging

Here we merge superfluous factors by inspecting the presence of common factors. It decreases the computational load, can reveal relationships of a common nature and — if a component pair will consist of common factors only — to reduce the number of components. We proceed similarly as with the merging of component pairs.

Candidates for merging

The pdf on parameters has the form (6.3)

$$f(\Theta|d(t)) \equiv Di_{\alpha}(\kappa_t) \prod_{c \in c^*} \prod_{i=1}^{\mathring{d}} f(\Theta_{ic} d(t)), \ t \in \{0, \mathring{t}\}.$$

For presentation simplicity, we assume that the order of items $d_{i;t}$ in the data record d_t is common for all components.

The pdf ${}^{\lfloor ic\tilde{c}}f(\Theta|d(t))$, $t \in \{0, \mathring{t}\}$, describing the parameter estimate — under the hypothesis that estimates of the ith factor in components $c \neq \tilde{c} \in c^*$ are equal — has the form

$${}^{\lfloor ic\tilde{c}}f(\Theta|d(\mathring{t})) \equiv f(\Theta|d(\mathring{t})) \frac{\tilde{f}(\Theta_{ic}|d(\mathring{t}))\tilde{f}(\Theta_{i\tilde{c}}|d(\mathring{t}))}{f(\Theta_{ic}|d(\mathring{t}))f(\Theta_{i\tilde{c}}|d(\mathring{t}))}, \tag{6.112}$$

where $f(\Theta|d(\mathring{t}))$ is the original pdf whose factors should be merged and $f(\Theta_{ic}|d(\mathring{t}))$, $f(\Theta_{i\tilde{c}}|d(\mathring{t}))$ are its marginal pdfs. The pdf $\tilde{f}(\Theta_{ic}|d(\mathring{t}))$ is the constructed common factor, i.e., $\tilde{f}(\Theta_{ic}|d(\mathring{t})) \equiv \tilde{f}(\Theta_{i\tilde{c}} = \Theta_{ic}|d(\mathring{t}))\ \forall \Theta_{ic} \in \Theta_{ic}^*$. Creation of a common factor makes sense only when the structures of Θ_{ic} and $\Theta_{i\tilde{c}}$ are the same.

The KL divergence of the mixture with the common factor (6.112) to the original mixture is

$${}^{\lfloor ic\tilde{c}}\mathcal{D} \equiv \mathcal{D}\left(\tilde{f}(\Theta_{ic}|d(\mathring{t}))\middle|\middle| f(\Theta_{ic}|d(\mathring{t}))\right) + \mathcal{D}\left(\tilde{f}(\Theta_{ic} = \Theta_{i\tilde{c}}|d(\mathring{t}))\middle|\middle| f(\Theta_{i\tilde{c}}|d(\mathring{t}))\right). \tag{6.113}$$

It is minimized by the geometric mean of the inspected factor estimates

$$\tilde{f}(\Theta_{ic}|d(\mathring{t})) \propto \sqrt{f(\Theta_{ic}|d(\mathring{t}))f(\Theta_{i\tilde{c}} = \Theta_{ic}|d(\mathring{t}))}. \tag{6.114}$$

Note that this merger is proposed in (6.102) without invoking this minimizing property.

Thus, for all $i = 1, \ldots, \mathring{d}$, we can independently evaluate distances ${}^{\lfloor ic\tilde{c}}\mathcal{D}$, $c \in c^*$, $\tilde{c} > c$, for the merger (6.114). The restriction to the upper triangular part only exploits the obvious symmetry of matrices ${}^{\lfloor i\cdot\cdot}\mathcal{D}$, $i = 1, \ldots, \mathring{d}$. The prospective common factors should be closest to those describing the original parameter estimate. This guides us in selecting the prospective merger.

Algorithm 6.22 (Selection and merging of pairs of factors)

Initial mode

- *Estimate the mixture with a sufficient number of components* $\mathring{c}$ *so that the pdfs* ${}^{\lfloor \mathring{c}}f(\Theta_{ic}|d(t))$ *and statistics* ${}^{\lfloor \mathring{c}}\kappa_t$, $t \in \{0, \mathring{t}\}$, $c \in c^*$, $i = 1, \ldots, \mathring{d}$, *are at our disposal.*

Evaluation mode

```
For   $i = 1, \ldots, \mathring{d}$
           Set the smallest value of the KL divergence $\underline{\mathcal{D}} = +\infty$
            and initiate indexes of the component pair
           $(\underline{c}, \underline{\tilde{c}}) = (0, 0)$ in which the merged factors are listed.
  For   $c = 1, \ldots, \mathring{c} - 1$
    For   $\tilde{c} = c + 1, \ldots, \mathring{c}$
           Go to the end of the cycle over $\tilde{c}$ if the structures of
           $\Theta_{ic}$ and $\Theta_{i\tilde{c}}$ differ.
           Create the trial merger (6.114)
           Evaluate the KL divergence $\tilde{\mathcal{D}} \equiv {}^{\lfloor ic\tilde{c}}\mathcal{D}$ according to (6.113).
           If $\tilde{\mathcal{D}} < \underline{\mathcal{D}}$
             Set $\underline{\mathcal{D}} = \tilde{\mathcal{D}}$, $(\underline{c}, \underline{\tilde{c}}) = (c, \tilde{c})$ and store the trial merger.
           end of the test on the value of the KL divergence
    end   of the cycle over $\tilde{c}$
  end   of the cycle over $c$
           Go to the end of cycle over $i$ if $(\underline{c}, \underline{\tilde{c}}) = (0, 0)$.
           Create a merger of prior pdfs corresponding to factors
            with indexes $i\underline{c}$ and $i\underline{\tilde{c}}$.
           Replace prior and posterior estimates of factors with indexes
           $i\underline{c}$, $i\underline{\tilde{c}}$ by the trial merger.
end   of the cycle over $i$
```

Merging of a group of factors

The need to perform full estimation after flattening the merged factors needed for judging improvement of the v-likelihood disqualifies the way described in the previous paragraph. Prediction of the v-likelihood in the same way done with the component provides the adequate remedy. The ability to merge two components only when all factors in them are merged is the main change with respect to the merging of a group of components.

The proposed merging provides both

$$ {}^{\lfloor ic\tilde{c}}f(\Theta|d(\mathring{t})) \text{ and } {}^{\lfloor ic\tilde{c}}f(\Theta) \equiv {}^{\lfloor ic\tilde{c}}f(\Theta|d(0)). $$

We also know $f(\Theta|d(\mathring{t}))$, $f(\Theta) \equiv f(\Theta|d(0))$ and the original v-likelihood $f(d(\mathring{t}))$. We know that for values $\Theta_{ic} = \Theta_{i\tilde{c}}$ the corresponding likelihood functions coincide. Assuming the ratio $f(\Theta|d(\mathring{t}))/\ {}^{\lfloor ic\tilde{c}}f(\Theta|d(\mathring{t}))$ at such points, using the Bayes rule and exploiting the merger form, we get

$$ \begin{aligned} \frac{f(\Theta|d(\mathring{t}))}{{}^{\lfloor ic\tilde{c}}f(\Theta|d(\mathring{t}))} &\equiv \frac{\int f(\Theta_{ic}|d(\mathring{t}))f(\Theta_{i\tilde{c}} = \Theta_{ic}|d(\mathring{t}))\, d\Theta_{ic}}{\int f(\Theta_{ic}|d(\mathring{t}))\, d\Theta_{ic} \int f(\Theta_{i\tilde{c}}|d(\mathring{t}))\, d\Theta_{i\tilde{c}}} \\ &= \frac{{}^{\lfloor ic\tilde{c}}f(d(\mathring{t}))}{f(d(\mathring{t}))} \frac{\int f(\Theta_{ic})f(\Theta_{i\tilde{c}} = \Theta_{ic})\, d\Theta_{ic}}{\int f(\Theta_{ic})\, d\Theta_{ic} \int f(\Theta_{i\tilde{c}})\, d\Theta_{i\tilde{c}}}. \end{aligned} $$

It gives the rule how to find the v-likelihood ${}^{\lfloor ic\tilde{c}}f(d(\mathring{t}))$ after merging

$$ \begin{aligned} {}^{\lfloor ic\tilde{c}}f(d(\mathring{t})) = f\left(d(\mathring{t})\right) & \\ \times \frac{\int f(\Theta_{ic}|d(\mathring{t}))\, d\Theta_{ic} \int f(\Theta_{i\tilde{c}}|d(\mathring{t}))\, d\Theta_{i\tilde{c}}}{\int f(\Theta_{ic}|d(\mathring{t}))f(\Theta_{i\tilde{c}} = \Theta_{ic}|d(\mathring{t}))\, d\Theta_{ic}} & \frac{\int f(\Theta_{ic})f(\Theta_{i\tilde{c}} = \Theta_{ic})\, d\Theta_{ic}}{\int f(\Theta_{ic})\, d\Theta_{ic} \int f(\Theta_{i\tilde{c}})\, d\Theta_{i\tilde{c}}}. \end{aligned} \tag{6.115} $$

The constructed estimate makes it possible to merge a group of factors while the v-likelihood is improving. Then, it remains to merge components. Obviously, the components $c, \tilde{c}$ with all factors common can be taken as a single one. It makes no sense to distinguish them and their weights. Thus, the statistics of their weights are simply summed. This leads to the following algorithm.

Algorithm 6.23 (Merging of a group of factors)

Initial mode

- *Perform estimation of a mixture with a sufficient number of components.*
- *Initialize the list of factors with rows $\rho = (i, c, \tilde{c}) \equiv i$th factor is common for components $c, \tilde{c}$.*
  Mostly, ρ is selected as the empty one.

Evaluation mode

```
For  $i = 1, \ldots, \mathring{d}$
  For  $c = 1, \ldots, \mathring{c} - 1$
```

```
    For   $\tilde{c} = c+1,\ldots,\mathring{c}$
        Go to the end of the cycle over $\tilde{c}$ if the structures of
        $\Theta_{ic}$ and $\Theta_{i\tilde{c}}$ differ.
        Create and store the trial merger (6.114) of posterior pdfs.
        Create the trial merger (6.114) of prior pdfs.
        Evaluate the factor-related change (6.115) of
        the v-likelihood predicted after merging.
        Go to the end of cycle over $\tilde{c}$ if no improvement is predicted.
        Replace prior and posterior factor estimates with indexes
        $i\underline{c}$, $i\underline{\tilde{c}}$ by the trial merger.
        Extend the list of common factors by $\rho = [\rho; (i, \underline{c}, \underline{\tilde{c}})]$ .
    end   of the cycle over $\tilde{c}$
  end   of the cycle over $c$
end   of the cycle over $i$
  For   $c = 1,\ldots,\mathring{c}-1$
    For   $\tilde{c} = c+1,\ldots,\mathring{c}$
        Set $\kappa_{\tilde{c};\mathring{t}} = \kappa_{\tilde{c};\mathring{t}} + \kappa_{c;\mathring{t}} - 1$ $\kappa_{\tilde{c};0} = \kappa_{\tilde{c};0} + \kappa_{c;0} - 1$ and cancel cth
        component if the components consist of common factors only.
    end   of the cycle over $\tilde{c}$
  end   of the cycle over $c$
```

Problem 6.16 (Influence of component weights) *The Dirichlet factor influences the final v-likelihood whenever components are merged. This fact is not taken into account in Algorithm 6.23. It would be necessary to evaluate the change of the v-likelihoods for the possibility that the components will be merged as well as for the case that merging will concern only a few of factors. The modification is worth inspecting.*

The above algorithm is now written for factors in the exponential family. It is enriched so that common factors in several components can be recognized.

Algorithm 6.24 (Systematic merging of factors in EF)

Initial mode

- *Perform estimation of a mixture with factors in the exponential family. The inspected part of the mixture estimate is described by the collection of statistics*

$$\{V_{ic;t}, \nu_{ic;t}\}_{c\in c^*, i=1,\ldots,\mathring{d}, t\in\{0,\mathring{t}\}} \, .$$

  The factors with the common i are supposed to describe the same entry of $d_{i;t}$ irrespective of the component number.

- *Initialize the list with rows $\rho = (i, c, \tilde{c}) \equiv$ ith factor, which is common for components $c, \tilde{c}$. Usually, ρ is initialized as the empty one.*
- *Evaluate logarithms of the individual normalization factors* $\ln(\mathcal{I}(V_{ic;t}, \nu_{ic:t}))$, $\forall c \in c^*, i = 1, \ldots, \mathring{d}, t \in \{0, \mathring{t}\}$.

```
Evaluation mode
   For   $i = 1, \ldots, \mathring{d}$
            Set pointers $c = 1, \tilde{c} = 2$ to trial components to be merged.

     Test of the common structure
            Set the indicator of the common structure to $cs = 0$.
            Set $cs = -1$ if the structures of $\Theta_{ic}$ and $\Theta_{i\tilde{c}}$ differ.
            Do if $cs = 0$
               Create the trial merger
               $\tilde{V}_{i;\mathring{t}} = V_{ic;\mathring{t}} + V_{i\tilde{c};\mathring{t}},\ \tilde{V}_{i;0} = V_{ic;0} + V_{i\tilde{c};0}$
               $\tilde{\nu}_{i;\mathring{t}} = \nu_{ic;\mathring{t}} + \nu_{i\tilde{c};\mathring{t}},\ \tilde{\nu}_{i;0} = \nu_{ic;0} + \nu_{i\tilde{c};0}.$
               Evaluate increment $\tilde{l}$ of the log-v-likelihood
                $\tilde{l} = +\left\{\ln(\mathcal{I}(\tilde{V}_{i;\mathring{t}}, \tilde{\nu}_{i;\mathring{t}})) - \ln(\mathcal{I}(V_{ic;\mathring{t}}, \nu_{ic;\mathring{t}})) - \ln(\mathcal{I}(V_{i\tilde{c};\mathring{t}}, \nu_{i\tilde{c};\mathring{t}}))\right\}$
                    $-\left\{\ln(\mathcal{I}(\tilde{V}_{i;0}, \tilde{\nu}_{i;0})) - \ln(\mathcal{I}(V_{ic;0}, \nu_{ic;0})) - \ln(\mathcal{I}(V_{i\tilde{c};0}, \nu_{i\tilde{c};0}))\right\}.$
            end of the test on $cs = 0$
            If $\tilde{l} \leq 0$ or $cs < 0$
               Set $\tilde{c} = \tilde{c} + 1$.
               Go to the Test of the common structure  if $\tilde{c} \leq \mathring{c}$.
                Otherwise continue.
               Set $c = c + 1$ and $\tilde{c} = c + 1$.
               Go to the beginning of Test of the common structure  if $c < \mathring{c}$.
               Otherwise go to the end of cycle over i.
            else
               replace prior and posterior estimates of factors with indexes
               $ic$ and $i\tilde{c}$ by the trial merger.
               Extend the list of common factors by $\rho = [\rho; (i, c, \tilde{c})]$.
            end of the test on improvement of v-likelihood and of $cs < 0$
   end   of the cycle over i
     Merging of components
For   $c = 1, \ldots, \mathring{c} - 1$
   For   $\tilde{c} = c + 1, \ldots, \mathring{c}$
            Set $\kappa_{\tilde{c};\mathring{t}} = \kappa_{\tilde{c};\mathring{t}} + \kappa_{c;\mathring{t}} - 1,\ \ \kappa_{\tilde{c};0} = \kappa_{\tilde{c};0} + \kappa_{c;0} - 1$ and cancel the
            cth component if the components consist of common factors only.
    end   of the cycle over $\tilde{c}$
end   of the cycle over $c$
```

Remark(s) 6.25
Solution of Problem 6.23 would improve the above algorithm, too.

Component cancelling

The technique used in connection with the merging of components allows us predict the influence of cancelling on the v-likelihood without re-estimation. Consequently, all cancelling possibilities can be checked with a relatively small computational load. The explicit solution is described here. It needs a bit of care since for $\alpha_{\mathring{c}} = 0$ the prior and posterior pdfs are zero.

The the chain rule $f(a|b) = f(a,b)/f(b)$, properties of the Dirichlet pdf, and the formula for marginal pdf (10.4) in Proposition 10.1, however, imply that for $f(\,^{\lfloor \mathring{c}}\alpha) = Di_{\,^{\lfloor \mathring{c}}\alpha}\left(\,^{\lfloor \mathring{c}}\kappa\right)$ the conditional pdf

$$^{\lfloor \mathring{c}}f(\alpha|\alpha_{\mathring{c}} = 0) = Di_{[\alpha_1,\ldots,\alpha_{\mathring{c}-1}]}\left(\left[\,^{\lfloor \mathring{c}}\kappa_1,\ldots,\,^{\lfloor \mathring{c}}\kappa_{\mathring{c}-1}\right]\right) \equiv \,^{\lfloor \mathring{c}-1}f(\alpha) \equiv Di_{\alpha}\left(\,^{\lfloor \mathring{c}-1}\kappa\right).$$

The posterior pdf on parameters before cancelling conditioned by $\alpha_{\mathring{c}} = 0$ and the posterior pdf after cancelling have the form

$$\begin{aligned} ^{\lfloor \mathring{c}}f(\Theta|d(\mathring{t}), \alpha_{\mathring{c}} = 0) &= Di_{\alpha}(\,^{\lfloor \mathring{c}-1}\kappa_{\mathring{t}}) \prod_{c=1}^{\mathring{c}} \frac{^{\lfloor \mathring{c}}f(\Theta_c|d(\mathring{t}))}{^{\lfloor \mathring{c}}\mathcal{I}_c(d(\mathring{t}))} \qquad (6.116)\\ &= \frac{^{\lfloor \mathring{c}}f(d(\mathring{t})|\Theta, \alpha_{\mathring{c}} = 0)}{^{\lfloor \mathring{c}}f(d(\mathring{t}))} Di_{\alpha}(\,^{\lfloor \mathring{c}-1}\kappa_0) \prod_{c=1}^{\mathring{c}} \frac{^{\lfloor \mathring{c}}f(\Theta_c|d(0))}{^{\lfloor \mathring{c}}\mathcal{I}_c(d(0))}\\ ^{\lfloor \mathring{c}-1}f(\Theta|d(\mathring{t})) &= Di_{\alpha}(\,^{\lfloor \mathring{c}-1}\kappa_{\mathring{t}}) \left(\,^{\lfloor \mathring{c}-1}\kappa_{\mathring{t}}\right) \prod_{c=1}^{\mathring{c}-1} \frac{^{\lfloor \mathring{c}-1}f(\Theta_c|d(\mathring{t}))}{^{\lfloor \mathring{c}-1}\mathcal{I}_c(d(\mathring{t}))}\\ &= \frac{^{\lfloor \mathring{c}-1}f(d(\mathring{t})|\Theta)}{^{\lfloor \mathring{c}-1}f(d(\mathring{t}))} Di_{\alpha}\left(\,^{\lfloor \mathring{c}-1}\kappa_0\right) \prod_{c=1}^{\mathring{c}-1} \frac{^{\lfloor \mathring{c}}f(\Theta_c|d(0))}{^{\lfloor \mathring{c}}\mathcal{I}_c(d(0))}. \end{aligned}$$

Similarly as above, $^{\lfloor \mathring{c}}\mathcal{I}_c(d(t)),\ t \in \{0, \mathring{t}\}$ denote normalizing integrals of the pdfs describing parameters Θ_c conditioned on $d(t)$ for the mixture with $\mathring{c}$ components.

The mixture form of the parameterized model implies that

$$^{\lfloor \mathring{c}}f\left(d_t\left|d(t-1), \{\Theta_c, \alpha_c\}_{c=1}^{\mathring{c}-1}, \Theta_{\mathring{c}}, \alpha_{\mathring{c}} = 0\right.\right)$$

does not depend on $\Theta_{\mathring{c}}$. Thus, the <u>exact</u> posterior pdf $^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}}|d(\mathring{t}), \alpha_{\mathring{c}} = 0)$ coincides with its prior counterpart $^{\lfloor \mathring{c}}f(\Theta_{\mathring{c}}|\alpha_{\mathring{c}} = 0)$. The evaluated approximate posterior pdf does not meet this condition. This makes us combine the relationships (6.116) at a specific point ${}^{\lfloor \mathring{c}}\Theta$ for which even approximate estimation provides an exact value likelihood.

We assume that there are parameter values ${}^{\lfloor \mathring{c}}\Theta_c,\ c = 1, \ldots, \mathring{c}$ such that

$$f\left(d_t|d(t-1),\, {}^{\lfloor \mathring{c}}\Theta_c, c\right) = g(d(t)) \text{ for } c = 1, \ldots, \mathring{c} \text{ and a function } g(d(t)). \tag{6.117}$$

Then, the likelihood values ${}^{\lfloor \mathring{c}}f\left(d(\mathring{t})|\, {}^{\lfloor \mathring{c}}\Theta,\, {}^{\lfloor \mathring{c}}\alpha_{\mathring{c}} = 0\right),\ {}^{\lfloor \mathring{c}-1}f\left(d(\mathring{t})|\, {}^{\lfloor \mathring{c}}\Theta\right)$ assigned to both parameterized models coincide as the component weights sum to unity. Taking the ratios of the posterior pdfs of both mixture models at such parameters, we get

$${}^{\lfloor \mathring{c}-1}f(d(\mathring{t})) = {}^{\lfloor \mathring{c}}f(d(\mathring{t})) \frac{{}^{\lfloor \mathring{c}}f\left(\, {}^{\lfloor \mathring{c}}\Theta_{\mathring{c}} \middle|\, d(\mathring{t})\right)}{{}^{\lfloor \mathring{c}}f\left(\, {}^{\lfloor \mathring{c}}\Theta_{\mathring{c}} \middle|\, d(0)\right)} \frac{\mathcal{I}_{\mathring{c}}(d(0))}{\mathcal{I}_{\mathring{c}}(d(\mathring{t}))}. \tag{6.118}$$

The value ${}^{\lfloor \mathring{c}}\Theta_{\mathring{c}}$ is fixed by the requirement (6.117); thus we can predict whether the cancelling will result in a higher v-likelihood. It directly gives the algorithm searched for. It is written for the exponential family.

Algorithm 6.25 (Systematic cancelling of components in EF)

Initial mode

- *Perform estimation of a mixture with factors in the exponential family. The relevant part of the mixture estimate is described by the statistics*

$$\left\{V_{ic;t}, \nu_{ic;t}\right\}_{c \in c^*, i=1,\ldots,\mathring{d}, t \in \{0,\mathring{t}\}}.$$

  The factors with the common i are supposed to describe the same entry of $d_{i;t}$ irrespective of the component number.
- *Evaluate logarithms of the normalization factors* $\ln(\mathcal{I}(V_{ic;t}, \nu_{ic;t})),\ \forall c \in c^*, i = 1, \ldots, \mathring{d}, t \in \{0, \mathring{t}\}$.
- *Select values* $\left\{{}^{\lfloor \mathring{c}}\Theta_c\right\}_{c \in c^*}$ *such that* $f\left(d_t\middle|\, d(t-1),\, {}^{\lfloor \mathring{c}}\Theta_c, c\right) = g(d(t)),\ \forall c \in c^*$, *with a positive finite value $g(d(t))$ independent of c.*
- *Set $c = 1$.*

Evaluation mode

```
Do while $c \le \mathring{c}$ and $\mathring{c} > 1$
    Set $l = \ln\left(f\left(\,{}^{\lfloor \mathring{c}}\Theta_c\middle|\, d(\mathring{t})\right)\right) - \ln\left(f\left(\,{}^{\lfloor \mathring{c}}\Theta_c\middle|\, d(0)\right)\right)$
    For   $i = 1, \ldots, \mathring{d}$
              $l = l + \ln\left(\mathcal{I}(V_{ic;0}, \nu_{ic;0})\right) - \ln\left(\mathcal{I}(V_{ic;\mathring{t}}, \nu_{ic;\mathring{t}})\right)$
    end   of the cycle over $i$
              If $l > 0$
                 Swap $c$ with $\mathring{c}$ and set $\mathring{c} = \mathring{c} - 1$,  i.e., cancel the component
              else
```

```
            Set $c = c + 1$
        end of the test on v-log-likelihood increase
end of the while cycle over c
```

Remark(s) 6.26

1. *Originally, components were cancelled using hypothesis testing with re-estimation after each trial cancelling. It worked but it was computationally expensive.*
2. *Intuitively, the components with very small estimates $\hat{\alpha}_{c;\mathring{t}}$ are candidates for cancelling. This check is made indirectly in the proposed algorithm as the increase of the statistics κ_c determining them is the same as the increase of statistics ν_{ic}; see Section 6.5. The presented solution respects, moreover, the values of pdfs defining the individual components. Consequently, the hard selection of the level defining the small values is avoided.*
3. *An alternative solution of the cancelling problem was proposed and tested in [157]. It estimates a background level by including a fixed flat pdf into the estimated mixture. Then, all components that have weights well below it are cancelled.*

Problem 6.17 (Predicted v-likelihood in model validation) *Both merging and cancelling exploits prediction of the v-likelihood. The quality of the prediction is strongly correlated with the quality of the model. It leads to the obvious idea: to exploit the discrepancy between predicted and evaluated v-likelihood for the model validation; see Section 6.7. The discrepancy can be potentially used in controlling the learning process, to be used as the main input in sequential stopping rules [127].*

Problem 6.18 (Removal of superfluous data items) *The final mixture found may model quantities that have, even indirectly, no influence on data determining quality markers of the problem, cf. Step 9 in Algorithm 5.2. Then, they should be removed completely from consideration. The adequate analysis of the graph describing nontrivial relationships among quantities is to be complemented by the tool set described in this text.*

6.7 Model validation

The complexity of the mixture estimation makes *model validation* a necessary part of the overall design of the p-system. A full-scale use of the p-system is a decisive validation test. However, a number of offline tests is needed in order to check the quality of its design while developing it. This aspect is the topic of this section.

Generally, we test the hypothesis $H \equiv$ estimated model is good. Tests differ in the specification of alternatives to this hypothesis. The validation art consists of a proper selection of alternatives that can compete with the tested hypothesis and can be evaluated with the limited computer power available.

Section 6.7.1 discusses the exploitation of additional externally supplied information on processed data for testing their homogeneity. The test serves for deciding whether or not to segment the data in subgroups before estimation. Learning results are inspected in Section 6.7.2. It refines the most usual tests that evaluate model quality on the subset of data unused in estimation.

The subsequent subsections represent samples of various heuristic inspections tried. They indicate how underdeveloped this area is and how many theoretical and algorithmic results are needed.

A nonstandard but quite efficient test based on comparing models estimated with different forgetting rates is in Section 6.7.3. Evaluation by designer is commented in Section 6.7.4 and links to the design part in Section 6.7.5.

6.7.1 Test of data homogeneity

Discussion on PCA in Section 6.2.1 indicates that learning is substantially simplified if the learning data are segmented into homogenous groups before processing. Such a segmentation is inspected in this subsection.

Often, raw data sets to be processed can be qualified by process experts. For instance, the expert distinguishes different classes of the input material and expects different behaviors of the o-system in their processing.

Here we consider an alternative situation when data segments are categorized as reflecting excellent, good, bad (possibly with a type of errors) behaviors. The categorization provides an important additional data item, say $e \in e^* \equiv \{1, \ldots, \mathring{e}\},\ \mathring{e} < \infty$, that is worth being exploited.

Some of the discussed labels might be directly measured. Then, they are simply discrete data. Even if they are nonmodelled, cf. Agreement 5.4, their presence in condition directly segments data. Markov-type factors are at our disposal when they are modelled and predicted. In both these cases, they do not differ from other data and require no special discussion.

Different situation arises when the data are classified <u>ex post</u>. In this case, a finer modelling is desirable as a reasonably managed system provides a few data sets with the label “bad” (errors are exceptional). They could get too small weight or could be completely overlooked when they are treated without taking into account their specific position. This danger is enhanced by the fact that such labels are constant over whole data blocks.

For the finer modelling of this case, we formulate adequate hypotheses. In their formulation, we assume that the standard and labelled data form blocks with their specific time counters.

$H_0 \equiv$ The observed difference in quality is caused by inseparable influence of external conditions and the way of operating. In technical terms, a single mixture describes the standard $d(\mathring{t}_s)$ as well as the labelled $d(\mathring{t}_e)$ data. Thus, for $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$,

$$f(d(\mathring{t})|H_0) = \int f(d(\mathring{t})|\Theta, H_0) f(\Theta|H_0)\, d\Theta. \tag{6.119}$$

The symbol H_0 in the conditions stresses that both the structure of a mixture and the prior pdf are chosen under this hypothesis.

$H_1 \equiv$ The observed difference in quality is caused by difference in operating. In technical terms, different mixtures should be used for the standard $d(\mathring{t}_s)$ and labelled $d(\mathring{t}_e)$ data, respectively. Thus, for $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$,

$$f(d(\mathring{t})|H_1) = \int f(d(\mathring{t}_s)|\Theta_s, H_1) f(\Theta_s, H_1)\, d\Theta_s \int f(d(\mathring{t}_e)|\Theta_e, H_1) f(\Theta_e|H_1)\, d\Theta_e. \tag{6.120}$$

The structure of the mixture and the prior parameter estimate for processing of $d(\mathring{t}_e)$ may differ from those used on $d(\mathring{t}_s)$. This fact is stressed by the indexes $_s$, $_e$.

With these elements and no prejudice, $f(H_0) = f(H_1)$, the Bayes rule provides the posterior pdf $f(H_0|d(\mathring{t}))$. The common model — hypothesis H_0 — can be accepted if this probability is high enough.

The conceptual testing algorithm that may test any suspicion on heterogeneity of data looks as follows.

Algorithm 6.26 (Test of data homogeneity)

1. *Run complete mixture estimations on the standard $d(\mathring{t}_s)$, labelled $d(\mathring{t}_e)$ and concatenated data $d(\mathring{t}) \equiv \big(d(\mathring{t}_s), d(\mathring{t}_e)\big)$.*
2. *Evaluate the corresponding v-likelihood*

$$\begin{aligned} f(d(\mathring{t}_s)|H_1) &= \int f(d(\mathring{t}_s)|\Theta_s, H_1) f(\Theta_s|H_1)\, d\Theta_s \\ f(d(\mathring{t}_e)|H_1) &= \int f(d(\mathring{t}_e)|\Theta_e, H_1) f(\Theta_e|H_1)\, d\Theta \\ f(d(\mathring{t})|H_0) &= \int f(d(\mathring{t}_s), d(\mathring{t}_e)|\Theta, H_0) f(\Theta|H_0)\, d\Theta. \end{aligned}$$

3. *Determine probability of the hypothesis H_0 that a single standard model should be used*

$$f(standard|d(\mathring{t})) \equiv f(H_0|d(\mathring{t})) = \frac{f(d(\mathring{t})|H_0)}{f(d(\mathring{t})|H_0) + f(d(\mathring{t}_s)|H_1) f(d(\mathring{t}_e)|H_1)}. \tag{6.121}$$

4. *Use the single model further on if $f(H_0|d(\mathring{t}))$ is close to one. Inspect the factors that were active on $d(\mathring{t}_e)$ as potentially dangerous.*
5. *Use both mixtures independently if $f(H_0|d(\mathring{t}))$ is close to zero. A danger connected with the situations labelled by e should be signaled whenever the model fitted to $d(\mathring{t}_e)$ makes better predictions than the model fitted to the standard data $d(\mathring{t}_s)$.*

Remark(s) 6.27

1. *Obviously, the excellent or bad outcomes of the o-system may be caused by conditions that are out of control of the operator. For instance, excellent output quality of the rolling may be attributed to the operator or to the excellent input quality of the rolled material. This dichotomy leads to the important warning.*
   Quality labels are insufficient to express the managing quality.
2. *It is worth stressing that the available labelling may help us in marking the factors active on bad data $d(\mathring{t}_e)$ as potentially dangerous even if inseparability of external conditions and quality of operating (hypothesis H_0) is accepted.*
3. *The inspected situation is often classified as learning with an imperfect teacher: the quality of the supplied labels is tested. This observation implies that we can use the same test also on segmentation recommended by an expert according to directly measured indicators. It will simply check whether the distinction supposed by the process is significant or not.*

6.7.2 Learning results

During the learning, basic assumptions used in the design of the p-system should be tested. The following questions have to be answered during the model validation; see Chapter 5.

- Can the o-system be described by a time invariant mixture model in practically all possible working conditions?
- Do learning data cover sufficiently all these working conditions?

Essentially, we are asking how good is the obtained model in extrapolation of the past to the future. In the offline mode, it can be tested by cutting the available data $d(\mathring{t})$ into *learning data* $d(\mathring{t}_l)$ and *validation data* $d(\mathring{t}_v)$. To check it, we assume that the test on labelled data, Algorithm 6.26, was performed. Thus, we are dealing with homogeneous data describable by a single mixture. For validation of learning results, we formulate hypotheses similar to those in Section 6.7.1 but inspecting the following validation aspect.

$H_0 \equiv$ All data — both learning and validation ones — $d(\mathring{t}) \equiv (d(\mathring{t}_l), d(\mathring{t}_v))$ are described by a single mixture model. The v-likelihood of this hypothesis is obtained by running of the quasi-Bayes or batch quasi-Bayes algorithms on all data

$$f(d(\mathring{t})|H_0) \equiv \int f(d(\mathring{t})|\Theta_l, H_0) f(\Theta_l|H_0)\, d\Theta_l. \tag{6.122}$$

In this case, both the structure of the model $f(d(\mathring{t})|\Theta_l, H_0)$ and the prior pdf used in learning phase $f(\Theta_l|H_0)$ are used also for the validation data $d(\mathring{t}_v)$.

$H_1 \equiv$ Learning data and validation data should be described by individual models. The v-likelihood of this hypothesis is obtained by independent runs of the quasi-Bayes or batch quasi-Bayes algorithms on both data collections giving

$$f(d(\mathring{t})|H_1) \equiv \int f(d(\mathring{t}_l)|\Theta_l, H_1) f(\Theta_l|H_1)\, d\Theta_l \times \int f(d(\mathring{t}_v)|\Theta_v, H_1) f(\Theta_v|H_1)\, d\Theta_v. \tag{6.123}$$

The structure of the used model $f(d(\mathring{t}_v)|\Theta_v, H_1)$ and the prior pdf $f(\Theta_v)$ for the validation data $d(\mathring{t}_v)$ may differ from those for learning data.

With these elements and no prejudice, $f(H_0) = f(H_1)$, the Bayes rule provides the posterior pdf $f(H_0|d(\mathring{t}))$. The learned model can be accepted as a good one if the posterior pf $f(H_0|d(\mathring{t}))$ is high enough. Otherwise, we have to search for reasons why the chosen model is not reliable enough. It gives the algorithmic solution that is formally identical with Algorithm 6.26 but with the processed data segmented in a different way.

Algorithm 6.27 (Model validation on homogenous data)

1. *Run complete mixture estimations on learning* $d(\mathring{t}_l)$, *validation* $d(\mathring{t}_v)$ *and full* $d(\mathring{t}) \equiv (d(\mathring{t}_l), d(\mathring{t}_v))$ *data.*
2. *Evaluate the corresponding v-likelihood values* $f\left(d(\mathring{t}_l)|H_1\right)$, $f\left(d(\mathring{t}_v)|H_1\right)$, $f\left(d(\mathring{t})|H_0\right)$.
3. *Determine the probability of successful learning*

$$f\left(success|d(\mathring{t})\right) \equiv f\left(H_0|d(\mathring{t})\right) = \frac{f\left(d(\mathring{t})|H_0\right)}{f\left(d(\mathring{t})|H_0\right) + f\left(d(\mathring{t}_l)|H_1\right) f\left(d(\mathring{t}_v)|H_1\right)}. \tag{6.124}$$

4. *The test is successfully passed if* $f(H_0|d(\mathring{t}))$ *is close to 1. Otherwise, measures for a better learning have to be taken.*

Results of the test depend, often strongly, on the way how the available data are cut into learning and validation parts. Surprisingly, this aspect is rarely discussed for dynamic systems. The extensive available results are almost exclusively focused on static problems [124].

Here, we inspect the choice of the *cutting moments* t_u in the dynamic case. The cutting should be obviously restricted to sufficiently long subsequences of consecutive records. The proper length of these subsequences is, however, unknown. Thus, it makes sense to validate learning for various cutting moments $t_u \in t_u^* \subset t^*$. The *cutting moment* controls the validation test as it identifies $d_{l;t} \equiv d_t$ for $t \le t_u$ and $d_{v;t} = d_t$ for $t > t_u$.

We are making a pair of decisions $(\hat{H}, t_u)$ based on the experience $\mathcal{P} \equiv d(\mathring{t})$. We select $t_u \in t_u^*$ and accept $(\hat{H} = H_0)$ that the learned model is valid or reject it ($\hat{H} = H_1$). We assume for simplicity that the losses caused by a wrong acceptance and rejection are identical, say $z > 0$.

We solve this static decision task and select the optimal decision ${}^{\lfloor o}\hat{H}$ on inspected hypotheses and optimal cutting time moment ${}^{\lfloor o}t_u$ as a minimizer of the expected loss

$$({}^{\lfloor o}\hat{H},\, {}^{\lfloor o}t_u) \in \mathrm{Arg} \min_{\hat{H} \in \{H_0, H_1\}, t_u \in t_u^*} \mathcal{E}\left[(1 - \delta_{\hat{H},H})z\right]. \tag{6.125}$$

Proposition 6.19 (Optimal cutting) *Let $0, \mathring{t} \in t_u^*$. Then, the optimal decision ${}^{\lfloor o}\hat{H}$ about the inspected hypotheses H_0, H_1 and the optimal cutting ${}^{\lfloor o}t_u$, that minimize the expected loss in (6.125) are given by the following rule*

$$\begin{aligned}
&\textit{Compute } {}^{\lfloor 0}t_u \in \mathrm{Arg}\max_{t \in t_u^*} f(H_0|d(\mathring{t}), t_u)\\
&\qquad\qquad {}^{\lfloor 1}t_u \in \mathrm{Arg}\min_{t \in t_u^*} f(H_0|d(\mathring{t}), t_u)\\
&\textit{Select}\, {}^{\lfloor o}\hat{H} = H_0,\ {}^{\lfloor o}t_u = {}^{\lfloor 0}t_u \textit{ if}\\
&\qquad f(H_0|d(\mathring{t}),\, {}^{\lfloor 0}t_u) \geq 1 - f(H_0|d(\mathring{t}),\, {}^{\lfloor 1}t_u)\\
&\quad {}^{\lfloor o}\hat{H} = H_1,\ \ {}^{\lfloor o}t_u = {}^{\lfloor 1}t_u \textit{ if}\\
&\qquad f(H_0|d(\mathring{t}),\, {}^{\lfloor 0}t_u) < 1 - f(H_0|d(\mathring{t}),\, {}^{\lfloor 1}t_u).
\end{aligned} \tag{6.126}$$

Proof. Let us take the cutting moments ${}^{\lfloor 0}t_u^* \equiv \left\{\tau \in t_u^* : f(H_0|d(\mathring{t}), t_u) \geq 0.5\right\}$. This finite set is nonempty, as for $t_u = 0$ $f(H_0|d(\mathring{t}), t_u) = 0.5$. For a fixed $t_u \in {}^{\lfloor 0}t_u^*$, the decision $\hat{H} = H_0$ leads to a smaller loss than the decision $\hat{H} = H_1$. The achieved minimum is expectation over $d(\mathring{t})$ of $1 - f(H_0|d(\mathring{t}),\, {}^{\lfloor 0}t_u)$. Thus, it is smallest for ${}^{\lfloor 0}t$ maximizing $f(H_0|d(\mathring{t}), t_u)$ on ${}^{\lfloor 0}t_u^*$.

For any fixed t_u in the set ${}^{\lfloor 1}t_u^* \equiv \{t_u \in t_u^* : f(H_0|d(\mathring{t}), t_u) \leq 0.5\}$, the decision $\hat{H} = H_1$ leads to a smaller loss than the decision $\hat{H} = H_0$. The achieved minimum is expectation over $d(\mathring{t})$ of $f(H_0|d(\mathring{t}), t_u)$. Thus, it is smallest for ${}^{\lfloor 1}t_u$ minimizing $f(H_0|d(\mathring{t}), t_u)$ on ${}^{\lfloor 1}\tau^*$. The smaller of the discussed pairs of minima determines the optimal decision pair. □

Remark(s) 6.28

1. *The choice of the prior pdf for estimation on the validation data is critical for a good performance of the test. The use of the posterior pdf obtained from the validation data and flattened as in branching (see Section 6.4.3) seems to be satisfactory.*
2. *A practical application of the above test depends strongly on the set t_u^* of the allowed cutting moments. The finest possible choice $t_u^* = t^*$. An exhaustive search is too demanding for the considered extensive data sets. A search for the minimizer by a version of the golden-cut rule, by a random choice or by systematic inspection on a proper subset on a small predefined grid can be applied. The predefined grid seems to be the simplest variant as minor changes in t_u make no physical sense.*

3. *Learning and validation data have to overlap in dynamic case in order to fill the richest regression vector occurring in the mixture. This fine point is neglected in analysis as it has negligible influence on the test performance.*

6.7.3 Forgetting-based validation

This subsection presents a relatively successful sample of many heuristic attempts to find efficient validation tests.

Forgetting techniques are proposed to cope with slow variations of parameters caused either by real changes or by under-modelling. It offers the following "natural" test of validity of the fixed estimated model.

The tested model is taken as an alternative in estimations with stabilized forgetting, Section 3.1, run in parallel for several forgetting factors. If the tested model is good then it can be expected that the additional adaptation brings no new quality in the description of data. Thus, the approximate learning is expected to describe the observed data the better, the higher weight is given to the valid tested model taken as an alternative in stabilized forgetting. It immediately gives the following validation algorithm.

Algorithm 6.28 (Forgetting-based validation)

Initial mode

- *Estimate completely the mixture model, i.e., both parameters and structure and denote the resulting pdf* $f(\Theta|d(\mathring{t}))$.
- *Apply flattening to* $f(\Theta|d(\mathring{t}))$ *in the branching version, Section 6.4.3, so that a good prior pdf* $f(\Theta)$ *is obtained.*
- *Select several forgetting factors* $0 \approx \lambda_1 < \lambda_2 < \cdots < \lambda_{\mathring{i}-1} < \lambda_{\mathring{i}} = 1$, $1 < \mathring{i} < \infty$.
- *Set prior pdfs* $f(\Theta|\lambda_i) = f(\Theta)$.

Validation mode

1. *Perform estimation with the stabilized forgetting, Section 3.1, for all* λ_i, *using the tested prior pdf* $f(\Theta|d(\mathring{t}))$ *as the alternative.*
2. *Evaluate values of the v-likelihood* $l_{i;\mathring{t}} = f(d(\mathring{t})|\lambda_i)$ *and compute MAP estimate* $\hat{\lambda}$ *of* λ.
3. *Take the model as a successful one if* $\lambda_1 = \hat{\lambda}$. *Otherwise search for its improvements.*

Remark(s) 6.29

Often, three alternative forgetting factors are sufficient for a good testing.

Problem 6.19 (Analysis of forgetting-based validation) *Algorithm 6.19 is intuitively appealing and practical experience with it is good. Appropriate insight is, however, missing and a deeper analysis is highly desirable.*

6.7.4 Inspection by a human designer

Complete visual inspections are mostly excluded due to the excessive dimensionality of data records. A partial visual inspection is possible and useful especially when the influence of various low-dimensional factors on the resulting quality is inspected. This situation happens repeatedly when various tools for designing the advisory system are being constructed or when these tools are tailored to a specific managed system. Human beings are able to grasp relationships that are hard to find in an algorithmic way. At the same time, this ability is restricted to 2 or 3 dimensional spaces. Proposition 7.2 shows how to get such low-dimensional projections in a computationally cheap way.

6.7.5 Operating modes

A stable model does not guarantee that we get efficient advisory system. As discussed in Section 5.1.2, the p-system makes sense if there are good and bad modes of operating that are reflected in the observation space of the p-system. These conditions have to be tested, too.

The distance of the observable behavior of the o-system to the user's ideal pdf ${}^{\lfloor U}f$ is measured by the KL divergence. Recall that we extended the true user's ideal pdf on the full data space of the p-system (5.7). The KL divergence has the form

$$\begin{aligned}\mathcal{D}\left(f \middle|\middle| {}^{\lfloor U}f\right) &= \int f(d(\mathring{t})) \ln\left(\frac{f(d(\mathring{t}))}{{}^{\lfloor U}f(d(\mathring{t}))}\right) dd(\mathring{t}) \\ &= \sum_{t\in t^*} \mathcal{E}\left[\int f(d_t|d(t-1)) \ln\left(\frac{f(d_t|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))}\right) dd_t\right].\end{aligned} \tag{6.127}$$

Let us assume, that $f(d_t|d(t-1)) = \sum_{c\in c^*} \alpha_c f(d_t|d(t-1), c)$ is a known, i.e., well estimated, mixture. Then, the Jensen inequality (2.14) and the inequality between weighted arithmetic and geometric means imply

$$\begin{aligned}&\mathcal{D}\left(f \middle|\middle| {}^{\lfloor U}f\right) \\ &\leq \sum_{c\in c^*} \alpha_c \sum_{t\in t^*} \mathcal{E}\left[\int f(d_t|d(t-1), c) \ln\left(\frac{f(d_t|d(t-1), c)}{{}^{\lfloor U}f(d_t|d(t-1))}\right) dd_t\right] \\ &\mathcal{D}\left(f \middle|\middle| {}^{\lfloor U}f\right) \\ &\geq \sum_{c\in c^*} \sum_{\tilde{c}\in c^*} \alpha_c \alpha_{\tilde{c}} \sum_{t\in t^*} \mathcal{E}\left[\int f(d_t|d(t-1), \tilde{c}) \ln\left(\frac{f(d_t|d(t-1), c)}{{}^{\lfloor U}f(d_t|d(t-1))}\right) dd_t\right].\end{aligned} \tag{6.128}$$

The inequalities (6.128) provide lower and upper bounds on the KL divergence to be optimized. The upper bound is determined by the KL divergences of individual components to the user's ideal pdf ${}^{\lfloor U}f$. For the weights α with a dominant entry $\alpha_c \approx 1$, the same distance dominates the lower bound. For

$\alpha_c = 1$, inequalities reduce to equalities. It supports an intuitive definition of good (bad) modes as components having small (large) KL divergence to ${}^{\lfloor U}f$. Obviously, advising has a chance to be successful if there is a component, with non-negligible weight, whose KL divergence to the user's ideal pdf is (significantly) smaller than that of other components with non-negligible weights. It makes the distribution of pairs (component weight, component KL divergence to ${}^{\lfloor U}f$) a significant indicator of the potential design success.

Problem 6.20 (Systematic model validation) *Above, various ideas on the model validation are outlined. Experimental results as well as the diversity of techniques proposed indicate that we are still dealing more with a bag of tricks than with a systematic approach. This status should be improved.*

We have at our disposal a wide supply of indicators when addressing the problem.

In learning, for instance, structure estimates should be stable over various data sets and no branching, merging or cancelling of components should improve the value of the v-likelihood, etc.

In design, for instance, prediction of quality markers, Section 5.1.4, as well as prediction of recognizable operator actions, Agreement 5.7, are good partial tests.

Also, having in mind the warning that the o-system has to be judged as a mapping of external conditions on the quality of results (see Section 6.7.1), we can proceed as follows.

- *Split available data on those with good and bad management results.*
- *Build independent mixture models on them.*
- *Judge whether the operator has a tendency to improve or spoil the quality of the overall behavior.*
- *Take operating modes leading to improvement (deterioration) as good (bad) ones.*
- *Search for similarity of good (bad) modes obtained on both data files.*

Remark(s) 6.30

Problem 6.20 is quite hard. Meanwhile, the collection of and stabilization of intuitively acceptable indicators like

- *comparison of data moments predicted by the estimated model with their sample counterparts*
- *analysis of components as dynamic mappings (structure of eigenvalues)*
- *comparison of theoretical and measured properties (whiteness of prediction errors)*
- *. . .*

are invaluable.

7

Solution and principles of its approximation: design part

Proposition 2.11 on fully probabilistic design, filled by elements described in Chapter 5, provides a complete formal solution of the design of the advisory system. It specifies the evaluation structure but cannot be practically used. The exact analytical solution is not available and the brute force numerical approach is inhibited by the "curse of dimensionality". Thus, the formal solution has to be complemented by approximate feasible evaluations. For the learning part of the advisory system, this is done in Chapter 6. The discussion related to the design part is discussed here.

The design is based on the model of the o-system. The observed behavior, tailored to the rate of operator actions, is described by the mixture (5.9)

$$f(d_t|\phi_{t-1},\Theta) = \sum_{c\in c^*} \alpha_c f(d_t|\phi_{c;t-1},\Theta_c,c).$$

Each component $f(d_t|\phi_{c;t-1},\Theta_c,c)$ is decomposed into the product of factors $f(d_{ic;t}|\psi_{ic;t},\Theta_{ic},c)$ (see Agreement 5.4)

$$f(d_t|\phi_{c;t-1},\Theta_c,c) = \prod_{i\in i^*} f(d_{ic;t}|\psi_{ic;t},\Theta_{ic},c),\ i^* \equiv \{1,\ldots,\mathring{d}\}.$$

The factors are parameterized by individual parameters Θ_{ic}. The collection of these parameters, together with probabilistic weights $\alpha_c \in \alpha^*$ (5.9), form the multivariate parameter Θ of the mixture. The structure of the mixture as well as this parameter are assumed to be known throughout this chapter. For the fixed advisory system, their reliable point estimates are supposed to be obtained during the offline learning phase. For the adaptive advisory system, the initial offline estimates are permanently corrected. In both cases, the certainty-equivalence strategy, Section 4.2.2, is adopted. Formally, parameters are taken as known quantities and thus can be omitted in conditions.

As outlined in Chapter 5, the estimated model of the unguided o-system provides the basic building blocks for creating the advising mixture that combines them with optional elements. The optional advising elements are optimized so that the resulting mixture is as close as possible to the user's ideal pdf. The result is then offered to the operator as the target to be followed. The design is performed in the fully probabilistic sense (Section 2.4.2), i.e.; proximity of the involved pdfs is judged via the KL divergence.

The chapter is organized as follows. Section 7.1 discusses common design tools like evaluation of the stationary behavior of the o-system and projections of the mixture model to low-dimensional spaces as well as construction of approximate predictors. The model, relating advices to responses of the o-system, is reviewed in Section 7.1.3. This model is restricted to the case of the fully cooperating operator. Then, the design part is prepared. Recursive formulas for the optimized KL divergence and its upper bounds of a Jensen type (2.14) are prepared in Sections 7.1.4 and 7.1.5. These feasible bounds contain no logarithms of mixtures or their ratios and serve as the optimized loss functions. The effort spent on the approximation instead of a direct invention of feasible loss functions pays back: significant relationships between various elements are “discovered”. This is most clear with the design of interactions with the operator in the academic case. Directly chosen feasible loss functions have failed to judge the presentation quality whenever components of the estimated and ideal models are identical.

The design of advising strategies is addressed in Section 7.2. Academic, industrial and simultaneous cases are distinguished; Agreements 5.6, 5.7. The presentation of advices and evaluation of the overall state are addressed in Section 7.3. Section 7.4 outlines validation of the design.

The basic evaluations performed in this chapter are simple. However, their description is sometimes cumbersome due to the use of integral expressions for expectations encountered. We prefer this way as it clearly distinguishes quantities in conditions, quantities integrated out and those optimized.

7.1 Common tools

Here, the tools used throughout this chapter are prepared.

7.1.1 Model projections in design

Let ϕ be the richest state vector of the considered mixture. For its given initial value, the time invariant mixture model defines a complete distribution of data $d(\mathring{t})$. Marginal and conditional pdfs derived from it, called *projections*, are used in analysis, design and exploitation of the p-system. They are discussed here.

Steady-state pdfs

Analysis of the steady-state behavior of individual, permanently active, components serves both for the model validation, Section 6.7, and approximate design, Section 7.2. In both cases, the steady-state moments of the quality marker are primarily judged. The choice of the marker depends on the specific application. The evaluation of the full steady state pdf described below postpones the selection of a specific marker and thus preserves the freedom in its choice.

Proposition 7.1 (Steady-state component)
Let a component be described by the pdf

$$f(d_t|d(t-1)) = f(d_t|d_{t-1},\ldots,d_{t-\partial}) = f(d_t|\phi_{t-1}),\ 1 \le \partial < \infty, \tag{7.1}$$

i.e., it has the state ϕ_{t-1} in the phase form $\phi'_{t-1} = [d'_{t-1},\ldots,d'_{t-\partial},1]$, Agreement 5.4. Let the steady state pdf $f_\infty(\phi) \equiv \lim_{t\to\infty} f(\phi_t = \phi)$ *exist on the domain $\phi^* \equiv \{[d'_{t-1} = \phi'_1,\ldots,d'_{t-\partial} = \phi'_\partial]'\}, \phi_k \in d^*,\ k \in \{1,\ldots,\partial\}$. Then, it solves the equation*

$$\begin{aligned} f_\infty\ &(\phi_1,\ldots,\phi_\partial) \\ &= \int f\left(d_t = \phi_1|\phi_{t-1} = \left[\phi_2,\ldots,\phi_\partial,\tilde{d}\right]\right) f_\infty\left(\phi_2,\ldots,\phi_\partial,\tilde{d}\right)\, d\tilde{d}. \end{aligned} \tag{7.2}$$

Proof. It is directly implied by the marginalization, the chain rule for pdfs, Proposition 2.4, and by the phase form of the state. □

Remark(s) 7.1

1. *Proposition 7.1 provides the necessary but not sufficient condition for the existence of the steady-state pdf. Thus, the solution of (7.2) has to be checked whether it represents the steady-state pdf f_∞. Also, uniqueness is not guaranteed.*
2. *For static components, $\partial = 0$, the steady-state pdf f_∞ coincides with the pdf describing the component. This makes the evaluation of individual static components trivial. It also indicates their descriptive weakness.*

Marginal and conditional pdfs

The important question is how to present the design results to the operator in a comprehensible manner. The probabilistic description of data by conditional pdfs offers an efficient presentation tool prepared here.

In the industrial case, Agreement 5.6, the advisory system directly provides the ideal randomized control strategy for the choice of the recognizable actions $u_{o;t}$; see Section 5.4.5. If the recognizable actions have a high dimension, marginal pdfs of the most important recognizable actions, conditioned on the known past data, are presented to the operator. Then, it is necessary to specify preference among them: either automatically, Section 7.3.1, or manually.

Similarly, in high-dimensional cases, only selected entries of the o-innovations $\mathit{\Delta}_{o;t}$, predicted by the optimized ideal pdf, can be shown to the operator in order to influence his actions.

In both cases, predictions by the mixture model of the o-system or by the ideal mixture model have to be done, and their relevant, low-dimensional, marginal pdfs presented. This makes conditioning and marginalization the universal operations needed for exploiting the mixture-based predictors.

Proposition 7.2 (Marginal and conditional pdfs of the mixture model) *Let the joint pdf of data $d(\mathring{t})$ be described by the known mixture model in the factorized form*

$$
\begin{aligned}
f(d_t|d(t-1)) &= \sum_{c \in c^*} \alpha_c \prod_{\iota \in \iota^*} f(d_{\iota c;t}|d_{(\iota+1)\cdots\mathring{\imath}c;t}, \phi_{c;t-1}, c) \\
&\equiv \sum_{c \in c^*} \alpha_c \prod_{i \in i^*} f(d_{ic;t}|\psi_{ic;t}, c),
\end{aligned}
$$

where $d_{ic;t}$ are permuted entries of d_t; $\phi_{c;t-1}$ are observable states of the individual components and $\psi_{ic;t}$ are regression vectors of respective factors.

Let $d_{\iota c;t}$, $\iota \in \iota^* \equiv \{\iota_1, \ldots, \iota_{\mathring{\iota}}\} \subset \{1, \ldots, \mathring{d}\} \equiv \iota^* \cup \overline{\iota^*}$, $\iota^* \cap \overline{\iota^*} = \emptyset$ be selected entries of $d_{c;t}$. Then, the predictor of these entries is the mixture

$$
f(d_{\iota_1\ldots\mathring{\iota}c;t}|d(t-1)) = \sum_{c \in c^*} \alpha_c \int \prod_{\iota \in \iota^*} f(d_{\iota c;t}|d_{(\iota+1)\cdots\mathring{\imath}c;t}, \phi_{c;t-1}, c)\, dd_{k \in \overline{\iota^*} c;t}, \tag{7.3}
$$

where the integration is performed over the nonpredicted entries $d_{\overline{\iota^*} c;t} \equiv \{d_{kc;t} : k \in \overline{\iota^*}\}$.

All factors predicting $d_{ic;t}$ with $i \in \overline{\iota^*}$ and $i < \min\{\iota \in \iota^*\}$ integrate to unity in (7.3). The remaining nonpredicted quantities have to be integrated out explicitly.

Let $\iota^* = \beta^* \cup \gamma^*$, $\beta^* \cap \gamma^* = \emptyset$. If $\beta^* \neq \emptyset$, $\gamma^* \neq \emptyset$, then, the predictor of the data entries $d_{\beta \in \beta^*;t}$, conditioned on the past data $d(t-1)$ and the data entries $d_{\gamma \in \gamma^*;t}$, is the ratio of mixtures

$$
\begin{aligned}
f(d_{\beta_{1\ldots\mathring{\beta}}c;t}|\ & d_{\gamma_1\ldots\mathring{\gamma}c;t}, d(t-1)) \\
&= \frac{\sum_{c \in c^*} \alpha_c \int \prod_{\iota \in \iota^*} f(d_{\iota c;t}|d_{(\iota+1)\cdots\mathring{\imath}c;t}, \phi_{c;t-1}, c)\, dd_{\overline{\iota^*} c;t}}{\sum_{c \in c^*} \alpha_c \int \prod_{\iota \in \iota^*} f(d_{\iota c;t}|d_{(\iota+1)\cdots\mathring{\imath}c;t}, \phi_{c;t-1}, c)\, dd_{\overline{\iota^* \cup \beta^* c;t}} c;t}.
\end{aligned} \tag{7.4}
$$

Proof. It is directly implied by the marginalization and the chain rules. □

Remark(s) 7.2

1. *Sometimes, the integrations in (7.3) or (7.4) can be approximated by inserting expectations of the appropriate quantities instead of the values to be integrated out.*
2. *The rational form (7.4) shows clearly the need to use predictors of the quantities with indexes in γ^* even if they are used in the condition only.*

3. *The condition used for predicting the presented quantities contains measured values, the values proposed by the p-system and values contemplated by the operator. The latter supports a question-and-answer mode of advising. The operator can inspect what happens if he chooses a specific value of the selected quantity. He also can select an appropriate value of a suitable quantity in order to move other quantities to a desired range.*

7.1.2 Dynamic predictors in advising

Long-horizon predictors

Generalized Bayesian estimation, Proposition 2.13, evaluates the posterior pdf $f(\Theta|d(\mathring{t}))$ on unknown parameters Θ. For the fixed advisory system with the pdf $f(\Theta|d(0))$ obtained on learning data $d(0)$, the parameter estimate serves for constructing fixed predictors $f(d_t|d(t-1)) = \int f(d_t|d(t-1),\Theta)f(\Theta|d(0))\,d\Theta$. They are used in the design of advices and applied to new data d_t measured in online mode. We assume that under-modelling is negligible and data are informative enough. Then, the predictors are approximated $f(d_t|d(t-1)) \approx f(d_t|d(t-1),\hat{\Theta})$, i.e., a point estimate $\hat{\Theta}$ of Θ is inserted into the parameterized model $f(d_t|d(t-1),\Theta)$.

During learning, we usually process normal operation records. Thus, information content of data may be poor. It does not harm the quality of short-horizon predictors but may produce rather bad long-horizon predictors. It is known that the ability to make long-term predictions significantly influences the quality of the design. The situation when the predictor behaves as an *unstable* dynamic mapping is the worst and the most visible case of a bad prediction, and has to be avoided. Recall that the predictor is unstable if it has no steady-state pdf.

We have to face this problem that may occur even if all measures increasing the information content of data and decreasing inevitable under-modelling are exhausted. A Bayesian solution is outlined here.

Let $\Theta_s^* \subset \Theta^*$ be set of "reasonable" parameters. Typically, Θ_s^* describes stable mappings. Prior information that parameters are expected to be "reasonable" implies that support of the prior pdf $f(\Theta)$ should be reduced from Θ^* to Θ_s^*. It does not influence evaluation of the likelihood and the "restricted" posterior pdf becomes simply proportional to $f(\Theta|d(\mathring{t}))\chi_{\Theta_s^*}(\Theta)$, where $\chi_{\Theta_s^*}(\cdot)$ is an indicator of the set Θ_s^*. The corresponding expected value $\hat{\Theta}_s \equiv \mathcal{E}\left[\Theta|d(\mathring{t})\right]$ is then inserted into the parameterized model. Thus, the fixed predictor gets the form $f(d_t|d(t-1),\hat{\Theta}_s)$. Evaluation of $\hat{\Theta}_s$ is the only trouble faced. Mostly, it must be done numerically. A straightforward use of Monte Carlo is often sufficient (for an exception and its treatment see Section 9.1.2). The following algorithm projects estimates on a "reasonable" set Θ_s^*.

Algorithm 7.1 (Choice of dynamic predictors)

Initial (offline) mode

- *Estimate the mixture model of the o-system; Chapter 6.*
- *Specify the set* $\Theta_s^* \subset \Theta^*$ *determining meaningful dynamic predictors, for instance, stable predictors or predictors having a given static gain.*
- *Test whether the unrestricted point estimate* $\hat{\Theta}$ *of* Θ *belongs to* Θ_s^*.
- *Set* $\hat{\Theta}_s = \hat{\Theta}$ *if* $\hat{\Theta} \in \Theta_s^*$ *and stop. Otherwise, continue.*
- *Select the upper bound* $\mathring{n}$ *on the number of iterations* n *and set* $n = 0$.
- *Select the required number* $\mathring{m}$ *of independent Monte Carlo samples to be observed in* Θ_s^* *in order to get reliable estimate* $\hat{\Theta}_s$ *and set* $m = 0$.
- *Set* $\hat{\Theta}_s = 0$*; the zero matrix* 0 *has the same dimensions as parameter* Θ.

Iterative mode

1. *Do while* $n < \mathring{n}$ & $m < \mathring{m}$.
2. *Set* $n = n + 1$.
3. *Take an independent sample* $\tilde{\Theta} \sim f\left(\Theta | d(\mathring{t})\right)$.
4. *Go to Step 1 if* $\tilde{\Theta} \notin \Theta_s^*$.
5. *Set* $\tilde{m} = m + 1$ *and* $\tilde{\Theta}_s = \hat{\Theta}_s + \frac{1}{\tilde{m}}(\hat{\Theta}_s - \tilde{\Theta})$.
6. *Set* $\hat{\Theta}_s = \tilde{\Theta}_s$ *and* $m = \tilde{m}$ *if* $\tilde{\Theta}_s \in \Theta_s^*$, *continue.*
7. *Go to Step 1.*

Accept the value $\hat{\Theta}_s$ *as the final estimate if* $m = \mathring{m}$*; otherwise take the evaluation as unsuccessful.*

Remark(s) 7.3

1. *The stopping when* $\hat{\Theta} \in \Theta_s^*$ *is justified by the narrow support of the posterior pdf that is expected after processing large number of data. This choice should not be applied whenever there are doubts in this respect.*
2. *The algorithm relies on the ability to efficiently determine whether a sample* $\tilde{\Theta}$ *belongs to* Θ_s^*.
3. *The exact Bayesian estimation is not influenced by the set* Θ_s^*. *The approximate estimation (see Section 6.5) is generally influenced by it as the set* Θ_s^* *influences one-step-ahead predictors used for weighting of the processed data. We conjecture that this influence can be neglected if one-step-ahead predictions are good. This conjecture is supported by experimental results.*
4. *The test whether the updated sample mean* $\hat{\Theta}_s$ *belongs to* Θ_s^* *is generally necessary as* Θ_s^* *need not be a convex set.*
5. *The portion of Monte Carlo samples falling into* Θ_s^* *may be rather small. Then, the way thay are generated has to be modified. In Section 9.1.2, an instance of this situation is described.*

Approximation of dynamic predictors

Discussion in Section 5.3 has shown that dynamic mixtures with data-dependent weights cannot be yet efficiently used. In the same section, a universal approach has been proposed that overcomes drawbacks of mixtures with constant weights. This solution requires, however, at least n-times (grouping rate $n > 1$; see Proposition 5.3) more evaluations than the standard estimation. Here, a generally applicable approximation is proposed that covers cases when a component is active for relatively long time intervals.

In Section 5.3, the data-dependent component weights $\tilde{\alpha}_c(d(t-1), \Theta)$ are interpreted through a projection and approximation of a more complex parameterized model. Alternatively, $\tilde{\alpha}_c(d(t-1), \Theta)$ can be seen as the probability that the data vector Ψ_t, Agreement 5.4, belongs to a set Ψ_c^* on which the cth component is the best approximation of the "objective" system description, cf.; Chapter 2. It is also the probability that the pointer $c_t = c$. This probability is evaluated during the approximate estimation and found to be proportional to $\alpha_c f(d_t|d(t-1), c)$, cf.; Section 6.5. Thus, we know it after observing Ψ_t.

For the discussed case, which properly models a rich class of applications, we can and will assume that

$$\Psi_t \in \Psi_c^* \text{ with a high probability if } \Psi_{t-1} \in \Psi_c^*. \tag{7.5}$$

This assumption leads to the following approximate dynamic predictor

$$f(d_t|d(t-1)) = \sum_{c \in c^*} w_c(d(t-1)) f(d_t|d(t-1), c) \quad \text{with} \tag{7.6}$$

$$w_c(d(t-1)) \equiv \frac{\alpha_c f(d_{t-1}|d(t-2), c)}{\sum_{\tilde{c} \in c^*} \alpha_{\tilde{c}} f(d_{t-1}|d(t-2), \tilde{c})}.$$

This is a simple and efficient dynamic predictor that, however, can be successful only when condition (7.5) is met.

7.1.3 Advices and their influence

We repeat the classification of actions of the p-system with an extended presentation of the adopted models of their influence; see Chapter 5. Recall that the superscript $^{\lfloor I}$ distinguishes the optimized elements.

Agreement 7.1 (Advices and their influence) *Advices, i.e., the actions of the p-system, are*

$$a_{p;t} \equiv (c_t, u_{o;t}, z_t, s_t). \tag{7.7}$$

The individual entries have the following interpretation.

- *Recommended pointers* $c_t \in c^*$ *to active components are actions in the academic design. With them, the optimized ideal pdf is*
$$^{\lfloor I}f(d_t, c_t|d(t-1)) \equiv f(d_t|d(t-1), c_t)\, ^{\lfloor I}f(c_t|d(t-1)). \tag{7.8}$$
In (7.8), the pdfs
$$\left\{ f(d_t|d(t-1), c_t) \equiv f(d_t|d(t-1), \hat{\Theta}_{c_t}, c_t) \right\}_{t \in t^*}$$
describe the c_t*th component of the estimated o-model. The point estimate* $\hat{\Theta}$*, based on offline data, replaces the unknown component parameters* Θ_c *in the parameterized component model.The collection of pdfs* $\left\{ ^{\lfloor I}f(c_t|d(t-1)) \right\}_{t \in t^*}$ *describes the optimized academic strategy generating the recommended pointers* $\{c_t\}_{t \in t^*}$*. The ideal pdf communicated to the operator has the form*
$$^{\lfloor I}f(d_{o;t}|d(t-1)) = \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|d(t-1)) f(d_{o;t}|d(t-1), c_t), \;\; t \in t^*. \tag{7.9}$$
The pdf $f(d_{o;t}|d(t-1), c_t)$ *is the marginal pdf of the joint* $f(d_t|d(t-1), c_t)$ *reduced on the o-data* $d_{o;t}$*.*
- *Recommended recognizable actions* $u_{o;t} \in u_o^*$ *guide the operator in selecting recognizable actions. The advising strategy* $\left\{ ^{\lfloor I}f(u_{o;t}|d(t-1)) \right\}_{t \in t^*}$ *influences the constructed ideal differently for industrial and simultaneous designs.*
  - *Industrial design assumes that the component weights* $f(c_t|d(t-1))$ *are fixed either at learned values* α_{c_t} *or at values* $^{\lfloor I}f(c_t|d(t-1))$ *resulting in the preceding academic design. The optimized ideal pdf is*
$$\begin{aligned} ^{\lfloor I}f\;\; & (d_t|d(t-1)) \\ & \equiv {}^{\lfloor I}f(u_{o;t}|d(t-1)) f(\Delta_t|u_{o;t}, d(t-1)) = {}^{\lfloor I}f(u_{o;t}|d(t-1)) \\ & \times \frac{\sum_{c_t \in c^*} f(c_t|d(t-1)) f(\Delta_t|u_{o;t}, d(t-1), c_t) f(u_{o;t}|d(t-1), c_t)}{\sum_{c_t \in c^*} f(c_t|d(t-1)) f(u_{o;t}|d(t-1), c_t)}. \end{aligned} \tag{7.10}$$
The pdfs $f(\Delta_t|u_{o;t}, d(t-1), c)$*,* $f(u_{o;t}|d(t-1), c)$ *are marginal pdfs of the estimated cth component. The ideal pdf communicated to the operator is then the marginal pdf reduced on the o-data*
$$\begin{aligned} & ^{\lfloor I}f(d_{o;t}|d(t-1)) = {}^{\lfloor I}f(u_{o;t}|d(t-1)) \\ & \times \frac{\sum_{c_t \in c^*} f(c_t|d(t-1)) f(\Delta_{o;t}|u_{o;t}, d(t-1), c_t) f(u_{o;t}|d(t-1), c_t)}{\sum_{c_t \in c^*} f(c_t|d(t-1)) f(u_{o;t}|d(t-1), c_t)}. \end{aligned} \tag{7.11}$$
  - *Simultaneous design selects the joint strategy* $\left\{ ^{\lfloor I}f(c_t, u_{o;t}|d(t-1)) \right\}_{t \in t^*}$*. The joint optimization allows us to make the recommended recognizable actions* $u_{o;t}$ *dependent on the recommended pointer* c_t*. It simplifies the optimized model to*

$$^{\lfloor I}f(d_t, c_t|d(t-1)) = f(\mathit{\Delta}_t|u_{o;t}, d(t-1), c_t)\, ^{\lfloor I}f(c_t, u_{o;t}|d(t-1)). \quad (7.12)$$

The conditional pdf $f(\mathit{\Delta}_t|u_{o;t}, d(t-1), c_t)$ is derived from the cth estimated component $f(d_t|d(t-1), c_t)$. The corresponding ideal pdf communicated to the operator becomes

$$^{\lfloor I}f(d_{o;t}|d(t-1)) = \sum_{c_t \in c^*} f(\mathit{\Delta}_{o;t}|u_{o;t}, d(t-1), c_t)\, ^{\lfloor I}f(c_t, u_{o;t}|d(t-1)). \quad (7.13)$$

- *<u>Priority actions</u> $z_t \in z^*$ consist of a $\mathring{z}-$vector ($\mathring{z} \le \mathring{d}_o$) of different indexes $z_{k;t} \in \{1, \ldots, \mathring{d}_o\}$, $k \in \{1, \ldots \mathring{z}\}$; z_t selects entries of $d_{o;t}$ to be shown to the operator. Given the ideal pdf $^{\lfloor I}f(c_t, u_{o;t}|d(t-1))$ describing the recommended pointers c_t and the recommended recognizable actions $u_{o;t}$, the vectors of priority actions z_t are generated by the optimized strategy $\left\{ ^{\lfloor I}f(z_t|d(t-1))\right\}_{t\in t^*}$. The priority actions are supposed to define the optimized ideal pdf*

$$\begin{aligned} ^{\lfloor I}f(d_t|z_t, d(t-1)) &= f(d_{\bar{z}_t;t}|d_{z_t;t}, d(t-1))\, ^{\lfloor I}f(d_{z_t;t}|d(t-1)) \\ &= f(d_t|d(t-1)) \frac{^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))}. \end{aligned} \quad (7.14)$$

  In (7.14), $d_{\bar{z}_t;t}$ denotes entries of d_t <u>not</u> presented to the operator. The pdf $f(d_t|d(t-1))$ is the estimated mixture, the pdf $f(d_{z_t;t}|d(t-1))$ is its marginal pdf on $d^*_{z_t;t}$ (also mixture!). The pdf $^{\lfloor I}f(d_{z_t;t}|d(t-1))$ is the marginal pdf of the ideal pdf designed in academic, industrial or simultaneous design. The ideal pdf presented to the operator is $^{\lfloor I}f(d_{\hat{z}_t;t}|d(t-1))$, where the pointer $\hat{z}_t$ to the shown entries $d_{\hat{z}_t;t}$ is a sample from $^{\lfloor I}f(z_t|d(t-1))$.
- *<u>Signaling actions</u> $s_t \in s^* \equiv \{0, 1\}$ stimulate the operator to take appropriate measures when behavior of the o-system significantly differs from the desired one. These actions, when respected by the operator, modify the behavior of the unguided o-system $f(d_t|s_t = 0, d(t-1)) \equiv f(d_t|d(t-1))$ to the behavior described by the ideal pdf $f(d_t|s_t = 1, d(t-1)) \equiv {}^{\lfloor I}f(d_t|d(t-1))$. The ideal pdf $^{\lfloor I}f(d_t|d(t-1))$ arises from the academic, industrial or simultaneous design of the p-system. The optimized ideal pdf has the form*

$$^{\lfloor I}f(d_t, s_t|d(t-1)) = {}^{\lfloor I}f(d_t|s_t, d(t-1))\, ^{\lfloor I}f(s_t|d(t-1)), \quad (7.15)$$

  where the probabilities $\left\{ ^{\lfloor I}f(s_t|d(t-1))\right\}_{t\in t^*}$ describe the signaling strategy. These optimized probabilities are presented to the operator. Typically, the probability $^{\lfloor I}f(s_t = 1|d(t-1))$ is converted into a traffic-light color that reflects the degree of the need for operator actions.

The above items determine the overall model of the connection of the o-system and p-system.

Remark(s) 7.4
The proposed model is used for designing the optimized advices and assumes the fully cooperating operator. It would be desirable to model an imperfect cooperation of a real operator. Yet, we have found no simple, sufficiently universal, way to do that. The behavior of the guided system seems to be reasonably robust to the degree of cooperation as extensive experiments indicate.

Problem 7.1 (Real operator) *An active correction for a real operator that modifies the recommended strategy can be achieved by extending the adaptive advisory system. Such an extended system has to relate explicitly response of the guided o-system on advices. This is a feasible but nontrivial possibility. At present, we can just passively check whether the operator cooperates: the proposed guided model is expected to provide a higher v-likelihood than the unguided model of the o-system. At least this test should be implemented in the online use of the advisory system.*

7.1.4 Fully probabilistic design in advising

The fully probabilistic design, Proposition 2.11, makes the conceptual framework we follow. It requires an additional inspection of the KL divergence that serves us as the loss function. It is done here.

Let us split the p-data into the p-innovations $\varDelta_t$ and the p-actions a_t

$$d(\mathring{t}) \equiv d_p(\mathring{t}) \equiv (d_o(\mathring{t}), d_{p+}(\mathring{t})) = (\varDelta_o(\mathring{t}), a(\mathring{t}), \varDelta_{p+}(\mathring{t})).$$

Innovation entries $\varDelta_o$ belong to the o-data space and $\varDelta_{p+}$ to the surplus data space of the p-system.

A recursive expression of the KL divergence for a fixed advising strategy suits for initializations of the design and for checking its results. It is described by the following proposition, which is a simplified version of Proposition 2.11.

Proposition 7.3 (Recursive evaluation of the KL divergence) *The KL divergence of the pdfs* ${}^{\lfloor I}f(d(\mathring{t}))$ *and* ${}^{\lfloor U}f(d(\mathring{t}))$ *can be expressed as follows.*

$$\mathcal{D}\left({}^{\lfloor I}f(d(\mathring{t})) \middle|\middle| {}^{\lfloor U}f(d(\mathring{t})) \right) = \mathcal{E}\left[\sum_{t \in t^*} \omega(d(t-1)) \right], \tag{7.16}$$

where the conditional KL divergence

$$\omega(d(t-1)) \equiv \int {}^{\lfloor I}f(d_t|d(t-1)) \ln\left(\frac{{}^{\lfloor I}f(d_t|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))} \right) dd_t.$$

The value of the KL divergence $\mathcal{D}\left({}^{\lfloor I}f(d(\mathring{t})) \middle|\middle| {}^{\lfloor U}f(d(\mathring{t})) \right) \equiv -\ln(\gamma(d(0)))$ *is generated recursively for* $t = \mathring{t}, \mathring{t}-1, \ldots, 1,$ *as follows.*

$$\begin{aligned} &-\ln(\gamma(d(t-1))) \\ &\equiv \mathcal{E}\left[\omega_\gamma(d(t))|d(t-1)\right] \equiv \int {}^{\lfloor I}f(d_t|d(t-1)) \ln\left(\frac{{}^{\lfloor I}f(d_t|d(t-1))}{\gamma(d(t))\, {}^{\lfloor U}f(d_t|d(t-1))} \right) dd_t, \end{aligned}$$

where the weighted conditional KL divergence *is defined by*

$$\omega_\gamma(d(t-1)) \equiv \int {}^{\lfloor I}f(d_t|d(t-1)) \ln\left(\frac{{}^{\lfloor I}f(d_t|d(t-1))}{\gamma(d(t))\,{}^{\lfloor U}f(d_t|d(t-1))}\right) dd_t.$$

The terminal value is $\gamma(d(\mathring{t})) \equiv 1$.

Proof. It holds

$$\mathcal{D}\left({}^{\lfloor I}f(d(\mathring{t}))\middle|\middle|\,{}^{\lfloor U}f(d(\mathring{t}))\right) \equiv \int {}^{\lfloor I}f(d(\mathring{t}))\ln\left(\frac{{}^{\lfloor I}f(d(\mathring{t}))}{{}^{\lfloor U}f(d(\mathring{t}))}\right) dd(\mathring{t})$$

$$\underbrace{=}_{\text{chain rule}} \int {}^{\lfloor I}f(d(\mathring{t}))\sum_{t\in t^*}\ln\left(\frac{{}^{\lfloor I}f(d_t|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))}\right) dd(\mathring{t}) \qquad \underbrace{=}_{\substack{\text{linearity, marginalization}\\ \text{chain rule, Fubini theorem}}}$$

$$= \sum_{t\in t^*}\underbrace{\int {}^{\lfloor I}f(d(t-1))\underbrace{\int {}^{\lfloor I}f(d_t|d(t-1))\ln\left(\frac{{}^{\lfloor I}f(d_t|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))}\right) dd_t}_{\omega(d(t-1))}\, dd(t-1)}_{\mathcal{E}[\omega(d(t-1))]}.$$

Defining

$$-\ln(\gamma(d(t-1))) \equiv \mathcal{E}\left[\sum_{\tau=t}^{\mathring{t}}\omega(d(\tau))\middle|\, d(t-1)\right] \underbrace{\Rightarrow}_{\text{chain rule for } \mathcal{E}}$$

$$-\ln(\gamma(d(t-1))) \equiv \mathcal{E}\left[\omega(d(t)) - \ln(\gamma(d(t)))|d(t-1)\right] \qquad \underbrace{\equiv}_{\text{definition of } \omega_\gamma}$$

$$\equiv \mathcal{E}\left[\omega_\gamma(d(t))|d(t-1)\right] \quad \underbrace{\Rightarrow}_{\mathcal{E}[\cdot]\equiv\mathcal{E}[\cdot|d(0)]} \quad \mathcal{D}\left({}^{\lfloor I}f\middle|\middle|\,{}^{\lfloor U}f\right) = -\ln(\gamma(d(0))).$$

□

The ideal pdf ${}^{\lfloor I}f(d(\mathring{t})) \equiv {}^{\lfloor I}f(\mathit{\Delta}(\mathring{t}), a(\mathring{t}))$ is constructed so that it reflects managing objectives expressed by the true user's ideal pdf ${}^{\lfloor U}f(d_{o;t}|d_o(t-1))$, extended on d_t^* by the pdf ${}^{\lfloor I}f(d_{p+;t}|d(t-1))$ as discussed in Section 5.1.5,

$${}^{\lfloor U}f(d(\mathring{t})) \underbrace{\equiv}_{(5.7)} \prod_{t\in t^*} {}^{\lfloor U}f(\mathit{\Delta}_{o;t}|d_o(t-1))\,{}^{\lfloor I}f(\mathit{\Delta}_{p+;t}|a_t, d(t-1))\,{}^{\lfloor U}f(a_t|d(t-1)). \tag{7.17}$$

This extension expresses

- lack of information (interest) of the operator concerning data out of d_o^*;
- the user's wishes with respect to advices a_t. They are either given by the true user's ideal pdf if $a_t = u_{o;t}$ = recognizable actions or specified by ${}^{\lfloor U}f(a_t|d(t-1))$ employed as a tuning knob of the design. Stability of advices is a typical implicit operator's wish respected with the help of this design knob.

The special form (7.17) of the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$ leads to a special form of Proposition 2.11.

Proposition 7.4 (Fully probabilistic design with a special target) *Let the joint pdf*

$$ {}^{\lfloor I}f(d(\mathring{t})) \equiv {}^{\lfloor I}f(\Delta(\mathring{t}), a(\mathring{t})) \equiv \prod_{t \in t^*} f(\Delta_t | a_t, d(t-1))\, {}^{\lfloor I}f(a_t | d(t-1)) $$

be influenced by the optional strategy $\left\{ {}^{\lfloor I}f(a_t|d(t-1)) \right\}_{t\in t^*}$.

Let the innovations $\Delta_t = (\Delta_{o;t}, \Delta_{p+;t})$ *be split into the part* $\Delta_{o;t}$ *belonging to the data space of the operator and the part* $\Delta_{p+;t}$ *belonging to the surplus data space of the advisory system. Then, the optimal strategy, minimizing the KL divergence*

$$ \mathcal{D}\left({}^{\lfloor I}f \middle|\middle| {}^{\lfloor U}f \right) \equiv \int {}^{\lfloor I}f(\Delta(\mathring{t}), a(\mathring{t})) \ln\left(\frac{{}^{\lfloor I}f(\Delta(\mathring{t}), a(\mathring{t}))}{{}^{\lfloor U}f(\Delta(\mathring{t}), a(\mathring{t}))} \right) d(\Delta(\mathring{t}), a(\mathring{t})) $$

to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$ *(7.17), has the form*

$$ {}^{\lfloor I}f(a_t|d(t-1)) = {}^{\lfloor U}f(a_t|d(t-1)) \frac{\exp[-\omega_\gamma(a_t, d(t-1))]}{\gamma(d(t-1))}, \quad \textit{where} \tag{7.18} $$

$$ \gamma(d(t-1)) \equiv \int {}^{\lfloor U}f(a_t|d(t-1)) \exp[-\omega_\gamma(a_t, d(t-1))]\, da_t $$

$$ \omega_\gamma(a_t, d(t-1)) \equiv \int f(\Delta_t|a_t, d(t-1)) \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, a_t, d(t-1))}{\gamma(d(t))\, {}^{\lfloor U}f(\Delta_{o;t}|a_t, d_o(t-1))} \right) d\Delta_t $$

$$ \gamma(d(\mathring{t})) = 1. $$

The solution is performed against the time course, starting at $t = \mathring{t}$.

Proof. Comparing to Proposition 2.11, the main change concerns the definition of the function $\omega_\gamma(a_t, d(t-1))$. It is implied directly by the assumption (7.17). Other changes are just typographical. The target pdf is distinguished by the superscript ${}^{\lfloor U}$, the optimized one by ${}^{\lfloor I}$ and experience is explicitly expressed in terms of the observed p-data $d(t-1)$. □

7.1.5 Approximations of the KL divergence

Some predictors forming the optimized pdf ${}^{\lfloor I}f(d(\mathring{t}))$ are mixtures. This makes evaluation of the KL divergence $\mathcal{D}\left({}^{\lfloor I}f \middle|\middle| {}^{\lfloor U}f \right)$ and consequently the application of Proposition 7.4 hard. To avoid this trouble, we use Jensen-type inequalities (2.14) for finding an upper bound on this divergence. It gives us a chance to optimize at least such an upper bound.

Proposition 7.5 (The J divergence of a mixture f to ${}^{\lfloor U}f$) *Let us consider the joint pdf on observed data $d(\mathring{t}) \in d^*(\mathring{t})$*

$$f(d(\mathring{t})) \equiv \prod_{t \in t^*} \sum_{c \in c^*} \alpha_{c;t} f(d_t|d(t-1), c),$$

where the components $f(d_t|d(t-1), c)$ as well as their probabilistic weights $\alpha_{c;t}$, possibly dependent on $d(t-1)$, are known. Let ${}^{\lfloor U}f(d(\mathring{t}))$ be another pdf on the same data space $d^*(\mathring{t})$. Then, the following inequality holds

$$\begin{aligned}
\mathcal{D}\left(f \middle|\middle| {}^{\lfloor U}f\right) &\le \mathcal{J}\left(f \middle|\middle| {}^{\lfloor U}f\right) \equiv \mathcal{E}\left\{\sum_{t \in t^*} \omega(d(t-1))\right\} \\
\omega(d(t-1)) &\equiv \sum_{c \in c^*} \alpha_{c;t} \omega(c, d(t-1)) \qquad (7.19) \\
\omega(c, d(t-1)) &\equiv \int f(d_t|d(t-1), c) \ln\left(\frac{f(d_t|d(t-1), c)}{{}^{\lfloor U}f(d_t|d(t-1))}\right) dd_t \ge 0.
\end{aligned}$$

The Jensen divergence (J divergence) $\mathcal{J}\left(f \,||\, {}^{\lfloor U}f\right)$ *is*

- *nonnegative,*
- *equal to zero iff $f(d_t|d(t-1), c) = {}^{\lfloor U}f(d_t|d(t-1))$ a.s. for all those $t \in t^*$ and $c \in c^*$ for which $\alpha_{c;t} \not\equiv 0$,*
- *infinite iff for some $t \in t^*$ and $c \in c^*$ the pdf ${}^{\lfloor U}f(d_t|d(t-1)) = 0$ and the pdf $f(d_t|d(t-1), c) > 0$ on a subset of d^* of a positive dominating measure while $\alpha_{c;t} \not\equiv 0$ on this subset.*

Proof. The the chain rule and definition of the KL divergence, Proposition 2.10, imply

$$\begin{aligned}
&\mathcal{D}\left(f||\, {}^{\lfloor U}f\right) \\
&= \sum_{t \in t^*} \int f(d(t-1)) \left[\int f(d_t|d(t-1)) \ln\left(\frac{f(d_t|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))}\right) dd_t\right] dd(t-1).
\end{aligned}$$

For a given $d(t-1)$, the inner integration over d_t^* concerns the function

$$\begin{aligned}
f(d_t|d(t-1)) \;\; &\ln\left(\frac{f(d_t|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))}\right) \\
&= \sum_{c \in c^*} \alpha_{c;t} f(d_t|d(t-1), c) \ln\left(\sum_{c \in c^*} \alpha_{c;t} f(d_t|d(t-1), c)\right) \\
&- \sum_{c \in c^*} \alpha_{c;t} f(d_t|d(t-1), c) \ln\left({}^{\lfloor U}f(d_t|d(t-1))\right) \\
&\underbrace{\le}_{(2.14)} \sum_{c \in c^*} \alpha_{c;t} f(d_t|d(t-1), c) \ln(f(d_t|d(t-1), c))
\end{aligned}$$

$$-\sum_{c\in c^*}\alpha_{c;t}f(d_t|d(t-1),c)\ln\left({}^{\lfloor U}f(d_t|d(t-1))\right)$$
$$=\sum_{c\in c^*}\alpha_{c;t}f(d_t|d(t-1),c)\ln\left(\frac{f(d_t|d(t-1),c)}{{}^{\lfloor U}f(d_t|d(t-1))}\right).$$

Integration of this inequality over d_t^* and the definitions of $\omega(d(t-1))$ and $\omega(c,d(t-1))$ imply (7.19). The non-negativity of $\omega(d(t-1))$ is obvious as, for any fixed $d(t-1)$, it is a convex combination of the nonnegative conditional KL divergences $\omega(c,d(t-1))$. Remaining properties of the J divergence follow from its form and properties of the KL divergence, Proposition 2.10. □

The inequality (7.19) between the KL and J divergences is proved for the mixture model. We have to deal with the ratio of mixture models, Proposition 7.4, whenever the surplus data space $d_{p+;t}^*$ of the p-system is nonempty and the special form of the user's ideal pdf (7.17) is used. The inequality (7.19) can be extended to this case, too.

Proposition 7.6 (J divergence of a pdf f to ${}^{\lfloor U}f$ (7.17)) *Let us consider the joint pdf on observed data $d(\mathring{t})\in d^*(\mathring{t})$, $d_t=(d_{o;t},d_{p+;t})$ with a nontrivial part $d_{p+;t}$ belonging to the surplus data space of the advisory system. Let us assume that*

$$f(d(\mathring{t}))\equiv\prod_{t\in t^*}\sum_{c\in c^*}\alpha_{c;t}f(d_t|d(t-1),c),$$

where the components $f(d_t|d(t-1),c)$ as well as their probabilistic weights $\alpha_{c;t}$, possibly dependent on $d(t-1)$, are known. Let

$${}^{\lfloor U}f(d(\mathring{t}))=\prod_{t\in t^*}{}^{\lfloor U}f(d_{o;t}|d_o(t-1))\sum_{c\in c^*}\alpha_{c;t}f(d_{p+;t}|d(t-1),c)$$

be the pdf defining the extended user's ideal pdf (7.17). Let for some constant $K\geq 1$ and some pdf $g(d_{p+;t}|d(t-1))$ hold

$$f(d_{p+;t}|d(t-1),c)\leq Kg(d_{p+;t}|d(t-1)),\forall c\in c^*\ \textit{and almost all data involved.}\quad(7.20)$$

Then, the following inequality holds

$$\begin{aligned}
&\mathcal{D}\left(f\middle|\middle|{}^{\lfloor U}f\right)\leq\mathcal{J}\left(f\middle|\middle|{}^{\lfloor U}f\right)+\ln(K)\ \ \textit{with}\\
&\mathcal{J}\left(f\middle|\middle|{}^{\lfloor U}f\right)\equiv\mathcal{E}\left\{\sum_{t\in t^*}\omega(d(t-1))\right\}\qquad(7.21)\\
&\omega(d(t-1))\equiv\sum_{c\in c^*}\alpha_{c;t}\omega(c,d(t-1))\\
&\omega(c,d(t-1))\equiv\int f(d_t|d(t-1),c)\ln\left(\frac{f(d_{o;t}|d_{p+;t},d(t-1),c)}{{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right)dd_t\geq 0.
\end{aligned}$$

Proof. Use of the chain rule for expressing the KL divergence implies that we have to inspect the difference

$$\begin{aligned}
&\sum_{t\in t^*}\mathcal{E}\Bigg\{\sum_{c\in c^*}\alpha_{c;t}\int f(d_t|d(t-1),c)\\
\times\ &\Bigg[\ln\left(\frac{\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c};t}f(d_t|d(t-1),\tilde{c})}{\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c};t}f(d_{p+;t}|d(t-1),\tilde{c})\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right)\\
-\ &\ln\left(\frac{Kf(d_t|d(t-1),c)}{f(d_{p+;t}|d(t-1),c)\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right)\Bigg]\,dd_t\Bigg\}\\
\underbrace{\le}_{(2.14)}\ &\sum_{t\in t^*,c\in c^*}\mathcal{E}\Bigg\{\alpha_{c;t}\int f(d_t|d(t-1),c)\\
\times\ &\ln\left(\frac{f(d_{p+;t}|d(t-1),c)}{K\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c};t}f(d_{p+;t}|d(t-1),\tilde{c})}\right)dd_t\Bigg\}\underbrace{\le}_{(7.20)}\sum_{t\in t^*}\\
&\mathcal{E}\left\{\sum_{c\in c^*}\alpha_{c;t}\int f(d_t|d(t-1),c)\ln\left(\frac{g(d_{p+;t}|d(t-1))}{\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c};t}f(d_{p+;t}|d(t-1),\tilde{c})}\right)dd_t\right\}\\
=\ &-\sum_{t\in t^*}\mathcal{E}\Bigg\{\sum_{c\in c^*}\alpha_{c;t}\int f(d_t|d(t-1),c)\\
\times\ &\ln\left(\frac{\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c};t}f(d_{p+;t}|d(t-1),\tilde{c})}{g(d_{p+;t}|d(t-1))}\right)dd_t\Bigg\}\\
=\ &-\sum_{t\in t^*}\mathcal{E}\left\{\mathcal{D}\left(\sum_{c\in c^*}\alpha_{c;t}f(d_{p+;t}|d(t-1),c)||g(d_{p+;t}|d(t-1))\right)\right\}\underbrace{\le}_{-\mathcal{D}(\cdot)\le 0}0.
\end{aligned}$$

□

In the industrial design and when designing presentation priorities, the KL divergences of more general ratios of finite mixtures to a given pdf are inspected. The following proposition shows how to approximate such distances.

Proposition 7.7 (The J divergence of a mixture ratio ${}^{\lfloor I}f$ to ${}^{\lfloor U}f$) *Let us consider the mixture*

$$f(x,y)=\sum_{c\in c^*}\alpha_c f(x,y|c)$$

on quantities x,y given by known components $f(x,y|c)$ and their probabilistic weights α_c, $c\in c^*$. Its marginal pdf is $f(y)=\sum_{c\in c^*}\alpha_c f(y|c)$. Let ${}^{\lfloor U}f(x,y)$, ${}^{\lfloor I}f(y)$ be given pdfs. Then, for

$${}^{\lfloor I}f(x,y)\equiv\frac{f(x,y)\,{}^{\lfloor I}f(y)}{f(y)}$$

$$\begin{aligned}
\mathcal{D}\left({}^{\lfloor I}f \middle|\middle| {}^{\lfloor U}f\right) &\equiv \int \frac{f(x,y)\,{}^{\lfloor I}f(y)}{f(y)} \ln\left(\frac{f(x,y)\,{}^{\lfloor I}f(y)}{f(y)\,{}^{\lfloor U}f(x,y)}\right) dxdy \\
&\leq \int {}^{\lfloor I}f(y) \left[\ln\left(\frac{{}^{\lfloor I}f(y)}{{}^{\lfloor U}f(y)}\right) + \omega(y)\right] dy \\
\omega(y) &\equiv \sum_{c\in c^*} f(c|y)\left[\omega(c,y) + \ln\left(\frac{f(c|y)}{\alpha_c}\right)\right], \quad \textit{where} \qquad (7.22) \\
\omega(c,y) &\equiv \int f(x|y,c)\ln\left(\frac{f(x|y,c)}{{}^{\lfloor U}f(x|y)}\right) dx \ \textit{and}\ f(c|y) \equiv \frac{\alpha_c f(y|c)}{\sum_{c\in c^*}\alpha_c f(y|c)}.
\end{aligned}$$

Proof. It holds that

$$\mathcal{D}\left({}^{\lfloor I}f \middle|\middle| {}^{\lfloor U}f\right) \equiv \int \frac{f(x,y)\,{}^{\lfloor I}f(y)}{f(y)} \ln\left(\frac{f(x,y)\,{}^{\lfloor I}f(y)}{f(y)\,{}^{\lfloor U}f(x,y)}\right) dxdy$$

$$\underbrace{\leq}_{(2.14)} \sum_{c\in c^*}\alpha_c \int \frac{f(x,y|c)}{f(y)}\,{}^{\lfloor I}f(y)\ln\left(\frac{f(x,y|c)\,{}^{\lfloor I}f(y)}{f(y)\,{}^{\lfloor U}f(x,y)}\right) dxdy \qquad \underbrace{=}_{\substack{\text{chain rule} \\ \text{marginalization} \\ f(y)=\sum_{c\in c^*}\alpha_c f(y|c)}}$$

$$= \int {}^{\lfloor I}f(y)\ln\left(\frac{{}^{\lfloor I}f(y)}{{}^{\lfloor U}f(y)}\right) dy + \sum_{c\in c^*}\int {}^{\lfloor I}f(y)\underbrace{\frac{\alpha_c f(y|c)}{f(y)}}_{f(c|y)}\ln\left(\frac{\alpha_c f(y|c)}{f(y)\alpha_c}\right) dy$$

$$+ \sum_{c\in c^*}\int {}^{\lfloor I}f(y)f(c|y)\underbrace{\int f(x|y,c)\ln\left(\frac{f(x|y,c)}{{}^{\lfloor U}f(x|y)}\right) dx}_{\omega(c,y)\geq 0}\, dy.$$

□

Problem 7.2 (Bounds on the KL divergence of mixture ratio) *There are surely variants of the derived upper bound. The problem should be inspected further on in order to get tighter upper bounds while preserving the possibility of evaluating them at least for normal and Markov mixtures.*

The approximations discussed up to now counteract evaluation problems caused by the form of the adopted models: the KL divergence is evaluated for mixtures or their ratios. The found bounds overcome them. In the adopted fully probabilistic design, Proposition 7.4, we have to evaluate two functions $\gamma(d(t))$ and $\omega_\gamma(d(t))$; see (7.18). This brings additional problems and calls for additional approximations. They are prepared here.

We search for an upper bound in order to guarantee that the strategy minimizing it bounds the KL divergence of interest.

We bound first the function $\gamma(d(t))$. The Bellman function of the solved design is $-\ln(\gamma(t))$; thus, if we find the lower bound on $\gamma(d(t))$, we get the desired upper bound on it. This result we call the *γ-bound.*

Proposition 7.8 (The γ-bound) *Let ${}^{\lfloor U}f(a_{t+1}|d(t))$ be a given pdf on a^*_{t+1} and $\omega_\gamma(a_{t+1}, d(t)) \geq 0$ be such a function that*

$$\omega_\gamma(d(t)) \equiv \int \omega_\gamma(a_{t+1}, d(t))\, {}^{\lfloor U}f(a_{t+1}|d(t))\, da_{t+1} < \infty \text{ a.s. on } d^*(t).$$

Then, (7.23)

$$\gamma(d(t)) \equiv \int {}^{\lfloor U}f(a_{t+1}|d(t)) \exp[-\omega_\gamma(a_{t+1}, d(t))]\, da_{t+1} \geq \exp[-\omega_\gamma(d(t))] > 0.$$

Proof. For fixed $d(t)$, $\gamma(d(t)) \equiv \mathcal{E}[\exp(-\omega(\cdot, d(t)))|d(t)]$. $\exp(-\omega(\cdot, d(t)))$ is a convex function of the argument ω. Thus, The Jensen inequality (2.14) is directly applicable. It gives the lower bound searched for. Its positivity is implied by the assumed finiteness of $\omega_\gamma(d(t))$. □

The estimate (7.23) can be refined if the considered actions have a finite number of possible values. It is the case of advices restricted to $a^* \subset (c^*, z^*, s^*)$; cf. Agreement 7.1. Essentially, bounds on individual functions $\omega_\gamma(a_{t+1}, d(t))$, $a_{t+1} \in a^*$ are constructed. These bounds lead to the upper bound on $-\ln(\gamma(d(t-1)))$. The corresponding bound is called the *ω-bound.*

Proposition 7.9 (The ω-bound) *For $\mathring{a} < \infty$ and $t = \mathring{t}, \mathring{t}-1, \ldots, 1$, let us consider sequence functions, cf. (7.18),*

$$\begin{aligned}
&\gamma(d(t)) \equiv \sum_{a_{t+1} \in a^*} {}^{\lfloor U}f(a_{t+1}|d(t)) \exp[-\omega_\gamma(a_{t+1}, d(t))]\\
&\omega_\gamma(a_t, d(t-1)) \equiv\\
&\equiv \int f(\Delta_t|a_t, d(t-1)) \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, a_t, d(t-1))}{\gamma(d(t))\, {}^{\lfloor U}f(\Delta_{o;t}|a_t, d_o(t-1))}\right) d\Delta_t\\
&\equiv \underbrace{\int f(\Delta_t|a_t, d(t-1)) \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, a_t, d(t-1))}{{}^{\lfloor U}f(\Delta_{o;t}|a_t, d_o(t-1))}\right) d\Delta_t}_{\omega(a_t, d(t-1))}\\
&\underbrace{-\int f(\Delta_t|a_t, d(t-1)) \ln(\gamma(d(t)))\, d\Delta_t}_{q(a_t, d(t-1))}.
\end{aligned}$$

The last decomposition separates the influence of $\gamma(d(t))$ on $\omega_\gamma(a_t, d(t-1))$. The first member of the sequence is $\omega(a_{\mathring{t}}, d(\mathring{t}-1)) \equiv \omega_{\gamma\equiv 1}(a_{\mathring{t}}, d(\mathring{t}-1))$.

For fixed t, $d(t-1)$, a_t, let us select arbitrary $a_{t+1} \equiv a_{a_t} \in a^*$ and let us define

$$\begin{aligned}
&\tilde{\omega}_\gamma(a_t, d(t-1)) \equiv \omega(a_t, d(t-1))\\
&+ \int f(\Delta_t|a_t, d(t-1)) \ln\Big({}^{\lfloor U}f(a_{a_t}|d(t)) \exp[-\tilde{\omega}_\gamma(a_{a_t}, d(t))]\Big)\, d\Delta_t
\end{aligned}$$

with zero second term for $t = \mathring{t}$. Then, $\forall t \in t^*$,

$$
\begin{aligned}
&-\ln[\gamma(d(t))] \equiv -\ln\left[\sum_{a_{t+1}\in a^*} {}^{\lfloor U}f(a_{t+1}|d(t))\exp[-\omega_\gamma(a_{t+1},d(t))]\right]\\
&\le -\ln[\tilde{\gamma}(d(t))] \equiv -\ln\left[\sum_{a_{t+1}\in a^*} {}^{\lfloor U}f(a_{t+1}|d(t))\exp[-\tilde{\omega}_\gamma(a_{t+1},d(t))]\right]. \qquad (7.24)
\end{aligned}
$$

Proof. Let us consider, for fixed $d(t-1)$, a_t, the second term $q(a_t, d(t-1))$ in the expression for $\omega_\gamma(a_t, d(t-1))$ that depends on $\gamma(d(t))$.

By definition, $\gamma(d(t))$ is a superposition of nonnegative terms. Omitting all of them except one, we get the lower bound on $q(a_t, d(t-1))$. Thus, for an arbitrary fixed $a_{t+1} \equiv a_{a_t} \in a^*$,

$$
\begin{aligned}
&\omega_\gamma(a_t, d(t-1)) \equiv \omega(a_t, d(t-1)) - q(a_t, d(t-1)) \le \omega(a_t, d(t-1))\\
&-\int f(\Delta_t|a_t, d(t-1)) \ln\left({}^{\lfloor U}f(a_{a_t}|d(t))\exp[-\omega_\gamma(a_{a_t}, d(t))]\right) d\Delta_t\\
&\equiv \tilde{\omega}(a_{a_t}, d(t-1)).
\end{aligned}
$$

Using the definition of $\gamma(d(t-1))$, we get

$$
\begin{aligned}
-\ln(\gamma(d(t-1))) &\equiv -\ln\left[\sum_{a_t} {}^{\lfloor U}f(a_t|d(t-1))\exp[-\omega_\gamma(a_t, d(t-1)]\right]\\
&\le -\ln\left\{\sum_{a_t} {}^{\lfloor U}f(a_t|d(t-1)\exp\left[-\tilde{\omega}(a_t, d(t-1))\right]\right\}.
\end{aligned}
$$

Thus, the use of the derived upper bound on a particular $\omega_\gamma(a_{t+1}, d(t))$ for the definition of $\omega_\gamma(a_t, d(t-1))$ guarantees that we get the upper bound on the Bellman function at time $t-1$. For $t = \mathring{t}$, inequality is guaranteed by the common starting condition $\gamma(d(\mathring{t})) = 1$. □

It remains to determine on a proper choice of a_{a_t} to get the tight version of the derived ω-bound. This decision is specific for the specific model and design type.

7.2 Design of advising strategies

7.2.1 Academic design

The recommended pointers $c_t \in c^*$ to mixture components form the actions of the academic p-system. They are generated by a causal strategy $d^*(t-1) \to c_t \in c^*$ described by pfs $\left\{ {}^{\lfloor I}f(c_t|d(t-1))\right\}_{t\in t^*}$. The strategy determines the optimized ideal pdfs

$$
{}^{\lfloor I}f(d_t, c_t|d(t-1)) = f(d_t|d(t-1), c_t)\, {}^{\lfloor I}f(c_t|d(t-1)).
$$

The extended user's ideal pdf, Section 5.1.5, has the form

$$ {}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t})) = \prod_{t \in t^*} {}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \, {}^{\lfloor I}f(d_{p+;t}|d(t-1), c_t) \, {}^{\lfloor U}f(c_t|d(t-1)). \tag{7.25} $$

The user's ideal pf ${}^{\lfloor U}f(c_t|d(t-1))$ is assumed to have support that excludes dangerous components; see Agreement 5.9 and Section 7.1.1. If a single nondangerous component remains, the optimal design of the academic p-system is solved by this exclusion. The optimal advising strategy has to be searched for if several nondangerous components exist.

Proposition 7.10 (Academic fully probabilistic design) *Let us consider the design of the academic advisory system with actions $c_t \in c^*$ restricted to those indexes in c^* that point to the nondangerous components only. Let us search for the optimal causal advising strategy $d^*(t-1) \to c_t \in c^*$ minimizing the KL divergence of ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t}))$ as given in (7.25). Then, this strategy is given by the following formulas, solved for $t = \mathring{t}, \mathring{t}-1, \ldots, 1$,*

$$ \begin{aligned} {}^{\lfloor I}f(c_t|d(t-1)) &= {}^{\lfloor U}f(c_t|d(t-1)) \frac{\exp[-\omega_\gamma(c_t, d(t-1))]}{\gamma(d(t-1))}, \quad \textit{where} \qquad (7.26) \\ \gamma(d(t-1)) &\equiv \sum_{c_t \in c^*} {}^{\lfloor U}f(c_t|d(t-1)) \exp[-\omega_\gamma(c_t, d(t-1))] \\ \omega_\gamma(c_t, d(t-1)) &\equiv \int f(d_t|d(t-1), c_t) \ln\left(\frac{f(d_{o;t}|d_{p+;t}, d(t-1), c_t)}{\gamma(d(t)) \, {}^{\lfloor U}f(d_{o;t}|d_o(t-1))} \right) dd_t \\ \gamma(d(\mathring{t})) &= 1. \end{aligned} $$

The needed pdf $f(d_{o;t}|d_{p+;t}, d(t-1), c)$ is the conditional pdf of the cth component $f(d_t|d(t-1), c)$ obtained in learning the o-system.

Proof. The proposition coincides with Proposition 7.4, see (7.18), for $d_t \equiv \mathit{\Delta}_t$ and the specific choice of actions of the p-system and their assumed influence (7.8) on the resulting ideal pdf. □

The above proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the state description with the state ϕ.

Algorithm 7.2 (Academic advising optimizing KL divergence)

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state ϕ_t; Chapter 6.*
- *Evaluate the steady-state behaviors of individual components; Section 7.1.1.*
- *Exclude dangerous components; Agreement 5.9.*
- *Return to the learning phase if all components are dangerous.*
- *Take the nondangerous component as the ideal pdf offered to the operator and stop if $\mathring{c} = 1$.*

- *Specify the user's ideal pdf* ${}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1})$ *on the response of the o-system.*
- *Specify the user's ideal pdf* ${}^{\lfloor U}f(c_t|d(t-1)) \equiv {}^{\lfloor U}f(c_t|\phi_{t-1})$ *on the recommended pointers* c_t *to the nondangerous components.*
- *Select the length of the receding horizon* $T \geq 1$.

Online (sequential) mode, *running for* $t = 1, 2, \ldots$

1. *Acquire the data record* d_t *and determine the state vector* ϕ_t.
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Initialize the iterative mode by setting* $\tau = t + T$ *and* $\gamma(\phi_\tau) = 1$.

   Iterative mode

   a) *Evaluate the* <u>*functions*</u>

$$\omega_\gamma(c_\tau, \phi_{\tau-1}) \equiv \int f(d_\tau|\phi_{\tau-1}, c_\tau) \ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau}, \phi_{\tau-1}, c_\tau)}{\gamma(\phi_\tau)\, {}^{\lfloor U}f(d_{o;\tau}|\phi_{\tau-1})}\right) dd_\tau, \;\; c_\tau \in c^*,$$

$$\gamma(\phi_{\tau-1}) \equiv \sum_{c_\tau \in c^*} {}^{\lfloor U}f(c_\tau|\phi_{\tau-1}) \exp[-\omega_\gamma(c_\tau, \phi_{\tau-1})]. \tag{7.27}$$

   b) *Continue if* $\tau = t + 1$. *Otherwise decrease* $\tau = \tau - 1$ *and go to Step 3a.*

4. *Evaluate the advising strategy*

$${}^{\lfloor I}f(c_{t+1}|\phi_t) = {}^{\lfloor U}f(c_{t+1}|\phi_t)\frac{\exp[-\omega_\gamma(c_{t+1}, \phi_t)]}{\gamma(\phi_t)}; \quad \textit{cf. (7.26).}$$

5. *Present to the operator projections of the ideal pdf (advisory mixture)*

$${}^{\lfloor I}f(d_{o;t+1}|\phi_t) = \sum_{c_{t+1} \in c^*} f(d_{o;t+1}|\phi_t, c_{t+1})\, {}^{\lfloor I}f(c_{t+1}|\phi_t),$$

   where the pdf $f(d_{o;t+1}|\phi_t, c_{t+1})$ *is marginal pdf of the cth estimated component* $f(d_{t+1}|\phi_t, c_{t+1})$.

6. *Go to the beginning of* Sequential mode.

Remark(s) 7.5

1. *The quasi-Bayes or quasi-EM algorithms, Section 6.5, possibly with the stabilized forgetting, Section 3.1, are suited for the online updating of the mixture model.*
2. *Other approximate strategies than the certainty-equivalence one are possible; see Section 4.2. They are not elaborated in this text.*
3. *Presentation of the design results is discussed in Section 7.3.1.*

Feasibility of Algorithm 7.2 depends on our ability to evaluate functions $\omega_\gamma(\cdot)$ and $\gamma(\cdot)$. It is hard even for normal and Markov mixtures due to the additive form (7.27) determining $\gamma(\cdot)$. In order to simplify the situation, we can use the freedom in the choice of the user's ideal pdf for advising actions ${}^{\lfloor U}f(c_{t+1}|\phi_t)$.

Proposition 7.11 (Academic design and the most probable advices) *Let us consider the design of the academic advisory system with actions $c_t \in c^*$ restricted to those indexes in c^* that point to the nondangerous components only. Let us search for the optimal causal advising strategy $d^*(t-1) \to c_t \in c^*$ that minimizes the KL divergence of ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ to the user ideal ${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t}))$ (7.25). The obtained minimum depends on the used ideal probabilities $\left\{ {}^{\lfloor U}f(c_t|d(t-1)) \right\}_{t \in t^*}$. Let us select these probabilities so that the minimum reached is the smallest one. Then, the optimal advising strategy is the deterministic feedback recommending the component with the index*

$$
{}^{\lfloor I}c_t \in \mathrm{Arg} \min_{c_t \in c^*} \omega_\gamma(c_t, d(t-1)) \tag{7.28}
$$

$$
\begin{aligned}
\omega_\gamma(c_t, d(t-1)) \equiv & \int f(d_t|d(t-1), c_t) \\
& \times \ln\left(\frac{f(d_{o;t}|d_{p+;t}, d(t-1), c_t)}{\exp(-\omega_\gamma({}^{\lfloor I}c_{t+1}, d(t)))\, {}^{\lfloor U}f(d_{o;t}|d_o(t-1))} \right) dd_t \\
\omega_\gamma\left({}^{\lfloor I}c_{\mathring{t}+1}, d(\mathring{t}) \right) = & \ 0.
\end{aligned}
$$

The pdf $f(d_{o;t}|d_{p+;t}, d(t-1), c_t)$ is the conditional pdf of the cth component $f(d_t|d(t-1), c_t)$ obtained by learning the mixture model of the o-system.

The chosen action selects the most probable academic advises among those given by Proposition 7.2.

Proof. The complexity of the academic design described by Proposition 7.2 stems from complexity of the Bellman function that equals to $-\ln(\gamma(d(t)))$ with

$$
\begin{aligned}
\gamma(d(t)) &\equiv \sum_{c_{t+1} \in c^*} {}^{\lfloor U}f(c_{t+1}|d(t)) \exp[-\omega_\gamma(c_{t+1}, d(t))] \\
\omega_\gamma(c_t, d(t-1)) &\equiv \int f(d_t|d(t-1), c_t) \ln\left(\frac{f(d_{o;t}|d_{p+;t}, d(t-1), c_t)}{\gamma(d(t))\, {}^{\lfloor U}f(d_{o;t}|d_o(t-1))} \right) dd_t.
\end{aligned}
$$

The pf ${}^{\lfloor U}f(c_{t+1}|d(t))$ is the optional design knob. We select it so that the Bellman function is minimized with respect to this pf. For that, we have to concentrate ${}^{\lfloor U}f(c_{t+1}|d(t))$ on ${}^{\lfloor I}c_{t+1} \in \mathrm{Arg}\min_{c_{t+1} \in c^*} \omega_\gamma(c_{t+1}, d(t))$. In this way, we get $\gamma(d(t)) = \exp\left[-\omega_\gamma\left({}^{\lfloor I}c_{t+1}, d(t)\right)\right]$ and the optimal strategy is deterministic

$$
{}^{\lfloor I}f(c_{t+1}|d(t)) = \delta_{c_{t+1},\, {}^{\lfloor I}c_{t+1}} \equiv \begin{cases} 1, \text{ if } c_{t+1} = {}^{\lfloor I}c_{t+1} \\ 0, \text{ if } c_{t+1} \neq {}^{\lfloor I}c_{t+1} \end{cases}.
$$

Moreover, this choice of ${}^{\lfloor U}f(c_{t+1}|d(t))$ maximizes the probability

$$
{}^{\lfloor U}f(c_{t+1}|d(t)) \exp[-\omega_\gamma(c_{t+1}, d(t))],
$$

i.e., it selects the most probable academic advice among those given by Proposition 7.2.

The terminal value $\omega_\gamma\left({}^{\lfloor I}c_{\mathring{t}+1}, d(\mathring{t})\right) = 0$ is implied by the general terminal value $\gamma(d(\mathring{t})) = 1$ valid for the fully probabilistic design.

The additional optimization replaces the random choice (7.26) of the recommended pointers c_t. It takes the single component $f\left(d_t|d(t-1), {}^{\lfloor I}c_t\right)$ of the learned model as the constructed ideal pdf. □

The proved proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the practically used state description with the state ϕ.

Algorithm 7.3 (The most probable academic advising)

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state ϕ_t; Chapter 6.*
- *Evaluate the steady state behaviors of individual components; Section 7.1.1.*
- *Exclude dangerous components; Agreement 5.9.*
- *Return to the learning phase if all components are dangerous.*
- *Take the nondangerous component as the ideal pdf offered to the operator and stop if $\mathring{c} = 1$.*
- *Specify the user's ideal pdf ${}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1})$ on the response of the o-system.*
- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Initialize the iterative mode by setting $\tau = t + T$ and $\omega_\gamma\left({}^{\lfloor I}c_{\tau+1}, \phi_\tau\right) = 0$.*

   Iterative mode

   a) Evaluate the functions, $c_\tau \in c^*$, $\omega_\gamma(c_\tau, \phi_{\tau-1})$

$$\equiv \int f(d_\tau|\phi_{\tau-1}, c_\tau) \ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau}, \phi_{\tau-1}, c_\tau)}{\exp\left[-\omega_\gamma\left({}^{\lfloor I}c_{\tau+1}, \phi_\tau\right)\right] {}^{\lfloor U}f(d_{o;\tau}|\phi_{\tau-1})}\right) dd_\tau$$

$${}^{\lfloor I}c_\tau(\phi_{\tau-1}) \in \operatorname{Arg}\min_{c_\tau \in c^*} \omega_\gamma(c_\tau, \phi_{\tau-1}).$$

   b) Continue if $\tau = t + 1$ otherwise decrease $\tau = \tau - 1$ and go to Step 3a.
4. *Present to the operator the ideal pdf ${}^{\lfloor I}f(d_{o;t+1}|\phi_t) = f\left(d_{o;t+1}|\phi_t, {}^{\lfloor I}c_{t+1}\right)$, which is a marginal pdf of the estimated component $f\left(d_{t+1}|d(t), {}^{\lfloor I}c_{t+1}\right)$.*
5. *Go to the beginning of* Sequential mode.

The feasibility of the design can be achieved also by minimizing an upper bound on the KL divergence. The following proposition provides such a solution that suits both normal components, for which $\omega_\gamma(c_{t+1}, d(t))$ are quadratic forms in data (see Chapter 9) as well as Markov chains, for which $\omega_\gamma(c_{t+1}, d(t))$ are finite-dimensional arrays; see Chapter 11.

Proposition 7.12 (Academic design with the γ-*bound*) *Let us consider the design of the academic advisory system with actions $c_t \in c^*$ restricted to those indexes in c^* that point to the nondangerous components only. Let us search for the optimal causal advising strategy $d^*(t-1) \to c_t \in c^*$ that approximately minimizes the KL divergence of ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t}))$; see (7.25). The strategy given by the following formulas, evaluated for $t = \mathring{t}, \mathring{t}-1, \ldots, 1$,*

$$\begin{aligned} {}^{\lfloor I}f(c_t|d(t-1)) &\propto {}^{\lfloor U}f(c_t|d(t-1)) \exp[-\omega_\gamma(c_t, d(t-1))] \qquad (7.29) \\ \omega_\gamma(c_t, d(t-1)) &\equiv \int f(d_t|d(t-1), c_t) \ln\left(\frac{f(d_{o;t}|d_{p+;t}, d(t-1), c_t)}{\gamma(d(t))\, {}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t \\ \gamma(d(t)) &\equiv \exp\left[-\sum_{c_{t+1}\in c^*} {}^{\lfloor U}f(c_{t+1}|d(t))\omega_\gamma(c_{t+1}, d(t))\right] \\ \gamma(d(\mathring{t})) &= 1 \end{aligned}$$

minimizes the γ-bound, Proposition 7.8, on the KL divergence.

The pdf $f(d_{o;t}|d_{p+;t}, d(t-1), c_t)$ is the conditional version of the cth component $f(d_t|d(t-1), c_t)$ of the learned mixture model of the o-system.

Proof. We apply formula (7.18) in Proposition 7.4, for the specific choice of actions of the p-system and their influence on the resulting ideal pdf. It gives

$$\begin{aligned} {}^{\lfloor I}f(c_t|d(t-1)) &\propto {}^{\lfloor U}f(c_t|d(t-1)) \exp[-\omega_\gamma(c_t, d(t-1))], \quad \text{where} \\ \omega_\gamma(c_t, d(t-1)) &\equiv \int f(d_t|d(t-1), c_t) \ln\left(\frac{f(d_{o;t}|d_{p+;t}, d(t-1), c_t)}{\gamma(d(t))\, {}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t \\ \gamma(d(\mathring{t})) &= 1 \text{ and for } t < \mathring{t} \\ \gamma(d(t)) &\equiv \sum_{c_{t+1}\in c^*} {}^{\lfloor U}f(c_{t+1}|d(t)) \exp[-\omega_\gamma(c_{t+1}, d(t))]. \end{aligned}$$

The Bellman function of this optimization equals to $-\ln(\gamma(d(t)))$. Thus, any replacement of $\gamma(d(t))$ by a lower value provides an upper bound on the optimized functional. We get it by using inequality between the weighted arithmetic and geometric means. Thus,

$$\gamma(d(t)) \geq \exp\left[-\sum_{c_{t+1}\in c^*} {}^{\lfloor U}f(c_{t+1}|d(t))\omega_\gamma(c_{t+1}, d(t))\right].$$

The use of the right-hand side of this inequality in the role of $\gamma(d(t))$ gives the claimed result. □

The proved proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the practically used state description with the state ϕ.

Algorithm 7.4 (Academic advising with the γ-bound) *Apply Algorithm 7.2 with the formula (7.27) replaced by*

$$\gamma(\phi_{\mathring{T}-1}) \equiv \exp\left[-\sum_{c_{\mathring{T}}\in c^*} {}^{\lfloor U}f(c_{\mathring{T}}|\phi_{\mathring{T}-1})\omega_\gamma(c_{\mathring{T}},\phi_{\mathring{T}-1})\right].$$

For the desirable receding horizon $T > 1$, all proposed variants depend heavily on our ability to evaluate the Bellman function $-\ln(\gamma(\phi))$. This calls for use of the strategy of iterations spread in time (IST); Section 4.2.1. The horizon $T = 1$ is expected to be sufficient in the majority of cases.

The common patch introducing IST strategy in all algorithms looks as follows.

Algorithm 7.5 (The common IST patch)

Initial (offline) mode

- $\ldots, \ldots$
- *Parameterize properly* $\gamma(\phi) = \gamma(\vartheta, \phi)$ *by a finite-dimensional parameter* ϑ.
- *Select the initial value* ϑ_0 *so that the general initial condition of the fully probabilistic design* $\gamma(\vartheta_0, \phi) = 1$ *is fulfilled.*

Sequential (online) mode, *running for* $t = 1, 2, \ldots$

$$\vdots$$

Omit resetting of $\gamma(\cdot)$ *before* Iterative mode.

Iterative mode

$$\vdots$$

Approximate the minimum reached after the tth step of Iterative mode *by* $\gamma(\vartheta_t, \phi)$.

Problem 7.3 (Extension of a set of approximate designs) *The list of approximate designs given in this chapter should be gradually completed. For instance, the following constraint on the target pdfs in advising*

$$\begin{aligned} {}^{\lfloor U}f\;\;(a_t|d(t-1)) &\propto {}^{\lfloor I}f(a_t|d(t-1))\exp\left(-0.5\left|\left|a_t - {}^{\lfloor U}a_t\right|\right|^2\right) \\ \Leftrightarrow {}^{\lfloor I}f(a_t|d(t-1)) &\propto {}^{\lfloor U}f(a_t|d(t-1))\exp\left(0.5\left|\left|a_t - {}^{\lfloor U}a_t\right|\right|^2\right) \quad (7.30) \\ \forall t \in t^*,\ a_t \in a_t^*,\ d(t-1) &\in d^*(t-1) \text{ and a user-specified } {}^{\lfloor U}a_t \end{aligned}$$

was considered. It reduces the choice of ${}^{\lfloor U}f(a_t|d(t-1))$ *to the choice of the appropriate weighted norm* $||\cdot||$ *and the specification of the target value* ${}^{\lfloor U}a_t$ *of the action* a_t. *The assumption (7.30) assigns low target probabilities to the actions far from a given* ${}^{\lfloor U}a_t$.

It can be shown that with this restriction, a deterministic strategy arises and a sort of quadratic programming has to be solved for normal mixtures.

Problem 7.4 (Use of the ω-bound in approximate designs) *The ω-bound, Proposition 7.9, opens the way for construction of a wide set of approximate designs that are expected to improve the elaborated use of the γ-bound for discrete-valued advices.*

7.2.2 Choice of user ideal on pointers

The pfs $\left\{ {}^{\lfloor U}f(c_t|d(t-1)) \right\}_{t\in t^*}$ have an important role in academic and simultaneous designs. They are out of the direct user interest and should be left to their fate; see Section 5.1.3. This leads to the choice ${}^{\lfloor U}f(c_t|d(t-1)) = \alpha_{c_t}$, where α are estimated component weights. This choice has the following interpretation in the γ-bounding of the Bellman function

$$\begin{aligned} & -\ln(\gamma(d(t))) \\ & \equiv -\ln\left\{ \sum_{c_{t+1}\in c^*} \alpha_{c_{t+1}} \right. \qquad (7.31) \\ & \left. \times \exp\left[-\int f(d_{t+1}|d(t),c_{t+1}) \ln\left(\frac{f(d_{t+1}|d(t),c_{t+1})}{\gamma(d(t+1))\, {}^{\lfloor U}f(d_{t+1}|d(t))} \right) d_{t+1} \right] \right\} \\ & \le \sum_{c_{t+1}\in c^*} \int f(d_{t+1},c_{t+1}|d(t)) \ln\left(\frac{f(d_{t+1},c_{t+1}|d(t))}{\gamma(d(t+1))\, {}^{\lfloor U}f(d_{t+1}|d(t))\alpha_{c_{t+1}}} \right) d_{t+1}. \end{aligned}$$

Thus, the bound describing the loss-to-go, Agreement 2.9, is derived from the objectively estimated pdf $f(d_{t+1},c_{t+1}|d(t))$ and it is expected to lead to a realistic but conservative approximation. At the same time, recommended pointers are hoped to improve the unguided situation. Thus, it makes sense to deviate from the complete resignation on the desired ${}^{\lfloor U}f(c_t|d(t-1))$. For instance, a constraint of the support of ${}^{\lfloor U}f(c_t|d(t-1))$ serves us for the exclusion of dangerous components. An active choice of ${}^{\lfloor U}f(c_t|d(t-1))$ opens other useful possibilities that are discussed here.

The low rate of operator actions calls for a low rate of changes of recommended components. This requirement can be respected by the specification

$$\begin{aligned} {}^{\lfloor U}f(c_\tau|d(\tau-1)) &= \delta_{c_\tau,c_{nt+1}} \text{ for } \tau = nt+1,\ldots,n(t+1) \qquad (7.32) \\ {}^{\lfloor U}f(c_{nt+1}|d(nt)) &\equiv \text{ a given pf.} \end{aligned}$$

The integer number $n \ge 1$ determines a *grouping rate* that should correspond to the rate with which the advices offered to the operator may change. Here, $t \in t^*$ counts the groups.

The choice (7.32) allows the recommended pointer to be changed at most each nth time moment. It leads to a special form of the academic design.

Proposition 7.13 (Academic design for ${}^{\lfloor U}f(c_t|d(t-1))$ (7.32)) *Let us consider the design of the academic advisory system with actions $c_t \in c^*$*

restricted to indexes in c^* that point to the nondangerous components only. Let us also assume that the user's ideal pf on recommended pointers is selected according to (7.32). Let us search for the optimal causal advising strategy $d^*(t-1) \to c_t \in c^*$ that minimizes the KL divergence of ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t}))$ as given in (7.25). Then, this strategy has the functional structure

$$
{}^{\lfloor I}f(c_\tau|d(\tau-1)) = \delta_{c_\tau, c_{nt+1}} \ \ \textit{for } \tau = nt+1, \ldots, n(t+1), \ \textit{where} \tag{7.33}
$$

${}^{\lfloor I}f(c_{nt+1}|d(nt))$ is given by the following formulas, solved for $t = \mathring{t}-1, \ldots, 1$,

$$
\begin{aligned}
{}^{\lfloor I}f(c_{nt+1}|d(nt)) &= {}^{\lfloor U}f(c_{nt+1}|d(nt)) \frac{\exp[-\omega_\gamma(c_{nt+1}, d(nt))]}{\gamma(d(nt))} \qquad (7.34)\\
\gamma(d(nt)) &\equiv \sum_{c_{nt+1} \in c^*} {}^{\lfloor U}f(c_{nt+1}|d(nt)) \exp[-\omega_\gamma(c_{tn+1}, d(nt))]\\
\omega_\gamma(c_{nt+1}, d(nt)) &\equiv -\frac{1}{n} \ln(\gamma(c_{nt+1}, d(nt))) \ \textit{with}
\end{aligned}
$$

$\gamma(c_{nt+1}, d(nt))$ being the last term of the sequence $\gamma(c_{nt+1}, d(\tau))$ with $\gamma(c_{nt+1}, d(n(t+1))) = \gamma(d(n(t+1)))$ for $t < \mathring{t}-1$ and $\gamma(d(n\mathring{t})) = 1$

$$
\begin{aligned}
&-\ln(\gamma(c_{nt+1}, d(\tau-1)))\\
&= \int f(d_\tau|d(\tau-1), c_{nt+1}) \ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau}, d(\tau-1), c_{nt+1})}{\gamma(c_{nt+1}, d(\tau)) \, {}^{\lfloor U}f(d_{o;\tau}|d_o(\tau-1), c_{nt+1})} \right) dd_\tau .
\end{aligned}
$$

The needed pdf $f(d_{o;\tau}|d_{p+;\tau}, d(\tau-1), c_\tau)$ is the conditional pdf of the cth component $f(d_\tau|d(\tau-1), c_\tau)$ obtained in learning of the o-system. The ideal pdf presented to the operator at moments $\tau = nt+1, \ldots, n(t+1)$ has the form

$$
{}^{\lfloor I}f(d_{\tau+1}|d(\tau)) = \sum_{c_{tn+1} \in c^*} {}^{\lfloor I}f(c_{nt+1}|d(nt)) f(d_{\tau+1}|d(\tau), c_{nt+1}), \tag{7.35}
$$

where the learned components $f(d_{\tau+1}|d(\tau), c_{\tau+1})$ are used.

Proof. The optimized KL divergence is infinite if the support of the optimized pdf ${}^{\lfloor I}f(c_\tau|d(\tau-1))$ is not fully included in the support of the user's ideal pf ${}^{\lfloor U}f(c_\tau|d(\tau-1))$. This determines the functional structure (7.33) of the optimal pf ${}^{\lfloor I}f(c_\tau|d(\tau-1))$. For such a pf, the optimized KL divergence can be given the form

$$
\begin{aligned}
&\mathcal{D}\left({}^{\lfloor I}f \middle|\middle| {}^{\lfloor U}f \right)\\
&\equiv \mathcal{E}\left\{ \sum_{t \in t^*} \sum_{c_{(tn+1)\cdots(t+1)n}} \int {}^{\lfloor I}f(d_{(tn+1)\cdots(t+1)n}, c_{(tn+1)\cdots(t+1)n}|d(tn), c(tn)) \right.\\
&\left. \times \ln\left(\frac{{}^{\lfloor I}f(d_{(tn+1)\cdots(t+1)n}, c_{(tn+1)\cdots(t+1)n}|d(tn), c(tn))}{{}^{\lfloor U}f(d_{(tn+1)\cdots(t+1)n}, c_{(tn+1)\cdots(t+1)n}|d(tn), c(tn))} \right) dd_{(tn+1)\cdots(t+1)n} \right\}
\end{aligned}
$$

$$= \mathcal{E}\left\{\sum_{t\in t^*}\sum_{c_{tn+1}}\sum_{\tau=tn+1}^{(t+1)n}\int f(d_\tau|d(\tau-1),c_{tn+1})\,{}^{\lfloor I}f(c_{tn+1}|d(tn))\right.$$

$$\left.\times\ln\left(\frac{f(d_\tau|d(\tau-1),c_{tn+1})\,{}^{\lfloor I}f(c_{tn+1}|d(tn))}{{}^{\lfloor U}f(d_\tau|d(\tau-1),c_{tn+1})\,{}^{\lfloor U}f(c_{tn+1}|d(tn))}\right)dd_\tau\right\}$$

$$= n\mathcal{E}\left\{\sum_{t\in t^*}\sum_{c_{tn+1}}{}^{\lfloor I}f(c_{tn+1}|d(tn))\left[\ln\left(\frac{{}^{\lfloor I}f(c_{tn+1}|d(tn))}{{}^{\lfloor U}f(c_{tn+1}|d(tn))}\right)+\omega(c_{tn+1},d(tn))\right]\right\}$$

$$\omega(c_{tn+1},d(tn)) = \frac{1}{n}\mathcal{E}\left[\sum_{\tau=tn+1}^{(t+1)n}\right.$$

$$\left.\int f(d_\tau|d(\tau-1),c_{tn+1})\ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau},d(\tau-1),c_{tn+1})}{{}^{\lfloor U}f(d_{o;\tau}|d_o(\tau-1),c_{tn+1})}\right)d_\tau|d(tn)\right].$$

The function $\omega(c_{tn+1},d(tn))$ is the conditional KL divergence that can be evaluated recursively in the way mimic to Proposition 7.3, i.e., $\omega(c_{tn+1},d(tn)) \equiv -1/n\ln(\gamma(c_{tn+1},d(tn)))$ with $\gamma(c_{\mathring{t}n+1},d(\mathring{t}n))=1$, $\gamma(c_{tn+1},d((t+1)n)) = \gamma(d((t+1)n))$, and

$$-\ln(\gamma(c_{tn+1},d(\tau-1)))$$
$$= \int f(d_\tau|d(\tau-1),c_{tn+1})\ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau},d(\tau-1),c_{tn+1})}{\gamma(c_{nt+1},d(\tau))\,{}^{\lfloor U}f(d_{o;\tau}|d_o(\tau-1),c_{tn+1})}\right)dd_\tau$$

for $\tau = (t+1)n,(t+1)n-1,\ldots,tn+1$. With the introduced function $\omega(c_{tn+1},d(tn))$, we have arrived to the standard expression for the KL divergence: the standard dynamic programming gives the claimed result. □

Let us present the algorithm corresponding to the proved proposition combined with the receding-horizon certainty-equivalence strategy. As above, it is written for the practically used state description with the state ϕ.

Algorithm 7.6 (Grouped academic advising)

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state ϕ_t; Chapter 6.*
- *Evaluate the steady state behaviors of individual components; Section 7.1.1.*
- *Exclude dangerous components; Agreement 5.9.*
- *Return to the learning phase if all components are dangerous.*
- *Take the nondangerous component as the ideal pdf offered to the operator and stop if $\mathring{c}=1$.*
- *Specify the user's ideal pdf ${}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1})$ on the response of the o-system.*
- *Select the grouping rate $n>1$ and specify the user's ideal pdf ${}^{\lfloor U}f(c_{nt+1}|d(nt)) \equiv {}^{\lfloor U}f(c_{nt+1}|\phi_{nt})$ on the recommended pointers to the nondangerous components.*

- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots$,*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Initialize the iterative mode by setting $\tau = n(t+T)$ and $\gamma(\phi_{n(t+T)}) = 1$.*

   Iterative mode

   a) *Evaluate the functions $\omega_\gamma(c_\tau, \phi_\tau)$, $c_\tau \in c^*$, as follows.*

   Set $\tilde{\tau} = \tau$, $\gamma(c, \phi_{\tilde{\tau}}) = \gamma(\phi_\tau)$

   For $\tilde{\tau} = \tau, \tau - 1, \ldots, \tau - n + 1$ evaluate

   $$-\ln(\gamma(c_{\tilde{\tau}}, \phi_{\tilde{\tau}-1})) = \int f(d_{\tilde{\tau}}|\phi_{\tilde{\tau}-1}, c_{\tilde{\tau}}) \ln\left(\frac{f(d_{o;\tilde{\tau}}|d_{p+;\tilde{\tau}}, \phi_{\tilde{\tau}-1}, c_{\tilde{\tau}})}{\gamma(c_{\tilde{\tau}}, \phi_{\tilde{\tau}}) f(d_{o;\tilde{\tau}}|\phi_{\tilde{\tau}-1})}\right) dd_{\tilde{\tau}}$$

   end of the cycle over $\tilde{\tau}$

   $$\omega_\gamma(c, \phi_{\tau n}) \equiv -\frac{1}{n} \ln(\gamma(c, \phi_{\tau n}))$$
   $$\gamma(\phi_{\tau n}) \equiv \sum_{c_{\tau n+1} \in c^*} {}^{\lfloor U}f(c_{\tau n+1}|\phi_{\tau n}) \exp[-\omega_\gamma(c_{\tau n+1}, \phi_{\tau n})]. \quad (7.36)$$

   b) *Continue if $\tau = t$ otherwise decrease $\tau = \tau - 1$ and go to Step 3a.*

4. *Evaluate the advising strategy*

   $${}^{\lfloor I}f(c_{nt+1}|\phi_{nt}) = {}^{\lfloor U}f(c_{nt+1}|\phi_{tn}) \frac{\exp[-\omega_\gamma(c_{nt+1}, \phi_{tn})]}{\gamma(\phi_{nt})}.$$

5. *Present to the operator projections of the ideal pdfs*

   $${}^{\lfloor I}f(d_{o;\tau}|\phi_{\tau-1}) = \sum_{c_{nt+1} \in c^*} f(d_{o;\tau}|\phi_{\tau-1}, c_{nt+1})\, {}^{\lfloor I}f(c_{nt+1}|\phi_{nt}), \ \tau = nt+1, \ldots, n(t+1).$$

   The pdf $f(d_{o;\tau}|\phi_{\tau-1}, c)$ is derived from the cth estimated components .

6. *Go to the beginning of* Sequential mode.

The introduced grouping can be used for an alternative, problem-tailored analysis of dangerous components. The nonweighted version of the conditional KL divergence, which assumes a fixed selection of the recommended pointers for n steps, expresses the expected quality when a single component is used for a long time. Thus, dangerous components are those for which

$$\lim_{n \to \infty} \omega(c_{tn+1} = c, \phi_{tn} = \phi) \equiv \omega_\infty(c, \phi)$$
$$\equiv \lim_{n \to \infty} \frac{1}{n} \mathcal{E}\left[\sum_{\tau=tn+1}^{(t+1)n} \int f(d_\tau|\phi_{\tau-1}, c) \ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau-1}, \phi_{\tau-1}, c)}{{}^{\lfloor U}f(d_{o;\tau}|d_o(\tau-1), c)}\right) d_\tau \,\middle|\, \phi_{tn} = \phi\right]$$

is too large. Moreover, the values $\omega_\infty(c,\phi)$ indicate preferences among components and guide in selecting the user's ideal pf ${}^{\lfloor U}f(c_t|\phi_{t-1})$ even when no grouping is applied. It is reasonable to choose

$$ {}^{\lfloor U}f(c_t|\phi_{t-1}) \in \mathrm{Arg}\min_{f(c_t|\phi_{t-1})} \left[\mathcal{D}\left(f||\,\mathring{c}^{-1}\right) + q\mathcal{E}(\omega_\infty(c_t,\phi_{t-1})\right], $$

i.e., to select the user's ideal pf close to the uniform $f(c) \equiv \mathring{c}^{-1}$ and having a small expected value of $\omega_\infty(c_t,\phi_{t-1}$. The positive optional weight q defines compromise between these requirements. Such a pf has the form

$$ {}^{\lfloor U}f(c_t|\phi_{t-1}) \propto \exp[-q\omega_\infty(c_t,\phi_{t-1})]. \tag{7.37} $$

Proposition 7.12 optimizes a computationally attractive upper bound on the KL divergence. Experiments, however, indicate that the resulting strategy is not sensitive enough to differences in quality of individual components. This sensitivity can be increased by applying the same approximation on grouped advices discussed above. This motivates the following proposition.

Proposition 7.14 (Grouped academic design with the γ-bound) *Let us consider the design of the academic advisory system with actions $c_t \in c^*$ restricted to indexes in c^* that point to the nondangerous components only. Let us also assume that the user's ideal pf on the recommended pointers is selected according to (7.32). Let us search for the optimal causal advising strategy $d^*(t-1) \to c_t \in c^*$ that minimizes the γ-bound, Proposition 7.8, on the KL divergence of ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t}))$ as given in (7.25). Then, such a strategy has the functional structure*

$$ {}^{\lfloor I}f(c_\tau|d(\tau-1)) = \delta_{c_\tau, c_{nt+1}}.\ \mathit{for}\ \tau = nt+1,\ldots,n(t+1). $$

The value c_{nt+1}, generated by the pf ${}^{\lfloor I}f(c_{nt+1}|d(nt))$, is given by the following formulas, solved for $t = \mathring{t}, \mathring{t}-1,\ldots,1$,

$$ \begin{aligned} {}^{\lfloor I}f(c_{nt+1}|d(nt)) &\propto {}^{\lfloor U}f(c_{nt+1}|d(nt))\exp[-\omega_\gamma(c_{nt+1},d(nt))] \qquad (7.38)\\ \omega_\gamma(c_{nt+1},d(nt)) &\equiv \frac{1}{n}\mathcal{E}\left[\sum_{\tau=nt+1}^{n(t+1)} \int f(d_\tau|d(\tau-1),c_{nt+1}) \right.\\ &\quad \left. \times\ \ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau},d(\tau-1),c_{nt+1})}{\gamma(d(n(t+1)))\,{}^{\lfloor U}f(d_{o;\tau}|d_o(\tau-1))}\right) dd_\tau \right| d(tn)\Bigg]\\ \gamma(d(nt)) &\equiv \exp\left[-\sum_{c_{nt+1}\in c^*} {}^{\lfloor U}f(c_{nt+1}|d(nt))\omega_\gamma(c_{tn+1},d(nt))\right]\\ \gamma(d(n\mathring{t})) &= 1. \end{aligned} $$

The needed pdf $f(d_{o;\tau}|d_{p+;\tau},d(\tau-1),c_\tau)$ is the conditional pdf of the cth component $f(d_\tau|d(\tau-1),c_\tau)$ obtained in learning of the o-system. The ideal pdf communicated to the operator has the form

$$^{\lfloor I}f(d_{\tau+1}|d(\tau)) = \sum_{c_{tn}\in c^*} {}^{\lfloor I}f(c_{nt+1}|d(nt))f(d_{\tau+1}|d(\tau), c_{nt+1}),$$

$\tau = nt+1,\ldots,n(t+1)$, the cth learned component $f(d_{\tau+1}|d(\tau), c_{\tau+1})$ are used.

Proof. The result is a straightforward extension of Proposition 7.13 when approximating the Bellman function from above through inequality between geometric and arithmetic means as it is done in the proof of Proposition 7.12 or in Proposition 7.8. □

This proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the practically used state description with the state ϕ.

Algorithm 7.7 (Grouped academic advising with the γ-bound)
Apply Algorithm 7.6 with the formula (7.36) replaced by

$$\gamma(\phi_{\tau n}) \equiv \exp\left[-\sum_{c_{\tau n+1}\in c^*} {}^{\lfloor U}f(c_{\tau n+1}|\phi_{\tau n})\omega_\gamma(c_{\tau n+1}, \phi_{\tau n})\right].$$

Problem 7.5 (Design and application of grouped advices) *Estimation in the adaptive version of the grouped design should run at the highest data collection rate. The grouped design can be repeated either after n real-time steps or repeated in every real time moment; cf. [158].*

Both possibilities have their pros and cons that need a detailed study.

7.2.3 Industrial design

The industrial design that influences components but leaves their weights unchanged deals with a rather complex and hardly manageable model. This leads to the temptation to omit this design completely and to rely on the simultaneous design solved in the next section. There is, however, an important class of problems in which component weights are objectively given. Thus, the industrial design inspected here is a necessary part of the design toolkit.

Proposition 7.15 (Industrial design with the bound (7.22)) *Let the optimized joint pdf*

$$\begin{aligned}&^{\lfloor I}f(\varDelta(\mathring{t}), u_o(\mathring{t}))\\ &\equiv \prod_{t\in t^*} \frac{\sum_{c\in c^*}\alpha_c f(\varDelta_t|u_{o;t}, d(t-1), c) f(u_{o;t}|d(t-1), c)}{\sum_{c\in c^*}\alpha_c f(u_{o;t}|d(t-1), c)}\, {}^{\lfloor I}f(u_{o;t}|d(t-1))\end{aligned}$$

be determined by the optional industrial advising strategy described by pdfs $\left\{{}^{\lfloor I}f(u_{o;t}|d(t-1))\right\}_{t\in t^*}$. The involved mixture with components, $c \in c^*$,

$$\{f(d_t|u_{o;t}, d(t-1), c) = f(\Delta_t|u_{o;t}, d(t-1), c)f(u_{o;t}|d(t-1), c)\}_{t\in t^*}$$

and their weights α_c are assumed to be known (well estimated).

Let us search for the strategy minimizing the upper bound of the type (7.22) on the KL divergence $\mathcal{D}\left({}^{\lfloor I}f\middle|\middle|\, {}^{\lfloor U}f\right)$ to the user's ideal pdf

$${}^{\lfloor U}f(d(\mathring{t})) \equiv \prod_{t\in t^*} {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1))\, {}^{\lfloor I}f(\Delta_{p+;t}|d(t-1)).$$

Let us denote

$$f(c|u_{o;t}, d(t-1)) \equiv \frac{\alpha_c f(u_{o;t}|d(t-1), c)}{\sum_{c\in c^*} \alpha_c f(u_{o;t}|d(t-1), c)}. \tag{7.39}$$

Then, the strategy searched for is described by the following formulas, solved for $t = \mathring{t}, \mathring{t}-1, \ldots, 1$,

$$\begin{aligned}
{}^{\lfloor I}f(u_{o;t}|d(t-1)) &= {}^{\lfloor U}f(u_{o;t}|d(t-1))\frac{\exp[-\omega_\gamma(u_{o;t}, d(t-1))]}{\gamma(d(t-1))}, \quad \textit{where}\\
\gamma(d(t-1)) &\equiv \int {}^{\lfloor U}f(u_{o;t}|d(t-1))\exp[-\omega_\gamma(u_{o;t}, d(t-1))]\, du_{o;t}\\
\omega_\gamma(u_{o;t}, d(t-1)) &\equiv \sum_{c\in c^*} f(c|u_{o;t}, d(t-1))\omega_\gamma(c, u_{o;t}, d(t-1)) \qquad (7.40)\\
\omega_\gamma(c, u_{o;t}, d(t-1)) &\equiv \ln\left(\frac{f(c|u_{o;t}, d(t-1))}{\alpha_c}\right)\\
&+ \int f(\Delta_t|u_{o;t}, d(t-1), c)\ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, d(t-1), c)}{\gamma(d(t))\, {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1))}\right) d\Delta_t\\
\gamma(d(\mathring{t})) &\equiv 1.
\end{aligned}$$

Proof. We set the correspondence with quantities used in Proposition 7.7: $d(t-1)$ is a fixed condition, $x \Leftrightarrow \Delta_t$, $y \Leftrightarrow u_{o;t}$. Then, we can write the optimized upper bound as the expected value of the loss function

$$\begin{aligned}
&\sum_{t\in t^*}\int {}^{\lfloor I}f(u_{o;t}|d(t-1))\left[\ln\left(\frac{{}^{\lfloor I}f(u_{o;t}|d(t-1))}{{}^{\lfloor U}f(u_{o;t}|d(t-1))}\right) + \omega(u_{o;t}, d(t-1))\right] du_{o;t}\\
&\omega_\gamma(u_{o;t}, d(t-1)) \equiv \sum_{c\in c^*} f(c|u_{o;t}, d(t-1))\omega_\gamma(c, u_{o;t}, d(t-1))\\
&\omega_\gamma(c, u_{o;t}, d(t-1)) \equiv \ln\left(\frac{f(c|u_{o;t}, d(t-1))}{\alpha_c}\right)\\
&\qquad + \int f(\Delta_t|u_{o;t}, d(t-1), c)\ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, d(t-1), c)}{{}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1))}\right) d\Delta_t.
\end{aligned}$$

The standard backward induction for $t = \mathring{t}$ gives

$${}^{\lfloor I}f(u_{o;\mathring{t}}|d(\mathring{t}-1)) \propto {}^{\lfloor U}f(u_{o;\mathring{t}}|d(\mathring{t}-1))\exp[-\omega(u_{o;\mathring{t}}, d(\mathring{t}-1))]$$

and the minimum value $-\ln(\gamma(d(\mathring{t}-1)))$ transferred to the further step with

$$\gamma(d(\mathring{t}-1)) = \int {}^{\lfloor U}f(u_{o;\mathring{t}}|d(\mathring{t}-1)) \exp[-\omega(u_{o;\mathring{t}}, d(\mathring{t}-1))]\, du_{o;\mathring{t}}.$$

For a generic t, we redefine

$$\begin{aligned}&\omega_\gamma(c, u_{o;t}, d(t-1))\\ &\equiv \int f(\mathit{\Delta}_t|u_{o;t}, d(t-1), c) \ln\left(\frac{f(\mathit{\Delta}_{o;t}|\mathit{\Delta}_{p+;t}, u_{o;t}, d(t-1), c)}{\gamma(d(t))\, {}^{\lfloor U}f(\mathit{\Delta}_{o;t}|u_{o;t}, d_o(t-1))}\right) d\mathit{\Delta}_t,\end{aligned}$$

in the description of $\omega(u_{o;t}, d(t-1))$. It shows that the computation has the same structure as for $t = \mathring{t}$. The full conformity is reached when defining $\gamma(d(\mathring{t})) = 1$. □

The proved proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the state description with the state ϕ.

Algorithm 7.8 (Industrial advising with the bound (7.22))

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state ϕ_t; Chapter 6.*
- *Specify the user's ideal pdf on the response of the o-system*

$$ {}^{\lfloor U}f(\mathit{\Delta}_{o;t}|u_{o;t}, d_o(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1)).$$

- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Initialize the iterative mode by setting $\tau = t + T$ and $\gamma(\phi_\tau) = 1$.*

   Iterative mode

   a) *Evaluate the functions*

$$\begin{aligned} f(c_\tau|u_{o;\tau}, \phi_{\tau-1}) &\equiv \frac{\alpha_{c_\tau} f(u_{o;\tau}|\phi_{\tau-1}, c_\tau)}{\sum_{c_\tau\in c^*} \alpha_{c_\tau} f(u_{o;\tau}|\phi_{\tau-1}, c_\tau)}\\ \omega_\gamma(c_\tau, u_{o;\tau}, \phi_{\tau-1}) &\equiv \ln\left(\frac{f(c_\tau|u_{o;\tau}, \phi_{\tau-1})}{\alpha_{c_\tau}}\right)\\ &+ \int f(\mathit{\Delta}_\tau|u_{o;\tau}, \phi_{\tau-1}, c_\tau) \ln\left(\frac{f(\mathit{\Delta}_{o;\tau}|\mathit{\Delta}_{p+;\tau}, u_{o;\tau}, \phi_{\tau-1}, c_\tau)}{\gamma(\phi_\tau)\, {}^{\lfloor U}f(\mathit{\Delta}_{o;\tau}|u_{o;\tau}, \phi_{\tau-1})}\right) d\mathit{\Delta}_\tau\\ \omega_\gamma(u_{o;\tau}, \phi_{\tau-1}) &\equiv \sum_{c_\tau\in c^*} f(c|u_{o;\tau}, \phi_{\tau-1}) \omega_\gamma(c_\tau, u_{o;\tau}, \phi_{\tau-1})\\ \gamma(\phi_{\tau-1}) &\equiv \int {}^{\lfloor U}f(u_{o;\tau}|\phi_{\tau-1}) \exp[-\omega_\gamma(u_{o;\tau}, \phi_{\tau-1})]\, du_{o;\tau}\end{aligned}$$

b) Continue if $\tau = t+1$, otherwise decrease $\tau = \tau - 1$ and go to Step 3a.

4. *Evaluate the industrial strategy*

$$^{\lfloor I}f(u_{o;t+1}|\phi_t) \propto \, ^{\lfloor U}f(u_{o;t+1}|\phi_t)\exp[-\omega_\gamma(u_{o;t+1},\phi_t)].$$

5. *Present to the operator projections of the ideal pdf (advisory mixture)*

$$^{\lfloor I}f(d_{o;t+1}|\phi_t) = \frac{\sum_{c\in c^*}\alpha_c f(\mathit{\Delta}_{o;t}|u_{o;t+1},\phi_t,c)f(u_{o;t+1}|\phi_t,c)}{\sum_{c\in c^*}\alpha_c f(u_{o;t+1}|\phi_t,c)} \, ^{\lfloor I}f(u_{o;t+1}|\phi_t).$$

The pdfs $f(\mathit{\Delta}_{o;t}|u_{o;t+1},\phi_t,c)$, $f(u_{o;t+1}|\phi_t,c)$ are derived from cth learned mixture component.

6. *Go to the beginning of* Sequential mode.

Remark(s) 7.6

1. *Various bounds have been and can be derived. The bound (7.22) seems to be the best one of those tried. Alternatives have to be considered in the future.*
2. *Use of γ- or ω-bounds, Propositions 7.8 and 7.9 is probably unnecessary as the used bound (7.22) includes them.*
3. *The maximum probable choice of the recognizable actions is possible.*

7.2.4 Simultaneous academic and industrial design

The separation of academic and industrial designs is often motivated pedagogically only. Whenever possible, the optimization of the recommended pointers c_t and recognizable actions $u_{o;t}$ should be performed simultaneously. It may give much better results. The recognizable actions may completely change the character of the analyzed components, for instance, to change dangerous components into nondangerous ones. Even less drastic changes reflected in the improved quality are worth considering. Moreover, the simultaneous design addressed here is simpler than the industrial one.

Proposition 7.16 (Simultaneous fully probabilistic design) *Let the optimized joint pdf*

$$^{\lfloor I}f(d(\mathring{t}),c(\mathring{t})) \equiv \prod_{t\in t^*} f(\mathit{\Delta}_t|u_{o;t},d(t-1),c_t) \, ^{\lfloor I}f(u_{o;t}|d(t-1),c_t) \, ^{\lfloor I}f(c_t|d(t-1))$$

be determined by the optional simultaneous advising strategy described by pdfs

$$\left\{ ^{\lfloor I}f(c_t,u_{o;t}|d(t-1)) = \, ^{\lfloor I}f(u_{o;t}|d(t-1),c_t) \, ^{\lfloor I}f(c_t|d(t-1))\right\}_{t\in t^*}.$$

The pdfs $\{f(\mathit{\Delta}_t|u_{o;t},d(t-1),c_t)\}_{t\in t^*}$ are determined by the components of the known (well estimated) mixture. Let us search for the strategy minimizing the KL divergence $\mathcal{D}\left(^{\lfloor I}f\middle|\middle| \, ^{\lfloor U}f\right)$ to the user's ideal pdf

$$
\begin{aligned}
{}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t})) \equiv \prod_{t \in t^*} & {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1)) \\
& \times {}^{\lfloor I}f(\Delta_{p+;t}|d(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1)).
\end{aligned}
$$

Then, the optimal strategy is described by the following formulas, solved for $t = \mathring{t}, \mathring{t}-1, \ldots, 1$,

$$
\begin{aligned}
& {}^{\lfloor I}f(c_t, u_{o;t}|d(t-1)) = {}^{\lfloor U}f(u_{o;t}|d_o(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1)) \\
& \times \frac{\exp[-\omega_\gamma(c_t, u_{o;t}, d(t-1))]}{\gamma(d(t-1))} \\
& \gamma(d(t-1)) \equiv \sum_{c_t \in c^*} \int {}^{\lfloor U}f(u_{o;t}|d_o(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1)) \\
& \times \exp[-\omega_\gamma(c_t, u_{o;t}, d(t-1))]\, du_{o;t} \\
& \omega_\gamma(c_t, u_{o;t}, d(t-1)) \\
& \equiv \int f(\Delta_t|u_{o;t}, d(t-1), c_t) \ln\left[\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, d(t-1), c_t)}{\gamma(d(t))\, {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d(t-1))}\right] d\Delta_t \\
& \gamma(d(\mathring{t})) \equiv 1. \qquad (7.41)
\end{aligned}
$$

Proof. Let us assume that the Bellman function in the addressed problem has the form $-\ln(\gamma(d(t)))$. For $t = \mathring{t}$, it holds with $\gamma(d(\mathring{t})) = 1$. A generic optimization step has to find

$$
\begin{aligned}
& -\ln(\gamma(d(t-1))) \\
\equiv & \min_{\{{}^{\lfloor I}f(c_t, u_{o;t}|d(t-1))\}} \sum_{c_t \in c^*} \int f(\Delta_t|u_{o;t}, d(t-1), c_t)\, {}^{\lfloor I}f(c_t, u_{o;t}|d(t-1)) \ln \\
& \frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, d(t-1), c_t)\, {}^{\lfloor I}f(c_t, u_{o;t}|d(t-1))}{\gamma(d(t))\, {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1))}\, dd_t \\
& \underbrace{=}_{(7.41)} \sum_{c_t \in c^*} \int {}^{\lfloor I}f(c_t, u_{o;t}|d(t-1)) \times \\
\times & \left[\omega_\gamma(c_t, u_{o;t}, d(t-1)) + \ln\left(\frac{{}^{\lfloor I}f(c_t, u_{o;t}|d(t-1))}{{}^{\lfloor U}f(u_{o;t}|d_o(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1))}\right)\right] du_{o;t}.
\end{aligned}
$$

The minimizing joint pdf of pointers and recognizable actions is

$$
\begin{aligned}
& {}^{\lfloor I}f(c_t, u_{o;t}|d(t-1)) \\
& = \frac{{}^{\lfloor U}f(u_{o;t}|d_o(t-1))\, {}^{\lfloor U}f(c_t|d(t-1)) \exp[-\omega_\gamma(c_t, u_{o;t}, d(t-1))]}{\gamma(d(t-1))}.
\end{aligned}
$$

The achieved minimum that coincides with $-\ln(\cdot)$ of the normalizing factor defines the achieved form of $\gamma(d(t-1))$, namely,

$$\gamma(d(t-1)) = \sum_{c_t \in c^*} \int {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1)) \times \exp[-\omega_\gamma(c_t, u_{o;t}, d(t-1))]\, du_{o;t}.$$

□

The proved proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the state description with the state ϕ.

Algorithm 7.9 (Simultaneous advising with the KL divergence)

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state ϕ_t; Chapter 6.*
- *Specify the user's ideal pdf on the response of the o-system*

$$\begin{aligned} {}^{\lfloor U}f\ &(\Delta_{o;t}|u_{o;t}, d_o(t-1), c_t)\, {}^{\lfloor U}f(u_{o;t}, c_t|d_o(t-1)) \\ &\equiv {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, \phi_{t-1})\, {}^{\lfloor U}f(u_{o;t}, c_t|\phi_{t-1}). \end{aligned}$$

- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Initialize the iterative mode by setting $\tau = t + T$ and $\gamma(\phi_\tau) = 1$.*

   Iterative mode

   a) *Evaluate the functions $\omega_\gamma(c_\tau, u_{o;\tau}, \phi_{\tau-1}) \equiv$*

$$\begin{aligned} &\omega_\gamma(c_\tau, u_{o;\tau}, \phi_{\tau-1}) \\ &\equiv \int f(\Delta_\tau|u_{o;\tau}, \phi_{\tau-1}, c_\tau) \ln\left(\frac{f(\Delta_{o;\tau}|u_{o;\tau}, \Delta_{p+;\tau}, \phi_{\tau-1}, c_\tau)}{\gamma(\phi_\tau)\, {}^{\lfloor U}f(\Delta_{o;\tau}|u_{o;\tau}, \phi_{\tau-1})}\right) d\Delta_\tau \\ &\gamma(\phi_{\tau-1}) \equiv \sum_{c_\tau \in c^*} \int {}^{\lfloor U}f(u_{o;\tau}|\phi_{\tau-1})\, {}^{\lfloor U}f(c_\tau|u_{o;\tau}, \phi_{\tau-1}) \qquad (7.42) \\ &\times \exp[-\omega_\gamma(c_\tau, u_{o;\tau}, \phi_{\tau-1})]\, du_{o;\tau}. \end{aligned}$$

   b) *Continue if $\tau = t+1$, otherwise decrease $\tau = \tau - 1$ and go to Step 3a.*

4. *Evaluate the optimal simultaneous strategy*

$$\begin{aligned} {}^{\lfloor I}f(c_{t+1}, u_{o;t+1}|\phi_t) \propto\ & {}^{\lfloor U}f(c_{t+1}|u_{o;t+1}, \phi_t)\, {}^{\lfloor U}f(u_{o;t+1}|\phi_t) \\ & \times \exp[-\omega_\gamma(c_{t+1}, u_{o;t+1}, \phi_t)]. \end{aligned}$$

5. *Present to the operator projections of the ideal pdf (advisory mixture)*

$${}^{\lfloor I}f(d_{o;t+1}|\phi_t) = \sum_{c_{t+1} \in c^*} f(\Delta_{o;t+1}|u_{o;t+1}, \phi_t, c_{t+1})\, {}^{\lfloor I}f(c_{t+1}, u_{o;t+1}|\phi_t),$$

   where the pdf $f(\Delta_{o;t+1}|u_{o;t+1}, \phi_t, c_{t+1})$ is derived from the c_{t+1}th learned mixture component.

6. *Go to the beginning of* Sequential mode*.*

The additional minimization of the achieved minimum of the KL divergence with respect to the optional ${}^{\lfloor U}f(c_t|d(t-1))$ does not simplify the evaluation. For this reason, no counterpart of Proposition 7.3 is presented. However, the replacement of $\gamma(d(t))$ in the definition of $\omega_\gamma(\cdot)$ by a lower bound on $\gamma(d(t))$ simplifies the design substantially.

Proposition 7.17 (Simultaneous design with the γ-bound) *Let the optimized joint pdf*

$$ {}^{\lfloor I}f(d(\mathring{t}),c(\mathring{t})) \equiv \prod_{t\in t^*} f(\Delta_t|u_{o;t},d(t-1),c_t)\, {}^{\lfloor I}f(u_{o;t}|d(t-1),c_t)\, {}^{\lfloor I}f(c_t|d(t-1)) $$

be determined by the optional simultaneous advising strategy described by pdfs

$$ \left\{ {}^{\lfloor I}f(c_t,u_{o;t}|d(t-1)) = {}^{\lfloor I}f(u_{o;t}|d(t-1),c_t)\, {}^{\lfloor I}f(c_t|d(t-1)) \right\}_{t\in t^*}. $$

The pdfs $\{f(\Delta_t|u_{o;t},d(t-1),c_t)\}_{t\in t^*}$ are derived from the components of the known (well-estimated) mixture. Let us search for the strategy approximately minimizing the KL divergence $\mathcal{D}\left({}^{\lfloor I}f\middle|\middle|\,{}^{\lfloor U}f\right)$ to the user's ideal pdf

$$ \begin{aligned} {}^{\lfloor U}f(d(\mathring{t}),c(\mathring{t})) \equiv \prod_{t\in t^*} & {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t},d_o(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1)) \\ & \times\, {}^{\lfloor I}f(\Delta_{p+;t}|d(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t},d(t-1)). \end{aligned} $$

Then, the following strategy minimizes the γ-bound, Proposition 7.8, on the KL divergence, $t=\mathring{t},\mathring{t}-1,\ldots,1$,

$$ \begin{aligned} & {}^{\lfloor I}f(c_t,u_{o;t}|d(t-1)) \\ & \propto\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1))\, {}^{\lfloor U}f(c_t|u_{o;t},d(t-1)) \exp[-\omega_\gamma(c_t,u_{o;t},d(t-1))] \\ & \omega_\gamma(c_t,u_{o;t},d(t-1)) \equiv \int f(\Delta_t|u_{o;t},d(t-1),c_t) \qquad (7.43) \\ & \times \ln\left[\frac{f(\Delta_{o;t}|\Delta_{p+;t},u_{o;t},d(t-1),c_t)}{\gamma(d(t))\, {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t},d(t-1))}\right] d\Delta_t \\ & \gamma(d(t-1)) \equiv \int {}^{\lfloor U}f(u_{o;t}|d_o(t-1)) \\ & \times \exp\left[-\sum_{c_t\in c^*} {}^{\lfloor U}f(c_t|u_{o;t},d(t-1))\omega_\gamma(c_t,u_{o;t},d(t-1))\right] du_{o;t} \\ & \gamma(d(\mathring{t})) \equiv 1. \end{aligned} $$

Proof. The optimal strategy is determined by Proposition 7.16 and its Bellman function equals to $-\ln(\gamma(d(t)))$. Thus, it is sufficient to find a lower bound on $\gamma(d(t))$ in order to care about minimization of an upper bound on the KL divergence. The inequality between the weighted arithmetic and geometric means is used here; similarly as in Proposition 7.12.

Notice that the γ-bound, Proposition 7.8, is applied to the marginal pf on pointers $c \in c^*$ only. $\square$

The above proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for systems with the state ϕ.

Algorithm 7.10 (Simultaneous advising with the γ-bound) *Apply Algorithm 7.9 with the formula (7.42) replaced by*

$$\gamma(\phi_{\tau-1}) \equiv \int {}^{\lfloor U}f(u_{o;\tau}|\phi_{\tau-1})$$
$$\times \exp\left[-\sum_{c_\tau \in c^*} {}^{\lfloor U}f(c_\tau|u_{o;\tau},\phi_{\tau-1})\omega_\gamma(c_\tau,u_{o;\tau},\phi_{\tau-1})\right] du_{o;\tau}.$$

Remark(s) 7.7

1. *The grouped variant, similar to Algorithm 7.7, can be simply derived. Its use seems to be desirable.*
2. *Other design variants with the γ-bound, like the maximum probable one, are possible and worth inspecting.*

Problem 7.6 (Industrial design as a restricted simultaneous one) *Industrial design can be interpreted as a simultaneous one restricted by the requirement ${}^{\lfloor I}f(c_t|u_{o;t},d(t-1)) = \alpha_{c_t} \equiv$ estimated component weight. It seems that the design tuning knob ${}^{\lfloor U}f(c_t|u_{o;t},d(t-1))$ could be chosen so that the discussed constraint is met at least approximately. This promising approximation is worth a deeper inspection.*

7.3 Interaction with an operator

Priority z_t and signaling s_t actions are the main tools for communication with the operator. Their approximate fully probabilistic designs are addressed here.

7.3.1 Assigning priorities

Actions z_t, called priorities, select the few, most important entries of $d_{o;t}$ that should be presented to the operator; see Agreement 7.1. Such a selection is necessary in a generic case when the dimension of $d_{o;t}$ does not allow us to jointly show all entries.

When assigning priorities, we assume that the operator has been given a full go for making recommended actions. It means that the signaling action

$s_t = 1$; see Section 7.3.2. Then, the model relating the priorities z_t to the response of the o-system reduces to

$$
\begin{aligned}
{}^{\lfloor I}f(d_t|z_t, d(t-1)) &\equiv f(d_t|d(t-1)) \frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))}, \quad \text{where} \qquad (7.44)\\
f(d_t|d(t-1)) &= \sum_{c\in c^*} \alpha_c f(d_t|d(t-1), c) \ \text{ is the learned mixture and}\\
{}^{\lfloor I}f(d_t|d(t-1)) &= \sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|d(t-1))\, {}^{\lfloor I}f(d_t|d(t-1), c_t)
\end{aligned}
$$

is the ideal pdf resulting from the academic, industrial or simultaneous design.

Data $d_{p+;t}$ are not influenced directly by priorities as we cannot show the surplus p-data to the operator. Thus, the extension of the true user's ideal pdf

$$
\begin{aligned}
{}^{\lfloor U}f(d_t|z_t, d(t-1)) &\equiv {}^{\lfloor I}f(d_{p+;t}|z_t, d(t-1))\, {}^{\lfloor U}f(d_{o;t}|d_o(t-1))\\
&= {}^{\lfloor I}f(d_{p+;t}|d(t-1))\, {}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_t|d(t-1))
\end{aligned}
$$

does not depend on the presentation actions z_t. Similarly to academic design, we introduce and use ${}^{\lfloor U}f(z_t|d(t-1))$ as a tuning knob of the presentation design that fixes the externally supplied priorities and damps the rate of z_t-changes.

The conditional KL divergence, the key quantity in the fully probabilistic design, has the form

$$
\omega_\gamma(z_t, d(t-1)) \equiv \int {}^{\lfloor I}f(d_t|z_t, d(t-1)) \ln\left[\frac{{}^{\lfloor I}f(d_t|z_t, d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))}\right] dd_t \underbrace{=}_{(7.44)}
$$

$$
= \int f(d_t|d(t-1)) \frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))} \ln \frac{f(d_t|d(t-1))\, {}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))\, {}^{\lfloor U}f(d_t|d_o(t-1))} dd_t.
$$

The rational form (7.44) of the model ${}^{\lfloor I}f(d_t|z_t, d(t-1))$ implies that an upper bound on the conditional KL divergence has to be constructed in order to get a loss function that can be effectively minimized. Essentially, Proposition 7.7 is tailored and complemented by a single use of the Jensen inequality, Proposition 7.6.

Proposition 7.18 (J bound on the KL divergence in presentation) *Let us consider that the academic, industrial or simultaneous design of the advisory system has provided the strategy $d^*(t-1) \to (c_t^*, u_{o;t}^*)$ determining the ideal pdf*

$$
{}^{\lfloor I}f(d_t|d(t-1)) = \sum_{c_t\in c^*} {}^{\lfloor I}f(d_t|d(t-1), c_t)\, {}^{\lfloor I}f(c_t|d(t-1)).
$$

The presentation strategy $\{f(z_t|d(t-1))\}_{t\in t^*}$ *is inspected, assuming the influence of the presentation advices z_t through the model (7.44). Let us denote*

$$f(c|d_{z_t;t}, d(t-1)) = \frac{\alpha_c f(d_{z_t;t}|d(t-1), c)}{\sum_{c \in c^*} \alpha_c f(d_{z_t;t}|d(t-1))}, \forall c \in c^*. \tag{7.45}$$

Let us assume that the basic assumption (7.20) of Proposition 7.6 is met with a constant $K \geq 1$. Let $\bar{z}_t$ consist of indexes of the data record d_t in $\{1, 2, \ldots, \mathring{d}\}$ not included in z_t and let us define

$$\begin{aligned} \omega(c, d_{z_t;t}, d(t-1)) &\equiv \ln\left(\frac{f(c|d_{z_t;t}, d(t-1))}{\alpha_c}\right) \qquad (7.46) \\ &\quad + \int f(d_{\bar{z}_t;t}|d_{z_t;t}, d(t-1), c) \ln\left(\frac{f(d_{\bar{z}_t;t}|d_{z_t;t}, d(t-1), c)}{{}^{\lfloor U}f(d_{\bar{z}_t;t}|d_{z_t;t}, d(t-1))}\right) dd_{\bar{z}_t;t} \\ \omega(z_t, d_{z_t;t}, d(t-1)) &\equiv \sum_{c \in c^*} f(c|d_{z_t;t}, d(t-1))\omega(c, d_{z_t;t}, d(t-1)). \end{aligned}$$

Then,

$$\begin{aligned} \omega(z_t, d(t-1)) &\equiv \int f(d_t|d(t-1)) \frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))} \qquad (7.47) \\ &\quad \times \ln\left[\frac{f(d_t|d(t-1))\, {}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1)) f(d_{z_t;t}|d(t-1))}\right] dd_t \leq \bar{\omega}(z_t, d(t-1)) \\ \bar{\omega}(z_t, d(t-1)) &\equiv \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|d(t-1)) \int {}^{\lfloor I}f(d_{z_t;t}|d(t-1), c_t) \\ &\quad \times \left[\ln\left(\frac{{}^{\lfloor I}f(d_{z_t;t}|d_{p+;t}, d(t-1), c_t)}{{}^{\lfloor U}f(d_{z_t;t}|d_o(t-1))}\right)\right. \\ &\quad \left. + \sum_{c \in c^*} f(c|d_{z_t;t}, d(t-1))\omega(c, d_{z_t;t}, d(t-1))\right] dd_{z_t;t}. \end{aligned}$$

Proof. For a fixed z_t, we use the following correspondence with quantities of Proposition 7.7

$$\begin{aligned} d(t-1) &\quad \text{is a fixed condition} \\ x &\Leftrightarrow d_{\bar{z}_t;t} \\ y &\Leftrightarrow d_{z_t;t} \\ f(x, y) &\Leftrightarrow f(d_t|d(t-1)) \\ f(y) &\Leftrightarrow f(d_{z_t;t}|d(t-1)) \\ {}^{\lfloor I}f(y) &\Leftrightarrow {}^{\lfloor I}f(d_{z_t;t}|d(t-1)) \\ {}^{\lfloor U}f(x, y) &\Leftrightarrow {}^{\lfloor U}f(d_t|d(t-1)) \underbrace{\equiv}_{(5.7)} {}^{\lfloor U}f(d_{o;t}|d_o(t-1))\, {}^{\lfloor I}f(d_{p+;t}|d(t-1)). \end{aligned}$$

It gives, cf. (7.22),

$$\omega(z_t, d(t-1)) \equiv \int f(d_t|d(t-1)) \frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{f(d_{z_t;t}|d(t-1))}$$

$$\times \ln\left(\frac{f(d_t|d(t-1))\,{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{{}^{\lfloor U}f(d_t|d(t-1))f(d_{z_t;t}|d(t-1))}\right) dd_t$$

$$\leq \int {}^{\lfloor I}f(d_{z_t;t}|d(t-1))\left[\ln\left(\frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{{}^{\lfloor U}f(d_{z_t;t}|d(t-1))}\right)+\omega(d_{z_t;t},d(t-1))\right] dd_{z_t;t}$$

$$\omega(d_{z_t;t},d(t-1)) \equiv \sum_{c\in c^*} f(c|d_{z_t;t},d(t-1))\omega(c,d_{z_t;t},d(t-1)).$$

$$\omega(c,d_{z_t;t},d(t-1)) \equiv \ln\left(\frac{f(c|d_{z_t;t},d(t-1))}{\alpha_c}\right)$$

$$+\int f(d_{\bar{z}_t;t}|d_{z_t},d(t-1),c)\ln\left(\frac{f(d_{\bar{z}_t;t}|d_{z_t;t},d(t-1),c)}{{}^{\lfloor U}f(d_{\bar{z}_t;t}|d_{z_t;t},d(t-1))}\right) dd_{\bar{z}_t;t}.$$

This upper bound still contains logarithms of the mixture ${}^{\lfloor I}f(d_{z_t;t}|d(t-1))$. The use of the Jensen inequality extended according to Proposition 7.6 implies

$$\int {}^{\lfloor I}f(d_{z_t;t}|\;\; d(t-1))\ln\left(\frac{{}^{\lfloor I}f(d_{z_t;t}|d(t-1))}{{}^{\lfloor U}f(d_{z_t;t}|d(t-1))}\right) dd_{z_t,t}$$

$$\leq \sum_{c_t\in c^*}\int {}^{\lfloor I}f(c_t|d(t-1))\,{}^{\lfloor I}f(d_{z_t;t}|d(t-1),c_t)$$

$$\times \ln\left(\frac{{}^{\lfloor I}f(d_{z_t;t}|d_{p+;t},d(t-1),c_t)}{{}^{\lfloor U}f(d_{z_t;t}|d(t-1))}\right) dd_{z_t,t}.$$

Inserting this estimate into the upper bound on $\omega(z_t, d(t-1))$, using the fact that $\sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|d(t-1)) = 1$ and the observation that the surplus p-data $d_{p+;t}$ are not presented, we get the claimed result

$$\omega(z_t,d(t-1)) \leq \sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|d(t-1))\int {}^{\lfloor I}f(d_{z_t;t}|d(t-1),c_t)$$

$$\times\left[\ln\left(\frac{{}^{\lfloor I}f(d_{z_t;t}|d_{p+;t},d(t-1),c_t)}{{}^{\lfloor U}f(d_{z_t;t}|d_o(t-1))}\right)\right.$$

$$\left.+\sum_{c\in c^*} f(c|d_{z_t;t},d(t-1))\omega(c,d_{z_t;t},d(t-1))\right] dd_{z_t;t}.$$

□

Using the found upper bound as the loss function in the fully probabilistic design, Proposition 7.4, we get the presentation strategy searched for.

Proposition 7.19 (Presentation design with the bound (7.47)) *Let us consider that the academic, industrial or simultaneous design of the advisory system has provided the strategy $d^*(t-1) \to (c_t^*, u_{o;t}^*)$ determining the ideal pdf*

$${}^{\lfloor I}f(d_t|d(t-1)) = \sum_{c_t\in c^*} {}^{\lfloor I}f(d_t|d(t-1),c_t)\,{}^{\lfloor I}f(c_t|d(t-1)).$$

Let the signaling strategy make the operator fully alert, i.e., signaling actions $s(\mathring{t}) \equiv 1$*. Let also assume that there is constant* $K \geq 1$ *and a pdf* $g(d_{p+;t}|d(t-1))$ *such that*

$$
{}^{\lfloor I}f(d_{p+;t}|d(t-1),c) \leq Kg(d_{p+;t}|d(t-1)), \forall c \in c^* \tag{7.48}
$$

and for almost all data in arguments; cf. Proposition 7.6.

Let us define for each $z_t \in z^*$

$$
f(c|d_{z_t;t},d(t-1)) = \frac{\alpha_c f(d_{z_t;t}|d(t-1),c)}{\sum_{c\in c^*} \alpha_c f(d_{z_t;t}|d(t-1))} \leq 1, \forall c \in c^*. \tag{7.49}
$$

Let us specify the user's ideal pf ${}^{\lfloor U}f(z_t|d(t-1))$ *on the set of possible priority actions* $z^* \equiv \{[z_1,\ldots,z_{\mathring{z}}]\}$ *with* z_j *being a unique index in the set of indexes of the data records* d_o*. Then, the presentation strategy that minimizes the upper bound on the KL divergence, implied by the inequality given in Proposition 7.7, is described by the following algorithm, in which* $\bar{z}_t$ *mark indexes of* d_t *in* $\{1,\ldots,\mathring{d}\}$ *excluded from* z_t*,*

$$
\begin{aligned}
{}^{\lfloor I}f(z_t|d(t-1)) &\equiv \frac{{}^{\lfloor U}f(z_t|d(t-1))\exp[-\omega_\gamma(z_t,d(t-1))]}{\gamma(d(t-1))} \qquad (7.50)\\
\gamma(d(t-1)) &\equiv \sum_{z_t\in z^*} {}^{\lfloor U}f(z_t|d(t-1))\exp[-\omega_\gamma(z_t,d(t-1))]\\
\omega_\gamma(z_t,d(t-1)) &\equiv \sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|d(t-1)) \int {}^{\lfloor I}f(d_{z_t;t}|d(t-1),c_t)\\
&\quad\times \left[\ln\left(\frac{{}^{\lfloor I}f(d_{z_t;t}|d_{p+;t},d(t-1),c_t)}{{}^{\lfloor U}f(d_{z_t;t}|d_o(t-1))}\right)\right.\\
&\quad\left. + \sum_{c\in c^*} f(c|d_{z_t;t},d(t-1))\omega(c,d_{z_t;t},d(t-1))\right] dd_{z_t;t}\\
\omega(c,d_{z_t;t},d(t-1)) &\equiv \ln\left(\frac{f(c|d_{z_t;t},d(t-1))}{\alpha_c}\right)\\
&\quad+\int f(d_{\bar{z}_t;t}|d_{z_t;t},d(t-1),c)\ln\left(\frac{f(d_{\bar{z}_t;t}|d_{z_t;t},d(t-1),c)}{\gamma(d(t))\,{}^{\lfloor U}f(d_{\bar{z}_t;t}|d_{z_t;t},d(t-1))}\right) dd_{\bar{z}_t;t}\\
\gamma(d(\mathring{t})) &\equiv 1.
\end{aligned}
$$

The solution is evaluated against the time course, starting at $t = \mathring{t}$*.*

Proof. The conditional KL divergence is just replaced by the weighted KL divergence arising in dynamic programming. It preserves the non-negativity of the original definition that was used in deriving the used upper bound. □

The above proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the practically used state description with the state ϕ.

Algorithm 7.11 (Presentation with the bound (7.47))

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state ϕ_t; Chapter 6.*
- *Specify the user's ideal pdf on $\Delta_{o;t}, u_{o;t}$ coinciding with the true user's ideal pdf and on c_t, z_t as the tuning knob of the presentation design.*
- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Design an academic, industrial or simultaneous strategy generating the optimal $c_t, u_{o;t}$ and thus generating on d_t^* the ideal pdf*

$$ {}^{\lfloor I}f(d_t|\phi_{t-1}) = \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|\phi_{t-1})\, {}^{\lfloor I}f(d_t|\phi_{t-1}, c_t). $$

4. *Initialize the iterative mode by setting $\tau = t + T$ and $\gamma(\phi_\tau) = 1$.*

   Iterative mode

   a) Evaluate for each $z_\tau \in \{1, \ldots, \mathring{d}_o\}^{\mathring{z}}$ the <u>*functions*</u>

$$ f(c|d_{z_\tau;\tau}, \phi_{\tau-1}) \equiv \frac{\alpha_c f(d_{z_\tau;\tau}|\phi_{\tau-1}, c)}{\sum_{c \in c^*} \alpha_c f(d_{z_\tau;\tau}|\phi_{\tau-1}, c)} $$

$$ \gamma(\phi_{\tau-1}) \equiv \sum_{z_\tau \in z^*} {}^{\lfloor U}f(z_\tau|\phi_{\tau-1}) \exp[-\omega_\gamma(z_\tau, \phi_{\tau-1})] $$

$$ \begin{aligned} \omega_\gamma(z_\tau, \phi_{\tau-1}) \equiv & \sum_{c_\tau \in c^*} {}^{\lfloor I}f(c_\tau|\phi_{\tau-1}) \int {}^{\lfloor I}f(d_{z_\tau;\tau}|\phi_{\tau-1}, c_\tau) \\ & \times \left[\ln\left(\frac{{}^{\lfloor I}f(d_{z_\tau;\tau}|d_{p+;\tau}, \phi_{\tau-1}, c_\tau)}{{}^{\lfloor U}f(d_{z_\tau;\tau}|\phi_{\tau-1})} \right) \right. \\ & \left. + \sum_{c \in c^*} f(c|d_{z_\tau;\tau}, \phi_{\tau-1}) \omega(c, d_{z_\tau;\tau}, \phi_{t-1}) \right] dd_{z_\tau;\tau} \end{aligned} $$

$$ \begin{aligned} \omega(c, d_{z_\tau;\tau}, \phi_{\tau-1}) \equiv & \ln\left(\frac{f(c|d_{z_\tau;\tau}, \phi_{\tau-1})}{\alpha_c} \right) \\ + \int & {}^{\lfloor I}f(d_{\bar{z}_\tau;\tau}|d_{z_\tau;\tau}, \phi_{\tau-1}, c) \ln\left(\frac{{}^{\lfloor I}f(d_{\bar{z}_\tau;\tau}|d_{z_\tau;\tau}\phi_{\tau-1}, c)}{\gamma(\phi_\tau)\, {}^{\lfloor U}f(d_{\bar{z}_\tau;\tau}|d_{z_\tau;\tau}, \phi_{\tau-1})} \right) dd_{\bar{z}_\tau;\tau}. \end{aligned} $$

   b) Continue if $\tau = t + 1$, otherwise decrease $\tau = \tau - 1$ and go to Step 4a.

5. *Generate a sample $\hat{z}_{t+1} \in z^*$ from the pf ${}^{\lfloor I}f(z_{t+1}|\phi_t)$, for instance, maximizing argument,*

$$ {}^{\lfloor I}f(z_{t+1}|\phi_t) \propto {}^{\lfloor U}f(z_{t+1}|\phi_t) \exp[-\omega_\gamma(z_{t+1}, \phi_t)]. $$

6. *Present the projections* ${}^{\lfloor I}f(d_{\hat{z}_{t+1};t+1}|\phi_t)$ *of the ideal pdf* ${}^{\lfloor I}f(d_{t+1}|\phi_t)$ *to the operator.*
7. *Go to the beginning of* Sequential mode.

In spite of many approximations made, even this algorithm may be infeasible. This makes us to design the presentation strategy that minimizes the γ-bound, Proposition 7.8. Moreover, the number of compared presentation values is quickly growing with $\mathring{\hat{z}}$. It equals to $\binom{\mathring{d}_o}{\mathring{\hat{z}}}$. Thus, it is almost necessary to select $\mathring{\hat{z}} = 1$ and to show the operator data entries with the highest values of ${}^{\lfloor I}f(z_t|d(t-1))$. This variant is summarized in the algorithm below that otherwise assumes the receding-horizon certainty-equivalence strategy and the state ϕ.

Algorithm 7.12 (Presentation with (7.47); changed order)

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state* ϕ_t*; Chapter 6.*
- *Specify the user's ideal pdf on* $\Delta_{o;t}, u_{o;t}, c_t, z_t$*.*
- *Select the length of the receding horizon* $T \geq 1$*.*
- *Specify the dimension* $\mathring{\hat{z}} \leq \mathring{d}_o$ *of the vector* $\hat{z}_t$ *of* $d_{o;t}$*-indexes to be shown to the operator. It is limited by just perceiving the abilities of the operator and may be much higher than the dimension* $\mathring{z}_t = 1$*.*

Sequential (online) mode, *running for* $t = 1, 2, \ldots,$

1. *Acquire the data record* d_t *and determine the state vector* ϕ_t*.*
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Design an academic, industrial or simultaneous strategy generating the optimal* $c_t, u_{o;t}$ *and thus determining on* d_t^* *the ideal pdf*

$$ {}^{\lfloor I}f(d_t|\phi_{t-1}) = \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|\phi_{t-1})\, {}^{\lfloor I}f(d_t|\phi_{t-1}, c_t). $$

4. *Initialize the iterative mode by setting* $\tau = t + T$ *and* $\gamma(\phi_\tau) = 1$*.*

   Iterative mode

   a) *Evaluate, for each* $z_\tau \in z^* \equiv \{1, \ldots, \mathring{d}_o\}$ *the* functions

$$ f(c|d_{z_\tau;\tau}, \phi_{\tau-1}) \equiv \frac{\alpha_c f(d_{z_\tau;\tau}|\phi_{\tau-1}, c)}{\sum_{c \in c^*} \alpha_c f(d_{z_\tau;\tau}|\phi_{\tau-1}, c)} $$

$$ \gamma(\phi_{\tau-1}) \equiv \exp\left[- \sum_{z_\tau \in z^*} {}^{\lfloor U}f(z_\tau|\phi_{\tau-1}) \omega_\gamma(z_\tau, \phi_{\tau-1}) \right] $$

$$ \omega_\gamma(z_\tau, \phi_{\tau-1}) \equiv \sum_{c_\tau \in c^*} {}^{\lfloor I}f(c_\tau|\phi_{\tau-1}) \int {}^{\lfloor I}f(d_{z_\tau;\tau}|\phi_{\tau-1}, c_\tau) $$

$$\times \left[\ln \left(\frac{{}^{\lfloor I}f(d_{z_\tau;\tau}|d_{p+;\tau}, \phi_{\tau-1}, c_\tau)}{{}^{\lfloor U}f(d_{z_\tau;\tau}|\phi_{\tau-1})} \right) \right.$$

$$\left. + \sum_{c \in c^*} f(c|d_{z_\tau;\tau}, \phi_{\tau-1}) \omega(c, d_{z_\tau;\tau}, \phi_{t-1}) \right] dd_{z_\tau;\tau}$$

$$\omega(c, d_{z_\tau;\tau}, \phi_{\tau-1}) \equiv \ln \left(\frac{f(c|d_{z_\tau;\tau}, \phi_{\tau-1})}{\alpha_c} \right)$$

$$+ \int {}^{\lfloor I}f(d_{\bar{z}_\tau;\tau}|d_{z_\tau;\tau}, \phi_{\tau-1}, c) \ln \left(\frac{{}^{\lfloor I}f(d_{\bar{z}_\tau;\tau}|d_{z_\tau;\tau}\phi_{\tau-1}, c)}{\gamma(\phi_\tau)\, {}^{\lfloor U}f(d_{\bar{z}_\tau;\tau}|d_{z_\tau;\tau}, \phi_{\tau-1})} \right) dd_{\bar{z}_\tau;\tau}.$$

b) *Continue if $\tau = t + 1$, otherwise decrease $\tau = \tau - 1$ and go to the beginning of* Iterative mode.

5. *Order the values ${}^{\lfloor I}f(z_{t+1}|\phi_t) \propto {}^{\lfloor U}f(z_{t+1}|\phi_t) \exp[-\omega_\gamma(z_{t+1}, \phi_t)]$.*
6. *Present projections of the ideal pdf ${}^{\lfloor I}f(d_{\hat{z}_{t+1};t+1}|\phi_t)$ to the operator. The vector $\hat{z}_{t+1}$ of indexes has the entries with $\mathring{\hat{z}}$ highest values of ${}^{\lfloor I}f(z_{t+1}|\phi_t)$.*
7. *Go to the beginning of* Sequential mode.

Problem 7.7 (Completion of the set of presentation strategies) *Several algorithms should be completed for presentation combining IST strategy, Section 4.2.1, grouping mimic to (7.32), the maximum probable design, etc.*

Problem 7.8 (An alternative design of presentation strategies) *Alternatively, it seems possible to reduce data d_o to a subselection given by the presentation action z and perform "ordinary", say simultaneous, design. By doing it for several z's, we can compare the achieved minima and present those quantities corresponding to the smallest one. It avoids the cumbersome upper bounds but needs several simultaneous designs to be performed in parallel. It is expected that a carefully proposed strategy of selecting the compared presentation variants will lead to a feasible algorithm.*

7.3.2 Stimulating the operator

When designing a signaling strategy, we assume that the academic, industrial or simultaneous design of the ideal pdf ${}^{\lfloor I}f(d(\mathring{t}))$ has been performed. Then, the model relating the signaling action s_t to the response of the o-system becomes

$$\begin{aligned}
{}^{\lfloor I}f(d_t, s_t|d(t-1)) &\equiv {}^{\lfloor I}f(d_t|s_t, d(t-1))\, {}^{\lfloor I}f(s_t|d(t-1)), \; s_t \in s^* \equiv \{0, 1\} \\
{}^{\lfloor I}f(d_t|s_t = 0, d(t-1)) &\equiv f(d_t|d(t-1)) \\
&\equiv \text{ learnt mixture describing the unguided o-system} \\
{}^{\lfloor I}f(d_t|s_t = 1, d(t-1)) &\equiv {}^{\lfloor I}f(d_t|d(t-1)) \\
&= \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|d(t-1))\, {}^{\lfloor I}f(d_t|d(t-1), c_t) \\
{}^{\lfloor I}f(d_t|d(t-1), c_t) &\equiv f(d_t = \mathit{\Delta}_t|d(t-1), c_t) \text{ for the academic design}
\end{aligned}$$

$$\begin{aligned}
{}^{\lfloor I}f(d_t|d(t-1),c_t) &\equiv f(\mathit{\Delta}_t|u_{o;t},d(t-1),c_t)\,{}^{\lfloor I}f(u_{o;t}|d(t-1))\\
&\quad\ \text{for the industrial design}\\
{}^{\lfloor I}f(d_t|d(t-1),c_t) &\equiv f(\mathit{\Delta}_t|u_{o;t},d(t-1),c_t)\,{}^{\lfloor I}f(u_{o;t}|d(t-1),c_t)\\
&\quad\ \text{for the simultaneous design}\\
f(\mathit{\Delta}_t|u_{o;t},d(t-1),c_t) &\equiv \text{the conditional pdf computed from}\\
&\quad\ c_t\text{th estimated component}\\
{}^{\lfloor I}f(u_{o;t}|d(t-1)) &\equiv \text{the conditional pdf from the industrial design}\\
{}^{\lfloor I}f(u_{o;t}|d(t-1),c_t) &\equiv \text{the conditional pdf from the simultaneous design}\\
{}^{\lfloor I}f(c_t|d(t-1)) &\equiv \alpha_{c_t} \equiv \text{ component weight for the industrial design}\\
{}^{\lfloor I}f(c_t|d(t-1)) &\equiv \text{the pf from the academic or simultaneous design.}
\end{aligned}$$

The exact fully probabilistic design of the optimal signaling strategy is hopeless at least because of the mixture form of the used models and the rational form needed when d^*_{p+} is nonempty. This makes us to design directly an approximate strategy. Simplicity of the action space $s^* = \{0,1\}$ allows us to use the tighter bound of the ω-type; Proposition 7.9.

Proposition 7.20 (Signaling design with the ω-bound) *Let us consider that the academic, industrial or simultaneous design of the advisory system has provided the strategy $d^*(t-1) \rightarrow (c^*_t, u^*_{o;t})$ determining the ideal pdf*

$$ {}^{\lfloor I}f(d_t|d(t-1)) = \sum_{c_t \in c^*} {}^{\lfloor I}f(d_t|d(t-1),c_t)\,{}^{\lfloor I}f(c_t|d(t-1)).$$

Let us assume that the conditions of Proposition 7.4 are met. Then, the following signaling strategy minimizes the ω-type bound, Proposition 7.9, on the KL divergence. The recursion run for $t = \mathring{t}, \mathring{t}-1, \ldots, 1$.

$$\begin{aligned}
f(s_t|d(t-1)) &\propto {}^{\lfloor U}f(s_t|d(t-1))\exp[-\omega_\gamma(s_t,d(t-1))], \qquad (7.51)\\
\omega_\gamma(s_t=0,d(t-1)) &\equiv \sum_{c\in c^*}\alpha_c\omega_\gamma(c,d(t-1))\\
\omega_\gamma(c,d(t-1)) &\equiv \int f(d_t|d(t-1),c)\\
&\times\left[\ln\left(\frac{f(d_{o;t}|d_{p+;t},d(t-1),c)}{{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right)+\omega_\gamma(d(t))\right]dd_t\\
\omega_\gamma(s_t=1,d(t-1)) &\equiv \sum_{c\in c^*}{}^{\lfloor I}f(c_t|d(t-1))\,{}^{\lfloor I}\omega_\gamma(c,d(t-1))\\
{}^{\lfloor I}\omega(c,d(t-1)) &\equiv \int {}^{\lfloor I}f(d_t|d(t-1),c)\\
&\times\left[\ln\left(\frac{{}^{\lfloor I}f(d_{o;t}|d_{p+;t},d(t-1),c)}{{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right)+\omega_\gamma(d(t))\right]dd_t\\
\omega_\gamma(d(t)) &\equiv {}^{\lfloor U}f(s_{t+1}=0|d(t))\omega_\gamma(s_{t+1}=0,d(t))+
\end{aligned}$$

$$
\begin{aligned}
&\quad + {}^{\lfloor U}f(s_{t+1}=1|d(t))\omega_{\gamma}(s_{t+1}=1,d(t))\\
&\omega_{\gamma}(d(\mathring{t}))=0.
\end{aligned}
$$

Proof. According to Proposition 7.4, the optimal strategy minimizing the KL divergence

$$
\mathcal{D}\left({}^{\lfloor I}f\middle|\middle|\,{}^{\lfloor U}f\right) \equiv \int {}^{\lfloor I}f(d(\mathring{t}),s(\mathring{t}))\ln\left(\frac{{}^{\lfloor I}f(d(\mathring{t}),s(\mathring{t}))}{{}^{\lfloor U}f(d(\mathring{t}),s(\mathring{t}))}\right) d(d(\mathring{t})s(\mathring{t}))
$$

to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}),s(\mathring{t}))$, fulfilling (7.17), has the form

$$
\begin{aligned}
{}^{\lfloor I}f(s_t|d(t-1)) &= {}^{\lfloor U}f(s_t|d(t-1))\frac{\exp[-\omega_{\gamma}(s_t,d(t-1))]}{\gamma(d(t-1))}, \quad \text{where}\\
\gamma(d(t-1)) &\equiv \sum_{s_t\in s^*} {}^{\lfloor U}f(s_t|d(t-1))\exp[-\omega_{\gamma}(s_t,d(t-1))]\\
\omega_{\gamma}(s_t=0,d(t-1)) &\equiv \int f(d_t|d(t-1))\ln\left(\frac{f(d_{o;t}|d_{p+;t},d(t-1))}{\gamma(d(t))\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t\\
\omega_{\gamma}(s_t=1,d(t-1)) &\equiv \int {}^{\lfloor I}f(d_t|d(t-1))\ln\left(\frac{{}^{\lfloor I}f(d_{o;t}|d_{p+;t},d(t-1))}{\gamma(d(t))\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t\\
\gamma(d(\mathring{t})) &= 1.
\end{aligned}
$$

If we find upper bounds on $\omega_{\gamma}(d(t-1))$ and insert them into definition of $\gamma(d(t))$ we get a lower bound on it and consequently the upper bound on the Bellman function $-\ln(\gamma(d(t)))$. Let us do that.

We use Proposition 7.6 and omit the term $\ln(K)$ that has no influence on optimization. It implies

$$
\begin{aligned}
&\omega_{\gamma}(s_t=0,d(t-1))\\
&\equiv \int f(d_t|d(t-1))\ln\left(\frac{f(d_{o;t}|d_{p+;t},d(t-1))}{\gamma(d(t))\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t\\
&\le \sum_{c\in c^*}\alpha_c\int f(d_t|d(t-1),c)\ln\left(\frac{f(d_{o;t}|d_{p+;t},d(t-1),c)}{\gamma(d(t))\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t\\
&\equiv \sum_{c\in c^*}\alpha_c\omega_{\gamma}(c,d(t-1)) \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
\omega_{\gamma}(s_t=1,d(t-1)) &\equiv \int {}^{\lfloor I}f(d_t|d(t-1))\\
&\times\ln\left(\frac{{}^{\lfloor I}f(d_{o;t}|d_{p+;t},d(t-1))}{\gamma(d(t))\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t\\
&\le \sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|d(t-1))\\
&\times\int {}^{\lfloor I}f(d_t|d(t-1),c_t)\ln\left(\frac{{}^{\lfloor I}f(d_{o;t}|d_{p+;t},d(t-1),c_t)}{\gamma(d(t))\,{}^{\lfloor U}f(d_{o;t}|d_o(t-1))}\right) dd_t \equiv
\end{aligned}
$$

$$\equiv \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|d(t-1))\, {}^{\lfloor I}\omega_\gamma(c_t, d(t-1)).$$

Substituting this upper bound into the definition of $\gamma(d(t))$ and using the inequality between weighted arithmetic and geometric means for a further bounding of $\gamma(d(t))$, we complete the last step giving the claimed result

$$\ln(\gamma(d(t))) \geq \sum_{c_t \in c^*} \left[{}^{\lfloor U}f(s_t = 0|d(t-1))\alpha_{c_t}\omega_\gamma(c_t, d(t-1)) \right.$$
$$\left. + \ {}^{\lfloor U}f(s_t = 1|d(t-1))\, {}^{\lfloor I}f(c_t|d(t-1))\, {}^{\lfloor I}\omega_\gamma(c_t, d(t-1)) \right].$$

□

The proved proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm written for the practically used state description with the state ϕ.

Algorithm 7.13 (Signaling with the ω-bound)

Initial (offline) mode

- *Estimate the mixture model* $f(d_t|d(t-1)) = \sum_{c \in c^*} \alpha_c f(d_t|\phi_{t-1}, c)$ *of the o-system with the state* ϕ_t*; Chapter 6.*
- *Specify the user's ideal pf* ${}^{\lfloor U}f(s_t|d(t-1)) = {}^{\lfloor U}f(s_t|\phi_{t-1})$ *on signaling actions* $s_t \in \{0, 1\}$.
- *Specify the user's ideal pdf on the recommended pointers* c_t*, o-innovations* $\Delta_{o;t}$ *and recognizable actions* $u_{o;t}$.
- *Select the length of the receding horizon* $T \geq 1$.

Sequential (online) mode, *running for* $t = 1, 2, \ldots,$

1. *Acquire the data record* d_t *and determine the state vector* ϕ_t.
2. *Update the estimates of the model parameters if you deal with the adaptive advisory system.*
3. *Perform the academic, industrial or simultaneous design generating the ideal pdf*
$$ {}^{\lfloor I}f(d_t|\phi_{t-1}) = \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|\phi_{t-1})\, {}^{\lfloor I}f(d_t|\phi_{t-1}, c_t).$$
4. *Initialize the iterative mode by setting* $\tau = t + T$ *and* $\omega_\gamma(\phi_\tau) = 0$.

   Iterative mode

   a) *Evaluate the* <u>*functions*</u>

   $${}^{\lfloor I}f(c_\tau, u_{o;\tau}|d(\tau-1)) = {}^{\lfloor I}f(c_\tau, u_{o;\tau}|\phi_{\tau-1})$$

   using the academic, industrial or simultaneous design that gives

   $${}^{\lfloor I}f(d_\tau|\phi_{\tau-1}) = \sum_{c_\tau \in c^*} {}^{\lfloor I}f(c_\tau|\phi_{\tau-1})\, {}^{\lfloor I}f(d_\tau|\phi_{\tau-1}, c_\tau)$$

   $${}^{\lfloor I}f(d_\tau|\phi_{\tau-1}, c_\tau) \equiv f(\Delta_\tau|u_{o;\tau}, \phi_\tau, c_\tau)\, {}^{\lfloor I}f(u_{o;\tau}|\phi_\tau, c_\tau).$$

b) Determine the following functions, $c \in c^*$,

$$\omega_\gamma(c, \phi_{\tau-1})$$

$$\equiv \int f(d_\tau|\phi_{\tau-1}, c)\left[\ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau}, \phi_{\tau-1}, c)}{{}^{\lfloor U}f(d_\tau|\phi_{\tau-1})}\right) + \omega_\gamma(\phi_\tau)\right] dd_\tau$$

$${}^{\lfloor I}\omega_\gamma(c_\tau, \phi_{\tau-1}) \equiv \int {}^{\lfloor I}f(d_\tau|\phi_{\tau-1}, c_\tau)$$

$$\times \left[\ln\left(\frac{{}^{\lfloor I}f(d_{o;\tau}|d_{p+;\tau}, \phi_{\tau-1}, c_\tau)}{{}^{\lfloor U}f(c_\tau, d_{o;\tau}|\phi_{\tau-1})}\right) + \omega_\gamma(\phi_\tau)\right] dd_\tau$$

$$\omega_\gamma(\phi_\tau) \equiv {}^{\lfloor U}f(s_{\tau+1} = 0|\phi_\tau)\omega_\gamma(s_{\tau+1} = 0, \phi_\tau)$$

$$+ {}^{\lfloor U}f(s_{\tau+1} = 1|\phi_\tau)\omega_\gamma(s_{\tau+1} = 1, \phi_\tau), \quad \textit{where}$$

$$\omega_\gamma(s_{\tau+1} = 0, \phi_\tau) \equiv \sum_{c_{\tau+1} \in c^*} \alpha_{c_{\tau+1}}\omega_\gamma(c_{\tau+1}, \phi_\tau)$$

$$\omega_\gamma(s_{\tau+1} = 1, \phi_\tau) \equiv \sum_{c_{\tau+1} \in c^*} {}^{\lfloor I}f(c_{\tau+1}|\phi_\tau \; {}^{\lfloor I}\omega_\gamma(c_{\tau+1}, \phi_\tau).$$

c) Continue if $\tau = t + 1$. Otherwise decrease $\tau = \tau - 1$ and go to the beginning of Iterative mode*.*

5. *Evaluate the optimal signaling strategy*

$${}^{\lfloor I}f(s_{t+1}|\phi_t) \propto {}^{\lfloor U}f(s_{t+1}|\phi_t)\exp[-\omega_\gamma(s_{t+1}, \phi_t).$$

6. *Make the operator alert to perform recommended actions if ${}^{\lfloor I}f(s_{t+1} = 0|\phi_t)$ is close to zero. Otherwise let him proceed as usual.*
7. *Go to the beginning of* Sequential mode*.*

Problem 7.9 (Order of presentation and signaling designs) *The design of the signaling strategy has to be done after finishing the academic, industrial or simultaneous design. It is, however, unclear whether to perform it after specification of priorities or whether priorities should be decided after the decision that the operator should make some actions.*

The former version is more realistic as it measures possible improvements limited by the restricted ability of human being to grasp information. The latter case is less conservative and more sensitive to possible problems that need not be manageable by the operator. The higher sensitivity and simplicity of the design make us select the latter case. The alternative solution has to be inspected in the future.

Problem 7.10 (Approximation of Bellman functions) *Practically, the evaluation of Bellman functions in all design variants has to be reduced to algebraic manipulations. It seems that, in addition to minimization of upper bounds of the KL divergence, it is possible to exploit upper bounds resulting from insertion of nonoptimal actions during dynamic programming; cf. Proposition 7.9. Some results of this type are also in Chapter 9. This promising methodology should be, however, inspected systematically.*

7.4 Design validation

The fullscale use of the advisory system is the only decisive validation test. Naturally, a collection of indicators is needed in the design phase for checking the chance of the final success and for early recognition of troubles. The following, definitely incomplete, list of ideas has been collected and partially elaborated up to now.

- The model validation has to be successful; see Section 6.7.
- The estimated model has to be reliable as a long horizon predictor; see Section 7.1.2. If need be, the model estimate has to be stabilized as described in Algorithm 7.1 and tested for the quality of predictions.
- The KL divergence of the estimated model of the o-system to the user's ideal pdf should be close to a sample version of the KL divergence obtained from the measured data. The construction of the sample version is a generally nontrivial task for continuous valued data since logarithms of Dirac delta functions should be dealt with. A plausible solution is, however, available for the normal case that is of our primary interest; see Section 9.1.4. Note that the analysis of the sample version of the KL divergence should be done for both components and the overall mixture; see item 8 of Remarks 9.8.
- The advisory system should behave properly when an artificial closed loop is created with the estimated model in the role of the o-system. This test can be implemented only when recognizable actions are present, when the industrial or simultaneous design is tested.
- Open-loop advices, generated by the p-system even if fed into the o-system, should be reasonably close to the good practice reflected in learning data. It has an important psychological aspect with respect to users: they surely will not be ready to change radically their current habits.
- All tests above should be robust with respect to small changes of tuning knobs used both in learning and design.

Problem 7.11 (Systematic design validation) *Design validation is even harder that the learning validation. Thus, it is not surprising that it still needs a systematic solution. Even partial progress is highly desirable.*

8

Learning with normal factors and components

Factors form basic building blocks of components. The feasibility of evaluations outlined in Chapter 6 singles out *normal factors*, *Markov-chain factors* (Chapter 10), and so-called *mean tracking (MT)* factors (Chapter 12) as suitable candidates. This chapter elaborates on the operations needed for normal factors in detail. Operations outlined in Chapter 6 are specialized to normal factors and to Gauss-inverse-Wishart pdf (GiW), that is their conjugate pdf. Numerically stable learning algorithms are developed based on $L'DL$ or LDL' decompositions of the extended information matrix forming a decisive part of the sufficient statistics describing normal factors. Aspects related to the *normal components* consisting solely of normal factors are also discussed.

The normal parameterized factor that we deal with predicts a real-valued quantity $d_t \equiv d_{i;t},\ i \in \{1, \ldots, \mathring{d}\}$ by the pdf

$$f(d_t|d_{(i+1)\cdots\mathring{d};t}, d(t-1), \Theta) = \mathcal{N}_{d_t}(\theta'\psi_t, r), \text{ where} \tag{8.1}$$

$\Theta\ \equiv [\theta, r] \equiv$ [regression coefficients, noise variance] $\in \Theta^*$
$\Theta^* \subset$ ($\mathring{\psi}$-dimensional, nonnegative) real quantities, and
ψ is the regression vector,

$$\begin{aligned}\mathcal{N}_d(\theta'\psi, r) &\equiv (2\pi r)^{-0.5} \exp\left\{-\frac{(d-\theta'\psi)^2}{2r}\right\} \\ &= (2\pi r)^{-0.5} \exp\left\{-\frac{1}{2r}\operatorname{tr}\left(\Psi\Psi'[-1,\theta']'[-1,\theta']\right)\right\},\end{aligned} \tag{8.2}$$

$\Psi \equiv [d, \psi']' \equiv$ data vector, tr is the matrix trace, $'$ denotes transposition.

The layout of Chapter 6 is more or less copied. Section 8.1 prepares common tools used throughout the chapter. Preprocessing of raw data observed by the p-system is described in Section 8.2. Specific elicitation of prior knowledge is treated in Section 8.3. Even with such knowledge available, a proper specification of the prior pdf is nontrivial. An extensive discussion of its construction is in Section 8.4. Then, specialization of the approximate estimation to the normal mixtures follows; Section 8.5. Structure estimation is elaborated in Section 8.6. The concluding Section 8.6 covers model validation. It forms a bridge to Chapter 9, which treats design of the advising strategies.

8.1 Common tools

This section collects common technical tools used for learning with mixtures consisting solely of normal factors. For these factors, learning reduces to algebraic recursions so that the tools concern mainly matrix operations.

8.1.1 Selected matrix operations

Proposition 8.1 (Some matrix formulas) *Let the involved matrices A, B, C have compatible dimensions and the used inversions exist. Then,*

$$\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA) \tag{8.3}$$

$$\operatorname{tr}(A) = \operatorname{tr}(A'),\ |A| = |A'|,\ |A^{-1}| = |A|^{-1} \tag{8.4}$$

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1} \tag{8.5}$$

$$\frac{\partial}{\partial A}\operatorname{tr}(AB) = B',\ \frac{\partial}{\partial A}\ln(|A|) = A^{-1} \tag{8.6}$$

Proof.

Eqn. (8.3): $\operatorname{tr}(ABC) \equiv \sum_{i\in i^*}(ABC)_{ii} = \sum_{i\in i^*}\sum_{k\in k^*}(AB)_{ik}C_{ki}$

$$= \sum_{k\in k^*}\left(\sum_{i\in i^*} C_{ki}(AB)_{ik}\right) = \sum_{k\in k^*}(CAB)_{kk} \equiv \operatorname{tr}(CAB).$$

Eqn. (8.4): $\operatorname{tr}(A) \equiv \sum_{i\in i^*} A_{ii} = \sum_{i\in i^*}(A')_{ii} \equiv \operatorname{tr}(A')$.

Let λ be eigenvalue of A, i.e., there is a vector $x \neq 0$ such that $Ax = \lambda x$. It implies that x' solves the equation $x'A' = \lambda x'$. Thus, eigenvalues of A and A' coincide and consequently $|A| = |A'|$. Similarly, for nonzero eigenvalues $\lambda(A)$ of the regular matrix A, we get $\lambda(A^{-1}) = \frac{1}{\lambda(A)}$. Thus, $|A^{-1}| = \frac{1}{|A|}$.

Eqn. (8.5): $(A+BCD)\left[A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}\right]$

$$\begin{aligned}
&= I + BCDA^{-1} - B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}\\
&\quad -BCDA^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1} \underbrace{=}_{DA^{-1}B=-C^{-1}+C^{-1}+DA^{-1}B}\\
&= I + BCDA^{-1} - B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}\\
&\quad -BCDA^{-1} + B(C^{-1}+DA^{-1}B)^{-1}DA^{-1} = I.
\end{aligned}$$

Eqn. (8.6): $\frac{\partial}{\partial A_{ij}}\operatorname{tr}(AB) = \frac{\partial}{\partial A_{ij}}\left[\sum_{i\in i^*,j\in j^*} A_{ij}B_{ji}\right] = B_{ji} = (B')_{ij}$.

Let ${}^{\lfloor kl}A$ be the complement to the entry A_{kl}, i.e., determinant of the matrix obtained from A by cancelling the kth row and lth column multiplied by $(-1)^{k+l}$. Then, $|A| = \sum_{kl} A_{kl} \, {}^{\lfloor kl}A$ and

$$\frac{\partial}{\partial A_{ij}} \ln |A| = \frac{{}^{\lfloor ij}A}{|A|} = \left(A^{-1}\right)_{ij} .$$

□

8.1.2 $L'DL$ decomposition

A positive definite matrix V, called the *extended information matrix*, forms the decisive part of the sufficient statistic in estimation of normal factors; see Section 8.1.5. Often, it is poorly conditioned and the use of its $L'DL$ *decomposition* is the only safe way how to counteract the induced numerical troubles. We collect the facts used explicitly in this text. For a broader view see [159].

Agreement 8.1 ($L'DL$ decomposition) *Let V be a positive definite symmetric matrix with $\mathring{\Psi}$ rows. Let us suppose that $V = L'DL$, where L is a* lower triangular matrix *with a unit diagonal, D is a* diagonal matrix *with positive diagonal entries. This expression is called $L'DL$* decomposition *of V.*

Proposition 8.2 (Existence and uniqueness of $L'DL$ decomposition) *Let V be a positive definite symmetric matrix with $\mathring{\Psi}$ rows. Then, its $L'DL$ decomposition exists and it is unique. It is constructed as follows.*

```
For $j = \mathring{\Psi}, \ldots, 1$
          Set $L_{jj} = 1$
  For $i = \mathring{\Psi}, \ldots, j+1$
          Set $s = V_{ij}$
    For $k = i+1, \ldots, \mathring{\Psi}$
          Set $s = s - L_{ki} D_{kk} L_{kj}$
    end of the cycle over $k$
          Set $L_{ij} = s/D_{ii}$
  end of the cycle over $i$
          Set $s = V_{jj}$
  For $k = j+1, \ldots, \mathring{\Psi}$
          Set $s = s - L_{kj}^2 D_{kk}$
  end of the cycle over $k$
          Set $D_{jj} = s$
end of the cycle over $j$
```

Proof. The algorithm is directly implied by the desired identity $V = L'DL$ written entrywise. The algorithm fails if some $D_{jj} \leq 0$. Let us assume that such D_{jj} occurs and denote ${}^{\lfloor j}L$ the part of the matrix L found up to this moment. Let ${}^{\lfloor j}x$ solve the triangular equation ${}^{\lfloor j}x'\,{}^{\lfloor j}L' = [\underbrace{1, 0, \ldots, 0}_{lenght\ j}]$. Then, the quadratic form $x'Vx$ is not positive for $x = \left[0', {}^{\lfloor j}x'\right]'$. This contradicts the assumed definiteness of V. □

Remark(s) 8.1
It is worth stressing that the derived algorithm is useful for proving existence and uniqueness. It may fail numerically in evaluating entries D_{ii} when the matrix is poorly conditioned. We do not use it practically.

Updating $L'DL$ decomposition

The meaningful use of the $L'DL$ decomposition depends on our ability to convert the recursion $V = \lambda V + w\Psi\Psi'$ into a direct, numerically robust, recursion for matrices L and D; Proposition 8.2. This recursion, with Ψ equal to the data vector and scalar positive weights λ, w, arises in estimation; cf. Proposition 8.11. The direct updating is based on the *algorithm dydr* [146] that makes it efficiently.

Proposition 8.3 (Algorithm dydr) *The positive semi-definite matrix of the rank 2 can be given the two equivalent forms defining the same kernels of a quadratic form*

$$[a, b] \begin{bmatrix} D_a & 0 \\ 0 & D_b \end{bmatrix} [a, b]' = [c, d] \begin{bmatrix} D_c & 0 \\ 0 & D_d \end{bmatrix} [c, d]'. \tag{8.7}$$

There, D_a, D_b, D_c, D_d are positive scalars and a, b, c, d are column vectors of the same length $\mathring{a} \geq 2$.

Let the following entries of c, d be fixed at values

$$c_j \equiv 1, \;\; d_j \equiv 0, \quad \textit{for a chosen } j \in \{1, \ldots, \mathring{a}\}. \tag{8.8}$$

Then, the right-hand side matrix in (8.7) is determined uniquely by the left-hand side through the following algorithm.

Algorithm 8.1 (Dyadic reduction: dydr)

1. *Set $D_c = a_j^2 D_a + b_j^2 D_b$.*
2. *Set $x = \frac{a_j D_a}{D_c}$, $y = \frac{b_j D_b}{D_c}$.*
3. *Set $D_d = \frac{D_a D_b}{D_c}$.*
4. *Evaluate the remaining entries of c, d as follows. $c_i = xa_i + yb_i$, $d_i = -b_j a_i + a_j b_i$ for $i = 1, \ldots, \mathring{a}$. It gives $c_j = 1$, $d_j = 0$.*

The Matlab *script of* dydr *may look as follows.*

```
function [Dc,Dd,c,d]=dydr(Da,Db,a,b,j)
%
% dydr makes dyadic reduction of the positive definite
%           quadratic form [a,b]diag(Da,Db)[a,b]' to the
%           quadratic form [c,d]diag(Dc,Dd)[c,d]' so that
%           jth entry of c equals 1, jth entry of d is zero
 aj=a(j,1); bj=b(j,1); Dc=aj^(2)*Da+bj^(2)*Db;
 if Dc<machine_precision, 'not positive definite', return, end
 x =aj*Da/Dc; y=bj*Db/Dc; Dd=Da*Db/Dc;
 c=x*a+y*b; d=bj*a+aj*b;
 c(j,1)=1; d(j,1)=0;
```

Proof. Let us take for simplicity $j = 1$ and let $T = \begin{bmatrix} x & -b_1 \\ y & a_1 \end{bmatrix}$ be a regular matrix determined by free scalars x, y. It has its second column orthogonal to the row vector $[a_1, b_1]$. Thus, $[c, d] \equiv [a, b]T$ has $d_1 = 0$. The remaining requirement in (8.8) is fulfilled if $a_1x + b_1y = 1$. It also makes the determinant of T equal to 1. The inspected decomposed form of the positive semi-definite matrix (8.7) is preserved after such a transformation of $[a, b]$ iff

$$\begin{bmatrix} D_c & 0 \\ 0 & D_d \end{bmatrix} = T^{-1} \begin{bmatrix} D_a & 0 \\ 0 & D_b \end{bmatrix} (T^{-1})', \; T^{-1} = \begin{bmatrix} a_1 & b_1 \\ -y & x \end{bmatrix}.$$

The diagonal form is preserved iff $-ya_1D_a + xb_1D_b = 0$. This together with the former condition $a_1x + b_1y = 1$ determines uniquely y, x and consequently all other elements

$$D_c = a_1^2 D_a + b_1^2 D_b, \; x = \frac{a_1 D_a}{D_c}, \; y = \frac{b_1 D_b}{D_c}, \; D_d = \frac{D_a D_b}{D_c}$$
$$c = xa + yb, \; d = -b_1 a + a_1 b.$$

□

With the algorithm dydr, the rank-one updating $V = \lambda V + w\Psi\Psi'$ of $L'DL$ decomposition of the extended information matrix V is straightforward.

Algorithm 8.2 (Rank-one updating of $L'DL = \lambda L'DL + w\Psi\Psi'$)

```
    Set $b = \Psi$, $D_d = w$
For $j = \mathring{\Psi}, \ldots, 1$
    $[D_j, D_d, jth\ column\ of\ L', b]$ = dydr $(\lambda * D_j, D_d, jth\ column\ of\ L', b, j)$
end  of the cycle over $j$
```

Conversion of $L'DL$ to LDL'

Learning uses mostly the $L'DL$ decomposition; Agreement 8.1. For some ready structure estimation algorithms [93], it is useful to deal with the LDL' decomposition. It is given, again uniquely, by the lower triangular matrix L with unit diagonal and by the diagonal matrix D with positive diagonal entries. The algorithm dydr serves well for the conversion $L'DL \to LDL'$.

Algorithm 8.3 (Conversion of $L'DL$ to LDL')

```
For  $j = \mathring{\Psi} - 1, \ldots, 1$
  For  $i = j + 1, \ldots, \mathring{\Psi}$
    $[D_j, D_i, ith\ column\ of\ L, jth\ column\ of\ L]$
          $= \mathsf{dydr}\,(D_j, D_i, ith\ column\ of\ L, jth\ column\ of\ L, j)$
  end  of the cycle over i
end  of the cycle over j
```

Permutation of variables in quadratic forms

We permute entries of regression vectors ψ when evaluating marginal pdfs and when estimating the structure of normal factors; see Section 8.6. This action induces permutations of regression coefficients θ. The decisive quadratic form $[-1,\, \theta']V[-1,\, \theta']'$ determining GiW pdf, Subsection 8.1.3, remains unchanged if the corresponding rows and columns of V are permuted, too. It spoils, however, LDL' or $L'DL$ decompositions of V. The following propositions describe simple algorithms that recover these decompositions after permuting adjacent entries of ψ and thus of θ. More complex permutations are obtained through a sequence of such elementary permutations.

Proposition 8.4 (Permutation of adjacent entries in LDL') *Let $V = LDL'$ be the decomposition of the extended information matrix corresponding to the data $\Psi = [d, \psi']'$ and parameter $[-1,\, \theta']'$ vectors. Let*

$$\psi = [h', a, b, c']',\ \tilde{\psi} = [h',\ \underbrace{b, a}_{permuted}\ , c']', \qquad \text{with scalar } a,\, b, \tag{8.9}$$

$$\theta = [\theta_h', \theta_a, \theta_b, \theta_c']',\ \tilde{\theta} = [\theta_h',\ \underbrace{\theta_b, \theta_a}_{permuted}\ , \theta_c']', \quad \text{with scalar } \theta_a,\, \theta_b.$$

Then, the LDL' decomposition of the matrix $\tilde{V} = \tilde{L}\tilde{D}\tilde{L}'$ preserving the quadratic form $[-1,\, \theta']V[-1,\, \theta']' = [-1,\, \tilde{\theta}']\tilde{V}[-1,\, \tilde{\theta}']'$ is obtained from the decomposition $V = LDL'$ by the following algorithm.

1. *Permute rows of L corresponding to regressors a and b up to the main subdiagonal entry of the shorter one.*
2. *Store $D_{ao} = D_a =$ entry of D corresponding to the regressor a.*
3. *Recompute $D_a = D_b + \omega^2 D_a$, where ω is the scalar on the subdiagonal of the longer vector of the permuted rows.*
4. *Compute auxiliary quantities $x = \omega D_{ao}/D_a,\ y = D_b/D_a$.*

5. *Recompute* $D_b = D_{ao}D_b/D_a =$ *entry of* D *corresponding to the regressor* b.
6. *Store ath column* L_a *of* L *into* L_{ao} *starting from the entry below* ω.
7. *Recompute* $L_a = xL_a + yL_b$, *where* L_b *contains entries in the adjacent entries of the column* b.
8. *Recompute* $L_b = L_{ao} - \omega L_b$.
9. *Correct* $\omega = x$.

Proof. The permutation of entries a, b in the regression vector implies the permutation of the corresponding pair of columns and rows in the symmetric matrix V. In the equivalent LDL' decomposition, it changes

$$D = \begin{bmatrix} D_h & & & \\ & D_a & & \\ & & D_b & \\ & & & D_c \end{bmatrix}, \; L = \begin{bmatrix} L_h & 0 & 0 & 0 \\ L_{ha} & 1 & 0 & 0 \\ L_{hb} & \omega & 1 & 0 \\ L_{hc} & L_a & L_b & L_c \end{bmatrix}, \quad \text{with scalars } D_a,\ D_b,\ \omega, \quad \text{to}$$

$$\underline{D} = D,\ \underline{L} = \begin{bmatrix} L_h & 0 & 0 & 0 \\ L_{hb} & \omega & 1 & 0 \\ L_{ha} & 1 & 0 & 0 \\ L_{hc} & L_a & L_b & L_c \end{bmatrix}. \tag{8.10}$$

We search for $\tilde{D}$, $\tilde{L}$ such that quadratic forms $[-1,\ \tilde{\theta}']\tilde{L}\tilde{D}\tilde{L}'[-1,\ \tilde{\theta}']'$ and $[-1,\ \theta']'LDL'[-1,\ \theta']'$ are equal. It is achieved when $\underline{D}$ and $\underline{L}$ coincide with $\tilde{D}$ and $\tilde{L}_h$ with exception of columns having indexes $_a$, $_b$. For them, it must hold

$$\begin{bmatrix} \omega & 1 \\ 1 & 0 \\ L_a & L_b \end{bmatrix} \begin{bmatrix} D_a & \\ & D_b \end{bmatrix} \begin{bmatrix} \omega & 1 \\ 1 & 0 \\ L_a & L_b \end{bmatrix}' = \begin{bmatrix} 1 & 0 \\ \tilde{\omega} & 1 \\ \tilde{L}_a & \tilde{L}_b \end{bmatrix} \begin{bmatrix} \tilde{D}_a & \\ & \tilde{D}_b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \tilde{\omega} & 1 \\ \tilde{L}_a & \tilde{L}_b \end{bmatrix}'. \tag{8.11}$$

We proceed similarly as in the derivation of dydr but we exploit the special form of entries a_j, b_j. Let us consider the regular (2,2)-matrices $T = \begin{bmatrix} x & 1 \\ y & -\omega \end{bmatrix}$ parameterized by scalars x, y. Their second column is orthogonal to the vector occurring in the first row of the first matrix in (8.11). It holds

$$\begin{bmatrix} \omega & 1 \\ 1 & 0 \\ L_a & L_b \end{bmatrix} T = \begin{bmatrix} x\omega + y & 0 \\ x & 1 \\ xL_a + yL_b & L_a - \omega L_b \end{bmatrix}.$$

The desired “tilde” form is obtained if

$$\tilde{L}_a = xL_a + yL_b,\ \tilde{L}_b = L_a - \omega L_b$$
$$x\omega + y = 1,\ T^{-1}\text{diag}[D_a, D_b]\left(T^{-1}\right)' = \text{ diagonal matrix.}$$

The operator *diag*$[\cdot]$ converts the vector argument into the diagonal matrix with the vector placed on the diagonal and the matrix argument into the vector containing its diagonal.

The unspecified values x, y have to be chosen so that the last two constraints are fulfilled. It gives

$$\tilde{D}_a = D_b + \omega^2 D_a,\ x = \frac{\omega D_a}{\tilde{D}_a},\ y = \frac{D_b}{\tilde{D}_a},\ \tilde{D}_b = \frac{D_a D_b}{\tilde{D}_a}.$$

□

We also need to recover $L'DL$ decomposition after permuting adjacent entries. For the sake of completeness, we summarize it.

Proposition 8.5 (Permutation of adjacent entries in $L'DL$) *Let $V = L'DL$ be the $L'DL$ decomposition of the information matrix V corresponding to the data $\Psi = [d, \psi']'$ and parameter $[-1,\ \theta']'$ vectors. Let*

$$\psi = [h', a, b, c']',\ \tilde{\psi} = [h', b, a, c']' \text{ and } \theta = [\theta_h', \theta_a, \theta_b, \theta_c']',\ \tilde{\theta} = [\theta_h', \theta_b, \theta_a, \theta_c']' \tag{8.12}$$

with scalars a, b, θ_a, θ_b. Then, the $L'DL$ decomposition of $\tilde{V} = \tilde{L}'\tilde{D}\tilde{L}$ describing the same quadratic form $[-1,\ \theta']V[-1,\ \theta']' = [-1,\ \tilde{\theta}']\tilde{V}[-1,\ \tilde{\theta}']'$ can be obtained from the decomposition $V = L'DL$ by the following algorithm.

1. *Permute a and b columns of L up to main subdiagonal entry of the shorter one.*
2. *Store the $D_{ao} = D_a =$ entry of D corresponding to the regressor a.*
3. *Recompute $D_a = D_b + \omega^2 D_a$; ω denotes the second nonzero entry in the longer nonzero part of the permuted columns.*
4. *Compute auxiliary quantities $x = \omega D_{ao}/D_a,\ y = D_b/D_a$.*
5. *Recompute the $D_b = D_{ao}D_b/D_a$ entry of D corresponding to the regressor b.*
6. *Store ath row L_a of L into L_{ao} until the entry before ω.*
7. *Recompute $L_a = xL_a + yL_b$, where L_b contains entries in the adjacent entries of the row b.*
8. *Recompute $L_b = L_{ao} - \omega L_b$.*
9. *Correct $\omega = x$.*

Proof. The idea is identical with the proof of Proposition 8.4. The roles of rows and columns are just reversed in corrections. □

Complete squares and minimization of quadratic forms

Often, we deal with quadratic forms $[-1,\ \theta']V[-1,\ \theta]'$ in a $\mathring{\psi}+1$-vector formed by -1 and regression coefficients θ. The kernel of this form is the extended information matrix V, which is by construction positive definite. We assume that the $L'DL$ decomposition of V is available, $V = L'DL$.

Let us split the matrix V and its $L'DL$ decomposition as follows.

$$V = \begin{bmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}V' \\ {}^{\lfloor d\psi}V & {}^{\lfloor \psi}V \end{bmatrix}, \ {}^{\lfloor d}V \text{ is scalar,}$$

$$L = \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix}, \quad D = \begin{bmatrix} {}^{\lfloor d}D & 0 \\ 0 & {}^{\lfloor \psi}D \end{bmatrix}, \ {}^{\lfloor d}D \text{ is scalar.} \qquad (8.13)$$

From here onwards, when working with quadratic forms, the *upper left index in* V, D, L indicates blocks obtained by splitting of the matrix according to the element ordering in the corresponding data vector.

Proposition 8.6 (Completion of squares) *It holds*

$$[-1,\, \theta']V[-1,\, \theta']' \equiv [-1,\, \theta']L'DL[-1\, \theta']' = (\theta - \hat{\theta})'\, {}^{\lfloor \psi}L'\, {}^{\lfloor \psi}D\, {}^{\lfloor \psi}L(\theta - \hat{\theta}) + {}^{\lfloor d}D$$

$$\hat{\theta} \equiv {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor d\psi}L \equiv \text{least-squares (LS) estimate } \textit{of}\, \theta, \qquad (8.14)$$

$${}^{\lfloor d}D \equiv \text{least-square remainder.}$$

This quadratic form is minimized by $\theta = \hat{\theta}$. ${}^{\lfloor d}D$ *is to the minimum reached.*

Proof. The straightforward vector-matrix multiplication for the split $L'DL$ decomposition proofs equality (8.14). It is known as *completion of squares.* Due to the positive definiteness of V, the term dependent of θ is nonnegative and becomes zero for $\theta = \hat{\theta}$. □

8.1.3 *GiW* pdf as a conjugate prior

Normal factors belong to the exponential family (see Section 3.2) so that they possess conjugate (self-reproducing) prior. The following correspondence to the general form of the exponential family (3.6) holds

$$\mathcal{N}_d(\theta'\psi, r) = A(\Theta)\exp[\langle B(\Psi), C(\Theta)\rangle] \ \text{ with} \qquad (8.15)$$
$$A(\Theta) \equiv (2\pi r)^{-0.5}, \ B(\Psi) \equiv \Psi\Psi', \ D(\Psi) = 0.$$
$$C(\Theta) = (2r)^{-1}[-1, \theta']'[-1, \theta'], \ \ \langle B, C\rangle \equiv \operatorname{tr}[B'C].$$

This correspondence determines the conjugate prior (3.13) in the form known as *Gauss-inverse-Wishart pdf (GiW)*,

$$GiW_\Theta(V, \nu)$$
$$\equiv GiW_{\theta,r}(V, \nu) \equiv \frac{r^{-0.5(\nu + \mathring{\psi} + 2)}}{\mathcal{I}(V, \nu)} \exp\left\{-\frac{1}{2r}\operatorname{tr}\left(V[-1, \theta']'[-1, \theta']\right)\right\}. \ (8.16)$$

The value of the normalization integral $\mathcal{I}(V, \nu)$ and constraints on statistics V, ν that guarantee finiteness of $\mathcal{I}(V, \nu)$ are described below, together with other properties of this important pdf.

The $(\mathring{\Psi}, \mathring{\Psi})$-dimensional *extended information matrix* V can be chosen to be symmetric. Its potential antisymmetric constituents disappear in the quadratic form in which it occurs. Moreover, the extended information matrix must be positive definite. Otherwise, there are unbounded combinations of θ entries for which the inspected function (8.16) does not fall to zero and thus is not integrable.

Basic properties of the *GiW* pdf exploit the $L'DL$ decomposition of the extended information matrix, Agreement 8.1, and splitting (8.13).

Proposition 8.7 (Basic properties and moments of the *GiW* pdf)

1. $GiW_{\Theta}(V,\nu)$ has the following alternative expressions

$$\begin{aligned} GiW_{\Theta}(V,\nu) \equiv GiW_{\Theta}(L,D,\nu) &= \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(L,D,\nu)} \qquad (8.17)\\ &\times \exp\left\{-\frac{1}{2r}\left[\left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L\right)' {}^{\lfloor\psi}D\left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L\right) + {}^{\lfloor d}D\right]\right\}\\ &\equiv \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(L,D,\nu)} \exp\left\{-\frac{1}{2r}\left[(\theta-\hat{\theta})'C^{-1}(\theta-\hat{\theta}) + {}^{\lfloor d}D\right]\right\}\\ &\equiv \frac{r^{-0.5(\nu+\mathring{\psi}+2)}}{\mathcal{I}(L,D,\nu)} \exp\left\{-\frac{1}{2r}\left[Q(\theta) + {}^{\lfloor d}D\right]\right\}, \quad \textit{where}\\ \hat{\theta} &\equiv {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor d\psi}L \equiv \textit{ least-squares (LS) estimate of } \theta \qquad (8.18)\\ C &\equiv {}^{\lfloor\psi}L^{-1}\,{}^{\lfloor\psi}D^{-1}\left({}^{\lfloor\psi}L'\right)^{-1} \equiv \textit{ covariance of LS estimate} \qquad (8.19)\\ {}^{\lfloor d}D &\equiv \textit{least-squares remainder} \qquad (8.20)\\ Q(\theta) &\equiv \left(\theta-\hat{\theta}\right)' C^{-1}\left(\theta-\hat{\theta}\right) \qquad (8.21)\\ &\equiv \left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L\right)' {}^{\lfloor\psi}D\left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L\right). \end{aligned}$$

2. The normalization integral is

$$\begin{aligned} \mathcal{I}(L,D,\nu) &= \varGamma(0.5\nu)\,{}^{\lfloor d}D^{-0.5\nu}\left|{}^{\lfloor\psi}D\right|^{-0.5} 2^{0.5\nu}(2\pi)^{0.5\mathring{\psi}} \qquad (8.22)\\ \varGamma(x) &\equiv \int_0^{\infty} z^{x-1}\exp(-z)\,dz < \infty \quad \textit{for } x > 0. \end{aligned}$$

Thus, it is finite iff $\nu > 0$ and V is positive definite $\Leftrightarrow$ $D > 0$ (entrywise).

3. The GiW pdf has the following marginal pdfs and moments

$$\begin{aligned} f(r|L,D,\nu) &= \frac{r^{-0.5(\nu+2)}}{\mathcal{I}\left({}^{\lfloor d}D,\nu\right)}\exp\left[-\frac{{}^{\lfloor d}D}{2r}\right] \equiv iW_r\left({}^{\lfloor d}D,\nu\right)\\ &\equiv \text{inverse Wishart pdf} \qquad (8.23)\\ \mathcal{I}\left({}^{\lfloor d}D,\nu\right) &\equiv \frac{\varGamma(0.5\nu)}{\left(0.5\,{}^{\lfloor d}D\right)^{0.5\nu}} \equiv \textit{normalization of } f(r|L,D,\nu).\\ \mathcal{E}[r|L,D,\nu] &= \frac{{}^{\lfloor d}D}{\nu-2} \equiv \hat{r}, \quad \mathrm{var}[r|L,D,\nu] = \frac{2\hat{r}^2}{\nu-4} \end{aligned}$$

$$\begin{aligned}
\mathcal{E}[r^{-1}|L,D,\nu] &= \frac{\nu}{{}^{\lfloor d}D} \\
\mathcal{E}[\ln(r)|L,D,\nu] &= \ln\left({}^{\lfloor d}D\right) - \ln(2) - \frac{\partial \ln\left(\varGamma(0.5\nu)\right)}{\partial(0.5\nu)} \\
f(\theta|L,D,\nu) &= \mathcal{I}^{-1}(D,\nu) \\
&\times \left[1 + \left({}^{\lfloor d}D\right)^{-1} \left(\theta - \hat{\theta}\right)' \, {}^{\lfloor \psi}L' \, {}^{\lfloor \psi}D \, {}^{\lfloor \psi}L \left(\theta - \hat{\theta}\right)\right]^{-0.5(\nu+\mathring{\psi})} \\
\mathcal{I}(D,\nu) &\equiv \frac{{}^{\lfloor d}D^{0.5\mathring{\psi}} \prod_{i=1}^{\mathring{\psi}} \varGamma(0.5(\nu+\mathring{\psi}-i))}{\left|\,{}^{\lfloor \psi}D\right|^{0.5}} \\
&\equiv \textit{normalization integral of } f(\theta|L,D,\nu) \\
\mathcal{E}[\theta|L,D,\nu] &= {}^{\lfloor \psi}L^{-1} \, {}^{\lfloor d\psi}L \equiv \hat{\theta} \\
\operatorname{cov}[\theta|L,D,\nu] &= \frac{{}^{\lfloor d}D}{\nu-2} \, {}^{\lfloor \psi}L^{-1} \, {}^{\lfloor \psi}D^{-1} \left({}^{\lfloor \psi}L'\right)^{-1} \equiv \hat{r}C.
\end{aligned}$$

Proof. The majority of evaluations can be found in [69]. Here we fix the most important steps and evaluate the nonstandard quantities $\mathcal{E}[\ln(r)|L,D,\nu]$ and $\mathcal{E}\left[r^{-1}|L,D,\nu\right]$ needed in Proposition 8.9.

1. The alternative form of *GiW* pdf relies on the completion of squares as described in Proposition 8.6.
2. Let us put substitutions used in compound brackets within the sequence of evaluations. Then, the normalization integral is evaluated as follows.

$$\begin{aligned}
\mathcal{I}(L,D,\nu) &\equiv \int r^{-0.5(\nu+\mathring{\psi}+2)} \\
&\times \exp\left\{-\frac{1}{2r}\left[\left({}^{\lfloor \psi}L\theta - {}^{\lfloor d\psi}L\right)' \, {}^{\lfloor \psi}D \left({}^{\lfloor \psi}L\theta - {}^{\lfloor d\psi}L\right) + {}^{\lfloor d}D\right]\right\} d\theta dr \\
&= \left\{x \equiv r^{-0.5} \, {}^{\lfloor \psi}D^{0.5}\left({}^{\lfloor \psi}L\theta - {}^{\lfloor d\psi}L\right), \; dx = r^{-0.5\mathring{\psi}} \left|\,{}^{\lfloor \psi}D\right|^{0.5} d\theta\right\} \\
&= (2\pi)^{0.5\mathring{\psi}} \left|\,{}^{\lfloor \psi}D\right|^{-0.5} \int_0^\infty r^{-0.5(\nu+2)} \exp\left[-0.5r^{-1} \, {}^{\lfloor d}D\right] dr \\
&= \left\{x = 0.5r^{-1} \, {}^{\lfloor d}D, \; r = 0.5 \, {}^{\lfloor d}Dx^{-1}, \; dr = -0.5 \, {}^{\lfloor d}Dx^{-2}dx\right\} \\
&= (2\pi)^{0.5\mathring{\psi}} \left|\,{}^{\lfloor \psi}D\right|^{-0.5} \left(0.5 \, {}^{\lfloor d}D\right)^{-0.5\nu} \int_0^\infty x^{0.5(\nu-2)} \exp(-x)\, dx \\
&\underbrace{=}_{\text{definition of } \varGamma(\cdot)} \varGamma(0.5\nu) \, {}^{\lfloor d}D^{-0.5\nu} \left|\,{}^{\lfloor \psi}D\right|^{-0.5} 2^{0.5\nu} (2\pi)^{0.5\mathring{\psi}}.
\end{aligned}$$

3. Integrating out θ in the previous step, we find the marginal pdf of r proportional to $r^{-0.5(\nu+2)} \exp\left[-\frac{{}^{\lfloor d}D}{2r}\right]$. The corresponding normalization integral is obtained as follows.

$$\begin{aligned}\mathcal{I}\left({}^{\lfloor d}D,\nu\right) &\equiv \int_0^\infty r^{-0.5(\nu+2)}\exp\left[-0.5r^{-1}\,{}^{\lfloor d}D\right]\,dr\\ &= \left\{x=0.5r^{-1}\,{}^{\lfloor d}D,\ r=0.5\,{}^{\lfloor d}Dx^{-1},\ dr=-0.5\,{}^{\lfloor d}Dx^{-2}dx\right\}\\ &= \left(0.5\,{}^{\lfloor d}D\right)^{-0.5\nu}\varGamma(0.5\nu).\end{aligned}$$

4. Using the definition

$$\mathcal{I}\left({}^{\lfloor d}D,\nu\right)\equiv\frac{\varGamma(0.5\nu)}{\left(0.5\,{}^{\lfloor d}D\right)^{0.5\nu}}$$

of the normalizing factor in the marginal pdf of r (see (8.23)), we get
a) $\mathcal{E}\left[r^i\right]=\mathcal{I}\left({}^{\lfloor d}D,\nu-2i\right)/\mathcal{I}\left({}^{\lfloor d}D,\nu\right),\ i=\ldots,-1,0,1,\ldots,0.5\nu-1,$

$$\text{For } i=1, \text{ it gives } \mathcal{E}[r]=\frac{0.5\,{}^{\lfloor d}D}{0.5(\nu-2)}=\frac{{}^{\lfloor d}D}{\nu-2}.$$

$$\text{For } i=2, \text{ it gives } \mathcal{E}\left[r^2\right]=\frac{\left(0.5\,{}^{\lfloor d}D\right)^2}{0.5^2(\nu-2)(\nu-4)}\Rightarrow$$

$$\operatorname{var}(r)=\mathcal{E}[r^2]-(\mathcal{E}[r])^2=\frac{\hat{r}^2}{\nu-4}[\nu-2-(\nu-4)]=\frac{2\hat{r}^2}{\nu-4}.$$

$$\text{For } i=-1, \text{ it gives } \mathcal{E}\left[r^{-1}\right]=\left(0.5\,{}^{\lfloor d}D\right)^{-1}(0.5\nu)=\frac{\nu}{{}^{\lfloor d}D}.$$

b)

$$\begin{aligned}&\mathcal{E}[\ln(r)|L,D,\nu]\equiv\frac{1}{\mathcal{I}\left({}^{\lfloor d}D,\nu\right)}\int_0^\infty \ln(r)r^{-0.5(\nu+2)}\exp\left(-\frac{{}^{\lfloor d}D}{2r}\right)\,dr\\ &=\frac{-2}{\mathcal{I}\left({}^{\lfloor d}D,\nu\right)}\int_0^\infty\frac{\partial}{\partial\nu}r^{-0.5(\nu+2)}\exp\left(-\frac{{}^{\lfloor d}D}{2r}\right)\,dr\\ &=\frac{-2}{\mathcal{I}\left({}^{\lfloor d}D,\nu\right)}\frac{\partial}{\partial\nu}\int_0^\infty r^{-0.5(\nu+2)}\exp\left(-\frac{{}^{\lfloor d}D}{2r}\right)\,dr\\ &=-2\frac{\partial}{\partial\nu}\ln\left(\mathcal{I}\left({}^{\lfloor d}D,\nu\right)\right)=\ln\left({}^{\lfloor d}D\right)-\ln(2)-\frac{\partial\ln\left(\varGamma(0.5\nu)\right)}{\partial(0.5\nu)}.\end{aligned}$$

5. Marginal pdf of θ and corresponding moments are obtained as follows.

$$\begin{aligned}f(\theta)\propto&\int r^{-0.5(\nu+\mathring{\psi}+2)}\\ &\times\exp\left\{-0.5r^{-1}\,{}^{\lfloor d}D[1+(\theta-\hat{\theta})'\,{}^{\lfloor d}D^{-1}\,{}^{\lfloor \psi}L'\,{}^{\lfloor \psi}D\,{}^{\lfloor \psi}L(\theta-\hat{\theta})]\right\}\,dr\\ =&\left\{r=0.5\,{}^{\lfloor d}D\left[1+\left(\theta-\hat{\theta}\right)'\,{}^{\lfloor d}D^{-1}\,{}^{\lfloor \psi}L'\,{}^{\lfloor \psi}D\,{}^{\lfloor \psi}L\left(\theta-\hat{\theta}\right)\right]/x\right\}\\ \propto&\left[1+\left(\theta-\hat{\theta}\right)'\,{}^{\lfloor d}D^{-1}\,{}^{\lfloor \psi}L'\,{}^{\lfloor \psi}D\,{}^{\lfloor \psi}L\left(\theta-\hat{\theta}\right)\right]^{-0.5(\nu+\mathring{\psi})}.\end{aligned}$$

Symmetry with respect to $\hat{\theta}$ implies that $\mathcal{E}[\theta] = \hat{\theta}$. Moreover, neither normalization integral nor covariance depend on $\hat{\theta}$, thus we can write without loss of generality the further formulas for the special case $\hat{\theta} = 0$. Substitution with the unit Jacobian $x = {}^{\lfloor\psi}L\theta$ implies that the normalizing integral does not depend on ${}^{\lfloor\psi}L$ and

$$
\begin{aligned}
\mathcal{I}(D,\nu) &= \int \left[1 + x' \, {}^{\lfloor d}D^{-1} \, {}^{\lfloor\psi}Dx\right]^{-0.5(\nu+\mathring{\psi})-2} dx \\
&= \left\{ y = \left[{}^{\lfloor d}D^{-1} \, {}^{\lfloor\psi}D \right]^{0.5} x \right\} \\
&= {}^{\lfloor d}D^{0.5\mathring{\psi}} \, {}^{\lfloor\psi}D^{-0.5} \int (1 + y'y)^{-0.5(\nu+\mathring{\psi}-2)} \, dy \\
&\underbrace{=}_{\text{see [156]}} {}^{\lfloor d}D^{0.5\mathring{\psi}} \, {}^{\lfloor\psi}D^{-0.5} \prod_{i=1}^{\mathring{\psi}} \Gamma(0.5(\nu + \mathring{\psi} - i)).
\end{aligned}
$$

Using the fact that $\mathcal{E}[\theta|r] = \mathcal{E}[\theta]$, the identity $\mathrm{cov}[\theta|r] = \mathcal{E}[\theta\theta'|r] - \mathcal{E}[\theta]\mathcal{E}[\theta]'$ and the chain rule for expectations, Proposition 2.6, we have

$$
\begin{aligned}
\mathrm{cov}[\theta] &= \mathcal{E}[\theta\theta'] - \mathcal{E}[\theta]\mathcal{E}[\theta]' = \mathcal{E}\left[\mathcal{E}[\theta\theta' - \mathcal{E}[\theta]\mathcal{E}[\theta]'|r]\right] = \mathcal{E}[\mathrm{cov}[\theta|r]] \\
&= \mathcal{E}\left[r \left({}^{\lfloor\psi}L' \, {}^{\lfloor\psi}D \, {}^{\lfloor\psi}L \right)^{-1} \right] = \frac{{}^{\lfloor d}D}{\nu - 2} \left({}^{\lfloor\psi}L' \, {}^{\lfloor\psi}D \, {}^{\lfloor\psi}L \right)^{-1}.
\end{aligned}
$$

□

We need also to evaluate marginal and conditional pdfs of regression coefficients described by *GiW* pdf.

Proposition 8.8 (Low-dimensional pdfs related to *GiW* pdf) *Let the $L'DL$ decomposition of the extended information matrix determining the pdf $GiW_{\left[{}^{\lfloor a}\theta',\, {}^{\lfloor b}\theta'\right]',r}(L,D,\nu)$ be split in accordance with the splitting of regression coefficients $\theta = \left[{}^{\lfloor a}\theta',\, {}^{\lfloor b}\theta'\right]'$*

$$
L \equiv \begin{bmatrix} 1 & & \\ {}^{\lfloor da}L & {}^{\lfloor a}L & \\ {}^{\lfloor db}L & {}^{\lfloor ab}L & {}^{\lfloor b}L \end{bmatrix}, \quad D \equiv \begin{bmatrix} {}^{\lfloor d}D & & \\ & {}^{\lfloor a}D & \\ & & {}^{\lfloor b}D \end{bmatrix}.
$$

Then,

$$
\begin{aligned}
f\left({}^{\lfloor a}\theta, r\right) &= GiW_{{}^{\lfloor a}\theta,r}\left(\begin{bmatrix} 1 & \\ {}^{\lfloor da}L & {}^{\lfloor a}L \end{bmatrix}, \begin{bmatrix} {}^{\lfloor d}D & \\ & {}^{\lfloor a}D \end{bmatrix}, \nu \right) \qquad (8.24) \\
f\left({}^{\lfloor b}\theta \middle| {}^{\lfloor a}\theta, r\right) &= \mathcal{N}_{{}^{\lfloor b}\theta}\left({}^{\lfloor b}L^{-1} \left[{}^{\lfloor db}L - {}^{\lfloor ab}L \, {}^{\lfloor a}\theta \right], r \left({}^{\lfloor b}L' \, {}^{\lfloor b}D \, {}^{\lfloor b}L \right)^{-1} \right).
\end{aligned}
$$

Proof.

$$
GiW_{\left[{}^{\lfloor a}\theta',\, {}^{\lfloor b}\theta'\right]',r}(L,D,\nu) \propto r^{-0.5(\nu+\mathring{\psi}+2)} \exp\left\{ -\frac{1}{2r} \times \right.
$$

$$\times \begin{bmatrix}-1\\ {}^{\lfloor a}\theta\\ {}^{\lfloor b}\theta\end{bmatrix}' \begin{bmatrix}1 & & \\ {}^{\lfloor da}L & {}^{\lfloor a}L & \\ {}^{\lfloor db}L & {}^{\lfloor ab}L & {}^{\lfloor b}L\end{bmatrix}' \begin{bmatrix}{}^{\lfloor d}D & & \\ & {}^{\lfloor a}D & \\ & & {}^{\lfloor b}D\end{bmatrix} \begin{bmatrix}1 & & \\ {}^{\lfloor da}L & {}^{\lfloor a}L & \\ {}^{\lfloor db}L & {}^{\lfloor ab}L & {}^{\lfloor b}L\end{bmatrix} \begin{bmatrix}-1\\ {}^{\lfloor a}\theta\\ {}^{\lfloor b}\theta\end{bmatrix}\Bigg\}$$

$$= \underbrace{r^{-\frac{-\nu+{}^{\lfloor a}\mathring{\psi}+2}{2}} \exp\left\{-\frac{1}{2r}\begin{bmatrix}-1\\ {}^{\lfloor a}\theta\end{bmatrix}' \begin{bmatrix}1 & \\ {}^{\lfloor da}L & {}^{\lfloor a}L\end{bmatrix}' \begin{bmatrix}{}^{\lfloor d}D & \\ & {}^{\lfloor a}D\end{bmatrix} \begin{bmatrix}1 & \\ {}^{\lfloor da}L & {}^{\lfloor a}L\end{bmatrix} \begin{bmatrix}-1\\ {}^{\lfloor a}\theta\end{bmatrix}\right\}}_{A}$$

$$\underbrace{\times\, r^{-0.5\,{}^{\lfloor b}\mathring{\psi}} \exp\left\{-\frac{1}{2r}\underbrace{\begin{bmatrix}-1\\ {}^{\lfloor a}\theta\\ {}^{\lfloor b}\theta\end{bmatrix}' \left[\,{}^{\lfloor db}L \;\; {}^{\lfloor ab}L \;\; {}^{\lfloor b}L\right]' \,{}^{\lfloor b}D \left[\,{}^{\lfloor db}L \;\; {}^{\lfloor ab}L \;\; {}^{\lfloor b}L\right] \begin{bmatrix}-1\\ {}^{\lfloor a}\theta\\ {}^{\lfloor b}\theta\end{bmatrix}}_{Q}\right\}}_{B}.$$

Integration over ${}^{\lfloor b}\theta$, with the substitution

$$x = r^{-0.5}\left[\,{}^{\lfloor db}L \;\; {}^{\lfloor ab}L \;\; {}^{\lfloor b}L\right] \begin{bmatrix}-1\\ {}^{\lfloor a}\theta\\ {}^{\lfloor b}\theta\end{bmatrix},$$

concerns only the factor B. The integration result is independent both of r and ${}^{\lfloor a}\theta$. Thus, the factor A, which has the GiW form, is the marginal pdf of $\left({}^{\lfloor a}\theta, r\right)$ searched for.

Consequently, the factor B has to be pdf $f\left({}^{\lfloor b}\theta \middle| {}^{\lfloor a}\theta, r\right)$. It is normal pdf. Its moments are easily found by completing squares in the quadratic form Q in the exponent

$$Q = \left({}^{\lfloor b}\theta - \mathcal{E}\left[{}^{\lfloor b}\theta \middle| {}^{\lfloor a}\theta, r\right]\right)' \,{}^{\lfloor b}L' \,{}^{\lfloor b}D \,{}^{\lfloor b}L \left({}^{\lfloor b}\theta - \mathcal{E}\left[{}^{\lfloor b}\theta \middle| {}^{\lfloor a}\theta, r\right]\right),$$
$$\mathcal{E}\left[{}^{\lfloor b}\theta \middle| {}^{\lfloor a}\theta, r\right] \equiv \mathcal{E}\left[{}^{\lfloor b}\theta \middle| {}^{\lfloor a}\theta\right] = {}^{\lfloor b}L^{-1}\left[{}^{\lfloor db}L - {}^{\lfloor ab}L\,{}^{\lfloor a}\theta\right].$$

The covariance is $\operatorname{cov}\left[{}^{\lfloor b}\theta \middle| {}^{\lfloor a}\theta, r\right] = \operatorname{cov}\left[{}^{\lfloor b}\theta \middle| r\right] = r\left({}^{\lfloor b}L' \,{}^{\lfloor b}D \,{}^{\lfloor b}L\right)^{-1}$. □

Note that the result of Proposition is simple due to the chosen variant of $L'DL$ decomposition. The alternative version LDL' is more complex.

Repeatedly, we need to evaluate the KL divergence of a pair of GiW pdfs.

Proposition 8.9 (The KL divergence of GiW pdfs) *Let $f(\Theta) = GiW_{\Theta}(L, D, \nu)$, $\tilde{f}(\Theta) = GiW_{\Theta}\left(\tilde{L}, \tilde{D}, \tilde{\nu}\right)$ be a pair of GiW pdfs of parameters $\Theta \equiv (\theta, r) =$ (regression coefficients, noise variance). Let D_{ii} stand for the diagonal element of the matrix D. Then, the KL divergence of f and $\tilde{f}$ is given by the formula*

$$\mathcal{D}\left(f \middle|\middle| \tilde{f}\right) = \ln\left(\frac{\Gamma(0.5\tilde{\nu})}{\Gamma(0.5\nu)}\right) - 0.5\ln\left(\prod_{i=2}^{\mathring{\Psi}} \frac{\tilde{D}_{ii}}{D_{ii}}\right) + 0.5\tilde{\nu}\ln\left(\frac{{}^{\lfloor d}D}{{}^{\lfloor d}\tilde{D}}\right) \qquad (8.25)$$

$$
\begin{aligned}
&+ 0.5(\nu - \tilde{\nu}) \frac{\partial}{\partial(0.5\nu)} \ln\left(\mathit{\Gamma}(0.5\nu)\right) - 0.5\mathring{\psi} \\
&+ 0.5\mathrm{tr}\left[{}^{\lfloor\psi}L^{-1}\, {}^{\lfloor\psi}D^{-1} \left({}^{\lfloor\psi}L' \right)^{-1}\, {}^{\lfloor\psi}\tilde{L}'\, {}^{\lfloor\psi}\tilde{D}\, {}^{\lfloor\psi}\tilde{L} \right] \\
&+ 0.5 \frac{\nu}{{}^{\lfloor d}D} \left[\left(\hat{\theta} - \hat{\tilde{\theta}}\right)'\, {}^{\lfloor\psi}\tilde{L}'\, {}^{\lfloor\psi}\tilde{D}\, {}^{\lfloor\psi}\tilde{L} \left(\hat{\theta} - \hat{\tilde{\theta}}\right) - {}^{\lfloor d}D + {}^{\lfloor d}\tilde{D} \right] \\
&\equiv \ln\left(\frac{\mathit{\Gamma}\left(0.5\tilde{\nu}\right)}{\mathit{\Gamma}\left(0.5\nu\right)} \right) - 0.5\ln\left| C\tilde{C}^{-1} \right| + 0.5\tilde{\nu}\ln\left(\frac{{}^{\lfloor d}D}{{}^{\lfloor d}\tilde{D}} \right) \\
&+ 0.5(\nu - \tilde{\nu}) \frac{\partial}{\partial(0.5\nu)} \ln\left(\mathit{\Gamma}(0.5\nu)\right) - 0.5\mathring{\psi} - 0.5\nu + 0.5\mathrm{tr}\left[C\tilde{C}^{-1} \right] \\
&+ 0.5 \frac{\nu}{{}^{\lfloor d}D} \left[\left(\hat{\theta} - \hat{\tilde{\theta}}\right)' \tilde{C}^{-1} \left(\hat{\theta} - \hat{\tilde{\theta}}\right) + {}^{\lfloor d}\tilde{D} \right].
\end{aligned}
$$

Proof. A substantial part of evaluations is prepared by Proposition 8.7. The definition of the KL divergence (2.25) reads

$$
\mathcal{D} \equiv \mathcal{D}\left(f \middle|\middle| \tilde{f}\right) = \int f(\mathit{\Theta}|L, D, \nu) \ln\left(\frac{f(\mathit{\Theta}|L, D, \nu)}{\tilde{f}\left(\mathit{\Theta}|\tilde{L}, \tilde{D}, \tilde{\nu}\right)} \right) d\mathit{\Theta}.
$$

We write down the *GiW* pdf in a bit simplified form, cf. (8.19), (8.21),

$$
\begin{aligned}
f(\mathit{\Theta}|L, D, \nu) \equiv GiW_{\mathit{\Theta}}(L, D, \nu) &= \frac{r^{-0.5(\nu + \mathring{\psi} + 2)}}{\mathcal{I}(L, D, \nu)} \exp\left\{ -\frac{1}{2r} \left[Q(\theta) + {}^{\lfloor d}D \right] \right\}, \\
Q(\theta) &\equiv \left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L \right)'\, {}^{\lfloor\psi}D \left({}^{\lfloor\psi}L\theta - {}^{\lfloor d\psi}L \right) \\
&= \left(\theta - {}^{\lfloor\psi}L^{-1}\, {}^{\lfloor d\psi}L \right)'\, {}^{\lfloor\psi}L'\, {}^{\lfloor\psi}D\, {}^{\lfloor\psi}L \left(\theta - {}^{\lfloor\psi}L^{-1}\, {}^{\lfloor d\psi}L \right) \\
&= \left(\theta - \hat{\theta}\right)' C^{-1} \left(\theta - \hat{\theta}\right),
\end{aligned}
$$

where $\hat{\theta} \equiv {}^{\lfloor\psi}L^{-1}\, {}^{\lfloor d\psi}L$ and $C^{-1} = {}^{\lfloor\psi}L'\, {}^{\lfloor\psi}D\, {}^{\lfloor\psi}L$. The same notation is used for $\tilde{f}\left(\mathit{\Theta}|\tilde{L}, \tilde{D}, \tilde{\nu}\right)$. If we recall that $\mathit{\Theta} = [\theta, r]$, we can write

$$
\begin{aligned}
\mathcal{D} &= \ln\left(\frac{\mathcal{I}\left(\tilde{L}, \tilde{D}, \tilde{\nu}\right)}{\mathcal{I}\left(L, D, \nu\right)} \right) \\
&+ \mathcal{E}\left[-0.5(\nu - \tilde{\nu})\ln(r) - \frac{1}{2r}\left(Q(\theta) + {}^{\lfloor d}D - \tilde{Q}(\theta) - {}^{\lfloor d}\tilde{D} \right) \middle| L, D, \nu \right] \\
&= \ln\left(\frac{\mathcal{I}\left(\tilde{L}, \tilde{D}, \tilde{\nu}\right)}{\mathcal{I}(L, D, \nu)} \right) - 0.5(\nu - \tilde{\nu}) \left[\ln\left({}^{\lfloor d}D \right) - \ln(2) - \frac{\partial \ln\left(\mathit{\Gamma}(0.5\nu)\right)}{\partial(0.5\nu)} \right] \\
&- 0.5 \underbrace{\mathcal{E}\left\{ r^{-1} \left[Q(\theta) - \tilde{Q}(\theta) \right] \middle| L, D, \nu \right\}}_{X} + \frac{\nu}{{}^{\lfloor d}D} \left[-\, {}^{\lfloor d}D + {}^{\lfloor d}\tilde{D} \right].
\end{aligned}
\tag{8.26}
$$

The the chain rule for expectations, Proposition 2.6, and the formula in (8.23) for $\mathcal{E}\left[r^{-1}\right]$ help us to evaluate the last term X.

$$X = \mathring{\psi} - \mathrm{tr}\left[C\tilde{C}^{-1}\right] + \frac{\nu}{{}^{\lfloor d}D}\left[-\left(\hat{\theta} - \hat{\tilde{\theta}}\right)' \tilde{C}^{-1}\left(\hat{\theta} - \hat{\tilde{\theta}}\right)\right].$$

We substitute this result and explicit formulas for $\mathcal{I}(L, D, \nu)$ and $\mathcal{I}(\tilde{L}, \tilde{D}, \tilde{\nu})$ into (8.26). Taking into account the form of $\mathcal{I}(V, \nu)$ and $\mathcal{I}(\tilde{V}, \tilde{\nu})$, the result reads

$$\begin{aligned}\mathcal{D} &= \ln\left(\frac{\Gamma(0.5\tilde{\nu})}{\Gamma(0.5\nu)}\right) - 0.5\ln\left(\frac{\left|{}^{\lfloor\psi}\tilde{D}\right|}{\left|{}^{\lfloor\psi}D\right|}\right) + 0.5\tilde{\nu}\ln\left(\frac{{}^{\lfloor d}D}{{}^{\lfloor d}\tilde{D}}\right)\\ &+ 0.5(\nu - \tilde{\nu})\frac{\partial}{\partial(0.5\nu)}\ln\left(\Gamma(0.5\nu)\right) - 0.5\mathring{\psi} + 0.5\mathrm{tr}\left[C\tilde{C}^{-1}\right]\\ &+ 0.5\frac{\nu}{{}^{\lfloor d}D}\left[\left(\hat{\theta} - \hat{\tilde{\theta}}\right)' \tilde{C}^{-1}\left(\hat{\theta} - \hat{\tilde{\theta}}\right) - {}^{\lfloor d}D + {}^{\lfloor d}\tilde{D}\right].\end{aligned}$$

This alternative relates the considered $L'DL$ decomposition to the LS quantities; see Proposition 8.7. □

8.1.4 KL divergence of normal pdfs

The KL divergence serves us as a good measure when judging distance of pairs of normal factors or components in the data space. It serves in the parameter space too when dealing with so called MT normal factors; see Chapters 12 and 13. Its value can be computed as follows.

Proposition 8.10 (KL divergence of normal pdfs) *The KL divergence (2.25) of the normal pdfs $f(d) = \mathcal{N}_d(M, R)$, $\tilde{f}(d) = \mathcal{N}_d(\tilde{M}, \tilde{R})$ is given by the formula*

$$\mathcal{D}\left(f\middle|\middle|\tilde{f}\right) = \frac{1}{2}\left[\ln\left|\tilde{R}R^{-1}\right| - \mathring{d} + \mathrm{tr}\left[R\tilde{R}^{-1}\right] + \left(M - \tilde{M}\right)' \tilde{R}^{-1}\left(M - \tilde{M}\right)\right]. \tag{8.27}$$

Let us consider a pair of normal pdfs in factorized forms, $i^* \equiv \{1, \ldots, \mathring{d}\}$,

$$f(d_t|d(t-1)) = \prod_{i \in i^*}\mathcal{N}_{d_i}(\theta_i'\psi_{i;t}, r_i),\quad {}^{\lfloor U}f(d_t|d(t-1)) = \prod_{i \in i^*}\mathcal{N}_{d_i}\left({}^{\lfloor U}\theta_i'\,{}^{\lfloor U}\psi_{i;t},\, {}^{\lfloor U}r_i\right)$$

with known regression vectors $\psi_{i;t}$, ${}^{\lfloor U}\psi_{i;t}$, regression coefficients θ_i', ${}^{\lfloor U}\theta_i'$ and variance r_i, ${}^{\lfloor U}r_i$, $i \in i^*$. Then, the KL divergence of ith factors, conditioned on $\psi_{i;t}$, ${}^{\lfloor U}\psi_{i;t}$, is

$$\begin{aligned}&\mathcal{D}\left(f(d_i|\psi_{i;t})\middle|\middle|\,{}^{\lfloor U}f\left(d_i|\,{}^{\lfloor U}\psi_{i;t}\right)\middle|\,\psi_{i;t},\,{}^{\lfloor U}\psi_{i;t}\right) \qquad (8.28)\\ &= \frac{1}{2}\left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_i}\right) - 1 + \frac{r_i}{{}^{\lfloor U}r_i} + \frac{\left(\theta_i'\psi_{i;t} - {}^{\lfloor U}\theta_i'\,{}^{\lfloor U}\psi_{i;t}\right)^2}{{}^{\lfloor U}r_i}\right].\end{aligned}$$

Proof. The definitions of the KL divergence, normal pdfs and the identity $\mathcal{E}[x'Qx] = [\mathcal{E}(x)]'Q\mathcal{E}[x] + \mathrm{tr}[\mathrm{cov}(x)Q]$ imply

$$\begin{aligned} 2\mathcal{D}\left(f\middle|\middle|\tilde{f}\right) &\equiv \ln\left(\left|\tilde{R}R^{-1}\right|\right) \\ &+ 2\int \mathcal{N}_d(M,R)\left[-(d-M)'R^{-1}(d-M) + \left(d-\tilde{M}\right)'\tilde{R}^{-1}\left(d-\tilde{M}\right)\right] dd \\ &= \ln\left(\left|\tilde{R}R^{-1}\right|\right) - \mathrm{tr}\left(RR^{-1}\right) + \left(M-\tilde{M}\right)'\tilde{R}^{-1}\left(M-\tilde{M}\right) + \mathrm{tr}\left(R\tilde{R}^{-1}\right). \end{aligned}$$

The factor version is obtained by a direct substitution of moments. □

8.1.5 Estimation and prediction with normal factors

The normal factors predicting a real-valued scalar d belong to the exponential family. Consequently, their estimation and the corresponding prediction reduce to algebraic operations; see Proposition 3.2.

Proposition 8.11 (Estimation and prediction of the normal factor) *Let natural conditions of decision making, Requirement 2.5, hold. The treated factor (8.1) is supposed to be normal. A conjugate prior pdf (3.13)* $GiW_\Theta(V_0,\nu_0)$ *as well as conjugate alternative* $GiW_\Theta(\,^{\lfloor A}V_t,\,^{\lfloor A}\nu_t)$ *in the stabilized forgetting are used; see Section 3.1. Then, the posterior pdf is* $GiW_\Theta(V_t,\nu_t)$ *and its sufficient statistics evolve according to the recursions*

$$\begin{aligned} V_t &= \lambda(V_{t-1} + \Psi_t\Psi_t') + (1-\lambda)\,^{\lfloor A}V_t, \quad V_0 \text{ given by the prior pdf} \\ \nu_t &= \lambda(\nu_{t-1}+1) + (1-\lambda)\,^{\lfloor A}\nu_t, \quad \nu_0 > 0 \text{ given by the prior pdf.} \end{aligned} \tag{8.29}$$

The used forgetting factor $\lambda \in [0,1]$.

The predictive pdf is Student pdf. *For any data vector* $\Psi = [d,\psi']'$, *its values can be found numerically as the ratio*

$$f(d|\psi,V,\nu) = \frac{\mathcal{I}\left(V+\Psi\Psi',\nu+1\right)}{\sqrt{2\pi}\mathcal{I}(V,\nu)}, \quad \text{or} \tag{8.30}$$

$$f(d|\psi,L,D,\nu) = \frac{\Gamma(0.5(\nu+1))\left[\,^{\lfloor d}D(1+\zeta)\right]^{-0.5}}{\sqrt{\pi}\Gamma(0.5\nu)\left(1+\frac{\hat{e}^2}{^{\lfloor d}D(1+\zeta)}\right)^{0.5(\nu+1)}}, \quad \text{where} \tag{8.31}$$

$$\hat{e} \equiv d - \hat{\theta}'\psi \equiv \text{prediction error}$$

$$\hat{\theta} = \,^{\lfloor\psi}L^{-1}\,^{\lfloor d\psi}L,\ \zeta \equiv \psi'\,^{\lfloor\psi}L^{-1}\,^{\lfloor\psi}D^{-1}\left(\,^{\lfloor\psi}L'\right)^{-1}\psi$$

cf. Proposition 8.7.

Proof. The first form is directly implied by Proposition 2.14. We use it while

- exploiting the form of the normalizing factor $\mathcal{I}(L,D,\nu)$, Proposition 8.7,
- respecting relationships among various forms of statistics, and basic algebraic formulas; Proposition 8.1.

It gives

$$
\begin{aligned}
f(d|\psi, L, D, \nu) &= \frac{\Gamma(0.5(\nu+1))}{\sqrt{\pi}\Gamma(0.5\nu)} \\
&\times \left|C^{-1}\right|^{0.5\nu} \left|C^{-1} + \psi\psi'\right|^{-0.5\nu} \left({}^{\lfloor d}D\right)^{-0.5} |V|^{0.5(\nu+1)} |V + \Psi\Psi'|^{-0.5(\nu+1)} \\
&= \frac{\Gamma(0.5(\nu+1))}{\sqrt{\pi}\Gamma(0.5\nu)} \\
&\times |C|^{-0.5\nu} (1+\zeta)^{-0.5\nu}\, {}^{\lfloor d}D^{-0.5} \left(1 + \Psi' L^{-1} D^{-1} \left(L^{-1}\right)' \Psi\right)^{-0.5(\nu+1)} \\
&= \frac{\Gamma(0.5(\nu+1))}{\sqrt{\pi}\Gamma(0.5\nu)\left[{}^{\lfloor d}D(1+\zeta)\right]^{-0.5}} \left(1 + \frac{\hat{e}^2}{{}^{\lfloor d}D(1+\zeta)}\right)^{-0.5(\nu+1)}.
\end{aligned}
$$

We also used the identity $\Psi' L^{-1} D^{-1} \left(L^{-1}\right)' \Psi = \hat{e}^2 / {}^{\lfloor d}D + \zeta$. Note that the majority of evaluations are in [69]. □

Remark(s) 8.2

1. *The recursion (8.29) is practically replaced by updating the $L'DL$ decomposition of the extended information matrix according to Algorithm 8.2. In order to cope with the stabilized forgetting,*
   a) the updated D is multiplied by λ,
   b) the data vector Ψ is replaced by $\sqrt{\lambda}\Psi$,
   c) the $L'DL$ decomposition of ${}^{\lfloor A}V_t$ is taken as the sum of weighted dyads formed by

$$\left(kth\_row\_of\_\,{}^{\lfloor A}L_t\right)' (1-\lambda)\, {}^{\lfloor A}D_{k;t} \left(kth\_row\_of\_\,{}^{\lfloor A}L_t\right), \; k = 1, \ldots, \mathring{\Psi}.$$

2. *The extensive use of the $L'DL$ decomposition is vital for numerical stability of computations. Experience confirms that without it, safe processing of real data records with target dimensions ($\mathring{\Psi}$ around hundreds) is impossible. Moreover, the evaluation of the determinants occurring in various pdfs becomes simple with the $L'DL$ decomposition.*
3. *The predictive pdf is the key element needed in approximate estimation, Section 8.5, and for comparing variants through v-likelihood, Section 6.1.2. It is thus worthwhile to check which of the versions (8.30), (8.31) is computationally simpler. The latter one has the added advantage that the prediction errors $\{\hat{e}_t\}_{t\in t^*}$ are suited for an intuitively plausible judgement of the predictive properties of the inspected factor. The form (8.31) of the Student pdf is suited for evaluating likelihood function when LS parameter estimates $\hat{\theta}$, ${}^{\lfloor d}D$ are fixed.*
4. *The performed evaluations are closely related to least squares and its recursive version; see e.g., [69].*

8.1.6 Estimation and prediction with log-normal factors

The applicability of the studied model can be simply extended by modelling and predicting unknown-parameter free, one-to-one, transformations of d_t; see Subsection 6.2.4 and [69]. Almost everything remains the same. Only moments of the original d_t have to be derived. *Log-normal parameterized factor* predicting $\ln(d_t)$ represents the most practical instance of this generalization. It allows us to deal with positive data having a relatively heavy-tailed pdf.

The log-normal parameterized factor models a positive scalar quantity d by the pdf

$$\begin{aligned}
& f(d|\psi,\Theta) = \mathcal{LN}_d(\theta'\psi, r), \text{ where} \qquad\qquad (8.32)\\
& \Theta \equiv [\theta, r] \equiv [\text{regression coefficients, noise variance}]\\
& \quad \in \Theta^* \subset (\mathring{\psi}\text{-dimensional, nonnegative) reals}\\
& \psi \equiv \text{regression vector}, \ \ \Psi = [\ln(d), \psi']' \equiv \text{data vector}\\
& \mathcal{LN}_d(\theta'\psi, r) \equiv (2\pi r)^{-0.5} d^{-1} \exp\left\{ -\frac{(\ln(d) - \theta'\psi)^2}{2r} \right\}\\
& = (2\pi r)^{-0.5} d^{-1} \exp\left\{ -\frac{1}{2r} \operatorname{tr}\left(\Psi\Psi'[-1,\theta']'[-1,\theta']\right) \right\}.
\end{aligned}$$

Log-normal parameterized factors belong to the exponential family (see Section 3.2) so that they possess conjugate (self-reproducing) prior. The following correspondence to (3.6) holds

$$\begin{aligned}
& \mathcal{LN}_d(\theta'\psi, r) = A(\Theta)\exp[\langle B(\Psi), C(\Theta) + D(\Psi)\rangle] \ \text{ with}\\
& \quad A(\Theta) \equiv (2\pi r)^{-0.5}, \ B(\Psi) \equiv \Psi\Psi', \ C(\Theta) \equiv 2r^{-1}\,[-1,\theta']'\,[-1,\theta']\\
& \quad D(\Psi) \equiv \ln\left(\frac{1}{d}\right), \ \ \langle B, C\rangle \equiv \operatorname{tr}\left[B'C\right]. \qquad\qquad (8.33)
\end{aligned}$$

The factor $\exp\left(D(\Psi)\right) \equiv d^{-1}$ could be included into $B(\Psi)$. It has, however, no influence on estimation. For this reason, it is kept separately.

The correspondence determines the conjugate prior (3.13) in the form of the Gauss-inverse-Wishart pdf (GiW); see (8.16). The estimation studied in connection with normal models remains obviously unchanged. The prediction is slightly influenced by the factor d^{-1} in (8.33).

Proposition 8.12 (Estimation and prediction of log-normal factors) *Let natural conditions of decision making, Requirement 2.5, hold. The treated factor (8.32) is log-normal and a conjugate prior (3.13) $GiW_\Theta(V_0, \nu_0)$ as well as a conjugate alternative $GiW_\Theta(\,{}^{\lfloor A}V_t,\, {}^{\lfloor A}\nu_t)$ in the stabilized forgetting are used; see Section 3.1. Then, the posterior pdf is $GiW_\Theta(V_t, \nu_t)$ and its sufficient statistics evolve according to the recursions*

$$\begin{aligned}
V_t &= \lambda(V_{t-1} + \Psi_t\Psi_t') + (1-\lambda)\,{}^{\lfloor A}V_t, \ V_0 \textit{ given by the prior pdf}\\
\nu_t &= \lambda(\nu_{t-1} + 1) + (1-\lambda)\,{}^{\lfloor A}\nu_t, \qquad \nu_0 \textit{ given by the prior pdf}.
\end{aligned}$$

The predictive pdf is log-Student pdf. *Its values can be numerically evaluated, for any data vector* $\Psi = [\ln(d), \psi']'$, *as the ratio*

$$f(d|\psi, V, \nu) = \frac{\mathcal{I}\left(V + \Psi\Psi', \nu + 1\right)}{\sqrt{2\pi} d \mathcal{I}(V, \nu)}, \quad \textit{see Proposition 2.14, or}$$

$$f(d|\psi, L, D, \nu) = \frac{\Gamma(0.5(\nu+1)) \left[{}^{\lfloor d}D(1+\zeta)\right]^{-0.5}}{d\sqrt{\pi}\Gamma(0.5\nu)\left(1 + \frac{\hat{e}^2}{{}^{\lfloor d}D(1+\zeta)}\right)^{0.5(\nu+1)}}, \quad \textit{where} \quad (8.34)$$

$$\hat{e} \equiv \ln(d) - \hat{\theta}'\psi \equiv \textit{ prediction error}$$

$$\hat{\theta} = {}^{\lfloor \psi}L^{-1} \, {}^{\lfloor d\psi}L, \; \zeta = \psi' \, {}^{\lfloor \psi}L^{-1} \, {}^{\lfloor \psi}D^{-1} \left({}^{\lfloor \psi}L' \right)^{-1} \psi.$$

Proof. It just copies Proposition 8.11 with d^{-1} included into the parameterized model and with $\ln(d)$ replacing d. □

8.1.7 Relationships of a component to its factors

We deal with the factor-based description of components $f(d_t|\phi_{t-1}, \Theta)$; see Agreement 5.4. They define the component through the chain rule

$$f(d_t|\phi_{t-1}, \Theta) = \prod_{i \in i^*} f(d_{i;t}|\psi_{i;t}, \Theta_i).$$

For specification of simulation examples and for interpretation of estimation results, it is useful to relate the normal component to its factors in a detail. This task is addressed here.

It is known that conditional pdfs of a normal pdf are again normal. Thus, it is sufficient to inspect first and second moments.

Proposition 8.13 (Normal component and its factors)
I. Let us consider the matrix form of a normal component [160]

$$f\left(d_t \middle| \phi_{t-1}, {}^{\lfloor M}\theta, {}^{\lfloor e}R\right) = \mathcal{N}_{d_t}\left({}^{\lfloor M}\theta'\phi_{t-1}, {}^{\lfloor e}L' \, {}^{\lfloor e}D \, {}^{\lfloor e}L \right), \quad \textit{where} \qquad (8.35)$$

${}^{\lfloor M}\theta$ *is* $(\mathring{\phi}, \mathring{d})$*-matrix of unknown regression coefficients,*
${}^{\lfloor e}L$, ${}^{\lfloor e}D$ *define the* $L'DL$ *decomposition of the unknown noise-covariance matrix* ${}^{\lfloor e}R$.

Then, the i*th parameterized factor is*

$$f(d_{i;t}|\psi_{i;t}, \Theta_i) = \mathcal{N}_{d_{i;t}}\left(\theta_i'\psi_{i;t}, {}^{\lfloor e}D_i\right), \quad \textit{where} \qquad (8.36)$$

$\psi_{i;t} = [d'_{(i+1)\cdots\mathring{i};t}, \phi'_{t-1}]' = [d_{i+1;t}, \psi'_{i+1;t}]'$, $i < \mathring{d}$, $\psi_{\mathring{d};t} \equiv \phi_{t-1}$ *and*
$\theta_i' = \left[\left({}^{\lfloor e}L' \right)^{-1} - I, \; \left({}^{\lfloor e}L' \right)^{-1} {}^{\lfloor M}\theta' \right]_i$, *where the operation* $[\;]_i$ *selects the* i*th row of the matrix in its argument and omits leading zeros.*

II. Let the normal parameterized factors

$$f(d_{i;t}|\psi_{i;t},\Theta_i) = \mathcal{N}_{d_{i;t}}\left(\theta_i'\psi_{i;t}, r_i\right) \tag{8.37}$$

be ordered so that $\psi_{i;t}$ *is a subvector of* $\left[d'_{(i+1)\cdots\mathring{i};t}, \phi'_{t-1}\right]'$. *Let us extend regression coefficients* θ_i *to* $\bar{\theta}_i$ *by inserting zeros at appropriate places so that* $\theta_i'\psi_{i;t} = \bar{\theta}_i'\left[d_t', \phi_{t-1}'\right]'$. *The vector* $\bar{\theta}_i'$ *has at least i leading zeros. Let us create*

$\left(\mathring{d},\mathring{d}\right)$*-upper triangular matrix* η_1 *having unit diagonal and* $-\bar{\theta}_{ij}$ *value on the position* $1 < i < j \le \mathring{d}$,

$\left(\mathring{d},\mathring{\phi}\right)$*-matrix* η_2 *with the value* $\bar{\theta}_{i(j+\mathring{d})}$ *on the position* $i \in \{1,\ldots,\mathring{d}\}$, $j \in \{1,\ldots,\mathring{\phi}\}$.

Then, these factors form normal component (8.35) with the moments

$$^{\lfloor e}L' = \eta_1^{-1},\ \ ^{\lfloor M}\theta' = {}^{\lfloor e}L'\eta_2 \quad \textit{and} \quad ^{\lfloor e}D_i = r_i. \tag{8.38}$$

Proof. By a direct comparison of moments. □

The derived relationships suit whenever uncertainty of parameters can be neglected. This is the case, for instance, when we specify simulation.

The situation is more complex when information on parameters Θ was obtained through a parameter estimation. Then, the predictive pdfs describing individual factors have Student pdf (8.31). It is known that this pdf resembles the normal pdf with the same moments whenever ν is high enough. The values above 20 make this approximation excellent. Thus, for a given factor, we get the correspondence required; see (8.23) and (8.31),

$$\hat{\theta} \leftrightarrow \theta \ \text{ and } \hat{r}(1+\zeta) \leftrightarrow r. \tag{8.39}$$

This observation gives the algorithm for creating an approximate normal component from incompletely estimated factors.

Algorithm 8.4 (Approximate component from uncertain factors)
For all factors $i \in i^*$ *let us have sufficient statistics* L_i, D_i, ν_i *determining the corresponding GiW pdfs.*

1. *For* $i \in i^*$,
   - *Compute vectors of parameter estimates* $\theta_i \equiv {}^{\lfloor\psi}L_i^{-1}\,{}^{\lfloor d\psi}L_i$ *and* $h_i = {}^{\lfloor\psi}L_i^{-1}\psi_{i;t}$.
   - *Define* $r_i \equiv \frac{^{\lfloor d}D_i}{\nu_i}\left(1 + \sum_{j=1}^{\mathring{d}} \frac{h_{ij}^2}{^{\lfloor\psi}D_{ij}}\right)$, *where* h_{ij}, $^{\lfloor\psi}D_{ij}$ *are jth entries of* h_i *and of* $^{\lfloor\psi}D_{ij}$ *diagonal.*
2. *Apply the transformation described in II. of Proposition 8.13.*

Remark(s) 8.3

1. *The variance of the factor output (cf. Algorithm 8.4) is greater than the estimate* $\hat{r}$ *of the noise variance. The increase by the factor* $1+\zeta_{i;t}$ *reflects a projection of parameter uncertainties to the direction of the regression vector* $\psi_{i;t}$. *The increase is obviously data dependent.*
2. *The needed inversion of the triangular matrix* ${}^{\lfloor\psi}L$ *is efficiently performed by a backward elimination with two-column right-hand side of the set of linear equations* ${}^{\lfloor\psi}L\left[\hat{\theta}, h\right] = \left[\,{}^{\lfloor d\psi}L, \psi\right]$.

Problem 8.1 (Exact relationships of factors to component) *It is worthwhile to find the studied relationships without the crude approximation adopted.*

8.1.8 Prediction and model selection

This subsection provides predictors based on normal mixtures. They are repeatedly used for selection of a model among several available alternatives. Proposition 8.11 leads immediately to a normal variant of Proposition 6.1.

Proposition 8.14 (Normal-mixture, *one-step-ahead predictor*) *Let the prior and posterior estimates of the normal mixture have the form (6.3)*

$$f(\Theta|d(t)) = Di_{\alpha}(\kappa_t)\prod_{c\in c^*}\prod_{i\in i^*} f(\Theta_{ic}|d(t)),\ t\in\{0\}\cup t^*.$$

The Dirichlet pdf $Di_{\alpha}(\kappa_t)$ *is the estimate of component weights* α *(Agreement 5.4) and has the expected value*

$$\hat{\alpha}_{c;t} = \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c};t}}.$$

Let the estimates of the individual normal factors be GiW pdfs (8.16)

$$f(\Theta_{ic}|d(t)) = GiW_{\theta_{ic},r_{ic}}(L_{ic;t}, D_{ic;t}, \nu_{ic;t}).$$

Then, the estimation within the adaptive advisory system provides the value of the one-step-ahead predictor at a possible data point d_{t+1} *in the form*

$$f(d_{t+1}|d(t)) = \sum_{c\in c^*}\hat{\alpha}_{c;t}\prod_{i\in i^*} f(d_{ic;t+1}|d(t),c) \quad \textit{with} \tag{8.40}$$

$$f(d_{ic;t+1}|d(t),c) = \frac{\mathcal{I}(V_{ic;t}+\Psi_{ic;t+1}\Psi'_{ic;t+1}, \nu_{ic;t}+1)}{\sqrt{2\pi}\mathcal{I}(V_{ic;t},\nu_{ic;t})}$$

$$\mathcal{I}(V_{ic;t},\nu_{ic;t}) \equiv \mathcal{I}\left(L_{ic;t}, D_{ic;t}, \nu_{ic;t}\right)$$

$$= \Gamma\left(0.5\nu_{ic;t}\right)\left({}^{\lfloor d}D_{ic;t}\right)^{-0.5\nu_{ic;t}}\left|{}^{\lfloor\psi_{ic}}D_{ic;t}\right|^{-0.5} 2^{0.5\nu_{ic;t}}(2\pi)^{0.5\mathring{\psi}_{ic}}$$

$$V_{ic;t} \equiv L'_{ic;t}D_{ic;t}L_{ic;t}$$

is the extended information matrix of the factor ic at time t

$\nu_{ic;t} \equiv$ *the number of degrees of freedom of the factor ic at time t.*

The predictive pdf for a factor ic can be given the alternative form (8.31). We show it explicitly in the following specification of the predictor in the fixed advisory system, when the measured data do not enter the condition of the pdf on parameters, $f(\Theta|d(t)) \equiv f(\Theta)$. The mixture predictor for the fixed advisory system is

$$
\begin{aligned}
&f(d_{t+1}|d(t)) \equiv \sum_{c\in c^*} \hat{\alpha}_{c;0} \prod_{i\in i^*} \underbrace{\int f(d_{ic;t+1}|\psi_{ic;t+1}, \Theta_{ic}, c) f(\Theta_{ic})\, d\Theta_{ic}}_{f(d_{ic;t+1}|\psi_{ic;t+1},c)}, \ \textit{where}\\
&f(d_{ic;t+1}|\psi_{ic;t+1}, c) = \frac{\Gamma(0.5(\nu_{ic;0}+1)) \left[{}^{\lfloor d}D_{ic;0}(1+\zeta_{ic;t+1}) \right]^{-0.5}}{\sqrt{\pi}\Gamma(0.5\nu_{ic;0}) \left(1 + \frac{\hat{e}^2_{ic;t+1}}{{}^{\lfloor d}D_{ic;0}(1+\zeta_{ic;t+1})}\right)^{0.5(\nu_{ic;0}+1)}}\\
&\hat{e}_{ic;t+1} = d_{ic;t+1} - \hat{\theta}'_{ic;0}\psi_{ic;t+1} \equiv \textit{prediction error of the fixed predictor}\\
&\zeta_{ic;t+1} = \psi'_{ic;t+1}\, {}^{\lfloor \psi_{ic}}L^{-1}_{ic;0}\, {}^{\lfloor \psi_{ic}}D^{-1}_{ic;0} \left({}^{\lfloor \psi_{ic}}L'_{ic;0} \right)^{-1} \psi_{ic;t+1}. \qquad (8.41)
\end{aligned}
$$

In both types of advisory systems, $D = \mathrm{diag}\left[{}^{\lfloor d}D,\ {}^{\lfloor \psi}D \right]$, $L = \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix}$, $\hat{\theta} = {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor d\psi}L$.

Proof. Omitted. □

8.1.9 Likelihood on variants

During the design phase of the p-system, predictors serve for selecting the best initialization, forgetting, structure, etc. For the adaptive advisory system, the v-likelihood is the product of one-step-ahead predictors. For the fixed advisory system, such a product is an approximate v-likelihood; see Section (6.1.2).

It is good news that in both cases one-step-ahead predictors of the mixture do not reflect the way statistics were collected. Thus, their product serves well as a v-likelihood even when comparing various estimation variants.

8.1.10 Branch-and-bound techniques

For normal parameterized factors with conjugate prior pdfs, various estimates of factors $GiW_{\theta,r}(V,\nu)$ are alternative descriptions inspected and compared. Thus, alternative sets of sufficient statistics V,ν, or their factorized and shifted equivalents L, D, ν, are branched and bounded as outlined in Section 6.1.3. The maximized functional F then becomes the function, namely, the v-likelihood $f(d(\mathring{t})|V,\nu)$ of data $d(\mathring{t})$ conditioned on the variant of sufficient statistics V,ν. These statistics determine the $GiW_{\theta,r}(V,\nu)$ pdf that is used for eliminating unknown parameters

$$f(d(\mathring{t})|V,\nu) = \int f(d(\mathring{t})|\theta,r) GiW_{\theta,r}(V,\nu)\, d\theta\, dr.$$

Evaluation of these v-likelihood values for normal mixtures is implied by Proposition 8.14. Specific versions of the branch-and-bound techniques are described in Section 8.4.

8.2 Data preprocessing

Data preprocessing is universally discussed in Chapter 6. Here, just some specific comments are made.

8.2.1 Use of physical boundaries

Normal pdfs are symmetric around the expected value and positive on the whole real line. The tails of this pdf fall down rather quickly. This can be exploited when respecting known physical boundaries. Individual signals should be scaled and shifted according to the physically expected range. For standardization purposes, it is useful to scale them so that the middle of this range is mapped to zero and 70% boundaries to unity. It simplifies the universal outlier detection.

8.2.2 Removal of high-frequency noise

Normal mixtures can be perceived as a collection of local normal regression models. Their estimation is known to be sensitive to the presence of a high-frequency measurement noise. It has to be removed before using these data for the normal mixture estimation. The general local filters exploiting oversampling of data are expected to suit this task. Use of normal ARMAX factors is worth being considered. Recall that the ARMAX model is an extension of the normal regression model that allows colored noise modelled by the moving-average process. At least the case with the known MA part, [145], is directly applicable. A promising extension to the unknown MA part is in [144].

8.2.3 Suppression of outliers

Outliers strictly contradict the nature of the normal model that has very light tails. Thus, their removal is crucial for obtaining reliable results. It corresponds to a well-known observation that least squares, which form the algorithmic core of estimation of normal mixtures, are sensitive to outlying data. Local filters [142] and simple mixtures [161] may serve for the discussed removal.

8.3 Use of prior knowledge at the factor level

Knowledge elicitation at the factor level (Section 6.3) is specialized here to the normal factors.

8.3.1 Internally consistent fictitious data blocks

Processing of internally consistent data blocks coincides with learning normal factors; see Section 8.1.5. Thus, we have to quantify individual knowledge items only and then specialize merging of the individual estimates $f(\Theta|K_k)$ expressing knowledge items $K(\mathring{k})$.

8.3.2 Translation of input–output characteristics into data

The quantified knowledge is assumed in the form of the initial moments $\hat{d}$, r_d (6.26) of the predictive pdf $f(d|\psi)$ given by a fixed regression vector ψ

$$\hat{d} = \int d\, f(d|\psi)\, dd,\ r_d = \int \left(d - \hat{d}\right)^2 f(d|\psi)\, dd. \tag{8.42}$$

It reflects the knowledge (6.25) with $h(\psi) = \left[\hat{d}, r_d\right]$, $H(\Psi) = \left[d, (d-\hat{d})^2\right]$. The function $g(\psi,\Theta)$ that serves the optimal quantification of this knowledge (see Proposition 6.4) has the form

$$g(\psi,\Theta) = \left[\theta'\psi, (\theta'\psi - \hat{d})^2 + r\right]. \tag{8.43}$$

With this function, the pdf $f(\Theta|K)$ constructed according to Proposition 6.4 falls out of the *GiW* class even with the pre-prior pdf chosen within it. For this reason, we perform the optimization described in the cited proposition under the additional requirement: the optimum is searched for within the *GiW* class.

Proposition 8.15 (Knowledge of input–output characteristics) *Let us consider GiW pdfs determined by least-squares statistics $\hat{\theta}, \hat{r}, C, \nu$; see Proposition 8.7. The GiW pdf that fulfills constraints (8.42) and minimizes the KL divergence to the flat pre-prior GiW — given by zero parameters estimate, noise variance $\varepsilon > 0$, covariance factor of LS estimate $1/\varepsilon^2 I$ and degrees of freedom $\nu = \varepsilon < 1$ — has the least-squares statistics*

$$\hat{\theta} = \frac{\hat{d}\psi}{\psi'\psi},\quad C^{-1} = \varepsilon^2(I + x\psi\psi'),\quad \hat{r} = r_d\frac{\varepsilon^2(1 + x\psi'\psi)}{\varepsilon^2(1 + x\psi'\psi) + \psi'\psi}, \tag{8.44}$$

where x is the unique (largest) solution of the equation

$$x + \frac{\varepsilon^2(1 + x\psi'\psi)}{\varepsilon^2(1 + x\psi'\psi) + \psi'\psi} = \frac{\frac{\hat{d}^2}{\psi'\psi} + 1/\varepsilon}{r_d}. \tag{8.45}$$

For $a \equiv (\psi'\psi)^{-1}$, $b \equiv r_d^{-1}\left[\hat{d}^2 a + \varepsilon^{-1}\right]$, it has the form

$$x = 0.5\left[-(a + \varepsilon^{-2} + 1 - b) + \sqrt{(a + \varepsilon^{-2} + 1 - b)^2 + 4a(b + ba - 1)}\right]. \tag{8.46}$$

Degrees of freedom ν should be unchanged.

Proof. The proof is straightforward but lengthy and error prone. For this reason, we present it in detail.

1. We express the optimized KL divergence (8.25) in terms of least-squares statistics. At the same time, we rearrange it, insert the considered specific form of the pre-prior pdf $\bar{f}$ and omit terms not influencing optimization. It gives the minimized function in the form
$$\begin{aligned}&F(\hat{\theta}, \hat{r}, C, \nu)\\ &= \underbrace{-\ln(\mathit{\Gamma}(0.5\nu)) + 0.5(\nu - \varepsilon)\frac{\partial}{\partial(0.5\nu)} \ln\left(\mathit{\Gamma}(0.5\nu)\right) + 0.5\varepsilon^2 \ln(\nu) - 0.5\nu}_{\tilde{F}(\nu)}\\ &+ 0.5\left(-\ln|C| + \varepsilon^2 \mathrm{tr}[C]\right) + 0.5\varepsilon^2 \ln\left(\hat{r}\right) + 0.5\frac{1}{\hat{r}}\left[\varepsilon^2\hat{\theta}'\hat{\theta} + \varepsilon\right].\end{aligned}$$
2. The constraints on moments of the predictor are
$$\hat{d} = \hat{\theta}'\psi, \quad r_d = \hat{r}\left(1 + \psi' C\psi\right) \equiv \hat{r}(1 + \zeta), \;\; \zeta \equiv \psi' C\psi.$$
3. The minimization of $F(\cdot)$ with respect to $\hat{\theta}$ gives directly $\hat{\theta} = \frac{\hat{d}\psi}{\psi'\psi}$ irrespective of other optimized quantities.
4. Inserting the found $\hat{\theta}$ into the optimized functional and using the second constraint for expressing $\hat{r}$, we get the following function to be minimized with respect to C.
$$\begin{aligned}F(C) &= 0.5\left(-\ln|C| + \varepsilon^2 \mathrm{tr}[C]\right) - 0.5\varepsilon^2 \ln(1 + \zeta) + 0.5\zeta\varepsilon^2 B\\ B &\equiv \frac{\frac{\hat{d}^2}{\psi'\psi} + \varepsilon^{-1}}{\lfloor d\, r}.\end{aligned}$$
Taking its derivatives with respect to C, using $\frac{\partial \ln|C|}{\partial C} = C^{-1}$ (see (8.6)), we get necessary condition for minimum
$$C^{-1} = \varepsilon^2\left[I + \left(B - \frac{1}{1+\zeta}\right)\psi\psi'\right].$$
Denoting $x = B - \frac{1}{1+\zeta}$, we get the optimal form of C.
5. We solve the above implicit ($\zeta = \psi' C\psi$) equation. Using the matrix inversion lemma, formula (8.5), and the definition of ζ, we get
$$\zeta = \varepsilon^{-2}\frac{\psi'\psi}{1 + x\psi'\psi}.$$
Definition of x in the previous step and an algebraic rearrangement give
$$\begin{aligned}x + 1/(1+\zeta) = B \;\Rightarrow\; x + \frac{1}{1 + \varepsilon^{-2}\frac{x\psi'\psi}{1+x\psi'\psi}} = B\\ \Rightarrow\; x + \frac{\varepsilon^2(1 + x\psi'\psi)}{\varepsilon^2(1 + x\psi'\psi) + x\psi'\psi} = B.\end{aligned}$$

This is the equation (8.45) to be proved. It can be converted into a quadratic equation for x and solved, but this form shows clearly that it has a solution. Its right-hand side is always positive. The left-hand side can be made arbitrarily large for positive x and when $x \rightarrow -(\psi'\psi)^{-1}$ is negative. Within the admissible range when $x > -(\psi'\psi)^{-1}$, which makes the constructed C positive definite, the left-hand side is continuous. Consequently, a real admissible solution exists. A direct inspection shows that just one formal solution guarantees positive definiteness of C, i.e., $x > -1/(\psi'\psi)$. Equation (8.46) gives the explicit form of the solution.

6. Expressing $\hat{r}$ in term of x, we get

$$\hat{r} = \frac{r_d}{1+\zeta} = \frac{r_d\varepsilon^2(1+x\psi'\psi)}{\varepsilon^2(1+x\psi'\psi)+\psi'\psi}.$$

7. It remains to find the minimizing argument of the part $\tilde{F}(\nu)$ that depends on ν only. The necessary condition for the extreme reads

$$1 = 0.5(\nu-\varepsilon)\frac{\partial^2 \ln(\Gamma(0.5\nu))}{\partial^2(0.5\nu)} + \frac{\varepsilon}{\nu}.$$

It is fulfilled for $\nu = \varepsilon$. The explicit evaluation of the factor at $\nu - \varepsilon$ and use of the Jensen inequality imply that the factor at $\nu = \varepsilon$ is positive. Thus, the right-hand side of this equation is 1 for $\nu = \varepsilon$ only.

□

8.3.3 Merging of knowledge pieces

We have a collection of factor estimates $f(\Theta|K_k) = GiW_{\theta,r}(V_k,\nu_k)$ $k \in k^* \equiv \left\{1,\ldots,\mathring{k}\right\}$ after processing individual, internally consistent, data blocks and individual knowledge items. We have to use them for constructing of a single posterior pdf $\hat{f}(\Theta|d(\mathring{t}),K(\mathring{k})) = GiW_{\theta,r}(V_{\mathring{t}},\nu_{\mathring{t}})$ that puts them together with available real data $d(\mathring{t})$. For this, Proposition 6.5 is applied. It leads directly to the GiW counterpart of Algorithm 6.3.

Algorithm 8.5 (Merging GiW information sources)

1. *Construct the prior factor estimates $GiW_{\theta,r}(V_k,\nu_k)$ reflecting internally consistent data blocks by applying ordinary Bayesian estimation or reflecting individual knowledge pieces and constructed according to Proposition 8.15. The least-squares representation of the latter case has to be transformed to the (V,ν)-description using the relationships (8.17).*
2. *Evaluate the likelihood function $\mathcal{L}(\Theta,d(\mathring{t})) = GiW_{\theta,r}(V_{0;\mathring{t}},\nu_{0;\mathring{t}})$ by applying Proposition 8.11 with zero initial conditions for V,ν.*

3. *Evaluate the posterior pdfs* $f(\Theta|d(t), K_k) = GiW_{\theta,r}(V_{0;\mathring{t}} + V_k, \nu_{0;\mathring{t}} + \nu_k)$ *and the v-likelihoods*

$$f(d(\mathring{t})|K_k) = \frac{\mathcal{I}(V_{0;\mathring{t}} + V_k, \nu_{0;\mathring{t}} + \nu_k)}{\mathcal{I}(V_k, \nu_k)}$$

corresponding to the prior pdfs $f(\Theta|K_k) = GiW_{\theta,r}(V_k, \nu_k),\ k \in k^*$. *The integral* $\mathcal{I}(V, \nu)$ *is given by the formula (8.22) expressed in terms of practically used* $L'DL$ *decomposition of information matrix* V. *The regularization by a flat pre-prior pdf* $GiW_{\theta,r}(\varepsilon I, \varepsilon)$, *given by small* $\varepsilon > 0$ *has to be considered if some* V_k *is singular.*

4. *Evaluate the weights*

$$\beta_{k|d(\mathring{t})} = \frac{f(d(\mathring{t})|K_k)}{\sum_{\tilde{k}\in k^*} f(d(\mathring{t})|K_{\tilde{k}})} = \frac{\frac{\mathcal{I}(V_{0;\mathring{t}}+V_k,\nu_{0;\mathring{t}}+\nu_k)}{\mathcal{I}(V_k,\nu_k)}}{\sum_{\tilde{k}\in k^*} \frac{\mathcal{I}(V_{0;\mathring{t}}+V_{\tilde{k}},\nu_{0;\mathring{t}}+\nu_{\tilde{k}})}{\mathcal{I}(V_{\tilde{k}},\nu_{\tilde{k}})}},\quad k \in k^*.$$

5. *Determine the merger as the posterior pdf to be used*

$$\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k})) = GiW_{\theta,r}\left(V_{0;\mathring{t}} + \sum_{k\in k^*} \beta_{k|d(\mathring{t})} V_k, \nu_{0;\mathring{t}} + \sum_{k\in k^*} \beta_{k|d(\mathring{t})} \nu_k\right)$$
$$= GiW_{\theta,r}(V_{0;\mathring{t}}, \nu_{0;\mathring{t}}) GiW_{\theta,r}\left(\sum_{k\in k^*} \beta_{k|d(\mathring{t})} V_k, \sum_{k\in k^*} \beta_{k|d(\mathring{t})} \nu_k\right).$$

Notice that the effectively used "prior pdf" is seen in the above formula.

8.4 Construction of the prior estimate

Efficient solution of the initialization problem is vital for a proper mixture estimation. At the general level, it is discussed in Section 6.4. Here, its relevant Gaussian counterpart is elaborated.

We apply the general branch-and-bound techniques (Section 6.1.3) for solving the initialization problem. Here and in the subsequent sections appropriate specializations are presented.

8.4.1 Common bounding mapping

As said in Section 6.4.2, the common bounding mapping just cancels the "least fit candidates". For normal mixtures, the posterior pdfs are described by the vector statistics κ, related to the component weights, and by the collection of the extended information matrices V_{ic} together with degrees of freedom ν_{ic}, $i \in i^* \equiv \{1, \ldots, \mathring{d}\}$, $c \in c^*$. These are finite-dimensional objects. They may contain, however, a huge number of entries. Taking, for instance, the simplest static case without common factors, the single alternative represents

$\mathring{c}[1 + (\mathring{d} + 1)(\mathring{d} + 2)/2]$ numbers. For the target dimensions $30 - 40$, it gives about 1000 numbers even for $\mathring{c} = 2$. It indicates that we face two contradictory requirements on the bounding mapping. We have to

- preserve as much information as possible from previous iteration steps,
- store and treat as few as possible alternatives.

We found no general guideline as to how to reach the adequate compromise. At present, we prefer the second item and bound the number of alternatives to two, at most to three.

8.4.2 Flattening mapping

Solution of the initialization problem is based on iterative processing of available learning data $d(\mathring{t})$. The danger of a plain repetitive use of the Bayes rule, discussed at general level in Section 6.4.1, becomes even more obvious when applied to normal factors. Within the class of conjugate *GiW* pdfs, the $\mathring{n}$th "naive" approximate evaluation of the posterior pdf gives

$$GiW_{\theta,r}\left(\sum_{n=1}^{\mathring{n}} V_{n;\mathring{t}}, \sum_{n=1}^{\mathring{n}} \nu_{n;\mathring{t}}\right).$$

Taking into account basic properties of the *GiW* pdf, Proposition 8.7, especially, second moments, it can be seen, that the point estimates of the unknown $\Theta = (\theta, r)$ are around nonweighted averages corresponding to the sum of the individual extended information matrices. Their apparent uncertainty, however, decreases proportionally to $\mathring{n}\mathring{t}$, i.e., quite quickly. This observation clarifies why the flattening may help; see Section 6.4.3. It acts as forgetting that prefers newer, hopefully better, values of $V_{n;\mathring{t}}$ and prevents the cumulative degrees of freedom $\nu = \sum_{n=1}^{\mathring{n}} \nu_{n;\mathring{t}}$ from growing too quickly.

Here, we specialize the flattening operation to the involved *GiW* pdfs.

Proposition 8.16 (Optimal flattening mapping for *GiW* pdfs) *Let* $\tilde{f} = GiW_{\theta,r}\left(\tilde{V}, \tilde{\nu}\right)$ *and* $\bar{f} = GiW_{\theta,r}\left(\bar{V}, \bar{\nu}\right)$ *be a pair of GiW pdfs defined on the common support* $\Theta^* = [\theta, r]^*$*. Then, the pdf* $\hat{f}$ *minimizing the functional* $\mathcal{D}\left(\hat{f}\middle|\middle|\tilde{f}\right) + q\mathcal{D}\left(\hat{f}\middle|\middle|\bar{f}\right)$, $q > 0$, *is* $GiW_{\theta,r}(\hat{V}, \hat{\nu})$ *with*

$$\hat{V} = \Lambda\tilde{V} + (1-\Lambda)\bar{V}, \ \hat{\nu} = \Lambda\tilde{\nu} + (1-\Lambda)\bar{\nu} \ \ \textit{and} \ \ \Lambda \equiv 1/(1+q) \in (0,1). \quad (8.47)$$

Proof. Omitted. □

The application of this result to whole mixtures is straightforward as the posterior pdf of its parameters is a product of the posterior estimates of weights α and of the individual factors; cf. Agreement 6.1.

Proposition 8.17 (Optimal flattening of normal-factor estimates) *Let parameters of a pair of normal mixtures be described by (6.3) with factor estimates* $\tilde{f}(\theta_{ic}, r_{ic}) = GiW_{\theta_{ic}, r_{ic}}\left(\tilde{V}_{ic}, \tilde{\nu}_{ic}\right)$, $\bar{f}(\theta_{ic}, r_{ic}) = GiW_{\theta_{ic}, r_{ic}}\left(\bar{V}_{ic}, \bar{\nu}_{ic}\right)$, $i \in i^*, c \in c^*$. *These factor estimates are assumed to have a common support* $[\theta_{ic}, r_{ic}]^*$, *i.e., the common structure of* θ_{ic}.

The corresponding estimates of the component weights are assumed to be described by $\tilde{f}(\alpha) = Di_\alpha(\tilde{\kappa})$ *and* $\bar{f}(\alpha) = Di_\alpha(\bar{\kappa})$ *of a common structure. Then, the pdf* $\hat{f}$ *minimizing the functional*

$$\mathcal{D}\left(\hat{f}\middle|\middle|\tilde{f}\right) + q\mathcal{D}\left(\hat{f}\middle|\middle|\bar{f}\right), \quad q > 0,$$

preserves the original form described in Agreement 6.1 with GiW factors. The resulting pdf is determined by the statistics, $i \in i^*$, $c \in c^*$,

$$\begin{aligned} \hat{V}_{ic} &= \Lambda\tilde{V}_{ic} + (1-\Lambda)\bar{V}_{ic}, \ \hat{\nu}_{ic} = \Lambda\tilde{\nu}_{ic} + (1-\Lambda)\bar{\nu}_{ic} \\ \hat{\kappa}_c &= \Lambda\tilde{\kappa}_c + (1-\Lambda)\bar{\kappa}_c, \ \text{with } \Lambda \equiv 1/(1+q) \in (0,1) \end{aligned} \tag{8.48}$$

Proof. It is omitted. Just notice that the flattening of the component-weight estimates is universal. It preserves the Dirichlet form with statistics being a convex combination of the individual statistics involved. □

The choice of the flattening rate is fully described by Propositions 6.7, 6.8 and 6.11. The *GiW* form brings nothing new in this respect.

8.4.3 Geometric mean as branching mapping

Geometric-mean branching mapping $\mathcal{A}$ (6.15) maps a pair of arguments $\hat{f}_n(\Theta)$, $n = 1, 2$, of the maximized functional (6.38) $\lfloor^h\mathcal{L}\left(f_n(d(\mathring{t})|\Theta)\hat{f}_n(\Theta), d(\mathring{t})\right) \equiv \hat{f}_n(d(\mathring{t})) \equiv \int f_n(d(\mathring{t})|\Theta)\hat{f}_n(\Theta)\, d\Theta$ on a new, hopefully better, candidate $\hat{f}_{n+1}(\Theta)$. We apply it, assuming that Agreement 6.1 is met for all involved pdfs and that *GiW* pdfs serve as the estimates of parameterized factors.

Proposition 8.18 (Geometric mean of a pair *GiW* pdfs) *Let, for* $n = 1, 2$, $t \in \{0\} \cup t^*$, *Agreement 6.1 be met*

$$\hat{f}_n(\Theta|d(t)) \equiv Di_\alpha(\kappa_{n;t}) \prod_{c \in c^*} \prod_{i \in i^*} GiW_{\theta_{ic}, r_{ic}}(L_{nic;t}, D_{nic;t}, \nu_{nic;t}).$$

These pdfs are obtained from a pair of different prior pdfs when estimating approximately, Section 8.5, the normal mixture

$$f(d_t|d(t-1), \Theta) \equiv \sum_{c \in c^*} \alpha_c \prod_{i \in i^*} \mathcal{N}_{d_{i;t}}\left(\theta'_{ic}\psi_{ic;t}, r_{ic}\right).$$

Recall, $Di_\alpha(\kappa)$ *denotes Dirichlet pdf of* $\alpha \in \alpha^* \equiv \left\{\alpha_c \geq 0 : \sum_{c \in c^*} \alpha_c = 1\right\}$ *of the form (6.3)*

$$Di_{\alpha}(\kappa) \equiv \frac{\prod_{c\in c^*} \alpha_c^{\kappa_c - 1}}{\mathcal{B}(\kappa)}, \ \mathcal{B}(\kappa) \equiv \frac{\prod_{c\in c^*} \Gamma(\kappa_c)}{\Gamma(\sum_{c\in c^*} \kappa_c)}.$$

The corresponding expected values of the component weights α_c, $c \in c^*$, are

$$\hat{\alpha}_{c;t} = \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c};t}}.$$

Then, the geometric means $\hat{f}_3(\Theta|d(t))$, $t \in t^*$, constructed according to Section 6.4.4, preserve the functional form (6.3). Its factors have statistics

$$\begin{aligned} V_{3ic;t} &= \lambda_t V_{1ic;t} + (1-\lambda_t) V_{2ic;t} \qquad (8.49)\\ \nu_{3ic;t} &= \lambda_t \nu_{1ic;t} + (1-\lambda_t)\nu_{2ic;t}\\ \kappa_{3c;t} &= \lambda_t \kappa_{1c;t} + (1-\lambda_t)\kappa_{2c;t}, \quad \textit{where}\\ \lambda_t &= \left[1 + \prod_{\tau=1}^{t} \frac{\sum_{c\in c^*} \hat{\alpha}_{2c;\tau-1} \prod_{i\in i^*} \frac{\mathcal{I}(L_{2ic;\tau}, D_{2ic;\tau}, \nu_{2ic;\tau})}{\mathcal{I}(L_{2ic;\tau-1}, D_{2ic;\tau}, \nu_{2ic;\tau-1})}}{\sum_{\tilde{c}\in c^*} \hat{\alpha}_{1\tilde{c};\tau-1} \prod_{\tilde{i}\in i^*} \frac{\mathcal{I}(L_{1\tilde{i}\tilde{c};\tau}, D_{1\tilde{i}\tilde{c};\tau}, \nu_{1\tilde{i}\tilde{c};\tau})}{\mathcal{I}(L_{1\tilde{i}\tilde{c};\tau-1}, D_{1\tilde{i}\tilde{c};\tau}, \nu_{1\tilde{i}\tilde{c};\tau-1})}}\right]^{-1}. \end{aligned}$$

The normalization integral $\mathcal{I}(L, D, \nu)$ is given by the formula (8.22).

Proof. Geometric bounding is directly applied with the explicitly expressed v-likelihood taken as the product of adaptive one-step-ahead predictors. □

8.4.4 Random branching of statistics

Random branching is a known safe way to reach global optimum. Computational complexity related to it prevents us from using the safe, completely random, repetitive search. Thus, we support a combined use of deterministic and random searches. Here, we discuss this possibility in connection with normal factors.

According to Agreement 6.1, we deal with the mixture estimate $f(\Theta|d(\mathring{t}))$ that is the product of the estimate Di of component weights and the estimates GiW of factor parameters. The number of factors may be relatively large as it is mainly determined by the dimension $\mathring{\Psi}$ of data vectors. Thus, random generating of variants should be and often can be reduced to their small subselection.

The GiW estimate of a factor is determined by the finite-dimensional, approximately sufficient, statistics L, D, ν. $L \in L^* \equiv$ set of lower triangular matrices with unit diagonal, $D \in D^* \equiv$ set of a diagonal matrices with positive diagonal entries and $\nu \in \nu^* \equiv$ set of positive real scalars.

Knowing this, we generate a random sample $\tilde{L}, \tilde{D}, \tilde{\nu}$ in L^*, D^*, ν^* for selected factors whose predictive ability is to be challenged. Even for a single factor, the possible extent of the random choice is usually prohibitive. Thus, it makes sense to exploit interpretation of parts of these statistics and change them partially only. It seems to be wise, cf. Proposition 8.7:

- To estimate the factor structure before generating new statistics.
- To preserve characteristics determining second moments, i.e., $\tilde{C}^{-1} \equiv C^{-1} \equiv {}^{\lfloor\psi}L' \, {}^{\lfloor\psi}D \, {}^{\lfloor\psi}L$ and $\tilde{\nu} \equiv \nu$.
- To change a point estimate $\hat{\theta}$ of regression coefficients

$$\tilde{\hat{\theta}} = \hat{\theta} + \rho\sqrt{\hat{r}}C^{0.5}e, \;\; f(e) = \mathcal{N}_e(0, I), \;\; \rho \in \rho^* \equiv (1,5), \;\; I \equiv \text{unit matrix} \;\Leftrightarrow$$
$${}^{\lfloor d\psi}\tilde{L} = {}^{\lfloor d\psi}L + \rho\sqrt{r}\, {}^{\lfloor\psi}D^{-0.5}e. \tag{8.50}$$

- To change a point estimate of noise covariance $\hat{r} \equiv {}^{\lfloor d}D/(\nu - 2)$, using normal approximation for its estimate, possibly cut to values greater than $\underline{r}$ specifying the smallest expected noise level. The level $\underline{r}$ is always strictly positive as at least numerical noise is always present. Specifically, with $f(e) = \mathcal{N}_e(0,1)$,

$${}^{\lfloor d}\tilde{D} = \max\left[{}^{\lfloor d}D\left(1 + \frac{\rho\sqrt{2}}{\sqrt{\nu - 4}}e\right), (\nu - 2)\underline{r}\right], \rho \in \rho^* \equiv (1,5). \tag{8.51}$$

Statistics κ determining the Dirichlet pdf of component weights should be changed so that components with $\hat{\alpha}_c \approx 0$ are given a chance to be modified. We use basic properties of the Dirichlet pdf, Proposition 10.1. For the considered purpose, we approximate it by a normal pdf. It is also cut so that the smallest κ_c does not decrease. It gives

$$\tilde{\kappa}_c = \max\left[\kappa_c\left(1 + \frac{\rho}{\sqrt{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c}}}}e_c\right), \min_{\tilde{c}\in c^*}\kappa_{\tilde{c}}\right], \;\; \rho \in (1,5), \;\; f(e_c) = \mathcal{N}_{e_c}(0,1). \tag{8.52}$$

Remark(s) 8.4

1. *The obtained result has to be flattened, Section 6.4.3.*
2. *Random generating of statistics can be used as a part of various compound algorithms for constructing the prior pdf.*
3. *The optional value ρ is tuning knob of generators. Its recommended range stems from standard properties of the normal pdf. Values of ρ in a more narrow range $\rho^* \equiv (1,2)$ seem to be preferable.*
4. *The adopted normal approximations are very simplified and can surely be improved. In the inspected context, it does not seem worth spending energy on this problem.*
5. *A proper selection of the factor to be randomly branched decides on the efficiency of the algorithm within which the branching is used. The problem is explicitly addressed in connection with branching by splitting; see Sections 6.4.8 and 8.4.7.*

8.4.5 Prior-posterior branching

The prior-posterior branching (Section 6.4.6) starts at some prior pdf, performs an approximate estimation (Section 8.5) and flattens the resulting posterior pdf so that it provides a new alternative to the prior pdf used. The specialization of the general results on flattening gives directly the following iterative Bayesian learning of normal mixtures.

Algorithm 8.6 (Prior-posterior branching)

Initial mode

- *Select an upper bound $\mathring{n}$ on the number n of iterations and set $n = 0$.*
- *Select a sufficiently rich structure of the mixture, i.e., specify the number of components $\mathring{c}$ and the ordered lists of factors allocated to considered components. The factor structure for each ic is determined by the structure of the corresponding data vector Ψ_{ic}.*
- *Select a flat pre-prior pdf $\bar{f}(\Theta)$ in the form (6.3) with GiW pdfs describing individual factors. It means, select the prior statistics $\bar{V}_{ic}, \bar{\nu}_{ic}, \bar{\kappa}_c$ determining GiW and Dirichlet pdfs with high uncertainties; cf. Propositions 8.7 and 10.1. The pre-prior pdf serves as an alternative in flattening.*
- *Select a prior pdf $f(\Theta)$ in the form (6.3) with GiW pdfs describing individual factors. It means, select the prior statistics $V_{ic}, \nu_{ic}, \kappa_c$ determining GiW and Dirichlet pdfs, cf. Propositions 8.7, 10.1. Note that generally $\bar{f}(\Theta) \neq f(\Theta)$.*
- *Set the first guess of the prior pdf in nth iteration $\hat{f}_{1n}(\Theta) \equiv f(\Theta)$, i.e., set $V_{1icn;0} = V_{ic}$, $\nu_{1icn;0} = \nu_{ic}, \kappa_{1cn;0} = \kappa_c$, with $c \in c^*$ marking component, and giving $\hat{f}_{1n}(\theta_{ic}, r_{ic}) = GiW_{\theta_{ic},r_{ic}}(V_{1icn;0}, \nu_{1icn;0})$ and $\hat{f}_{1n}(\alpha) = Di_\alpha(\kappa_{1n;0})$.*
- *Compute the posterior pdf*

$$\tilde{f}_{1n}(\Theta|d(\mathring{t})) = Di_\alpha(\kappa_{1n;\mathring{t}}) \prod_{i\in i^*\, c\in c^*} GiW_{\theta_{ic},r_{ic}}(V_{1icn;\mathring{t}}, \nu_{1icn;\mathring{t}})$$

  using an approximate Bayesian estimation that starts at $\hat{f}_{1n}(\Theta)$; see Section 8.5.
- *Evaluate the v-likelihood l_{1n} resulting from the use of $\hat{f}_{1n}(\Theta)$.*
- *Apply the flattening operation to $\tilde{f}_{1n}(\Theta|d(\mathring{t}))$ according to Propositions 6.7 (on Dirichlet marginal pdfs $Di_\alpha(\kappa_{1n;\mathring{t}})$) and 6.8 (on the GiW pdfs $GiW_{\theta_i,r_i}(V_{1icn;\mathring{t}}, \nu_{icn;\mathring{t}})$ describing factors). Denote the resulting pdf $\hat{f}_{2n}(\Theta)$. For*

$$\Lambda_D \equiv \frac{\sum_{c\in c^*}(\kappa_c - \bar{\kappa}_c)}{\sum_{\tilde{c}\in c^*}(\kappa_{1\tilde{c}n;\mathring{t}} - \bar{\kappa}_{\tilde{c}})}, \quad \Lambda_G \equiv \frac{\sum_{c\in c^*}\sum_{i\in i^*}(\nu_{ic} - \bar{\nu}_{ic})}{\sum_{\tilde{c}\in c^*}(\nu_{1i\tilde{c}n;\mathring{t}} - \bar{\nu}_{i\tilde{c}})},$$

  it preserves the form (6.3) with

$$\kappa_{2cn;0} = \Lambda_D \kappa_{1cn;\mathring{t}} + (1-\Lambda_D)\bar{\kappa}_c, \quad V_{2icn;0} = \Lambda_G V_{1icn;\mathring{t}} + (1-\Lambda_G)\bar{V}_{ic}$$
$$\nu_{2icn;0} = \Lambda_G \nu_{1icn;\mathring{t}} + (1-\Lambda_G)\bar{\nu}_{ic}.$$

- *Compute the posterior pdf*

$$\tilde{f}_{2n}(\Theta|d(\mathring{t})) = Di_{\alpha}(\kappa_{2n;\mathring{t}}) \prod_{i\in i^*} GiW_{\theta_{ic},r_{ic}}\left(V_{2icn;\mathring{t}}, \nu_{2icn;\mathring{t}}\right)$$

using an approximate Bayesian estimation that starts at $\hat{f}_{2n}(\Theta)$*; see Section 8.5.*
- *Evaluate the v-likelihood* l_{2n} *resulting from the use of* $\hat{f}_{2n}(\Theta)$*.*
- *Set* $\bar{l}_n = \max(l_{1n}, l_{2n})$*.*

Iterative mode

1. *Apply geometric branching to the pair* $\tilde{f}_{1n}(\Theta|d(\mathring{t}))$, $\tilde{f}_{2n}(\Theta|d(\mathring{t}))$ *with the v-likelihood* l_{1n}, l_{2n}*, respectively. For* $\lambda \equiv \frac{l_{1n}}{l_{1n}+l_{2n}}$*, it gives* $\tilde{f}_{3n}(\Theta|d(\mathring{t}))$ *of the form (6.3) with the GiW description of factors. It is determined by the statistics*

$$\begin{aligned} \kappa_{3cn;\mathring{t}} &= \lambda\kappa_{1cn;\mathring{t}} + (1-\lambda)\kappa_{2cn;\mathring{t}}, \ \kappa_{3cn} = \lambda\kappa_{1cn;0} + (1-\lambda)\kappa_{2cn;0} \\ \nu_{3icn;\mathring{t}} &= \lambda\nu_{1icn;\mathring{t}} + (1-\lambda)\nu_{2icn;\mathring{t}}, \ \nu_{3icn} = \lambda\nu_{1icn;0} + (1-\lambda)\nu_{2icn;0} \\ V_{3icn;\mathring{t}} &= \lambda V_{1icn;\mathring{t}} + (1-\lambda)V_{2icn;\mathring{t}}. \end{aligned}$$

2. *Apply the flattening operation to* $\tilde{f}_{3n}(\Theta|d(\mathring{t}))$ *according to Propositions 6.7 (on the Dirichlet marginal pdfs* $Di_{\alpha}(\kappa_{3;\mathring{t}})$ *and 6.8 (on the GiW pdfs* $GiW_{\theta_{ic},r_{ic}}(V_{3icn;\mathring{t}}, \nu_{3icn;\mathring{t}})$ *describing factors). Denote the resulting pdf* $\hat{f}_{3n}(\Theta)$*. For*

$$\Lambda_D \equiv \frac{\sum_{c\in c^*}(\kappa_{3cn} - \bar{\kappa}_c)}{\sum_{\tilde{c}\in c^*}(\kappa_{3\tilde{c}n;\mathring{t}} - \bar{\kappa}_{\tilde{c}})}, \quad \Lambda_G \equiv \frac{\sum_{c\in c^*}\sum_{i\in i^*}(\nu_{3ic} - \bar{\nu}_{ic})}{\sum_{i\in i^*}\sum_{\tilde{c}\in c^*}(\nu_{3i\tilde{c}n;\mathring{t}} - \bar{\nu}_{i\tilde{c}})},$$

it preserves the GiW version of (6.3) with

$$\begin{aligned} \kappa_{3cn;0} &= \Lambda_D\kappa_{3cn;\mathring{t}} + (1-\Lambda_D)\bar{\kappa}_c \\ V_{3icn;0} &= \Lambda_G V_{3icn;\mathring{t}} + (1-\Lambda_G)\bar{V}_{ic}, \ \nu_{3icn;0} = \Lambda_G\nu_{3icn;\mathring{t}} + (1-\Lambda_G)\bar{\nu}_{ic}. \end{aligned}$$

3. *Evaluate the v-likelihood* l_{3n} *resulting from the use of* $\hat{f}_{3n}(\Theta)$ *as the prior pdf in an approximate estimation; Section 8.5. The approximation prepared for the fixed advisory system may be used; see Section 6.1.2.*
4. *Choose among* $\tilde{f}_{\iota n}(\Theta|d(\mathring{t}))$, $\iota \in \{1,2,3\}$ *the pair with the highest v-likelihood values and call them* $\tilde{f}_{1(n+1)}(\Theta|d(\mathring{t}))$, $\tilde{f}_{2(n+1)}(\Theta|d(\mathring{t}))$ *with the v-likelihood values* $l_{1(n+1)}$, $l_{2(n+1)}$*.*
5. *Go to the beginning of* Iterative mode *with* $n = n+1$ *if*

$$\bar{l}_{n+1} \equiv \max(l_{1(n+1)}, l_{2(n+1)}) > \bar{l}_n$$

or if $\bar{l}_{n+1}, \bar{l}_n$ *are the same according to Proposition 6.2 and* $n < \mathring{n}$*.*
6. *Stop and select among* $\hat{f}_{\iota n}(\Theta)$, $\iota = 1, 2$ *that leading to the higher value of* $l_{\iota n}$ *and take it as the prior pdf constructed.*

Remark(s) 8.5

1. *Prediction of the v-likelihood related to the alternative gained by forgetting can be made in the way used in connection with merging and cancelling of components; cf. Section 6.3.3. It can spare one complete approximate estimation.*
2. *Notice that the flattening relevant to branching (Propositions 6.7 and 6.8) is applied as the geometric branching applied at time* $\mathring{t}$ *generates a new alternative to be tested.*
3. *No improvement can be expected when the value of* λ*, used for the geometric branching at the beginning of the iterative mode, is around 0.5. This observation may serve as an additional stopping rule.*
4. *The structure of the estimated mixture does not change during iterations. Thus, it has to be sufficiently reflected even in the prior pdf* $f(\Theta)$*.*

8.4.6 Branching by forgetting

Branching by forgetting (see Section 6.4.7) makes parallel recursive estimations without forgetting and with a fixed forgetting with the forgetting factor smaller than 1. The alternative pdfs are compared according to their v-likelihood. At the moment, when one of them is a sure winner they are bounded to a single pdf and the whole process is repeated. Here, the algorithm is specialized to normal mixtures.

Algorithm 8.7 (Online branching with forgetting)

Initial mode

- *Select a sufficiently rich structure of the mixture, i.e., specify the number of components* $\mathring{c}$ *and the ordered lists of factors allocated to the considered components. The factor structure for each ic is determined by the structure of the corresponding data vector* Ψ_{ic}*.*
- *Set a constant* $\rho \approx 3-5$ *defining the significant difference of v-log-likelihood values.*
- *Set the record counter* $t = 0$ *and select the statistics of the guessed prior pdf*

$$\hat{f}(\theta_{ic}, r_{ic}) \equiv \hat{f}(\theta_{ic}, r_{ic}|d(0)) = GiW_{\theta_{ic},r_{ic}}(V_{ic;0}, \nu_{ic;0}),\ i \in i^*,\ c \in c^*$$

  for factors as well as the statistics κ_0 *determining the pdf* $\hat{f}(\alpha) = Di_\alpha(\kappa_0)$ *of the component weights* α*.*
- *Choose a fixed relatively low forgetting factor* $\lambda < 1$*, say* $\lambda = 0.6$*.*
- *Select a fixed pre-prior alternative pdf used in the stabilized forgetting; see Proposition 3.1. The alternative is usually taken as a flat pre-prior pdf in the GiW version of (6.3). It is given by small* ${}^{\lfloor A}\kappa$*,* ${}^{\lfloor A}\nu_{ic}$ *and* ${}^{\lfloor A}V_{ic} = \varepsilon I$*. Here,* $\varepsilon > 0$*,* $\varepsilon \approx 0$ *and* I *is a unit matrix of an appropriate dimension.*

Data processing mode

1. *Set* $\hat{f}_1(\Theta|d(t)) = \hat{f}_\lambda(\Theta|d(t)) = \hat{f}(\Theta|d(t))$*, i.e.,*
$$\begin{aligned} \kappa_{1c;0} &= \kappa_{\lambda c;0} = \kappa_c, c \in c^* \\ V_{1ic;0} &= V_{\lambda ic;0} = V_{ic;0},\ i \in \{1, \ldots, \mathring{d}\},\ c \in c^*, \\ \nu_{1ic;0} &= \nu_{\lambda ic;0} = \nu_{ic;0},\ i \in \{1, \ldots, \mathring{d}\},\ c \in c^*. \end{aligned}$$
2. *Initialize the v-log-likelihood values assigned to the considered alternatives* $l_{1;t} = 0$*,* $l_{\lambda;t} = 0$*.*
3. *Collect new data* d_{t+1} *and construct the data vector* Ψ_{t+1}*.*
4. *Update* $\hat{f}_1(\Theta|d(t))$ *to* $\hat{f}_1(\Theta|d(t+1))$ *using an approximate estimation, Section 8.5, with the forgetting factor 1.*
5. *Recompute the v-log-likelihood* $l_{1;t}$ *to* $l_{1;t+1}$ *by adding the logarithm of the mixture prediction* $\ln(f(d(t+1)|d(t)))$ *obtained for the "prior" pdf* $\hat{f}_1(\Theta|d(t))$*.*
6. *Update* $\hat{f}_\lambda(\Theta|d(t))$ *to* $\hat{f}_\lambda(\Theta|d(t+1))$ *using approximate estimation with the forgetting factor* λ *and the chosen alternative. Thus, after obtaining statistics* $\kappa_{\lambda 1;t+1}$*,* $V_{ic\lambda 1;t+1}$*,* $\nu_{ic\lambda 1;t+1}$ *through updating of* $\kappa_{\lambda;t}$*,* $V_{ic\lambda;t}$*,* $\nu_{ic\lambda;t}$ *with forgetting 1, forget*
$$\begin{aligned} \kappa_{\lambda;t+1} &= \lambda \kappa_{\lambda 1;t+1} + (1-\lambda)\, {}^{\lfloor A}\kappa, \quad V_{ic\lambda;t+1} = \lambda V_{ic\lambda 1;t+1} + (1-\lambda)\, {}^{\lfloor A}V_{ic} \\ \nu_{ic\lambda;t+1} &= \lambda \nu_{ic\lambda 1;t+1} + (1-\lambda)\, {}^{\lfloor A}\nu_{ic}. \end{aligned}$$
7. *Add logarithm of the mixture prediction* $\ln(f(d(t+1)|d(t)))$*, obtained for the "prior" pdf* $\hat{f}_\lambda(\Theta|d(t))$*, to the v-log-likelihood* $l_{\lambda;t}$ *get* $l_{\lambda;t+1}$*.*
8. *Go to Step 3 with* $t = t+1$ *if* $|l_{1;t+1} - l_{\lambda;t+1}| < \rho$*.*
9. *Set* $\hat{f}(\Theta|d(t+1)) = \hat{f}_1(\Theta|d(t+1))$ *if* $l_{1;t+1} > l_{\lambda;t}$*. Otherwise set* $\hat{f}(\Theta|d(t+1)) = \hat{f}_\lambda(\Theta|d(t+1))$*. It consists of the assignment of the appropriate sufficient statistics, i.e., for* $l_{1;t+1} > l_{\lambda;t}$*,* $i \in i^*$*,* $c \in c^*$*,*
$$\kappa_{\lambda c;t+1} = \kappa_{1c;t+1},\ V_{\lambda ic;t+1} = V_{1ic;t+1},\ i \in i^*,\ \nu_{\lambda ic;t+1} = \nu_{1ic;t+1}$$
and similarly in the opposite case.
10. *Increase* $t = t+1$ *and go to the beginning of* Data processing mode *if* $t \leq \mathring{t}$ *otherwise stop and take* $\hat{f}_1(\Theta|d(\mathring{t}))$ *given by* $\kappa_{1c;\mathring{t}}$*,* $V_{1ic;\mathring{t}}$ *and* $\nu_{1ic;\mathring{t}}$*,* $i \in i^*$*,* $c \in c^*$*, as the final estimate.*

Remark(s) 8.6

1. *Speeding up of the learning is the main expectation connected with this algorithm. The model with no forgetting is expected to be the winner in a long run. Current experience indicates that the estimator with forgetting wins only for initial several tens of data records. The estimation with forgetting can be switched off when the estimation without forgetting is better for a majority of time.*

2. *The technique can be directly combined with prior-posterior branching. Use of the stabilized forgetting is vital in this case. Without it, flattening rates computed as recommended in Section 6.4.3 often go out of the admissible range (0,1].*
3. *Use of the general version of stabilized forgetting in Algorithm 8.7 is computationally expensive. The option that cares only about estimates of the noise variance and component weights seems to be acceptable to the given purpose. It has the form*

$$^{\lfloor A}V_{ic} = \begin{bmatrix} ^{\lfloor Ad}V_{ic} & 0 \\ 0 & 0 \end{bmatrix}, \quad ^{\lfloor A}\nu_{ic} > 0, \ \kappa_c > 0, i \in i^*, \ c \in c^*.$$

The scalars $^{\lfloor Ad}V_{ic}$, $^{\lfloor A}\nu_{ic}$, κ_c *are chosen so that the estimate of the noise variance and component weight as well as their precision are fixed at prior values.*
4. *The experience with a choice of the fixed forgetting factor* λ *indicates that the option* $\lambda \approx 0.6$ *is satisfactory.*
5. *It would be desirable to formulate the selection of the better model as sequential decision-making. It should prevent the situations when the decision is made using too uncertain probabilities of the compared hypotheses.*

8.4.7 Branching by factor splitting

Branching by factor splitting splits factors suspicious for hiding more modes and makes a new learning attempt. The general solution is described in Section 6.4.8. It covers important cases when we have no clue on structure of the mixture. Basic steps described by Algorithm 6.8 apply to the normal case without change. Normality adds nothing new to the selection of the very initial mixture and to the construction of a new mixture. Structure estimation aspects are specialized in Section 8.6. Thus, it remains to specialize the selection of factors to be split and the splitting algorithms. These aspects are discussed in this section.

Optimization-based splitting of factors

The preferable hierarchical selection of split factors, Section 6.4.8, is often too costly to be directly used. Thus, we have to orient ourselves on optimization-based splitting. Significant symmetry of GiW pdfs in regression coefficients θ calls for an explicit shift of factors to be split, for the application of Proposition 6.15. It enforces a nonzero shift of the optional function $g(\Theta)$ that determines the resulting functional form of the modified pdf as $f(\Theta)\exp[-\langle \mu, g(\Theta)\rangle]$, (6.71). It is easy to see that the result remains within GiW class if

$$g'(\Theta) \equiv g'(\theta, r) = \left[\ln(r), \frac{1}{r}, \frac{\theta'}{r}\right] M, \tag{8.53}$$

where M is an arbitrary fixed matrix with $(\mathring{\Psi}+1)$ rows. The application of Proposition 6.15 to this most general form does not provide a closed form solution. The problem can be avoided as follows.

First, we observe that the shifted pdf has to have $C^{-1} = {}^{\lfloor\psi}L' \, {}^{\lfloor\psi}D \, {}^{\lfloor\psi}L$ identical with its pattern as no shift of the quadratic term in θ is done. Second, we recall that splitting is done with the hope to improve predictive abilities of the factor. The combination of these facts implies that the noise variance r of the shifted factors is expected to be smaller than it appears currently. The variance of a convex mixture of a pair of normal pdfs is greater than a convex combination of individual ones. We suppose that the estimated variance resulted from such a combination and we have no reasons to assume their asymmetric role. Thus, we expect that for the shifted factors it will be smaller than the original value. Assuming that the error caused by merged first and second moments are similar, we expect that the noise variance will fall to its half after splitting. Use of this bound gives the rule how to replace ${}^{\lfloor d}D$ with ${}^{\lfloor d}\hat{D}$. We set

$$ {}^{\lfloor d}\hat{D} = s \, {}^{\lfloor d}D \ \text{ with the optional } s \in (0,5). \tag{8.54}$$

This determines shift in "noise-variance direction". Thus, we can consider the following simplified shift determining function $g(\Theta) = \theta/r$. This choice of $g(\Theta)$ guarantees that the (6.71) stay within *GiW* class with no changes in ν, ${}^{\lfloor\psi}L$ and ${}^{\lfloor\psi}D$; cf. Proposition 6.15. Thus, ${}^{\lfloor d\psi}\hat{L}$ is the only optional quantity. In other words, we can deal with a simpler, explicitly solvable case, that cares about shift of the regression coefficients only.

Proposition 8.19 (Shifted *GiW* factor) *Let us consider a fixed pdf* $f = GiW_{\theta,r}(V,\nu) \equiv GiW_{\theta,r}(L,D,\nu)$ *and the approximating pdfs* $\hat{f} = GiW_{\theta,r}(\hat{V},\nu) \equiv GiW_{\theta,r}(\hat{L},\hat{D},\nu)$ *with*

$$V = \begin{bmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}V' \\ {}^{\lfloor d\psi}V & {}^{\lfloor\psi}V \end{bmatrix}, \quad \hat{V} = \begin{bmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}\hat{V}' \\ {}^{\lfloor d\psi}\hat{V} & {}^{\lfloor\psi}V \end{bmatrix}$$

with scalar ${}^{\lfloor d}V$. *The optional* ${}^{\lfloor d\psi}\hat{V}$ *column is restricted by the requirement*

$$ {}^{\lfloor d}V - {}^{\lfloor d\psi}\hat{V}' \, {}^{\lfloor\psi}V^{-1} \, {}^{\lfloor d\psi}\hat{V} = s\left({}^{\lfloor d}V - {}^{\lfloor d\psi}V' \, {}^{\lfloor\psi}V^{-1} \, {}^{\lfloor d\psi}V \right) \Leftrightarrow {}^{\lfloor d}\hat{D} = s \, {}^{\lfloor d}D \tag{8.55}$$

with $s \in (0, 0.5)$ *in (8.54) being tuning knob of this shifting. Then, the pdf* $\hat{f}$ *minimizing the KL divergence* $\mathcal{D}\left(\hat{f}\middle|\middle|f\right)$ *is determined by*

$$\hat{L} \equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L + \rho \, {}^{\lfloor\psi}Lx & L_{\psi} \end{bmatrix} \Leftrightarrow \hat{\hat{\theta}} = \hat{\theta} + \rho x, \quad {}^{\lfloor\psi}\hat{D} = {}^{\lfloor\psi}D, \ \hat{\nu} = \nu. \tag{8.56}$$

The unit $\mathring{\psi}$*-vector* x *is the eigenvector of the covariance matrix of parameters* ${}^{\lfloor\psi}L^{-1} \, {}^{\lfloor\psi}D^{-1} \left({}^{\lfloor\psi}L'\right)^{-1}$ *corresponding to its largest eigenvalue* η. *The signed norm* ρ *of the shift of regression-coefficients estimate is given by the formula*

$$\rho = \pm\sqrt{(\hat{\theta}'x)^2 + (1-s)\,{}^{\lfloor d}D\eta} - \hat{\theta}'x. \tag{8.57}$$

Proof. The part of the KL divergence $\mathcal{D}\left(\hat{f}\middle|\middle| f\right)$ that depends on the optimal estimates of regression coefficients is positively proportional to

$$\left(\hat{\hat{\theta}} - \hat{\theta}\right)' C^{-1} \left(\hat{\hat{\theta}} - \hat{\theta}\right), \quad \text{where } \hat{\hat{\theta}} \text{ is the LS estimate of } \theta \text{ assigned to } \hat{f}.$$

A nonzero shift is necessary in order to reach the prescribed decrease of the least squares remainder ${}^{\lfloor d}D$ to $s\,{}^{\lfloor d}D$.

Obviously, $\hat{\theta}-\hat{\theta} = \pm\rho x$ is minimizing argument searched for, if x is the unit eigenvector corresponding to the smallest eigenvalue η^{-1} of ${}^{\lfloor \psi}L'\,{}^{\lfloor \psi}D\,{}^{\lfloor \psi}L$. This is the largest eigenvalue of its inversion. The signed norm of the eigenvector ρ has to be chosen so that

$$\begin{aligned}&{}^{\lfloor d}V = s\,{}^{\lfloor d}D + \left(\hat{\theta}+\rho x\right)'\,{}^{\lfloor \psi}L'\,{}^{\lfloor \psi}D\,{}^{\lfloor \psi}L\left(\hat{\theta}+\rho x\right) = {}^{\lfloor d}V + \hat{\theta}'\,{}^{\lfloor \psi}L'\,{}^{\lfloor \psi}D\,{}^{\lfloor \psi}L\hat{\theta} \Rightarrow\\ &\quad 0 = -(1-s)\,{}^{\lfloor d}D\eta + 2\rho x'\hat{\theta} + \rho^2\end{aligned}$$

This has the pair of claimed solutions $\rho = \pm\sqrt{(\hat{\theta}'x)^2 + (1-s)\,{}^{\lfloor d}D\eta} - \hat{\theta}'x$. □

Remark(s) 8.7

1. *In an over-parameterized case, the maximizing eigenvector points to "superfluous" directions. Thus, structure estimation and corresponding reduction have to be performed before splitting the factor; cf. item 1 in Remarks 6.14.*
2. *The following simple iterative algorithm for determining the maximum eigenvalue and corresponding eigenvector x is used*

$$\tilde{x}_n \equiv Cx_{n-1},\ x_n = \frac{\tilde{x}_n}{\sqrt{\eta_n}},\ \eta_n \equiv \tilde{z}_n'\tilde{x}_n,\ x_0 \neq 0. \tag{8.58}$$

   Its properties can be seen in the space rotated so that C is a diagonal matrix. In this space, x_n concentrates its unit mass at the entry corresponding to the largest eigenvalue and thus it selects the searched eigenvector in the original space. The scalar η_n converges to the corresponding eigenvalue. This simple case happens if the searched eigenvalue is strictly the largest one. If there is more such eigenvalues then the algorithm concentrates x_n to the subspace of the corresponding eigenvectors. This is, however, exactly what we need for the considered purpose.
3. *The above iterations have to be performed at an appropriate numerical level as the considered kernel C may be low due to the proportionality to the number of data records $\mathring{t}^{-1}$.*

4. *Experiments indicate sensitivity of splitting quality on the optional factor s in (8.54). For normalized data, values $s \approx 0.2$ seem to be reasonable but sometimes much better results can be gained with other options. It calls for experimenting with this tuning knob. Related computational demands make this feature highly undesirable.*

We have to carefully select the function $g(\Theta)$ in order to stay within the *GiW* class after shifting. The following proposition offers a variant that requires the class preservation explicitly. It is based on the alternative formulation of the splitting problem; cf. Proposition 6.16.

Proposition 8.20 (Shift of a *GiW* factor to a *GiW* factor) *Let us consider the set of GiW pdfs $f^* \equiv \{f = GiW_{\theta,r}(L, D, \nu)\}$ and a fixed pdf $f = GiW_{\theta,r}(L, D, \nu)$ in it.*

Let scalar $\beta \in (0, 1)$ be chosen and we search for $\hat{f} \in f^*$ that

1. *minimizes the KL divergence $\mathcal{D}\left(\hat{f}\middle|\middle| f\right)$,*
2. *has the prescribed shift in expectations $\hat{\theta}$, $\hat{r}$ (for f) and $\hat{\hat{\theta}}$, $\hat{\hat{r}}$ (for $\hat{f}$) of θ, r^{-1}*
$$f\left(\hat{\hat{\Theta}}\right) = \beta f\left(\hat{\Theta}\right) \quad \textit{with} \quad \hat{\hat{\Theta}} = \left[\hat{\hat{\theta}}, \hat{\hat{r}}\right] \equiv \left[{}^{\lfloor\psi}\hat{L}^{-1}\, {}^{\lfloor d\psi}\hat{L}, \frac{{}^{\lfloor d}\hat{D}}{\hat{\nu}}\right]$$
$$\hat{\Theta} = \left[\hat{\theta}, \hat{r}\right] \equiv \left[{}^{\lfloor\psi}L^{-1}\, {}^{\lfloor d\psi}L, \frac{{}^{\lfloor d}D}{\nu}\right]$$
computed for $\hat{f}(\Theta) \equiv GiW_{\theta,r}\left(\hat{L}, \hat{D}, \hat{\nu}\right)$ and $f(\Theta) \equiv GiW_{\theta,r}(L, D, \nu)$,
3. *has the prescribed ratio of peaks*
$$\hat{f}\left(\hat{\hat{\Theta}}\right) = \beta^{-1} f\left(\hat{\Theta}\right), \tag{8.59}$$
4. *has the largest Euclidean norm of the shift in $\hat{\theta}$, both as a consequence of the optimization and of the appropriate choice of the factor β.*

Then, the statistics determining $\hat{f}$ are given by the formulas

$$\hat{L} \equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L + s\rho\, {}^{\lfloor\psi}Lx & {}^{\lfloor\psi}L \end{bmatrix} \Leftrightarrow \hat{\hat{\theta}} = \hat{\theta} + s\rho x, \ s \in \{-1, 1\}, \tag{8.60}$$

$$\begin{aligned} x &= \textit{unit-length eigenvector of } C = \left({}^{\lfloor\psi}L'\, {}^{\lfloor\psi}D\, {}^{\lfloor\psi}L\right)^{-1} \\ &\qquad \textit{corresponding to the maximum eigenvalue } \eta, \\ {}^{\lfloor\psi}\hat{D} &= {}^{\lfloor\psi}D, \ \hat{\nu} = \nu, \ {}^{\lfloor d}\hat{D} = \hat{\hat{r}}\hat{\nu} \ \textit{with} \\ \hat{\hat{r}} &= \hat{r}/z. \end{aligned} \tag{8.61}$$

The scalars $z > 1, \rho > 0$ are determined by expressions

$$z \equiv \beta^{-\frac{2}{\mathring{\psi}+2}} \tag{8.62}$$

$$\rho^2 = \gamma \left\{ -1 + \frac{1 - 2/\nu \ln(\beta) + \left(1 + \frac{\mathring{\Psi}+2}{\nu}\right) \ln(z)}{z} \right\}, \quad \textit{where}$$

$$\gamma \equiv \nu \eta \hat{r} \ \textit{and} \ \beta = \exp\left(-0.5 \frac{(\mathring{\Psi}+2)^2}{2(\mathring{\Psi}+2)+\nu} \right).$$

Proof. The optimization is split in several steps using repeatedly the specific form of the GiW pdf and Propositions 8.7, 8.1.

1. The first constraint $f\left(\hat{\hat{\Theta}}\right) = \beta f\left(\hat{\Theta}\right)$, enforcing an explicit shift of $\hat{\hat{\Theta}}$, has for the pdf $f(\Theta) = GiW(\hat{\theta}, C, \hat{r}, \nu)$ the form

$$\hat{r}^{-0.5\left(\nu+\mathring{\Psi}+2\right)} \exp[-0.5\nu]$$
$$= \beta^{-1} \hat{\hat{r}}^{-0.5(\nu+\mathring{\Psi}+2)} \exp\left[-0.5\nu \frac{\left(\hat{\hat{\theta}} - \hat{\theta}\right)' (\nu C)^{-1} \left(\hat{\hat{\theta}} - \hat{\theta}\right) + \hat{r}}{\hat{\hat{r}}} \right].$$

   Let us introduce auxiliary variables

$$z \equiv \frac{\hat{r}}{\hat{\hat{r}}} > 0, \quad \rho x \equiv \hat{\hat{\theta}} - \hat{\theta}, \ x'x = 1,$$
$$\omega \equiv \rho^2 x' (\hat{r} \nu C)^{-1} x > 0, \ \delta \equiv \frac{\mathring{\Psi}+2}{\nu} > 0, \quad q \equiv 1 - \frac{2}{\nu} \ln(\beta) > 1.$$

   Then, the above constraint gets the form $(1+\delta)\ln(z) + q = z(\omega+1)$. Note that this constraint specifies a nonempty set of admissible solutions as the pair $z = 1$ and $\omega = -2/\nu \ln(\beta)$ solves it.

2. The KL divergence expressed in terms of the introduced variables has the form

$$\mathcal{D}\left(\hat{f} \middle|\middle| f\right) = \ln\left(\frac{\Gamma(0.5\nu)}{\Gamma(0.5\hat{\nu})}\right) - 0.5 \ln\left(\left|\hat{C} C^{-1}\right|\right) - 0.5\nu \ln(\nu) - 0.5\nu \ln(z)$$
$$+ 0.5\nu \ln(\hat{\nu}) + 0.5(\hat{\nu} - \nu) \frac{\partial}{\partial(0.5\hat{\nu})} \ln(\Gamma(0.5\hat{\nu})) - 0.5\mathring{\psi} - 0.5\hat{\nu}$$
$$+ 0.5 \mathrm{tr}\left[\hat{C} C^{-1}\right] + 0.5\nu z(\omega+1).$$

3. The part of the KL divergence optimized with respect to $0.5\hat{\nu}$ reads

$$-\ln\left(\Gamma(0.5\hat{\nu})\right) + 0.5\nu \ln(0.5\hat{\nu}) + 0.5(\hat{\nu} - \nu) \frac{\partial}{\partial(0.5\hat{\nu})} \ln(\Gamma(0.5\hat{\nu})) - 0.5\hat{\nu}.$$

   Setting its derivative with respect to $\hat{\nu}$ to zero, we get

$$1 = \nu/\hat{\nu} + 0.5(\hat{\nu} - \nu)\frac{\partial^2}{\partial(0.5\hat{\nu})^2}\ln\left(\mathit{\Gamma}(0.5\hat{\nu})\right).$$

Using an asymptotic formula [156], page 250, for derivatives of $\ln(\mathit{\Gamma}(\cdot))$,

$$\frac{\partial^2}{\partial(0.5\hat{\nu})^2}\ln\left(\mathit{\Gamma}(0.5\hat{\nu})\right) \approx (0.5\hat{\nu})^{-1} + 0.5(0.5\hat{\nu})^{-2}, \text{ we get, for } y = \nu/\hat{\nu},$$

$$1 = y + (1-y)(1+y/\nu) = 1 + (1-y)y/\nu \;\Rightarrow\; y = 1 \;\Rightarrow \hat{\nu} = \nu.$$

4. Respecting the difference of normalizing factors, the second constraint $\hat{f}\left(\hat{\hat{\mathit{\Theta}}}\right) = \beta^{-1} f(\hat{\mathit{\Theta}})$ becomes

$$|\hat{C}|^{-0.5}\hat{\hat{r}}^{-0.5\left(\mathring{\mathit{\Psi}}+2\right)} = \beta^{-1}|C|^{-0.5}\hat{r}^{-0.5\left(\mathring{\mathit{\Psi}}+2\right)}.$$

   This condition provides

$$z = \left[\beta^{-2}\left|\hat{C}C^{-1}\right|\right]^{\left(\mathring{\mathit{\Psi}}+2\right)^{-1}}.$$

   Note that $\left|\hat{C}C^{-1}\right| \geq 1$ and $\beta < 1$ imply that $z > 1$.

5. By minimizing the KL divergence unconditionally with respect to $\hat{C}$, we get $\hat{C} = C$, so that z is uniquely determined

$$z = \beta^{-\frac{2}{\mathring{\mathit{\Psi}}+2}}.$$

   It will be optimum value if the remaining free parameters stay within the admissible range.

6. Given z, the value of ω is uniquely determined. We require the largest shift. Thus, x must be an eigenvector of C corresponding to the largest eigenvalue η of C. Then, $\omega = \rho^2/\gamma$ with $\gamma = \eta\nu\hat{r}$ and the first constraint implies

$$\rho^2 = \gamma\left[-1 + \frac{(1+\delta)\ln(z) + q}{z}\right].$$

7. It remains to maximize the right-hand side of the above equation and check whether it is positive for the maximizing $\beta \in (0,1)$. We can maximize the expression in question with respect to z as there is its one-to-one mapping to β. The necessary condition for extreme and form of z give

$$1 - \frac{1 - \frac{2}{\nu}\ln(\beta)}{1+\delta} = \ln(z) = -\frac{2}{\mathring{\mathit{\Psi}}+2}\ln(\beta)$$
$$\Rightarrow \beta = \exp\left(-0.5\frac{\left(\mathring{\mathit{\Psi}}+2\right)^2}{2\left(\mathring{\mathit{\Psi}}+2\right)+\nu}\right).$$

□

8.4.8 Hierarchical and optimization-based splitting

As already stated, the preferable hierarchical selection of the split factors, Section 6.4.8, is too costly to be directly used. Here, we use the specific form of the split factor and combine the hierarchical and optimization-based splitting.

Hierarchical splitting exploits the fact that during approximate estimation the ith factor within the cth component is assigned a weight $w_{ic;t} \in [0,1]$. It expresses the degree to which the current data item $d_{ic;t}$ can be thought of as generated by this factor. The correspondingly modified parameterized factor

$$f_w(d_{ic;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1)) \equiv f_w(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) \propto [f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic})]^{w_{ic;t}}$$

is used in the "ordinary" Bayes rule for updating parameter estimates of this factor. Thus, we can inspect parameter estimation of this factor without considering the others. Working with a fixed factor allows us to drop temporarily the subscripts w, i, c. We can formulate hypotheses H_0, H_1 that the objective pdf is two and one component mixture, respectively,

$$\begin{aligned}{}^{\llcorner o}f(d_t|\psi_t, \Theta_0, H_0) &= \beta\mathcal{N}_{d_t}(\theta_1'\psi_t, r_1) + (1-\beta)\mathcal{N}_{d_t}(\theta_2'\psi_t, r_2),\ \beta \in (0,1),\\ {}^{\llcorner o}f(d_t|\psi_t, \Theta, H_1) &= \mathcal{N}_{d_t}(\theta'\psi_t, r). \qquad (8.63)\end{aligned}$$

H_0 corresponds to the hypothesis that the factor should be split into the two-component mixture characterized by the parameter $\Theta_0 \equiv (\beta, \theta_1, r_1, \theta_2, r_2)$. H_1 takes the estimated factor with parameters $\Theta = (\theta, r)$ as the correct one. We assume that structures of all regression coefficients coincide; cf. Requirement 6.2. Moreover, we assume that the pdf $f(\psi)$ brings no information on parameters and hypotheses involved, i.e., natural conditions of decision making (2.36) apply

$$f(d_t, \psi|\Theta_0, H_0) = f(d_t|\psi, \Theta_0, H_0)f(\psi),\quad f(d_t, \psi|\Theta, H_1) = f(d_t|\psi, \Theta, H_1)f(\psi).$$

According to Proposition 2.15, the posterior pdf $f(\Theta|d(t))$ concentrates on minimizers of the entropy rate (2.48). We assume that it converges to its expected value. Thus, asymptotically, the evaluated posterior pdf $f(\Theta|d(t))$ estimating the normal factor $\mathcal{N}_d(\theta'\psi, r)$ concentrates on

$$\begin{aligned}\lim_{\mathring{t}\to\infty} \operatorname{supp}\left[f(\Theta|d(\mathring{t}))\right] &= \operatorname{Arg}\min_{\theta,r}\left\{0.5\ln(r) + 0.5\,{}^{\llcorner o}\mathcal{E}\left[(d-\underline{\theta}'\psi)^2/r\right]\right\}\\ &\equiv \operatorname{Arg}\min_{\theta,r}\mathcal{H}_\infty\left({}^{\llcorner o}f\middle|\middle|\,\theta, r\right). \qquad (8.64)\end{aligned}$$

The expectation ${}^{\llcorner o}\mathcal{E}$ is an expectation assigned to the objective pdf ${}^{\llcorner o}f(d,\psi)$. Note that a constant term non-influencing the minimization is dropped.

Under the hypothesis H_0 on the objective pdf (8.63), we have ${}^{\llcorner o}\mathcal{E}[\cdot] = \mathcal{E}[\cdot|\Theta_0, H_0]$ we get, for $\mathcal{H} \equiv 2\mathcal{H}_\infty\left({}^{\llcorner o}f\middle|\middle|\,\Theta \equiv [\theta, r]\right)$,

$$\begin{aligned}\mathcal{H} &= \ln(r) + r^{-1}\mathcal{E}[(d-\theta'\psi)^2|H_0, \Theta_0]\\ &= \ln(r) + \beta r^{-1}\mathcal{E}[(d-\theta'\psi)^2|\theta_1, r_1] + (1-\beta)r^{-1}\mathcal{E}[(d-\theta'\psi)^2|\theta_2, r_2]\\ &= \ln(r) + \beta\frac{r_1}{r} + (1-\beta)\frac{r_2}{r} + r^{-1}\mathcal{E}\left\{\beta[(\theta_1-\theta)'\psi]^2 + (1-\beta)[(\theta_2-\theta)'\psi]^2\right\}.\end{aligned}$$

Let us denote $C^{-1} = \mathcal{E}[\psi\psi']$. Then, points r, θ in $\mathrm{supp}\left[f(\Theta|d(\mathring{t}))\right]$ minimizing $2\mathcal{H}$ are

$$\begin{aligned} \theta &= \beta\theta_1 + (1-\beta)\theta_2 \\ r &= \beta r_1 + (1-\beta) r_2 + \beta(1-\beta)(\theta_1 - \theta_2)' C^{-1} (\theta_1 - \theta_2). \end{aligned} \tag{8.65}$$

Let us consider a point r, θ in $\mathrm{supp}\left[f(\Theta|d(\mathring{t}))\right]$ and introduce the quantities $e_t = d_t - \theta'\psi_t = \beta(d_t - \theta_1'\psi_t) + (1-\beta)(d_t - \theta_2')\psi_t$. They are generated by the <u>static</u> mixture

$$f(e_t|d(t-1)) = \beta\mathcal{N}_{e_t}(0, r_1) + (1-\beta)\mathcal{N}_{e_t}(0, r_2).$$

The best point estimate of the unknown e_t is the *filtering error*

$$\hat{e}_t \equiv \mathcal{E}[e_t|d(t)] \equiv d_t - \hat{\theta}_t'\psi_t, \; \hat{\theta}_t \equiv \mathcal{E}[\theta|d(t)]. \tag{8.66}$$

Asymptotically, it is expected to behave as e_t. Thus, we can model the filtering error $\hat{e}_t$ by a similar static mixture as e_t. In order to cope with the adopted approximations, we include constant *offsets* $\mu_1, \; \mu_2$ into the two-component static mixture describing the filtering error

$$f(\hat{e}_t|d(t-1), r_1, r_2, \mu_1, \mu_2, \beta) = \beta\mathcal{N}_{\hat{e}_t}(\mu_1, r_1) + (1-\beta)\mathcal{N}_{\hat{e}_t}(\mu_2, r_2). \tag{8.67}$$

It is estimated jointly with the factor in question. The estimation of the model (8.67) gives us the estimates $\hat{\beta} \equiv \mathcal{E}[\beta|d(\mathring{t})]$, $\hat{r}_1 \equiv \mathcal{E}[r_1|d(\mathring{t})]$, $\hat{r}_2 \equiv \mathcal{E}[r_2|d(\mathring{t})]$ of $\beta, \; r_1, \; r_2$ and v-likelihood $f(\hat{e}(\mathring{t}))$. The relationships (8.65) determine approximately the same relationships for expectations

$$\begin{aligned} \hat{\theta} &= \hat{\beta}\hat{\theta}_1 + (1-\hat{\beta})\hat{\theta}_2 \\ \hat{r} &= \hat{\beta}\hat{r}_1 + (1-\hat{\beta})\hat{r}_2 + \hat{\beta}(1-\hat{\beta})(\hat{\theta}_1 - \hat{\theta}_2)' C^{-1} (\hat{\theta}_1 - \hat{\theta}_2). \end{aligned} \tag{8.68}$$

In it, the conditional independence of involved parameters, implied by the product form of the approximately evaluated posterior pdfs (6.3) in the special case of two-component mixture, is used. Moreover, the covariance of the term $\beta(1-\beta)$ is neglected.

The second equality in (8.68) determines the value of the quadratic form

$$q'C^{-1}q = \frac{\hat{r}}{\hat{\beta}(1-\hat{\beta})} - \frac{\hat{r}_1}{1-\hat{\beta}} - \frac{\hat{r}_2}{\hat{\beta}} \equiv \rho, \; q \equiv \hat{\theta}_1 - \hat{\theta}_2. \tag{8.69}$$

Recalling that $C^{-1} = \mathcal{E}[\psi\psi']$, we see, (8.19), that we have at disposal its sampled version while estimating θ, r.

We make the splitting unique by selecting the most distant $\hat{\theta}_1, \; \hat{\theta}_2$, by selecting q that fulfills (8.69) with the largest norm $q'q$. The vector $\hat{q} = \sqrt{\zeta\rho}q_0$, where q_0 is the unit eigenvector of C corresponding to the largest eigenvalue ζ of C can be shown to be vector of this type. Having it, we can use the first equality in (8.68) and specify

$$\hat{\theta}_1 = \hat{\theta} + (1-\hat{\beta})\hat{q},\ \hat{\theta}_2 = \hat{\theta} - \hat{\beta}\hat{q}. \tag{8.70}$$

The data attributed originally to a single factor should be distributed between newly created factors. It leads to the option

$$C_1 = \hat{\beta}^{-1}C,\ C_2 = (1-\hat{\beta})^{-1}C \text{ and } \nu_1 = \hat{\beta}\nu,\ \ \nu_2 = (1-\hat{\beta})\nu. \tag{8.71}$$

It concludes the specification of the splitting step.

In order to judge whether splitting is necessary, it is sufficient to evaluate the v-likelihood under the hypothesis H_1, i.e., to estimate during estimation also a one-component counterpart of model (8.63), i.e., $\mathcal{N}_{\hat{e}_t}(\mu, r)$. The test is described by the following algorithm that also generates as a byproduct quantities needed for performing the split. It is prepared for the normalized data that lie predominantly in the range $[-1, 1]$.

Algorithm 8.8 (Test on call for a factor split)

Initial mode

- *Select the decision level* $\gamma \in (0,1),\ \gamma \to 1$.
- *Attach to each factor* ic, $i \in i^*$, $c \in c^*$, *the two-component, one-dimensional, static mixture given by the sufficient LS statistics (8.17)*

$$\begin{aligned}
\hat{\theta}_{1ic;0} &= -\hat{\theta}_{2ic;0} \in (1,5),\ \ \textit{(significant shifts from the expected value 0)},\\
\hat{r}_{1ic;0} &= \hat{r}_{2ic;0} \in (10^{-6}, 10^{-4})\\
&\qquad\qquad \textit{(the noise level in the range (0.1,1)\%) of the data range},\\
C_{1ic;0} &= C_{2ic;0} = 1/\hat{r}_{1ic;0}\\
&\qquad\qquad \textit{(covariance of parameter estimates expected to be order} \approx 1\textit{)},\\
\nu_{1ic;0} &= \nu_{2ic;0} = 2(2 + 1/\varepsilon^2),\ \ \varepsilon \in (0.1, 1),\\
&\qquad\qquad \left(\textit{the ratio } \frac{\sqrt{\mathrm{var}(r)}}{\hat{r}} \textit{ expected in (10,100)\%, cf. (8.23)}\right),\\
\kappa_{1ic;0} &= \kappa_{2ic;0} = 0.05\mathring{t},\ \ \textit{(the standard option)}.
\end{aligned}$$

- *Attach to each factor* ic, $i \in i^*$, $c \in c^*$, *the one-component static mixture given by the sufficient LS statistics; cf. (8.17). It can be formally chosen as the two-component mixture described above but with completely overlapping components. Thus, all options of the real two-component mixture are copied but point estimates of positions are selected to be identical* $\hat{\theta}_{1ic;0} = -\hat{\theta}_{2ic;0} = 0$.
- *Set the* v*-likelihood values assigned to one- and two-component mixtures* ${}^{\lfloor 1}l_{ic;0} = 0$ *and* ${}^{\lfloor 2}l_{ic;0} = 0$.

Learning mode

```
For   $t = 1, \ldots, \mathring{t}$
            Perform a step of the QB algorithm
```

```
          estimating parameters of the original mixture.
    For  $c = 1, \ldots, \mathring{c}$
      For  $i = 1, \ldots, \mathring{d}$
          Weight the data vectors $[\hat{e}_{ic;t}, 1] \equiv$ [filtering error, $1]$ by
          $\sqrt{w_{ic;t}} = \sqrt{\text{weight of the factor in question assigned by QB}}$.
          Run in parallel the QB algorithm on the weighted data
          fed in the one- and two-component mixtures.
          Update the v-likelihood values $^{\lfloor 1}l_{ic;t}$, $^{\lfloor 2}l_{ic;t}$.
      end  of the cycle over i
    end  of the cycle over c
  end  of the cycle over t
Decision mode
    For  $i = 1, \ldots, \mathring{d}$
      For  $c = 1, \ldots, \mathring{c}$
        Denote the factor ic as ready for splitting if $\frac{^{\lfloor 2}l_{ic;\mathring{t}}}{^{\lfloor 2}l_{ic;\mathring{t}} + {}^{\lfloor 1}l_{ic;\mathring{t}}} > \gamma$
        and store the terminal values of the LS statistics.
      end  of the cycle over c
    end  of the cycle over i
```

Estimation results obtained for two-component mixture on prediction errors allow us to summarize the overall factor-splitting algorithm. The fixed indexes $ic; \mathring{t}$ are dropped in it.

Algorithm 8.9 (Splitting of factors with LS statistics $\hat{\theta}, \hat{r}, C, \nu$)

1. *Define LS statistics of newly created factors as follows.*
   - $\hat{r}_1, \hat{r}_2 \equiv$ *estimates of noise covariance of respective components of the two-component mixture estimated on the weighted filtering errors,*
   - $\nu_1 = \hat{\beta}\nu$, $\nu_2 = (1 - \hat{\beta})\nu$, *where* $\hat{\beta}$ *is the weight corresponding of the first component in the two-component mixture on filtering errors, cf. (8.71),*
   - $C_1 = \hat{\beta}^{-1}C$, $C_2 = (1 - \hat{\beta})^{-1}C$, *cf. (8.71).*
2. *Reduce temporarily factor structure to significant one; see Algorithm 8.13.*
3. *Evaluate the unit eigenvector* q_0 *corresponding to the maximum eigenvalue* ζ *of the LS covariance factor of parameters* C *(8.23) using (8.58).*
4. *Compute the norm of the shift in point estimates of regression coefficient*

$$\rho = \sqrt{\nu\zeta\left(\frac{\hat{r}}{\hat{\beta}(1-\hat{\beta})} - \frac{\hat{r}_1}{1-\hat{\beta}} - \frac{\hat{r}_2}{\hat{\beta}}\right)}.$$

The multiplying factor ν respects that inversion of the LS covariance factors is a <u>*non-normalized version*</u> *of the* <u>*sampling covariance*</u> *of regression vectors.*

The inspected factor is to be denoted as not ready for splitting if the argument of the above square root is negative.

5. *Complement the definition of the LS statistics by* $\hat{\theta}_1 = \hat{\theta} + s\rho q_0$, $\hat{\theta}_2 = \hat{\theta} - s\rho q_0, \quad s \in \{-1, 1\}$. *The sign s is selected randomly.*

Problem 8.2 (On a systematic choice of the combined factors) *The random choice in the last step of the above algorithm can be replaced by the choice attempting to maximize the resulting v-likelihood. Algorithm 12.5 used in so called shadow cancelling problem may serve to this purpose.*

8.4.9 Techniques applicable to static mixtures

Normality of the mixture brings nothing specific to this counterpart of Chapter 6. The following problem expresses the only conjecture worth mentioning.

Problem 8.3 (Learning of dynamic mixtures via static ones) *We conjecture that under rather weak conditions Proposition 6.17 is applicable to normal mixtures even for the considered continuous-valued data. This should be proved formally.*

8.5 Approximate parameter estimation

The approximate estimation described in Section 6.5 is specialized here to normal components. The most general variants are only described.

8.5.1 Quasi-Bayes estimation

A direct application of Algorithm 6.13 gives the specialized algorithm.

Algorithm 8.10 (Quasi-Bayes algorithm with common factors)
Initial (offline) mode

- *Select the structure of the mixture, i.e., specify the number of components* $\mathring{c}$ *and the ordered lists of factors allocated to the considered components. The factor structure for each ic is determined by the structure of the corresponding data vector* Ψ_{ic}.
- *Select the statistics* $L_{ic;0}, D_{ic;0}, \nu_{ic;0}$ *determining prior pdfs of the individual factors*

$$GiW_{\theta_{ic},r_{ic}}(V_{ic;0}, \nu_{ic;0}) \equiv GiW_{\theta_{ic},r_{ic}}(L_{ic;0}, D_{ic;0}, \nu_{ic;0}).$$

- *Select the initial values* $\kappa_{c;0} > 0$ *determining the Dirichlet pdf* $Di_{\alpha}(\kappa_0)$ *of the component weights* α, *say, about* $0.1\mathring{t}/\mathring{c}$.

- *Select forgetting factor $\lambda \in (0,1]$ if you intend to use forgetting.*
- *Specify the $L'DL$ decomposition of statistics used in alternative GiW pdf GiW $\left({}^{\lfloor A}V_{ic;t},\ {}^{\lfloor A}\nu_{ic;t} \right)$ pdf if you intend to use forgetting. Typically, ${}^{\lfloor A}V_{ic;t} = V_{ic;0}$, ${}^{\lfloor A}\nu_{ic;t} = \nu_{ic;0}$, $\forall t \in t^*$.*
- *Specify the alternative to component weights ${}^{\lfloor A}f(\alpha) = Di_{\alpha}\left({}^{\lfloor A}\kappa_{c;t} \right)$. Again, ${}^{\lfloor A}\kappa_{c;t} = {}^{\lfloor A}\kappa_{c;0}$ as a rule.*
- *Compute initial estimates of the component weights $\hat{\alpha}_{c;0} = \frac{\kappa_{c;0}}{\sum_{\tilde{c} c^*} \kappa_{\tilde{c};0}}$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots$,*

1. *Acquire the data record d_t.*
2. *Perform, for each individual factor $i \in i^* = \{1, \ldots, \mathring{d}\}$, $c \in c^*$, a* trial up-dating *of $L_{ic;t-1}$, $D_{ic;t-1}$ by the data vector $\Psi_{ic;t}$ using Algorithm 8.2 with the forgetting factor $\lambda = 1$. These operations give values of the predictive pdfs*

$$
\begin{aligned}
f(d_{ic;t}|d(t-1), c) &= \frac{\mathcal{I}(d(t)|ic)}{\sqrt{2\pi}\mathcal{I}(d(t-1)|ic)},\ (2.47),\quad \textit{with}\\
\mathcal{I}(d(t)|ic) &\equiv \mathcal{I}(L_{ic;t}, D_{ic;t}, \nu_{ic;t})\\
&= \Gamma(0.5\nu_{ic;t})\, {}^{\lfloor d}D_{ic;t}^{-0.5\nu_{ic;t}} \left| {}^{\lfloor \psi}D_{ic;t} \right|^{-0.5} 2^{0.5\nu_{ic;t}} (2\pi)^{0.5\mathring{\psi}}.
\end{aligned}
$$

3. *Compute values of the predictive pdfs, for each component $c \in c^*$,*

$$f(d_t|d(t-1), c) = \prod_{i \in i^*} f(d_{ic;t}|d(t-1), c).$$

4. *Compute the probabilistic weights $w_{c;t}$, using the formula*

$$w_{c;t} = \frac{\hat{\alpha}_{c;t-1} f(d_t|d(t-1), c)}{\sum_{\tilde{c} \in c^*} \hat{\alpha}_{\tilde{c};t-1} f(d_t|d(t-1), \tilde{c})},\ c \in c^*.$$

5. *Update the scalars*

$$\kappa_{c;t} = \lambda(\kappa_{c;t-1} + w_{c;t}) + (1-\lambda)\, {}^{\lfloor A}\kappa_{c;t},\ c \in c^*;$$

cf. (6.88).

6. *Update Bayesian parameter estimates of different factors, i.e., update the corresponding statistics*

$$L_{ic;t-1}, D_{ic;t-1}, \nu_{ic;t-1} \to L_{ic;t}, D_{ic;t}, \nu_{ic;t}.$$

The updating is equivalent to the weighted version of (8.29)

$$
\begin{aligned}
V_{ic;t} &= \lambda(V_{ic;t-1} + w_{ic;t}\Psi_{ic;t}\Psi'_{ic;t}) + (1-\lambda)\, {}^{\lfloor A}V_{ic;t},\ V_{ic;0} \textit{ given}\\
\nu_{ic;t} &= \lambda(\nu_{ic;t-1} + w_{ic;t}) + (1-\lambda)\, {}^{\lfloor A}\nu_{ic;t},\ \nu_{ic;0} \textit{ given} \qquad (8.72)\\
w_{ic;t} &= \sum_{\tilde{c} \in c_i^*} w_{\tilde{c};t} \qquad (8.73)
\end{aligned}
$$

with c_i^* being a set of pointers to components that contain an ith factor. Numerically, $D_{ic;t-1}$ are first multiplied by $\lambda = forgetting\_factor$ and then real updating of $L_{ic;t-1}$, $D_{ic;t-1}$ by data vectors $\Psi_{ic;t}$ is performed using Algorithm 8.2 with $\lambda = (forgetting\_factor) \times w_{ic;t}$. The alternative is added "dyad-wise" with weights $(1-(forgetting\_factor))\, {}^{\lfloor A}D_{ic;t}$, where ${}^{\lfloor A}D_{ic;t}$ is the corresponding diagonal of the $L'DL$ decomposition of ${}^{\lfloor A}V_{ic;t}$.

7. *Evaluate the point estimates of the mixing weights*

$$\mathcal{E}[\alpha_c|d(t)] = \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c};t}} \equiv \hat{\alpha}_{c;t}$$

and, if need be, characteristics of the pdf $f(\Theta_{ic}|d(t)) = GiW_{\theta_{ic},r_{ic}}(V_{ic;t}, \nu_{ic;t})$ describing other parameters Θ_{ic}.

8. *Go to the beginning of* Sequential mode *while data are available, while $t \leq \mathring{t}$.*

Remark(s) 8.8

1. *The predictive pdf can be computed in its Student form (8.31). It makes sense especially when prediction errors are explicitly needed.*
2. *The stabilized forgetting could be applied to the statistics $\kappa_{c;t}$ with its specific forgetting factor.*

8.5.2 EM estimation

A direct application of EM algorithm 6.15 to normal factors looks as follows.

Algorithm 8.11 (EM mixture estimation with common factors)

Initial mode

- *Select the upper bound $\mathring{n}$ on the number n of iterations and set $n = 0$.*
- *Select the structure of the mixture, i.e., specify the number of components $\mathring{c}$ and the ordered lists of factors allocated to the considered components. The factor structure for each ic is determined by the structure of the corresponding data vector Ψ_{ic}.*
- *Select the point estimates $\hat{\Theta}_{icn} \equiv \left[\hat{\theta}_{icn}, \hat{r}_{icn}\right]$ of parameters $\Theta_{ic} \equiv [\theta_{ic}, r_{ic}]$ characterizing the ith normal factor within the cth normal component.*
- *Select $\kappa_{cn;0} = 1$ that corresponds with the point estimates $\hat{\alpha}_{cn} = 1/\mathring{c}$ of components weights α_c.*

Iterative mode

1. *Use the current point estimate $\hat{\Theta}_n$ in the following evaluations.*
2. *Fill $L'DL$ decomposition of the extended information matrix related to the ith factor within the cth component as follows.*

$$L_{icn;0} = \begin{bmatrix} 1 & \\ \hat{\theta}_{icn}\varepsilon & I \end{bmatrix}, \; D_{icn;0} = \begin{bmatrix} \hat{r}_{icn}\varepsilon & \\ 0 & \varepsilon I \end{bmatrix}, \; \varepsilon > 0, \; \varepsilon \to 0. \qquad (8.74)$$

Initialize the corresponding $\nu_{icn;0} = \varepsilon$.

Sequential mode, *running for* $t = 1, 2, \ldots,$

a) *Construct the data vectors* $\Psi_{ic;t}$.

b) *Compute for this* $\Psi_{ic;t}$ *values of predictive pdfs* $f\left(d_{ic;t}|\psi_{ic;t}, \hat{\Theta}_{icn}, c\right) \equiv \mathcal{N}_{d_{ic;t}}\left(\hat{\theta}'_{icn}\psi_{ic;t}, \hat{r}_{icn}\right)$ *for each individual factor* $i \in i^* = \{1, \ldots, \mathring{d}\}$ *in all components* $c \in c^*$ *using the parameter estimates* $\hat{\Theta}_{icn}$ *that are constant during the time cycle of the sequential mode. Thus, the approximate, fixed, one-step-ahead predictor with certainty-equivalence approximation is used.*

c) *Compute the values of the predictive pdfs*

$$f\left(d_t|\psi_{c;t}, \hat{\Theta}_{cn}, c\right) \equiv \prod_{i \in i^*} \mathcal{N}_{d_{ic;t}}\left(\hat{\theta}'_{icn}\psi_{ic;t}, \hat{r}_{icn}\right)$$

for each component $c \in c^*$.

d) *Compute the probabilities* $w_{cn;t}$ *approximating* δ_{c,c_t}

$$w_{cn;t} = \frac{f\left(d_t|\psi_{c;t}, \hat{\Theta}_{cn}, c\right)\hat{\alpha}_{cn}}{\sum_{\tilde{c} \in c^*} f\left(d_t|\psi_{\tilde{c};t}, \hat{\Theta}_{cn}, \tilde{c}\right)\hat{\alpha}_{\tilde{c}n}}.$$

e) *Update the statistics determining the log-likelihood functions describing different factors*

$$V_{icn;t} = V_{icn;t-1} + w_{icn;t}\Psi_{ic;t}\Psi'_{ic;t}, \ \nu_{icn;t} = \nu_{icn;t-1} + w_{icn;t}$$
$$w_{icn;t} = \sum_{\tilde{c} \in c_i^*} w_{\tilde{c}n;t}. \tag{8.75}$$

The set c_i^* *includes pointers to components that contain an ith factor. The recursive step (8.75) is numerically performed by updating* $L'DL$ *decompositions of extended information matrices with data vectors* $\Psi_{ic;t}$ *weighted by* $w_{icn;t}$ *(using Algorithm 8.2).*

f) *Update* $\kappa_{c;t} = \kappa_{c;t-1} + w_{c;t}, \ c \in c^*$.

g) *Go to the beginning of* Sequential mode *if* $t \leq \mathring{t}$. *Otherwise continue.*

3. *Find the new point estimates* $\hat{\Theta}_{ic(n+1)} = \left[{}^{\lfloor\psi}L^{-1}_{icn;\mathring{t}} \, {}^{\lfloor d\psi}L_{icn;\mathring{t}}, \, {}^{\lfloor d}D_{icn;\mathring{t}}/\nu_{icn;\mathring{t}}\right]$ *of* $\Theta_{ic} \equiv [\theta_{ic}, r_{ic}]$ *and* $\hat{\alpha}_{c(n+1)} \propto \kappa_{cn;\mathring{t}}$.
These values are maximizing arguments of the nth approximate likelihood.
4. *Stop if the log-likelihood value*

$$\sum_{c \in c^*} \kappa_{cn;\mathring{t}}\left[\ln(\hat{\alpha}_{cn;\mathring{t}}) - 0.5\sum_{i \in i^*}\ln\left({}^{\lfloor d}D_{icn;\mathring{t}}/\nu_{icn;\mathring{t}}\right)\right] \tag{8.76}$$

is not increasing any more or $n = \mathring{n}$. *Otherwise set* $n = n + 1$ *and go to the beginning of* Iterative mode.

8.5.3 Batch quasi-Bayes estimation

Here, Algorithm 6.16 is specialized to normal mixtures. In this way, processing-order-independent, batch quasi-Bayes estimation of normal mixtures is gained.

Algorithm 8.12 (BQB mixture estimation with common factors)

Initial mode

- *Select the upper bound* $\mathring{n}$ *on the number* n *of iterations and set* $n = 0$.
- *Select the structure of the mixture, i.e., specify the number of components* $\mathring{c}$ *and the ordered lists of factors allocated to the considered components. The factor structure for each* ic *is determined by the structure of the corresponding data vector* Ψ_{ic}.
- *Set the maximum of the v-log-likelihood* $\bar{l} = -\infty$.
- *Select the statistics* $\bar{L}_{ic}, \bar{D}_{ic}, \bar{\nu}_{ic}$ *determining (flat) pre-prior pdfs of the individual factors* $GiW_{\theta_{ic},r_{ic}}\left(\bar{V}_{ic}, \bar{\nu}_{ic}\right) \equiv GiW_{\theta_{ic},r_{ic}}\left(\bar{L}_{ic}, \bar{D}_{ic}, \bar{\nu}_{ic}\right)$.
- *Select the values* $\bar{\kappa}_c > 0$ *determining a flat pre-prior Dirichlet pdf on component weights. These values serve for flattening.*
- *Select the statistics* $L_{icn}, D_{icn}, \nu_{icn}$ *determining prior pdfs of the individual factors*
$$GiW_{\theta_{ic},r_{ic}}(V_{icn}, \nu_{icn}) \equiv GiW_{\theta_{ic},r_{ic}}(L_{icn}, D_{icn}, \nu_{icn}).$$
- *Select the initial values* $\kappa_{cn} > 0$ *determining a prior Dirichlet pdf on component weights.*
- *Make copies* $L_{ic;0} = L_{icn}$, $D_{ic;0} = D_{icn}$, $\nu_{ic;0} = \nu_{icn}$ *and* $\kappa_{c;0} = \kappa_{cn}$.

Iterative mode

1. *Use the current prior pdf*
$$f_n(\Theta) = Di_\alpha(\kappa_n) \prod_{c\in c^*} \prod_{i\in i^*} GiW_{\theta_{ic},r_{ic}}(L_{icn}, D_{icn}, \nu_{icn})$$
in the following evaluations.
2. *Set the value of v-log-likelihood* $l_{n;0} = 0$.
3. *Compute the point estimates of the components weights* $\hat{\alpha}_{cn} = \frac{\kappa_{cn}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c}n}}$.

   Sequential mode, *running for* $t = 1, 2, \ldots,$

   a) *Construct the data vectors* $\Psi_{ic;t}$.

   b) *Compute for these data vectors the values of the predictive pdfs*
$$f_n(d_{ic;t}|\psi_{ic;t}, c) = \int f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c) f_n(\Theta_{ic})\, d\Theta_{ic}$$
$$= \frac{\Gamma(0.5(\nu_{icn}+1))\left[{}^{\lfloor d}D_{icn}(1+\zeta_{icn;t})\right]^{-0.5}}{\sqrt{\pi}\Gamma(0.5\nu_{icn})\left(1+\frac{\hat{e}^2_{icn;t}}{{}^{\lfloor d}D_{icn}(1+\zeta_{ic;t})}\right)^{0.5(\nu_{icn}+1)}},$$
$$\hat{e}_{icn;t} \equiv d_t - {}^{\lfloor d\psi}L'_{icn}\left({}^{\lfloor\psi}L'_{icn}\right)^{-1}\psi_{ic;t}$$
$$\zeta_{icn;t} = \psi'_{ic;t}\,{}^{\lfloor\psi}L^{-1}_{icn}\,{}^{\lfloor\psi}D^{-1}_{icn}\left({}^{\lfloor\psi}L'_{icn}\right)^{-1}\psi_{ic;t}$$

for each factor $i \in i^*$ in all components $c \in c^*$ using the prior pdfs $GiW_{\theta_{ic},r_{ic}}(L_{icn}, D_{icn}, \nu_{icn})$ that are constant during time cycle.

c) *Compute the values of predictive pdfs*

$$f_n(d_t|\psi_{c;t}, c) \equiv \prod_{i \in i^*} f_n(d_{ic;t}|\psi_{ic;t}, c)$$

for each component $c \in c^*$.

d) *Update the v-log-likelihood $l_{n;t} = l_{n;t-1} + \ln\left(\sum_{c \in c^*} \hat{\alpha}_{cn} f_n(d_t|\psi_{c;t}, c)\right)$.*

e) *Compute the probabilistic weights $w_{cn;t}$ approximating δ_{c,c_t} by*

$$w_{cn;t} = \frac{\hat{\alpha}_{cn} f_n(d_t|\psi_{c;t}, c)}{\sum_{\tilde{c} \in c^*} \hat{\alpha}_{\tilde{c}n} f_n(d_t|\psi_{\tilde{c};t}, \tilde{c})}.$$

f) *Update the statistics determining the posterior pdfs evolving from copies of the prior statistics $L_{ic;0}$, $D_{ic;0}$, $\nu_{ic;0}$ and $\kappa_{c;0}$. The factorized equivalent of the update of the extended information matrix is performed*

$$\begin{aligned} V_{icn;t} &= V_{icn;t-1} + w_{icn;t} \Psi_{ic;t} \Psi'_{ic;t}, \quad V_{icn;t} \equiv L'_{icn;t} D_{icn;t} L_{icn;t} \\ \kappa_{c;t} &= \kappa_{c;t-1} + w_{cn;t} \\ \nu_{icn;t} &= \nu_{icn;t-1} + w_{icn;t} \text{ with } w_{icn;t} = \sum_{\tilde{c} \in c_i^*} w_{\tilde{c}n;t}. \end{aligned} \tag{8.77}$$

The set c_i^* includes the pointers to components that contain the ith factor.

g) *Go to the beginning of* Sequential mode *if $t \leq \mathring{t}$. Otherwise continue.*

4. *Stop if the v-likelihood of the mixture does not increase $l_{n;\mathring{t}} < \bar{l}$, or the compared values are perceived as the same in the vein of Proposition 6.2, or $n = \mathring{n}$. Otherwise set $\bar{l} = l_{n;\mathring{t}}$, increase the iteration counter $n = n + 1$, apply flattening operation to $f(\Theta_{ic}|d(\mathring{t}))$ and $Di_{\alpha}(\kappa_{c;\mathring{t}})$*

$$\begin{aligned} D_{icn} &= \Lambda_n D_{ic;\mathring{t}} + (1 - \Lambda_n)\bar{D}_{ic}, \quad \nu_{icn} = \Lambda_n \nu_{ic;\mathring{t}} + (1 - \Lambda_n)\bar{\nu}_{ic} \\ \kappa_{cn} &= \Lambda_n \kappa_{c;\mathring{t}} + (1 - \Lambda_n)\bar{\kappa}_c, \quad \Lambda_n \to 0.5 \text{ according to Proposition 6.11.} \end{aligned}$$

5. *Go to the beginning of* Iterative mode.

Remark(s) 8.9

1. *As discussed in Remarks 6.22, the BQB algorithm uses Bayesian predictors for estimating δ_{c,c_t}. They respect uncertainty of the current estimates of unknown parameters. Predictions become too cautious if this uncertainty is initialized to high values. This may break down the algorithm completely. In the context of normal pdfs, it happens if the factor $\zeta = \psi' C \psi$ becomes too high. Then, the influence of prediction errors $\{\hat{e}\}$ that predominantly indicates membership of the data vector to the specific components is suppressed too much. Knowing the danger, the remedy is simple. Essentially,*

predictions used in the EM algorithm that ignore these uncertainties have to be used in several initial iterative steps of the algorithm, i.e., the value $\zeta = 0$ is enforced.

2. Quasi-EM algorithm can be simply specialized to normal mixtures. It can be used whenever use of quasi-Bayes algorithm is time-critical.

8.6 Structure estimation

Efficiency of the advisory system depends strongly on the quality of the model of the o-system. A good choice of the used structure of the mixture (see Agreement 5.4) is quite important in this respect. It includes an appropriate selection of

- quantities among those measured on the o-system that should be used by the p-system;
- the structure of individual factors; it mostly means data entries and delayed data entries used in the state vector in the phase form;
- the structure of components, i.e., the order of factors used;
- the structure of the mixture determined by the number of components and specification of the common factors in them.

All these tasks can be formally solved as the Bayesian estimation of discrete-valued pointers to alternative structures. The excessive number of possible alternatives makes the problem nontrivial. Details needed for implementing the strategy outlined in Section 6.6 are described here.

8.6.1 Estimation of factor structure

Here, we summarize facts that make a basis for the tailored choice of the neighborhood for normal factors. Publications [93, 95, 162] serve as references for details.

Agreement 8.2 (Nested normal factors) *Let regression vectors $\psi, \tilde{\psi}$ of a pair of normal factors predicting the same data item d fulfill*

$$\psi' = \left[\tilde{\psi}', \bullet\right] \tag{8.78}$$

with $\bullet$ marking here and below arbitrary arrays of appropriate dimensions. Let also the statistics V_0, $\tilde{V}_0$, ${}^{\lfloor A}V$, ${}^{\lfloor A}\tilde{V}$ of the conjugate prior pdfs $f(\Theta) = GiW_{\Theta}(V_0, \nu_0)$, $f\left(\tilde{\Theta}\right) = GiW_{\tilde{\Theta}}\left(\tilde{V}_0, \tilde{\nu}_0\right)$ and alternative pdfs, also in the GiW form, ${}^{\lfloor A}f(\Theta) = GiW_{\Theta}\left({}^{\lfloor A}V, {}^{\lfloor A}\nu\right)$, ${}^{\lfloor A}f(\tilde{\Theta}) = GiW_{\tilde{\Theta}}\left({}^{\lfloor A}\tilde{V}, {}^{\lfloor A}\tilde{\nu}\right)$ (8.16) fulfill

$$V_0 = \begin{bmatrix} \tilde{V}_0 & \bullet \\ \bullet & \bullet \end{bmatrix}, \quad {}^{\lfloor A}V = \begin{bmatrix} {}^{\lfloor A}\tilde{V} & \bullet \\ \bullet & \bullet \end{bmatrix}. \tag{8.79}$$

Then, we say that the factor $f(d|\tilde{\psi}, \tilde{\Theta})$ is nested into the factor $f(d|\psi, \Theta)$.

Proposition 8.21 (Estimation and prediction with nested factors) *Let natural conditions of decision making, Requirement 2.5, hold. The considered parameterized factor (8.1) is normal, given by the regression vector* ψ*. Moreover, let a conjugate prior (3.13) and a conjugate alternative (see Section 3.1) be used in the stabilized forgetting. Let another normal factor* $f\left(d\middle|\tilde{\psi},\tilde{\Theta}\right)$*, determined by the regression vector* $\tilde{\psi}$*, be nested into the factor* $f(d|\psi,\Theta)$*. Then, the statistic* V_t *describing estimates of the factor* $f(d|\psi,\Theta)$ *evolves according to the recursion*

$$V_t = \lambda(V_{t-1} + \Psi\Psi') + (1-\lambda)\,{}^{\lfloor A}V_t,\ V_0 \text{ given.}$$

It provides also updating of the statistics $\tilde{V}_t$ *as*

$$V_t = \begin{bmatrix} \tilde{V}_t & \bullet \\ \bullet & \bullet \end{bmatrix},\ \forall t \in t^*. \tag{8.80}$$

Let us decompose all symmetric positive definite V*-matrices into* LDL' *decomposition (it differs from the decomposition* $L'DL$ *used otherwise in learning!)*

$$V = LDL',\ L \text{ is lower triangular matrix with unit diagonal} \tag{8.81}$$

$$D \text{ is diagonal matrix with nonnegative entries}$$

$$L = \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix},\quad D = \begin{bmatrix} {}^{\lfloor d}D & 0 \\ 0 & {}^{\lfloor \psi}D \end{bmatrix},\quad {}^{\lfloor d}D \text{ is scalar.}$$

Then,
$$L = \begin{bmatrix} \tilde{L} & 0 \\ \bullet & \bullet \end{bmatrix},\ D = \begin{bmatrix} \tilde{D} & 0 \\ 0 & \bullet \end{bmatrix}. \tag{8.82}$$

The normalization integral defining the likelihood function of data for the given structure is

$$\mathcal{I}(V,\nu) = \Gamma(0.5\nu)\left({}^{\lfloor d}D\right)^{-0.5\nu}\left|{}^{\lfloor \psi}D\right|^{-0.5} 2^{0.5(\nu-2)}(2\pi)^{0.5\mathring{\psi}} \tag{8.83}$$

$$\Gamma(x) = \int_0^{\infty} z^{x-1}\exp(-z)\,dz < \infty \text{ for } x > 0.$$

The v-likelihood corresponding to the considered structure is

$$f(d(\mathring{t})) = \frac{\mathcal{I}(V_{\mathring{t}},\nu_{\mathring{t}})}{\mathcal{I}(V_0,\nu_0)}. \tag{8.84}$$

Proof. Detailed evaluations can be found in [162]. Nesting of the extended information matrix follows directly from assumptions on nesting of the initial and alternative values together with the nesting of data vectors and the form of updating. The definition of LDL' (!) decomposition implies its nesting. The form of the normalizing integral $\mathcal{I}(V,\nu)$ given in Proposition 8.7 is preserved in spite of the change of the decomposition type as it depends on the determinant of matrices V and ${}^{\lfloor \psi}V$ that can be unambiguously computed from diagonals D and ${}^{\lfloor \psi}D$ only. The formula (8.84) results from (2.47) and the chain rule for pdfs, Proposition 2.4. □

Remark(s) 8.10

1. *The version with $L'DL$ decomposition is prepared, too. It is nested at the tail position of ψ and L, D.*
2. *Note that the normal pdf belongs to the exponential family (3.6) with $B(\Psi) = \Psi\Psi'$. The nested normal model is obtained by cancelling entries in the regression vector at its tail. This can obviously be interpreted as a linear operator defining a nesting mapping on $B(\Psi)$; see Proposition 3.3. Thus, Proposition 8.21 is mostly implied by Agreement 8.2. The cancellation at tails only and the choice of the decomposition LDL' make elements of LDL' nested, too.*
3. *The nesting property can be and is used for an efficient evaluation of $f(d(\mathring{t})|s)$ for a rich neighborhood of some structure s_0. It contains*
   - *all regression vectors that can be gained by deleting one entry from that determined by s_0 or by adding one regressor from the richest regression vector ψ_r;*
   - *all regression vectors that are nested into regression vectors specified under the previous item.*

   By evaluating $\mathring{\psi}$ different LDL' decompositions of V we get much higher number of values of likelihood functions for nested structures. Moreover, recomputation of these LDL' decomposition that guarantee appropriate nesting is computationally cheap. It can be performed by the repetitive permutation of adjacent entries in the regression vector and corresponding restoration of the LDL' decomposition of the extended information matrix V; see Proposition 8.4.
4. *Nesting of the prior statistics is guaranteed for weakly informative prior pdfs. The nesting is violated when a nontrivial physical knowledge is used; see Section 8.3. An efficient algorithm coping with the induced problems is described in [64].*

8.6.2 Structure estimation in factor splitting

We address the problem that arises in connection with branching by factor splitting; see Section 8.4.7. Essentially, we need to perform splitting in the space of significant parameters and to extend it to the original parameter space. Particular steps and their solutions are given in the following algorithm.

Algorithm 8.13 (Factor splitting steps in a significant subspace)

1. Estimation of the factor structure $f(\theta, r) = GiW_{\theta,r}(L, D, \nu)$ by analysis of the given $L'DL$ decomposition of the extended information matrix. *Using Algorithm 8.3, we transform $L'DL$ to $\tilde{L}\tilde{D}\tilde{L}'$ and apply the structure estimation algorithm outlined in Section 8.6.1.*
2. Analysis of the marginal $f({}^{\llcorner a}\theta, r) = GiW_{{}^{\llcorner a}\theta,r}(\cdot)$ pdf on significant regression coefficients ${}^{\llcorner a}\theta$. *The needed marginal pdf*

$$f\left({}^{\lfloor a}\theta, r\right) = GiW_{{}^{\lfloor a}\theta, r}\left(\begin{bmatrix} 1 & \\ {}^{\lfloor da}L & {}^{\lfloor a}L \end{bmatrix}, \begin{bmatrix} {}^{\lfloor d}D & \\ & {}^{\lfloor a}D \end{bmatrix}, \nu\right)$$

is obtained directly from the original pdf whose statistics are split as follows. $f(\theta, r) = GiW_{\theta,r}(L, D, \nu)$

$$L \equiv \begin{bmatrix} 1 & & \\ {}^{\lfloor da}L & {}^{\lfloor a}L & \\ {}^{\lfloor db}L & {}^{\lfloor ab}L & {}^{\lfloor b}L \end{bmatrix}, \quad D \equiv \begin{bmatrix} {}^{\lfloor d}D & & \\ & {}^{\lfloor a}D & \\ & & {}^{\lfloor b}D \end{bmatrix}.$$

This simple relationship holds if the regression coefficients ${}^{\lfloor a}\theta$ *are at a leading position of all regression coefficients; see Proposition 8.8. This configuration is achieved by their permutation with the corresponding restoration of* $L'DL$ *according to Proposition 8.5. The conditional pdf of insignificant coefficients* $f\left({}^{\lfloor b}\theta | {}^{\lfloor a}\theta, r\right)$ *is obtained as a byproduct. The statistics determining it are in the part of the permuted extended information matrix that is unexploited in splitting. The mentioned analysis deals with the least-squares version of the found marginal pdf. One-to-one relationships (8.18), (8.19), (8.20) provide it.*

3. Splitting of the factor reduced on ${}^{\lfloor a}\theta, r$. *The solution is presented in Section 8.4.7 and modifies the reduced decomposition of the extended information matrix to the other one given by, say,*

$$\begin{bmatrix} 1 & \\ {}^{\lfloor da}\underline{L} & {}^{\lfloor a}\underline{L} \end{bmatrix}, \begin{bmatrix} {}^{\lfloor d}\underline{D} & \\ & {}^{\lfloor a}\underline{D} \end{bmatrix}.$$

4. Extension of the modified pdf

$$\underline{f}({}^{\lfloor a}\theta, r) = GiW_{{}^{\lfloor a}\theta, r}\left(\begin{bmatrix} 1 & \\ {}^{\lfloor da}\underline{L} & {}^{\lfloor a}\underline{L} \end{bmatrix}, \begin{bmatrix} {}^{\lfloor d}\underline{D} & \\ & {}^{\lfloor a}\underline{D} \end{bmatrix}, \nu\right)$$

to a new pdf on the original space of parameters. *It is done according to the chain rule*

$$\underline{f}(\theta, r) \equiv f\left({}^{\lfloor b}\theta | {}^{\lfloor a}\theta, r\right) \underline{f}({}^{\lfloor a}\theta, r) = GiW_{\theta,r}(\underline{L}, \underline{D}, \nu), \quad \textit{where}$$

$$\underline{L} \equiv \begin{bmatrix} 1 & & \\ {}^{\lfloor da}\underline{L} & {}^{\lfloor a}\underline{L} & \\ {}^{\lfloor db}L & {}^{\lfloor ab}L & {}^{\lfloor b}L \end{bmatrix}, \quad \underline{D} \equiv \begin{bmatrix} {}^{\lfloor d}\underline{D} & & \\ & {}^{\lfloor a}\underline{D} & \\ & & {}^{\lfloor b}D \end{bmatrix}.$$

In words, the original ${}^{\lfloor a}$*-parts are simply replaced by the new* ${}^{\lfloor a}$*-parts.*

8.6.3 Estimation of component structure

A practical universal search for the best order of factors forming the component is unsolved even for the normal components. Their specific form allows us to get a deeper insight, which will hopefully lead to practical algorithms.

A normal component can be written in the matrix "equation form"

$$d_t = {}^{\lfloor M}\theta' \phi_{t-1} + G' e_t, \tag{8.85}$$

where ${}^{\lfloor M}\theta$ is a matrix of regression coefficients complemented by zeros so that we can use the state vector ϕ_{t-1} that is common for all entries of d_t. $G' \equiv {}^{\lfloor e}L' \, {}^{\lfloor e}D^{0.5}$ is the *Choleski square root* of the noise covariance and e_t is normal white zero-mean noise with uncorrelated entries and unit variance.

The chosen triangular form of G uniquely defines the decomposition of this component into factors as well as their parameterizations; see Proposition 8.13. If we permute two entries of d_t, we have to permute corresponding columns of ${}^{\lfloor M}\theta$ and G. In this way, the triangular form of G is spoiled and has to be recovered by using invariance of the noise covariance to an orthogonal transformation applied directly to G'. Such a recovered G has generally different number of zero elements. Its inversion combines columns of ${}^{\lfloor M}\theta$ in a different way, too. Thus, the number of zero entries in parameters of new factors differs from the former one.

The result depends in a complex way on linear dependencies of modified rows of ${}^{\lfloor M}\theta$, rows of G and G^{-1}. Up to now, we have found no operational way how to describe them and thus, we have no guideline how to search for the "permutation" candidates. The brute-force solution considering permutations of all factors is of course formally possible.

8.6.4 Merging and cancelling of components

The normality of components brings a specific form of statistics and the normalizing integral $\mathcal{I}(V, \nu)$. It allows us to present directly the final algorithms of merging and cancelling.

Merging of a group of normal components

Here, the normal counterpart of Algorithm 6.21 is given.

Algorithm 8.14 (Merging of a group of normal components)

Initial mode

- *Estimate the mixture with a sufficient number of components* $\mathring{c}$ *so that statistics* ${}^{\lfloor \mathring{c}}\kappa_t$, ${}^{\lfloor \mathring{c}}\nu_{ic;t}$, ${}^{\lfloor \mathring{c}}L_{ic;t}$, ${}^{\lfloor \mathring{c}}D_{ic;t}$ $t \in \{0, \mathring{t}\}$, $c \in c^*$, $i \in i^* \equiv \{1, \ldots, \mathring{d}\}$ *are at disposal.*
- *Set pointers* $c = 1, \tilde{c} = 2$ *to trial components to be merged.*

Evaluation mode

Set the indicator of the common structure $cs = 0$.

For $i = 1, \ldots, \mathring{d}$

```
    Set $cs = -1$ & break the cycle over $i$ if the structures of $\theta_{ic}$, $\theta_{i\tilde{c}}$ differ.
end of the cycle over $i$
    Do if $cs = 0$
        Evaluate the common part of the trial merger
        $\tilde{\kappa}_{\mathring{t}} = \kappa_{c;\mathring{t}} + \kappa_{\tilde{c};\mathring{t}} - 1,\ \tilde{\kappa}_0 = \kappa_{c;0} + \kappa_{\tilde{c};0} - 1.$
        Evaluate and store the factor-related parts of the trial merger
For $i = 1, \ldots, \mathring{d}$
        $\tilde{\nu}_{i;\mathring{t}} = \nu_{ic;\mathring{t}} + \nu_{i\tilde{c};\mathring{t}}$
        $\tilde{L}'_{i;\mathring{t}} \tilde{D}_{i;\mathring{t}} \tilde{L}_{i;\mathring{t}} = L'_{ic;\mathring{t}} D_{ic;\mathring{t}} L_{ic;\mathring{t}} + L'_{i\tilde{c};\mathring{t}} D_{i\tilde{c};\mathring{t}} L'_{i\tilde{c};\mathring{t}}.$
end of the cycle over $i$
        Evaluate the change $\tilde{l}$ of log-v-likelihood expected after the merging
  $\tilde{l} = +\left\{ -\ln\left(\Gamma(\kappa_{c;\mathring{t}})\right) - \ln\left(\Gamma(\kappa_{\tilde{c};\mathring{t}})\right) + \ln\left(\Gamma(\tilde{\kappa}_{\mathring{t}})\right) - \ln\left(\left(\sum_{c=1}^{\mathring{c}-1} \kappa_{c;\mathring{t}}\right) - 1\right)\right\}$
      $-\left\{ -\ln\left(\Gamma(\kappa_{c;0})\right) - \ln\left(\Gamma(\kappa_{\tilde{c};0})\right) + \ln\left(\Gamma(\tilde{\kappa}_0)\right) - \ln\left(\left(\sum_{c=1}^{\mathring{c}} \kappa_{c;0}\right) - 1\right)\right\}.$
For $i = 1, \ldots, \mathring{d}$
        (factor parts)
        $\tilde{l} = \tilde{l}$
$+\left\{ \ln(\mathcal{I}(\tilde{L}_{i;\mathring{t}}, \tilde{D}_{i;\mathring{t}}, \tilde{\nu}_{i;\mathring{t}})) - \ln(\mathcal{I}(L_{ic;\mathring{t}}, D_{ic;\mathring{t}}, \nu_{ic;\mathring{t}})) - \ln(\mathcal{I}(L_{i\tilde{c};\mathring{t}}, D_{i\tilde{c};\mathring{t}}, \nu_{i\tilde{c};\mathring{t}}))\right\}$
$-\left\{ \ln(\mathcal{I}(\tilde{L}_{i;0}, \tilde{D}_{i;0}, \tilde{\nu}_{i;0})) - \ln(\mathcal{I}(L_{ic;0}, D_{ic;0}, \nu_{ic;0})) - \ln(\mathcal{I}(L_{i\tilde{c};0}, D_{i\tilde{c};0}, \nu_{i\tilde{c};0}))\right\}.$
end of the cycle over $i$
    end of the condition $cs = 0$
    Do if $\tilde{l} \leq 0$ or $cs < 0$
        Set $\tilde{c} = \tilde{c} + 1$.
        Go to the beginning of Evaluation mode if $\tilde{c} \leq \mathring{c}$. Otherwise continue.
        Set $c = c + 1$ and $\tilde{c} = c + 1$.
        Go to the beginning of Evaluation mode if $c < \mathring{c}$. Otherwise stop.
    else replace statistics related to the component $c$ by
        $\tilde{\kappa}_{\mathring{t}},\ \tilde{\kappa}_0,\ \left\{\tilde{L}_{i;\mathring{t}},\ \tilde{D}_{i;\mathring{t}},\ \tilde{L}_{i;0},\ \tilde{D}_{i;0},\ \tilde{\nu}_{i;\mathring{t}},\ \tilde{\nu}_{i;0}\right\}_{i=1}^{\mathring{d}}.$
        Swap the components $\mathring{c}$ and $\tilde{c}$.
        Decrease $\mathring{c} = \mathring{c} - 1$, i.e., omit the component $\tilde{c}$.
        Set $\tilde{c} = c + 1$ if $\tilde{c} > \mathring{c}$.
    end of the test on improvement of v-likelihood and of $cs < 0$
        Stop if $\mathring{c} = 1$. Otherwise go to the beginning of Evaluation mode.
```

Normal factor-based merging

The final, most promising Algorithm 6.24 is specialized only.

Algorithm 8.15 (Systematic merging of normal factors)

Initial mode

- *Estimate a mixture with normal factors. The mixture estimate is described by the collection of statistics*

$$\{L_{ic;t}, D_{ic;t}, \nu_{ic;t}\}_{c\in c^*, i=1,\ldots,\mathring{d}, t\in\{0,\mathring{t}\}}\,.$$

  The factors with the common i are supposed to describe the same entry of $d_{i;t}$ irrespective of the component number.
- *Initializing the list with rows $\rho = (i, c, \tilde{c})$ with meaning that the ith factor is common for components $c, \tilde{c}$. Usually, the list ρ is initialized as the empty one.*
- *Evaluate the individual normalization factors, $\forall c \in c^*, i = 1, \ldots, \mathring{d}, t \in \{0, \mathring{t}\}$,*

$$\mathcal{I}(L_{ic;t}, D_{ic;t}\nu_{ic:t}) = \Gamma(0.5\nu_{ic;t})\, {}^{\lfloor d}D_{ic;t}^{-0.5\nu_{ic;t}} \left| {}^{\lfloor \psi}D_{ic;t} \right|^{-0.5} 2^{0.5\nu_{ic;t}} (2\pi)^{0.5\mathring{\psi}}.$$

Evaluation mode

```
    For   $i = 1, \ldots, \mathring{d}$
            Set pointers $c = 1, \tilde{c} = 2$ to trial components.
      Test of the common structure
            Set the indicator of the common structure $cs = 0$.
            Set $cs = -1$ if the structures of $\theta_{ic}$ and $\theta_{i\tilde{c}}$ differ.
         Do if $cs = 0$
            Create $L'DL$ decomposition of the trial merger
            $\tilde{L}'_{i;\mathring{t}}\tilde{D}_{i;\mathring{t}}\tilde{L}_{i;\mathring{t}} = L'_{ic;\mathring{t}}D_{ic;\mathring{t}}L_{ic;\mathring{t}} + L'_{i\tilde{c};\mathring{t}}D_{i\tilde{c};\mathring{t}}L_{i\tilde{c};\mathring{t}}$
            $\tilde{L}'_{i;0}\tilde{D}_{i;0}\tilde{L}_{i;0} = L'_{ic;0}D_{ic;0}L_{ic;0} + L'_{i\tilde{c};0}D_{i\tilde{c};0}L_{i\tilde{c};0}$
            using Algorithm 8.2 on columns of the added matrices.
            Set $\tilde{\nu}_{i;\mathring{t}} = \nu_{ic;\mathring{t}} + \nu_{i\tilde{c};\mathring{t}}$, $\tilde{\nu}_{i;0} = \nu_{ic;0} + \nu_{i\tilde{c};0}$.
            Evaluate increment $\tilde{l}$ of the log-v-likelihood
            using prepared values of normalization integrals and (8.22)
            $\tilde{l} = \ln(\mathcal{I}(\tilde{L}_{i;\mathring{t}}, \tilde{D}_{i;\mathring{t}}, \tilde{\nu}_{i;\mathring{t}})) -$
              $- \ln(\mathcal{I}(L_{ic;\mathring{t}}, D_{ic;\mathring{t}}, \nu_{ic;\mathring{t}})) - \ln(\mathcal{I}(L_{i\tilde{c};\mathring{t}}, D_{i\tilde{c};\mathring{t}}, \nu_{i\tilde{c};\mathring{t}})) -$
              $- \ln(\mathcal{I}(\tilde{L}_{i;0}, \tilde{D}_{i;0}, \tilde{\nu}_{i;0}))$
              $+ \ln(\mathcal{I}(L_{ic;0}, D_{ic;0}, \nu_{ic;0})) + \ln(\mathcal{I}(L_{i\tilde{c};0}, D_{i\tilde{c};0}, \nu_{i\tilde{c};0}))$.
            end of the test on $cs = 0$
            Do if $\tilde{l} \leq 0$ or $cs < 0$
                Set $\tilde{c} = \tilde{c} + 1$.
                Go to the Test of the common structure if $\tilde{c} \leq \mathring{c}$.
                Otherwise continue.
                Set $c = c + 1$ and $\tilde{c} = c + 1$.
                Go to the beginning of Test of the common structure if $c < \mathring{c}$.
                Otherwise go to the end of cycle over $i$.
            else replace prior and posterior factors with indexes $ic$ and $i\tilde{c}$
                by the trial merger.
                Extend the list of common factors by $\rho = [\rho; (i, c, \tilde{c})]$.
            end of the test on improvement of v-likelihood and of $cs < 0$
    end    of the cycle over $i$
          Merging of components
  For    $c = 1, \ldots, \mathring{c} - 1$
    For    $\tilde{c} = c + 1, \ldots, \mathring{c}$
      Set $\kappa_{\tilde{c};\mathring{t}} = \kappa_{\tilde{c};\mathring{t}} + \kappa_{c;\mathring{t}}$,   $\kappa_{\tilde{c};0} = \kappa_{\tilde{c};0} + \kappa_{c;0}$ and cancel the component $c$
      if the components $c$, $\tilde{c}$ consist of common factors only.
    end    of the cycle over $\tilde{c}$
  end    of the cycle over $c$
```

Component cancelling

Cancelling is based on the specialized version of Algorithm 6.25. For it, it is necessary to find parameter values fulfilling (6.117). It is straightforward in the normal case. The choice $\theta_{ic} = 0$, $r_{ic} = 1$, $c \in c^*$, $i = 1, \ldots, \mathring{d}$, guarantees that all parameterized factors in all components coincide.

The following value is the specific item needed in the cancelling algorithm

$$\ln\left(\frac{{}^{\lfloor\mathring{c}}f\left({}^{\lfloor\mathring{c}}\Theta_{\mathring{c}}|d(\mathring{t})\right)}{{}^{\lfloor\mathring{c}}f\left({}^{\lfloor\mathring{c}}\Theta_{\mathring{c}}|d(0)\right)}\right) = 0.5\sum_{i=1}^{\mathring{d}}\left({}^{\lfloor d}V_{i\mathring{c};0} - {}^{\lfloor d}V_{i\mathring{c};\mathring{t}}\right)$$

The identity ${}^{\lfloor d}V = {}^{\lfloor d}D + {}^{\lfloor d\psi}L' \, {}^{\lfloor\psi}D \, {}^{\lfloor d\psi}L$ is used within the algorithm, .

Algorithm 8.16 (Systematic cancelling of normal components)

Initial mode

- *Estimate of the mixture with normal factors. The mixture estimate is described by the collection of statistics* $\{L_{ic;t}, D_{ic;t}, \nu_{ic;t}\}_{c \in c^*, i=1,\ldots,\mathring{d}, t \in \{0,\mathring{t}\}}$.

The factors with the common i are supposed to describe the same entry of $d_{i;t}$ irrespective of the component number.

- *Evaluate the individual normalization factors $\mathcal{I}(L_{ic;t}, D_{ic;t}, \nu_{ic:t})$, $\forall c \in c^*, i = 1, \ldots, \mathring{d}, t \in \{0, \mathring{t}\}$, using formula (8.22).*
- *Set $c = 1$.*

Evaluation mode

```
        Do while $c \le \mathring{c}$ and $\mathring{c} > 1$
          Set  $l = 0$
For   $i = 1, \ldots, \mathring{d}$
          $l = l + 0.5 \Big( {}^{\lfloor d}D_{ic;0} - {}^{\lfloor d}D_{ic;\mathring{t}}$
            $+ \; {}^{\lfloor d\psi}L'_{ic;0} \, {}^{\lfloor \psi}D_{ic;0} \, {}^{\lfloor d\psi}L_{ic;0} - {}^{\lfloor d\psi}L'_{ic;\mathring{t}} \, {}^{\lfloor \psi}D_{ic;\mathring{t}} \, {}^{\lfloor d\psi}L_{ic;\mathring{t}} \Big)$
            $+ \ln\left(\mathcal{I}(L_{ic;0}, D_{ic;0}, \nu_{ic;0})\right) - \ln\left(\mathcal{I}(L_{ic;\mathring{t}}, D_{ic;\mathring{t}}, \nu_{ic;\mathring{t}})\right)$
end   of the cycle over $i$
          If $l > 0$
            Swap $c$ with $\mathring{c}$ and set $\mathring{c} = \mathring{c} - 1$,  i.e., cancel the component
            Stop if $\mathring{c} = 1$
          else
            Set $c = c + 1$
          end of the test on v-log-likelihood increase
        end of the while cycle over $c$
```

8.7 Model validation

Section applies the general model validation given in Section 6.7 to normal components.

8.7.1 Test of data homogeneity

This part specializes Section 6.7.1.

Construction of the advisory system is simplified if learning data are qualified by experts. It provides an important additional data item, say $e \in e^* \equiv \{1, \ldots, \mathring{e}\}$, $\mathring{e} < \infty$. If the discussed labels are directly measured, then they can be treated as other discrete data, i.e., modelled by Markov-chain factors and require no special discussion.

A different situation arises with labels that classify data ex post. In this case, a finer modelling is desirable at least because of the inherently unbalanced scenario of the considered learning. It stems from the fact that any reasonably managed process leads to a few types of outcomes. Deviations from a good state are exceptional. As such, they can get too small weight or be completely overlooked when treated without taking into account their specific role. Quasi-Bayes estimation increases this danger. These labels are constant over whole data blocks and thus approximate processing-order-dependent estimation may provide quite misleading results.

In Section 6.7.1, the following hypotheses are formulated.

$H_0 \equiv$ The difference in observed consequences is due to the inseparable influence of external conditions and management way. A single mixture describes the standard $d(\mathring{t}_s)$ as well as another block of the labelled $d(\mathring{t}_e)$ data, i.e., for $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$

$$f(d(\mathring{t})|H_0) = \int f(d(\mathring{t})|\Theta, H_0) f(\Theta|H_0)\, d\Theta. \tag{8.86}$$

$H_1 \equiv$ The difference in observed consequences is caused by different ways of management. Thus, different mixtures should be used for the standard $d(\mathring{t}_s)$ and the labelled $d(\mathring{t}_e)$ data, i.e., for $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$

$$f(d(\mathring{t})|H_1) = \int f(d(\mathring{t}_s)|\Theta_s, H_1) f(\Theta_s|H_1)\, d\Theta_s \int f(d(\mathring{t}_e)|\Theta_e, H_1) f(\Theta_e|H_1)\, d\Theta_e. \tag{8.87}$$

The structures of both mixtures in (8.87) may differ.

Assuming no prejudice, $f(H_0) = f(H_1)$, the Bayes rule provides the posterior pf $f(H_0|d(\mathring{t}))$. The common model is accepted as a good one if this probability is high enough. Then, the labelling just helps in recognition of the factors active on $d(\mathring{t}_e)$ as potentially dangerous.

If $f(H_0|d(\mathring{t}))$ is small, H_1 is accepted and both models should be used separately. The factors of the predictor obtained from $f(d(\mathring{t}_s)|\Theta, H_0)$ near to the predictive factors obtained from $f(d(\mathring{t}_e)|\Theta_e, H_1)$ should be treated as potentially dangerous.

The corresponding algorithm is as follows.

Algorithm 8.17 (Test of normal data homogeneity)

1. *Run the complete model estimation on standard $d(\mathring{t}_s)$, labelled $d(\mathring{t}_e)$ and concatenated $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$ data. This provides the posterior pdfs*

$$f(\Theta|d(\mathring{t}_\iota)) = Di_\alpha(\kappa_{\mathring{t}_\iota}) \prod_{c\in c^*} \prod_{i \in i^*} GiW_{\theta_{ic}, r_{ic}}(L_{ic;\mathring{t}_\iota}, D_{ic;\mathring{t}_\iota} \nu_{ic;\mathring{t}_\iota}),\ \iota \in \iota^* \equiv \{s, e, \emptyset\}.$$

2. *The corresponding v-likelihood values indexed by $\iota \in \iota^*$ are obtained as a byproduct of approximate estimations. They have the form (8.40)*

$$f(d(\mathring{t}_\iota)|\iota) = \prod_{t \in t^*_\iota} \sum_{c \in c^*} \prod_{i \in i^*} \frac{\mathcal{I}(L_{ic;t}, D_{ic;t}, \nu_{ic;t})}{\mathcal{I}(L_{ic;t-1}, D_{ic;t-1}, \nu_{ic;t-1})} \frac{\kappa_{c;t-1}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c};t-1}}$$

$$\mathcal{I}(L_{ic;t}, D_{ic;t}, \nu_{ic;t}) = \Gamma(0.5\nu_{ic;t})\, {}^{\lfloor d}D_{ic;t}^{-0.5\nu_{ic;t}} \left| {}^{\lfloor \psi_{ic}}D_{ic;t} \right|^{-0.5} 2^{0.5\nu_{ic;t}} (2\pi)^{0.5\mathring{\psi}_{ic}}.$$

3. *Determine the probability that a single standard model should be used*

$$f(standard|d(\mathring{t})) \equiv f(H_0|d(\mathring{t})) = \frac{f(d(\mathring{t}))}{f(d(\mathring{t})) + f(d(\mathring{t}_s)|s)f(d(\mathring{t}_e)|e)}. \quad (8.88)$$

4. *Use the single model further on if $f(standard|d(\mathring{t}))$ is close to 1. The factors that were active on $f(d(\mathring{t}_e))$ are potentially dangerous.*
5. *Use both mixtures independently if $f(standard|d(\mathring{t}))$ is close to 0. The danger of provoking the situation labelled by e should be signaled whenever the model fitted to $d(\mathring{t}_e)$ makes better predictions than the model fitted to the standard data.*

8.7.2 Learning results

Let us cut the data available in offline mode $d(\mathring{t})$ into the learning $d(\mathring{t}_l) = d(t_u)$ and validation data $d(\mathring{t}_v) \equiv d(t_u - \partial, \ldots, t)$ for a cutting moment $t_u \in t_u^* \subset \{0\} \cup t^*$ and the highest model order ∂. We want to test whether the model fitted on the learning data is suited the validation set, too. We test the hypotheses

$H_0 \equiv$ All recorded data $d(\mathring{t}) \equiv (d(\mathring{t}_l), d(\mathring{t}_v))$ are described by a single mixture model.

$H_1 \equiv$ The learning data $d(\mathring{t}_l)$ set and the validation data set $d(\mathring{t}_v)$ should be described by individual models.

At the same time, we select the best cutting moment minimizing the expected loss $\mathcal{E}[\delta_{\hat{H},H}]$; see Proposition 6.19. The following algorithm elaborates this solution for normal mixtures using the flattened results gained on $d(\mathring{t}_l)$ as the prior pdf for the learning on $d(\mathring{t}_v)$.

Algorithm 8.18 (Model validation on homogenous data)

Initial phase

- *Select the structure of the mixture. Specify the collection of prior statistics*

$$\mathcal{M}_0 \equiv \{V_{ic;0},\, \nu_{ic;0}\; \kappa_{c;0}\}_{i=1,\ldots,\mathring{d},\, c\in c^*}\,.$$

- *Select a grid of cutting moments $t_u^* \equiv \{0 = t_{u;1} < t_{u;2} < \cdots < t_{u;\mathring{t}_u} \equiv \mathring{t}\}$.*

Collection of statistics for $t \in t_u^*$

- *Estimate the mixture using data $d(t_u)$ starting from the collection $\mathcal{M}_0$ This provides a collection of learned statistics ${}^{\lfloor l}\mathcal{M}_{t_u}$ and their v-likelihood ${}^{\lfloor l}\mathcal{L}_{t_u}$.*
- *Flatten the collection of statistics ${}^{\lfloor l}\mathcal{M}_{t_u}$ using Propositions 6.7 and 6.8. Denote the result ${}^{\lfloor v}\mathcal{M}_0$.*

- *Estimate the mixture of the same structure using data $d(t_u - \partial, \ldots, \mathring{t})$ and starting from the prior statistics ${}^{\lfloor v}\mathcal{M}_0$. Evaluate the corresponding v-likelihood ${}^{\lfloor v}\mathcal{L}_{t_u}$.*

Evaluation of the probabilities of hypotheses for $t_u \in t_u^*$

$$f(H_0|d(\mathring{t}), t_u) = \left(1 + \frac{{}^{\lfloor l}\mathcal{L}_{t_u}\ {}^{\lfloor v}\mathcal{L}_{t_u}}{{}^{\lfloor l}\mathcal{L}_{\mathring{t}_u}}\right)^{-1}$$

Decision-making on model validity

Accept the model learned on $d(\mathring{t})$(!) if

$$1 - \max_{t \in t_u^*} f(H_0|d(\mathring{t}), t_u) < \min_{t \in t_u^*} f(H_0|d(\mathring{t}), t_u).$$

Otherwise reject it and search for a better model.

8.7.3 Forgetting-based validation

Application of forgetting-based validation to normal mixtures is straightforward. We present it here for completeness only.

Algorithm 8.19 (Forgetting-based validation)

Initial mode

- *Estimate both parameters and structure of the normal mixture model.*
- *Apply flattening in the branching version so that a good prior pdf*

$$f(\Theta) = Di_{\alpha}(\kappa_0) \prod_{c \in c^*} \prod_{i \in i^*} GiW_{\theta_{ic}, r_{ic}}(L_{ic;0}, D_{i;0}, \nu_{ic;0}), \quad \textit{is obtained.}$$

- *Select several forgetting factors*

$$0 \approx \lambda_1 < \lambda_2 < \cdots \lambda_{\mathring{i}-1} < \lambda_{\mathring{i}} = 1,\ 1 < \mathring{i} < \infty.$$

- *Set $f(\Theta|\lambda_i) = f(\Theta)$.*

Validation mode

1. *Perform estimation with the stabilized forgetting, Section 3.1 for all λ_i, using $f(\Theta|d(\mathring{t}))$ as the alternative.*
2. *Evaluate values of v-likelihood $l_{i;\mathring{t}} = f(d(\mathring{t})|\lambda_i)$ as the product of the one-step-ahead adaptive predictors (8.40). Compute MAP estimate $\hat{\lambda}$ of λ .*
3. *Take the model as successful one if $\lambda_1 = \hat{\lambda}$. Otherwise search for its improvements.*

8.7.4 Inspection by a human designer

Low-dimensional projections of the estimated pdfs are computed using Proposition 8.8. Evaluation of the low-dimensional projections of normal predictors is described by Propositions 9.2, 9.3. Using these projections, it is possible to exploit human ability to grasp features hardly accessible in an algorithmic way.

8.7.5 Operating modes

This validation, described generally in Section 6.7.5, judges the KL divergence of sufficiently probable components to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$. The (worst) best of them should coincide with those operation modes that according to the expert judgement are taken as the (worst) best ones.

We assume that the estimated mixture is precise enough, i.e. it passed successfully validations described above, when making this test. Then, we can assume that components are normal with known parameters and specify the user's ideal pdf as a normal pdf, too. Then, we can measure distance of components to the user's ideal using Proposition 8.10.

Moreover, the data record d can be split into

d_m quantities determining directly markers that consist of the decisive outputs of the o-system and recognizable actions u_o of the operator,

d_e external and/or informative quantities that cannot or can be influenced by the operator but whose specific values do not influence the user's ideal pdf.

In other words, the user ideal is expressed in terms of the nonvoid part d_m of d.

The above discussion allows us allows us to specify the following validation algorithm.

Algorithm 8.20 (Analysis of normal operating modes)

1. *Split data records d into marker-defining d_m and remaining external quantities d_e.*
2. *Determine the user's ideal pdf $\prod_{i=1}^{\mathring{d}_m} \mathcal{N}_{d_i}\left({}^{\lfloor U}\theta_i' \, {}^{\lfloor U}\psi_i, \, {}^{\lfloor U}r_i\right)$ on d_m.*
3. *Take the estimated normal mixture that passed successfully learning tests.*
4. *Select a lower bound $\underline{\alpha} \in [0, 1)$ on non-negligible component weights.*
5. *Take gradually all components with weights $\alpha_c \geq \underline{\alpha}$.*
6. *Evaluate marginal pdfs $f(d_m|d(t-1))$ from the inspected component using Proposition 7.2 and if need be use the results of Section 9.1.1.*
7. *Evaluate the KL divergence of the resulting normal components to the user's ideal pdf using Proposition 8.10.*
8. *Check whether the KL divergences are decreasing the functions of quality assigned by an expert to operating modes around the respective components.*
9. *The model passes this test successfully if no visible discrepancy is recognized in the previous step. Otherwise, the model has to be improved.*

Remark(s) 8.11

1. *The algorithm can be easily extended to markers that are a one-to-one known image of d_m. The corresponding substitution in the integral defining the KL divergence does not change its value.*
2. *The conditional version of divergence is evaluated. It gives a local view on the achieved quality. If we make the user's ideal pdf of external quantities d_e equal to the pdf $f(d_e|d_m, d(t-1))$, if let them to their fate (cf. Section 5.1.5) then the overall divergence of a component to the user's ideal pdf reads*

$$\mathcal{D}\left(f \middle|\middle| {}^{\lfloor U}f\right) = \mathcal{E}\left[\sum_{t \in t^*} \int f(d_m|d(t-1)) \ln\left(\frac{f(d_m|d(t-1))}{{}^{\lfloor U}f(d_m|d(t-1))}\right) dd_m\right].$$

   Its sample version can be evaluated using the result of Section 9.1.4, giving the formula of the (9.23) type.
3. *This validation part can be substantially extended by using results of the design; see Chapter 9. Other tests, like comparison of the optimal recommendations and real actions, can be and should be added.*

9

Design with normal mixtures

Normal models with known parameters and the state in the phase form, in conjunction with unrestricted optimization of quadratic or exponential-quadratic loss function, lead to feasible Bellman functions; Agreement 2.9. For them, the optimal design reduces to the numerically tractable manipulations with quadratic forms. Here, we inspect how this property can be extended to mixtures made of the *normal components*

$$f(d_t|d(t-1),\Theta_c,c)=\mathcal{N}_{d_t}(\theta_c'\phi_{c;t-1},r_c), \tag{9.1}$$

with parameters $\Theta_c=[\theta_c,r_c]'$ consisting of matrix regression coefficients θ_c and covariance matrix r_c. The state vector $\phi_{c;t-1}$ has the phase form $\phi_{c;t-1}=[d'_{(t-1)\cdots(t-\partial_c)},1]'$, $\partial_c\geq 0$. When recognizable actions $u_{o;t}$ are at our disposal, the innovations Δ_t available to the p-system are modelled by the components

$$f(\Delta_t|u_{o;t},d(t-1),\Theta_c,c)=\mathcal{N}_{\Delta_t}\left({}^{\lfloor\Delta}\theta_c'\psi_t,\ {}^{\lfloor\Delta}r_c\right), \tag{9.2}$$

${}^{\lfloor\Delta}\Theta_c=\left[{}^{\lfloor\Delta}\theta_c,\ {}^{\lfloor\Delta}r_c\right]$, ${}^{\lfloor\Delta}\theta_c$ contains matrix regression coefficients and ${}^{\lfloor\Delta}r_c$ is the noise covariance. The regression vector $\psi_{c;t}'\equiv\left[u_{o;t}',d'_{(t-1)\cdots(t-\partial_c)},1\right]\equiv\left[u_{o;t}',\phi_{c;t-1}'\right]$, $\partial_c\geq 0$. The generating of recognizable actions is modelled by

$$f\left(u_{o;t}|d(t-1),\ {}^{\lfloor u}\Theta_c,c\right)=\mathcal{N}_{u_{o;t}}\left({}^{\lfloor u}\theta_c'\phi_{c;t-1},\ {}^{\lfloor u}r_c\right). \tag{9.3}$$

${}^{\lfloor u}\Theta_c=\left[{}^{\lfloor u}\theta_c,\ {}^{\lfloor u}r_c\right]$, ${}^{\lfloor u}\theta_c$ contains matrix regression coefficients and ${}^{\lfloor u}r_c$ is the corresponding covariance matrix. The coefficient $\theta_c\equiv\left[{}^{\lfloor\Delta}\theta_c,\ {}^{\lfloor u}\theta_c\right]$ and covariance matrices ${}^{\lfloor\Delta}r_c$, ${}^{\lfloor u}r_c$ are obtained from the factorized version of the normal mixture as described in Section 8.1.7. The design description is simplified when adopting various organizational agreements; the common state vector $\phi_{t-1}\equiv\phi_{c;t-1}$ is used. For reference, let us fix them.

Agreement 9.1 (Design conditions) *The system is modelled by a mixture with normal components (9.1) having known parameters Θ and the state*

ϕ_t in the phase form. The regression coefficients are complemented by zeros so that all factors within a single component have a common state vector. In the academic design, the data record $d_t = \Delta_t \equiv$ innovations. Regression vectors of individual factors are nested as follows, $i = 1, \ldots, \mathring{d} - 1$,

$$\psi_{i;t} \equiv [d'_{(i+1)\cdots\mathring{d};t}, \phi'_{t-1}]' \equiv [d_{i+1;t}; \psi'_{i+1;t}]' \qquad (9.4)$$

$$\psi_{\mathring{d};t} \equiv \phi_{t-1} \equiv [d'_{(t-1)\cdots(t-\partial)}, 1]',\ \partial \geq 0, \quad \psi_{0;t} \equiv \Psi_t \equiv [d'_{t\cdots(t-\partial)}, 1]'.$$

In the industrial or simultaneous design, the data record

$$d_t = (\Delta', u'_{o;t})' \equiv (\textit{innovations, recognizable actions})$$

and regression vectors are nested in the following way

$$\psi_{i;t} \equiv [\Delta'_{(i+1)\cdots\mathring{\Delta};t}, u'_{o;t}, \phi'_{t-1}]' \equiv [\Delta_{i+1;t}, \psi'_{i+1;t}]' = \Psi_{i+1;t},\ i < \mathring{\Delta} \qquad (9.5)$$

$$\psi_{\mathring{\Delta};t} \equiv [u'_{o;t}, \phi'_{t-1}]' \equiv [u'_{o;t}, d'_{(t-1)\cdots(t-\partial)}, 1]',\ \partial \geq 0, \quad \psi_{0;t} \equiv \Psi_t \equiv [d'_{t\cdots(t-\partial)}, 1]'.$$

Graphical expression of this nesting looks as follows.

$$\Psi_t \equiv \psi_{0;t} \left\{ \begin{array}{r} \left[\begin{array}{c} d_t \\ --- \end{array}\right. \\ \phi_{t-1} \equiv \psi_{\mathring{d};t} \equiv \left\{ \begin{array}{c} d_{t-1} \\ --- \\ \cdots \\ --- \\ d_{t-\partial} \\ --- \\ 1 \end{array}\right] \end{array} \right. \equiv \begin{bmatrix} d_{1;t} \\ d_{2;t} \\ \cdots \\ d_{\mathring{d};t} \\ --- \\ d_{1;t-1} \\ d_{2;t-1} \\ \cdots \\ d_{\mathring{d};t-1} \\ --- \\ \cdots \\ --- \\ d_{1;t-\partial} \\ d_{2;t-\partial} \\ \cdots \\ d_{\mathring{d};t-\partial} \\ --- \\ 1 \end{bmatrix} \equiv \left[\begin{array}{l} \begin{array}{c} \Delta_{1;t} \\ \Delta_{2;t} \\ \cdots \\ \Delta_{\mathring{\Delta};t} \end{array} \\ \left. \begin{array}{c} u_{o1;t} \\ u_{o2;t} \\ \cdots \\ u_{o\mathring{u}_o;t} \\ --- \\ \Delta_{1;t-1} \\ \Delta_{2;t-1} \\ \cdots \\ \Delta_{\mathring{\Delta};t-1} \\ u_{o1;t-1} \\ u_{o2;t-1} \\ \cdots \\ u_{o\mathring{u}_o;t-1} \\ --- \\ \cdots \\ --- \\ \Delta_{1;t-\partial} \\ \Delta_{2;t-\partial} \\ \cdots \\ \Delta_{\mathring{\Delta};t-\partial} \\ u_{o1;t-\partial} \\ u_{o2;t-\partial} \\ \cdots \\ u_{o\mathring{u}_o;t-\partial} \\ --- \\ 1 \end{array} \right\} \psi_{\mathring{\Delta};t}. \end{array} \right]$$

$$\left.\begin{bmatrix} d_{1;t} \\ \ldots \\ d_{i;t} \\ d_{i+1;t} \\ \ldots \\ d_{\mathring{d};t} \\ --- \\ d_{1;t-1} \\ d_{2;t-1} \\ \ldots \\ d_{\mathring{d};t-1} \\ --- \\ \ldots \\ --- \\ d_{1;t-\partial} \\ d_{2;t-\partial} \\ \ldots \\ d_{\mathring{d};t-\partial} \\ --- \\ 1 \end{bmatrix}\right\} \equiv \psi_{i;t} \equiv \begin{bmatrix} d_{i+1;t} \\ --- \\ \psi_{i+1;t} \end{bmatrix} \equiv \begin{bmatrix} d_{i+1;t} \\ \ldots \\ d_{\mathring{d};t} \\ --- \\ \phi_{t-1} \end{bmatrix}.$$

This chapter starts with common tools, Section 9.1, that cover model projections, discussion of practical presentation aspects, evaluations of feasible upper bounds of a Jensen type on the KL divergence and evaluation of expected quadratic forms. Then, academic, industrial and simultaneous designs of the advisory system are elaborated in detail; Section 9.2. Designs of presentation and signaling strategies conclude the chapter; Section 9.3.

9.1 Common tools

9.1.1 Model projections in design

Steady-state pdfs

Steady-state pdfs of the observed data are needed for recognizing dangerous components; Agreement 5.9. The large number of data available justifies the assumption that the uncertainty of parameters is negligible after estimation. It implies that we can evaluate the steady-state pdf of data assuming that the parameters $\Theta = [\theta, r]$, characterizing the normal component (9.1) with matrix regression coefficients

$$\theta' \equiv [\tilde{\theta}', \mu] \equiv [A_1, \ldots, A_\partial, \mu],\ A_i \text{ are square matrices,} \tag{9.6}$$

and noise covariance

$$r \equiv L_e' D_e L_e,\ L_e, D_e \text{ are lower triangular and diagonal matrices}$$

are known. The column $\mathring{d}$-vector μ in θ describes the data offset. The modelled o-system is supposed to be stable without a significant loss of generality. Thus, a good model has to have A's describing a stable auto-regression.

The steady-state pdf of the state ϕ is normal since the relationships among the observed data are linear and the driving white noise is normal. Thus, it is sufficient to evaluate the steady-state expected value $\phi_\infty \equiv \mu_\infty \otimes \mathbf{1}_\partial$ and the steady-state covariance matrix $C_\infty = L'_\infty D_\infty L_\infty$. Recall that $\otimes$ denotes the Kronecker product and $\mathbf{1}_\partial$ is the ∂-vector of units.

The value μ_∞ can be computed recursively as the limiting value of the sequence $\{\mu_n\}_{n=1-\partial}^{\infty}$. The sequence starts from zero initial conditions $\mu_n = 0$, for $n \leq 0$, and its other members are generated as follows.

$$\mu_n = \sum_{i=1}^{\partial} A_i \mu_{n-i} + \mu. \tag{9.7}$$

This evaluation is preferable against an explicit solution. The convergence of (9.7) tests the stability of the estimated auto-regressive model; it tests whether the component is dangerous.

The $L'DL$ decomposition of the steady-state covariance C_∞ of ϕ without constant entry 1 can be found also recursively

$$L'_{n+1} D_{n+1} L_{n+1} = \left[L_n \tilde{\theta}\ \tilde{L}_n \right]' D_n \left[L_n \tilde{\theta}\ \tilde{L}_n \right] + [L_e\ 0]' D_e [L_e\ 0], \tag{9.8}$$

where $L_1 = D_1 = I =$ unit matrix, $\tilde{\theta}$ denotes θ with the omitted offset μ and $\tilde{L}_n$ coincides with initial $\mathring{d}(\partial - 1)$ columns of L_n. The full algorithm looks as follows.

Algorithm 9.1 (Moments of the steady-state normal pdf)

Initial mode

- *Select the upper bound $\mathring{n}$ on the number n of recursive steps and set $n = 0$.*
- *Select a small $\varepsilon > 0$ used for stopping.*
- *Set $[\mu_{-1}, \ldots, \mu_{-\partial}] = 0$.*
- *Set $L_n = D_n = I = (\mathring{d}\partial, \mathring{d}\partial)$-unit matrix.*

Iterative mode (expectation)

1. *Update $\mu_n = \sum_{i=1}^{\partial} A_i \mu_{n-i} + \mu$.*
2. *Stop completely and announce the instability of the model and label the component as a potentially dangerous one if $||\mu_n|| > 1/\varepsilon$ for some norm $||\cdot||$.*
3. *Set $n = 0$ and go to the beginning of* Iterative mode (covariance) *if $n > \mathring{n}$ or $||\mu_n - \mu_{n-1}|| < \varepsilon$.*
4. *Increase the counter $n = n + 1$ and go to the beginning of* Iterative mode (expectation).

Iterative mode (covariance)

1. *Make the updating (9.8) using a variant of Algorithm 8.2.*
2. *Stop if* $n > \mathring{n}$ *or* $||D_n - D_{n-1}|| < \varepsilon$ *and take* μ_n *and* $L_n' D_n L_n$ *as steady-state moments.*
3. *Increase the counter* $n = n + 1$ *and go to the beginning of* Iterative mode (covariance).

The unstable components are obvious candidates to be labelled dangerous, but those with stationary moments far from the target area may be classified as dangerous, too; Agreement 5.9.

Remark(s) 9.1
The presence of unstable components does not automatically mean that the mixture with such components is unstable. The following simple example indicates it. Let the scalar d_t *be generated by the normal two-component mixture*

$$f(d_t|d(t-1)) = \alpha \mathcal{N}_{dt}(\theta_1 d_{t-1}, 1) + (1-\alpha)\mathcal{N}_{dt}(\theta_2 d_{t-1}, 1).$$

Then, $\rho_t \equiv \mathcal{E}[d_t^2] \geq 0$ *evolves according to the recursion*

$$\rho_t = \underbrace{\left[\alpha\theta_1^2 + (1-\alpha)\theta_2^2\right]}_{\lfloor m \theta^2} \rho_{t-1} + 1$$

and has the finite solution if $^{\lfloor m}\theta^2 < 1$. *This condition is met even for combinations of stable and unstable components, for instance,* $\alpha = 0.8,\ \theta_1 = 1.1$ *(!) and* $\theta_2 = 0.3$ *giving* $^{\lfloor m}\theta^2 = 0.986 < 1$. *With it, the second noncentral and consequently first moments are finite even for* $t \to \infty$. *The conclusion can be expressed in the appealing form: the presence of, even rarely active, stable component has stabilizing effect on the mixture.*

Marginal and conditional pdfs

Low-dimensional marginal and conditional pdfs provide the main technical tool for presenting the results of the design of the p-system. The algorithmic description of their evaluation is given here for a known (well-estimated) normal mixture.

In the industrial design with quadratic loss function and in simultaneous design, the pdf of the optimal recognizable actions is also normal. Thus, the results of this section serve for all versions of advisory systems. Essentially, Proposition 7.2 is specialized here. The specialization is split in Proposition 9.2, dealing with marginal pdfs, and Proposition 9.3, addressing the conditioning. Both of them rely on a specific structure of components; Agreement 5.4. Thus, we need a tool for changing it.

Proposition 9.1 (Marginal pdf of adjacent normal factors) *Let us consider a pair of* adjacent normal factors *(see, Agreement 5.4) with known parameters*

$$f(\mathit{\Delta}_1,\mathit{\Delta}_2|\psi_1,\psi_2) = \mathcal{N}_{\mathit{\Delta}_1}\left([\beta,\theta_1'][\mathit{\Delta}_2,\psi_1']',r_1\right)\mathcal{N}_{\mathit{\Delta}_2}\left(\theta_2'\psi_2',r_2\right). \quad (9.9)$$

$$\textit{Then,}\quad f(\mathit{\Delta}_1,\mathit{\Delta}_2|\psi_1,\psi_2) = \mathcal{N}_{\mathit{\Delta}_2}\left([\tilde{\beta},\tilde{\theta}_2'][\mathit{\Delta}_1,\psi']',\tilde{r}_2\right)\mathcal{N}_{\mathit{\Delta}_1}\left(\tilde{\theta}_1'\psi,\tilde{r}_1\right),$$

where ψ contains union of entries in ψ_1,ψ_2. The entry, which is at least in one of them, is put on the corresponding position in ψ.

The obtained normal factor for $\mathit{\Delta}_1$ in the second version coincides with the marginal pdf of $\mathit{\Delta}_1$. The quantities marked by $\tilde{}$ are generated by the following algorithm.

- *Extend vectors θ_i by zeros to the vectors $\bar{\theta}_i$ of a common length so that $[\beta,\theta_1'][\mathit{\Delta}_2,\psi_1']' = [\beta,\bar{\theta}_1'][\mathit{\Delta}_2,\psi']'$ and $\theta_2'\psi_2 = \bar{\theta}_2'\psi$.*
- *Permute the entries $\mathit{\Delta}_1,\mathit{\Delta}_2$ in the quadratic form Q*

$$\begin{aligned} Q &\equiv [\mathit{\Delta}_1,\mathit{\Delta}_2,\psi'] \begin{bmatrix} -1 & \beta & \bar{\theta}_1 \\ 0 & -1 & \bar{\theta}_2 \end{bmatrix}' \mathrm{diag}[r_1^{-1},r_2^{-1}] \begin{bmatrix} -1 & \beta & \bar{\theta}_1 \\ 0 & -1 & \bar{\theta}_2 \end{bmatrix} [\mathit{\Delta}_1,\mathit{\Delta}_2,\psi']' \\ &= [\mathit{\Delta}_2,\mathit{\Delta}_1,\psi'] \begin{bmatrix} 0 & -1 & \bar{\theta}_2 \\ -1 & \beta & \bar{\theta}_1 \end{bmatrix}' \mathrm{diag}[r_1^{-1},r_2^{-1}] \begin{bmatrix} 0 & -1 & \bar{\theta}_2 \\ -1 & \beta & \bar{\theta}_1 \end{bmatrix} [\mathit{\Delta}_2,\mathit{\Delta}_1,\psi']' \end{aligned}$$

and recover $L'DL$ decomposition with -1(!) on the diagonal of L of its kernel using Proposition 8.4

$$Q = [\mathit{\Delta}_2,\mathit{\Delta}_1,\psi'] \begin{bmatrix} -1 & \tilde{\beta} & \tilde{\theta}_2 \\ 0 & -1 & \tilde{\theta}_1 \end{bmatrix}' \mathrm{diag}[\tilde{r}_2^{-1},\tilde{r}_1^{-1}] \begin{bmatrix} -1 & \tilde{\beta} & \tilde{\theta}_2 \\ 0 & -1 & \tilde{\theta}_1 \end{bmatrix} [\mathit{\Delta}_2,\mathit{\Delta}_1,\psi']'.$$

Proof. Omitted. □

Any desired change of the component structure can be achieved by a sequence of pairwise *permutations on adjacent factors.*

If the component consists of normal factors only, then it can be written in the matrix form (9.1) and the discussed permutation can be made on it.

Proposition 9.2 (Marginal predictors for normal mixtures) *Under Agreement 9.1, let us consider the known normal mixture model in the factorized form*

$$f(\mathit{\Delta}_t|u_{o;t},d(t-1)) = \sum_{c\in c^*}\alpha_c\prod_{i\in i^*}\mathcal{N}_{\mathit{\Delta}_{ic;t}}(\theta_{ic}'\psi_{ic;t},r_{ic}), \quad \textit{where}$$

$\psi_{ic;t} = [\mathit{\Delta}'_{(i+1)\cdots\mathring{\mathit{\Delta}}c;t}, u'_{o;t}, \phi'_{c;t-1}]' \equiv [\mathit{\Delta}'_{(i+1)\cdots\mathring{\mathit{\Delta}}c;t}, \psi'_{\mathring{\mathit{\Delta}}c;t}]'$ *are regression vectors, $i \in i^* \equiv \{1,\ldots,\mathring{\mathit{\Delta}}-1\}$, $c \in c^*$,*

$\psi'_{\mathring{\mathit{\Delta}}c;t}$ *is the common part of regression vectors forming the component c,*

$\mathit{\Delta}_{ic;t}$ *are entries of innovations $\mathit{\Delta}_{c;t}$ and $u_{o;t}$ are recognizable o-actions,*

θ_{ic} *are regression coefficients, with entries ordered accordingly to the corresponding regression vector* ψ_{ic}*, i.e.,* $\theta_{ic} = \left[\theta_{1ic}, \ldots, \theta_{(\mathring{\Delta}-i)ic}, {}^{\lfloor\psi}\theta'_{ic}\right]'$*, where* θ_{jic}*,* $j = 1, \ldots, \mathring{\Delta} - i$*, are scalars and* ${}^{\lfloor\psi}\mathring{\theta}_{ic} \equiv \mathring{\psi}_{\mathring{\Delta}c;t}$*; zeros are inserted into the coefficients* θ_{ic} *in order to get the common regression vectors* $\psi_{\mathring{\Delta}c;t} = \psi_{\mathring{\Delta};t}$*,*

r_{ic} *is the noise variance of the corresponding factor.*

The considered normal component is determined by the matrices G_c, H_c *and* F_c*, Proposition 8.13,*

$$F_c \equiv \operatorname{diag}\left[r_{1c}^{-1}, \ldots, r_{\mathring{\Delta}c}^{-1}\right] \qquad (9.10)$$

$$[G_c,\, H_c] \equiv \left[\underbrace{\begin{matrix} -1 & \theta_{11c} & \theta_{21c} & \ldots & \theta_{(\mathring{\Delta}-1)1c} \\ 0 & -1 & \theta_{12c} & \ldots & \theta_{(\mathring{\Delta}-2)2c} \\ & & \ddots & & \\ 0 & 0 & \ldots & -1 & \theta_{1(\mathring{\Delta}-1)c} \\ 0 & 0 & \ldots & 0 & -1 \end{matrix}}_{G_c} \underbrace{\begin{matrix} {}^{\lfloor\psi}\theta'_{1c} \\ {}^{\lfloor\psi}\theta'_{2c} \\ \vdots \\ {}^{\lfloor\psi}\theta'_{(\mathring{\Delta}-1)c} \\ {}^{\lfloor\psi}\theta'_{\mathring{\Delta}c} \end{matrix}}_{H_c}\right], \; G_c \text{ is square matrix.}$$

Let the marginal pdf be computed for the selected entries of Δ_t*, say,* $\Delta_{\iota;t}$*,* $\iota \in \iota^* \equiv \{k_1, \ldots, k_{\mathring{\iota}}\} \subset i^* \equiv \iota^* \cup \overline{\iota^*}$*,* $\iota^* \cap \overline{\iota^*} = \emptyset$*, i.e., the marginal pdf of entries* $\Delta_{\iota^*;t} \equiv [\Delta_{k_1;t}, \ldots, \Delta_{k_{\mathring{\iota}};t}]'$ *is computed. Let the permutation matrix* $\mathcal{T}'_c$ *permute* $\Delta_{c;t}$ *so that the entries* $\Delta_{\iota^*;t}$ *are placed at the end of the permuted vector of innovations, i.e.* $\mathcal{T}'_c \Delta_{c;t} = \left[\bullet, \Delta'_{\iota^*;t}\right]'$*.*

Let us transform matrices F_c *and* $[G_c\mathcal{T}_c,\, H_c]$ *to* $\tilde{F}_c$ *and* $[\tilde{G}_c,\, \tilde{H}_c]$ *so that* $\tilde{F}_c$ *is diagonal matrix,* $\tilde{G}_c$ *has the same form as* G_c *and the following quadratic form is preserved*

$$[G_c\mathcal{T}_c,\, H_c]' F_c [G_c\mathcal{T}_c,\, H_c] = [\tilde{G}_c,\, \tilde{H}_c]' \tilde{F}_c [\tilde{G}_c,\, \tilde{H}_c]. \qquad (9.11)$$

Then, the predictor of the quantities of interest is the mixture

$$\begin{aligned} f(\Delta_{\iota^*;t}|u_{o;t}, d(t-1)) &\equiv f(\Delta_{k_1;t}, \ldots, \Delta_{k_{\mathring{\iota}};t}|u_{o;t}, d(t-1)) \\ &= \sum_{c\in c^*} \alpha_c \prod_{\iota=1}^{\mathring{\iota}} \mathcal{N}_{\Delta_{k_\iota;t}}(\theta'_{k_\iota c}\psi_{k_\iota c;t}, r_{k_\iota c}), \end{aligned}$$

where the regression coefficients $\theta'_{k_\iota c}$ *are in the last* $\mathring{\iota}$ *rows of* $[\tilde{G}_c,\, \tilde{H}_c]$ *(with* -1 *omitted) and variance inversions* $r_{k_\iota c}^{-1}$ *are at corresponding positions in* $\tilde{F}_c$*.*

Proof. Let us take a fixed component c. Then, for $\Delta_{c;t} = [\Delta_{1c;t}, \ldots, \Delta_{\mathring{\Delta}c;t}]'$, its pdf is

$$f(\Delta_{c;t}|u_{o;t},d(t-1),c) = \prod_{i\in i^*} \mathcal{N}_{\Delta_{ic;t}}\left(\theta'_{ic}\psi_{ic;t}, r_{ic}\right) \propto \exp$$

$$-\frac{X'}{2}\underbrace{\begin{bmatrix} r_{1c}^{-1} & & & \\ & r_{2c}^{-1} & & \\ & & \ddots & \\ & & & r_{\mathring{\Delta}c}^{-1}\end{bmatrix}}_{F_c}\underbrace{\begin{bmatrix} -1 & \theta_{11c} & \theta_{21c} & \ldots & \theta_{(\mathring{\Delta}-1)1c} & {}^{\lfloor\psi}\theta'_{1c} \\ 0 & -1 & \theta_{12c} & \ldots & \theta_{(\mathring{\Delta}-2)2c} & {}^{\lfloor\psi}\theta'_{2c} \\ & & \ddots & & & \vdots \\ 0 & 0 & \ldots & -1 & \theta_{1(\mathring{\Delta}-1)c} & {}^{\lfloor\psi}\theta'_{(\mathring{\Delta}-1)c} \\ 0 & 0 & \ldots & 0 & -1 & {}^{\lfloor\psi}\theta'_{\mathring{\Delta}c}\end{bmatrix}\begin{bmatrix}\Delta_{c;t} \\ \psi_{\mathring{\Delta};t}\end{bmatrix}}_{X}$$

$$\equiv \exp\left\{-0.5[\Delta'_{c;t}, \psi'_{\mathring{\Delta};t}]\begin{bmatrix} G'_c \\ H'_c\end{bmatrix} F_c[G_c\ H_c]\begin{bmatrix}\Delta_{c;t} \\ \psi_{\mathring{\Delta};t}\end{bmatrix}\right\} \equiv \exp\{-0.5Q\},$$

where F_c is diagonal $(\mathring{d},\mathring{d})$-matrix, G_c is upper triangular matrix with -1 on its diagonal. The rectangular $(\mathring{\Delta}, \mathring{\psi}_{\mathring{\Delta}c})$ matrix H_c contains regression coefficients corresponding to the common part $\psi_{\mathring{\Delta};t}$ of regression vectors.

Let J' be a permutation matrix, which places entries $\Delta_{\iota^*;t}$ whose marginal pdfs should be computed to the end of the transformed vector of innovations Δ_t. Taking another matrix T_c that performs the permutation $T_c\Delta_t = \Delta_{c;t} \Leftrightarrow \Delta_t = T'_c\Delta_{c;t}$, the following equality is obtained

$$[\bullet, \Delta'_{\iota^*;t}]' \equiv J'\Delta_t = J'T'_c\Delta_{c;t} \equiv \mathcal{T}'_c\Delta_{c;t}.$$

Orthogonality of permutation matrices implies that the quadratic form Q in the exponent of the normal component can be expressed as

$$Q = \left[[\bullet, \Delta'_{\iota^*;t}], \psi'_{\mathring{\Delta}c;t}\right][G_c\mathcal{T}_c,\ H_c]'F_c[G_c\mathcal{T}_c,\ H_c]\left[[\bullet, \Delta'_{\iota^*;t}], \psi'_{\mathring{\Delta}c;t}\right]'.$$

Using nonuniqueness of the kernel decomposition, we find diagonal $\tilde{F}_c$, upper triangular $\tilde{G}_c$ and rectangular $\tilde{H}_c$ matrices, which satisfy the equality $[G_c\mathcal{T}_c,\ H_c]'F_c[G_c\mathcal{T}_c,\ H_c] = [\tilde{G}_c,\ \tilde{H}_c]'\tilde{F}_c[\tilde{G}_c,\ \tilde{H}_c]$. In algorithmic terms, LDL' decomposition, with -1 on diagonal L, is re-created. With it, the integration of superfluous quantities reduces to the omission of leading rows and columns of the matrices $\tilde{F}_c$, $\tilde{G}_c$ together with leading rows of $\tilde{H}_c$. □

Proposition 9.3 (Conditional pdfs of the normal mixture model) *Under Agreement 9.1, let us consider the known normal mixture model in the factorized form*

$$f(\Delta_t|u_{o;t},d(t-1)) = \sum_{c\in c^*}\alpha_c\prod_{i\in i^*}\mathcal{N}_{\Delta_{ic;t}}(\theta'_{ic}\psi_{ic;t}, r_{ic}), \quad \textit{where}$$

$\psi_{ic;t} = [\Delta'_{(i+1)\cdots\mathring{\Delta}c;t}, u'_{o;t}, \phi'_{c;t-1}]' \equiv [\Delta_{(i+1)c;t}, \psi'_{(i+1)c;t}]'$ *are regression vectors,* $i \in \{1,\ldots,\mathring{\Delta}-1\}$, $c \in c^*$, *with the common part* $\psi_{\mathring{\Delta};t} \equiv [u'_{o;t}, \phi'_{t-1}]'$, $\Delta_{ic;t}$ *are entries of innovations* $\Delta_{c;t}$, $u_{o;t}$ *are recognizable o-actions and* $\phi_{t-1} \equiv \phi_{c;t-1}$ *is the common observable state of individual components,*

$\theta_{ic} = \left[\theta_{1ic}, \ldots, \theta_{(\mathring{\Delta}-i)ic}, {}^{\lfloor\psi}\theta'_{ic}\right]'$, ${}^{\lfloor\psi}\mathring{\theta}_{ic} \equiv \mathring{\psi}_{\mathring{\Delta}c;t}$, *are regression coefficients split in accordance with the corresponding regression vector; if need be, zeros are inserted to get the common part* $\psi'_{\mathring{\Delta};t}$,

r_{ic} *is the noise variance of the factor ic.*

For simplicity, let the order of factors for all components be common. In other words, a common structure of components, Agreement 5.4, implying $\Delta_{c;t} = \Delta_t$ *is considered. Suppose that indexes* $\iota \in \iota^* \equiv \{\underline{\iota}, \ldots, \mathring{i}\}$, $\underline{\iota} \le \mathring{i}$, *point to selected entries of* Δ_t. *Their marginal pdfs are already computed (see Proposition 9.2) while index* $\underline{k} \in (\underline{\iota}, \mathring{i})$ *determines the splitting of* $\Delta_{\iota^*;t}$ *into two parts. Then, the predictor of* $\Delta_{\underline{\iota}\cdots(\underline{k}-1);t}$ *conditioned on* $u_{o;t}, d(t-1)$ *and the remaining* $\Delta_{\underline{k}\cdots\mathring{i};t}$ *is the* ratio of normal mixtures

$$f\left(\Delta_{\underline{\iota}\cdots(\underline{k}-1);t} | \Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1)\right) = \frac{\sum_{c\in c^*} \alpha_c \prod_{\iota=\underline{\iota}}^{\mathring{i}} \mathcal{N}_{\Delta_{\iota;t}}(\theta'_{\iota c}\psi_{\iota;c}, r_{\iota c})}{\sum_{c\in c^*} \alpha_c \prod_{\iota=\underline{k}}^{\mathring{i}} \mathcal{N}_{\Delta_{\iota;t}}(\theta'_{\iota c}\psi_{\iota;c}, r_{\iota c})}. \tag{9.12}$$

When all values in the condition are fixed, the resulting predictor is the normal mixture

$$f(\Delta_{\underline{\iota}\cdots(\underline{k}-1);t} | \Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1))$$
$$= \sum_{c\in c^*} \tilde{\alpha}_c(\Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1)) \prod_{\iota=\underline{\iota}}^{\underline{k}-1} \mathcal{N}_{\Delta_{\iota;t}}(\theta'_{\iota c}\psi_{\iota;c}, r_{\iota c}), \ \textit{with}$$
$$\tilde{\alpha}_c\left(\Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1)\right) \propto \alpha_c \prod_{\iota=\underline{k}}^{\mathring{i}} \mathcal{N}_{\Delta_{\iota;t}}(\theta'_{\iota c}\psi_{\iota;c}, r_{\iota c}). \tag{9.13}$$

Proof. It is implied by Proposition 9.2 and by the formula $f(\beta|\gamma) = \frac{f(\beta,\gamma)}{f(\gamma)} = \frac{f(\beta,\gamma)}{\int f(\beta,\gamma)\, d\beta}$. □

Remark(s) 9.2

1. *The dependence of weights (9.13) on data indicates that the model obtained through the estimation of the full predictor followed by marginalization is richer than the directly estimated, low-dimensional predictor. This very practical observation is easily overlooked.*
2. *Permutations of adjacent factors according to Proposition 9.1 have to be made if the considered entries are not at the assumed positions. This need also implies that the assumption of the c-invariant ordering of* $\Delta_{c;t} = \Delta_t$ *is not restrictive: we have to reach this situation anyway, at least for the innovation entries of interest.*

9.1.2 Dynamic predictors in advising

Here, the general results of Section 7.1.2 are specialized to normal mixtures. For these mixtures, individual components are multivariate *auto-regression* models (AR). Their stability is the key property we need for good long-horizon predictors.

Stable auto-regression

Normal multivariate AR models serve well as short-horizon predictors in a range of applications including those covered by this text. They are often unsuitable for long-term predictions as the dynamic system represented by the estimated coefficients is too close to the stability boundary or even unstable.

The remedy should be searched in data preprocessing (Chapter 6), use of continuous-time model [138, 163] or model with moving average part ARMAX factors [26]. None of them, however, can guarantee the full success. Then, the reduction of the support of the prior pdf on the set, cf. (9.6),

$$\Theta_s^* \equiv \left\{ \left([A_1, \ldots, A_{\partial}, \mu], r\right), \mathcal{A} = \begin{bmatrix} A_1 & \ldots & A_{\partial - 1} & A_{\partial} \\ I & \ldots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & I & 0 \end{bmatrix} \begin{matrix} \text{has spectral} \\ \text{radius} < 1 \end{matrix} \right\} \tag{9.14}$$

as discussed in Section 7.1.2, is the only possibility. Technically, Monte Carlo evaluation of the expectation over the set of stable AR models is applied. The samples of Θ are generated from the estimated factorized version of the posterior pdf

$$f(\Theta|d(\mathring{t})) \propto \prod_{i=1}^{\mathring{d}} GiW_{\theta_i, r_i}\left(L_{i;\mathring{t}}, D_{i;\mathring{t}}, \nu_{i;\mathring{t}}\right) \chi_{\Theta_s^*}(\Theta). \tag{9.15}$$

The involved statistics are updated irrespective of the constraint used. For the approximate mixture estimation, Section 6.5, weights of individual data vectors should be modified compared to the unrestricted case. They are formed, however, by one-step-ahead predictors that are weakly influenced by the information about stability. Thus, this dependence can be and will be neglected. Consequently, the evaluation of the posterior likelihood function is uninfluenced by the stability restriction even for mixtures. Having the posterior pdf, we need to evaluate a point estimate $\hat{\Theta}$ of the parameter Θ. The expected value

$$\hat{\Theta} \equiv \mathcal{E}\left[\Theta|d(\mathring{t})\right] = \frac{\int \Theta \chi_{\Theta_s^*}(\Theta) \prod_{i=1}^{\mathring{d}} GiW_{\theta_i, r_i}\left(L_{i;\mathring{t}}, D_{i;\mathring{t}}, \nu_{i;\mathring{t}}\right) d\Theta}{\int \chi_{\Theta_s^*}(\Theta) \prod_{i=1}^{\mathring{d}} GiW_{\theta_i, r_i}\left(L_{i;\mathring{t}}, D_{i;\mathring{t}}, \nu_{i;\mathring{t}}\right) d\Theta} \tag{9.16}$$

is taken as the required estimate. Its analytical evaluation is impossible. A Monte Carlo evaluation is to be used. It has to take into account the non-convex shape of the integration domain. The evaluations are performed by the following algorithm that specializes Algorithm 7.1 to a normal component. For a mixture, it is applied componentwise.

Algorithm 9.2 (Point estimation of the stable AR component)

Initial mode

- *Perform the standard estimation (Chapter 8) that gives the collection of statistics* $L_{i;\mathring{t}}, D_{i;\mathring{t}}, \nu_{i;\mathring{t}},\ i = 1, \ldots, \mathring{d}$.
- *Convert the collected statistics into their least-squares (LS) counterparts*

$$\begin{aligned} \hat{\theta}_{i;\mathring{t}} &\equiv {}^{\lfloor\psi}L^{-1}_{i;\mathring{t}}\ {}^{\lfloor d\psi}L_{i;\mathring{t}} \equiv \textit{LS estimate of regression coefficients} \\ \hat{r}_{i;\mathring{t}} &\equiv \frac{{}^{\lfloor d}D_{i;\mathring{t}}}{\nu_{i;\mathring{t}} - 2} \equiv \textit{LS estimate of noise variance} \\ C_{i;\mathring{t}} &\equiv {}^{\lfloor\psi}L^{-1}_{i;\mathring{t}}\ {}^{\lfloor\psi}D^{-1}_{i;\mathring{t}} \left({}^{\lfloor\psi}L'_{i;\mathring{t}}\right)^{-1} \equiv \textit{LS covariance factor.} \end{aligned}$$

- *Stop and take the unrestricted LS point estimates as the approximation of (9.16) if they define a stable AR model, i.e., if they belong to* Θ^*_s *(9.14).*
- *Select the upper bound* $\mathring{n}$*, say on the order of* $10^4 - 10^6$*, on the number* n *of samples and set* $n = 0$*.*
- *Select the bound* $\mathring{m}$*, say on the order of* $100 - -1000$*, on the necessary number* m *of stable samples and set* $m = 0$*.*
- *Initialize the generated stable point estimates* $\hat{\theta}_{is} = 0,\ \hat{r}_{is} = 0,\ i = 1, \ldots, \mathring{d}$*.*

Iterative mode

1. *Do while* $n < \mathring{n}\ \&\ m < \mathring{m}$*.*
2. *Set* $n = n + 1$*.*
3. *Generate, for* $i = 1, \ldots, \mathring{d}$*, random samples*

$$\theta_{in},\ r_{in} \sim GiW_{\theta_i, r_i}\left(\hat{\theta}_{i;\mathring{t}}, \hat{r}_{i;\mathring{t}}, C_{i;\mathring{t}}, \nu_{i;\mathring{t}}\right)$$

   as follows.

$$\begin{aligned} r_{in} &= \hat{r}_{i;\mathring{t}}(1 + e_r),\ e_r \sim \mathcal{N}_{e_r}\left(0, \sqrt{\frac{2}{\nu_{i;\mathring{t}} - 4}}\right);\ cf.(8.23), \\ r_{in} &= \hat{r}_{i;\mathring{t}} \text{ if the sample } r_{in} \le 0 \\ \theta_{in} &= \hat{\theta}_{i;\mathring{t}} + \sqrt{\tilde{r}_i}\ {}^{\lfloor\psi}L^{-1}_{i;\mathring{t}}\ {}^{\lfloor\psi}D^{-0.5}_{i;\mathring{t}} e, \quad e \sim \mathcal{N}_e(0, I_{\mathring{\theta}_i}). \end{aligned}$$

   Above, the normal, moments-preserving, approximation of the marginal pdf $f(r_i|d(\mathring{t}))$ *is adopted.*
4. *Go to the beginning of* Iterative mode*, if* $\Theta_n \notin \Theta^*_s$*, i.e., if the multivariate AR model given by* $\theta_{in},\ i = 1, \ldots, \mathring{d}$*, is unstable.*

5. *Set* $\tilde{m} \equiv m+1$ *and* $\tilde{\theta}_{is} \equiv \hat{\theta}_{is} + \frac{1}{\tilde{m}}\left(\hat{\theta}_{is} - \theta_{in}\right)$, $\tilde{r}_{is} = \hat{r}_{is} + \frac{1}{\tilde{m}}(\hat{r}_{is} - r_{in})$, $i = 1, \ldots, \mathring{d}$.
6. *Set* $m = \tilde{m}$ *and* $\hat{\Theta}_s = \tilde{\Theta}_s$ *if* $\tilde{\Theta}_s \in \Theta_s^*$, *i.e., if* $\tilde{\theta}_{is}$, $i = 1, \ldots, \mathring{d}$, *define a stable AR model.*
7. *Go to the beginning of* Iterative mode.

Accept $\hat{\Theta}_s$ *as the final estimate if* $m = \mathring{m}$. *Otherwise, take the search as unsuccessful.*

Remark(s) 9.3

1. *Better stopping rules can be created using simple estimates of covariance matrices of the constructed expected value and inequality of the Chebyshev type [164]. Also, Bayesian stopping rules [94, 128] are at our disposal.*
2. *It is useful to remember that* Θ_s^* *may be restricted by another prior knowledge, for instance, by an expected range of static gains.*

The restricted set Θ_s^* is often much smaller than Θ^*. Then, the number of discarded samples is prohibitive. This makes us design an alternative sampling strategy. Essentially, samples are taken from Θ_s^* and the expected value is computed as their weighted mean value. The weights express the degree of how much the given sample can be attributed to the posterior pdf $f(\Theta|d(\mathring{t}))$. The need to preserve the structure of individual factors is the key obstacle faced.

The following steps have to be discussed to design such an improved sampling strategy. They use more or less textbook results. Their presentation complexity stems from the effort to stay with decisive evaluations at the factor level.

Let us inspect the eigenvalues of the *state matrix*

$$\mathcal{A} \equiv \begin{bmatrix} A_1 & A_2 & \ldots & A_{\partial-1} & A_\partial \\ I & 0 & \ldots & 0 & 0 \\ 0 & I & \ldots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \ldots & I & 0 \end{bmatrix} \tag{9.17}$$

corresponding to the multivariate AR model without offset μ. Let ρ be an eigenvalue of $\mathcal{A}$ assigned to an eigenvector ϕ, i.e., $\mathcal{A}\phi = \rho\phi$. Let us split ϕ into ∂ $\mathring{d}$-blocks $\phi' = [\phi_1', \ldots, \phi_\partial']$ corresponding to the size $\mathring{d}$ of the data record d_t, i.e., to the size of the multivariate regression coefficients A_i. The blocks ϕ_i are related by the recursion $\phi_i\rho = \phi_{i-1}$ valid for $i > 1$. This recursion and the identity for the block ϕ_1 give, for $\rho \neq 0$ of interest,

$$\phi_i = \rho^{-i+1}\phi_1,\ i = 2, \ldots, \partial, \text{ and } \rho\phi_1 = \sum_{i=1}^{\partial} A_i\phi_i = \sum_{i=1}^{\partial} A_i\rho^{-i+1}\phi_1$$

$$\Rightarrow\ 0 = \left(I - \sum_{i=1}^{\partial} A_i \rho^{-i}\right)\phi_1 \ \Rightarrow \rho \text{ is root of } \left|I - \sum_{i=1}^{\partial} A_i\rho^{-i}\right| = 0. \tag{9.18}$$

Let us consider the <u>modified</u> multivariate AR model with matrix coefficients $\tilde{A}_i = x^i A_i,\ i = 1, \ldots, \partial$, where x is an optional positive scalar. Obviously, if ρ is an eigenvalue corresponding to the coefficients $\{A_i\}_{i=1}^{\partial}$ then $x\rho$ is eigenvalue corresponding to the coefficients $\{\tilde{A}_i\}_{i=1}^{\partial}$.

Thus, if ρ is an eigenvalue with the largest module $|\rho|$ greater than one, then the AR model modified by the factor $x = g/|\rho|$ is stable for an arbitrary $g \in (0,1)$,.

Let us inspect individual components. Using widely available algorithms for evaluation of matrix eigenvalues, the component is tested whether or not the expected value of its regression coefficients define a stable state matrix. It is taken as the point estimate searched for, if the answer is positive. Otherwise, we select $g \in (0,1)$, and transform expected values $\hat{\theta}_i$, of regression coefficients θ_i of the ith factor, $i = 1, \ldots, \mathring{d}$, as follows.

$$\tilde{\theta}_i = \begin{bmatrix} I_{\mathring{d}-i} & & & \\ & \frac{g}{|\rho|} I_{\mathring{d}} & & \\ & & \ddots & \\ & & & \left(\frac{g}{|\rho|}\right)^{\partial} I_{\mathring{d}} \end{bmatrix} \theta_i \equiv T_i \hat{\theta}_i. \tag{9.19}$$

Let us consider the pdf $GiW_{\theta_i, r_i}\left(\tilde{\theta}_i, \tilde{C}_i, \tilde{r}_i, \tilde{\nu}_i\right)$. By construction, the expected value of the component described by such factors has spectral radius equal to $g < 1$.

We select the optional statistics $\tilde{C}_i, \tilde{r}_i, \tilde{\nu}_i$ so that the product of factors is as close as possible to the posterior pdf $\prod_{i=1}^{\mathring{d}} GiW_{\theta_i, r_i}\left(\hat{\theta}_i, C_i, \hat{r}_i, \nu_i\right)$. The following simple proposition holds.

Proposition 9.4 (The nearest shifted *GiW* pdf) *Let us consider some set of GiW pdfs* $\left\{\tilde{f}(\theta, r) = GiW_{\theta,r}(\tilde{\theta}, \tilde{C}, \tilde{r}, \tilde{\nu}),\ \tilde{\theta} \text{ is fixed}\right\}$, *and a given pdf* $f(\theta, r) = GiW_{\theta,r}(\hat{\theta}, C, \hat{r}, \nu)$. *Then,*

$$GiW_{\theta,r}\left(\tilde{\theta}, C + \frac{\nu\left(\tilde{\theta} - \hat{\theta}\right)\left(\tilde{\theta} - \hat{\theta}\right)'}{(\nu - 2)\hat{r}}, \hat{r}, \nu\right) = \arg \min_{\tilde{C} > 0, \tilde{r} > 0, \tilde{\nu} > 0} \mathcal{D}(f || \tilde{f}). \tag{9.20}$$

Proof. Using the explicit form of the KL divergence of a pair *GiW* pdfs, (8.25), it is straightforward to find the minimizing values $\tilde{r} = \hat{r}$ and $\tilde{\nu} = \nu$. The part depending on the remaining optional $\tilde{C}$ has the form

$$-0.5 \ln \left|C\tilde{C}^{-1}\right| + 0.5 \mathrm{tr}\left[C\tilde{C}^{-1}\right] + 0.5 \frac{\nu}{(\nu - 2)\hat{r}} \left(\hat{\theta} - \tilde{\theta}\right)' \tilde{C}^{-1} \left(\hat{\theta} - \tilde{\theta}\right).$$

Taking the derivative with respect to $\tilde{C}^{-1}$ and using the formulas (8.3), (8.6), we get the claimed form of $\tilde{C}$. $\square$

Using a direct inspection or formalism of the Radon–Nikodým derivative, [72], we find that the expected value $\hat{\theta}_i$ of the pdf $GiW_{\theta_i,r_i}(\hat{\theta}_i, C_i, \hat{r}_i, \nu_i)$ can be expressed in terms of the expectation $\tilde{\mathcal{E}}[\cdot]$ assigned to $GiW_{\theta_i,r_i}\left(\tilde{\theta}_i, \tilde{C}_i, \hat{r}_i, \nu_i\right)$ with $\tilde{\theta}_i,\ \tilde{C}_i$ given by formulas (9.19), (9.20). It has the following form.

$$\hat{\theta}_i = \mathcal{E}[\chi_{\theta_s^*}(\theta_i)\theta_i] = \frac{\tilde{\mathcal{E}}[\chi_{\theta_s^*}(\theta_i)\theta_i q(\theta_i)]}{\tilde{\mathcal{E}}[\chi_{\theta_s^*}(\theta_i)q(\theta)]} \quad \text{with} \tag{9.21}$$

$$q(\theta_i) \equiv \frac{GiW_{\theta_i,r_i}(\hat{\theta}_i, C, \hat{r}_i, \nu_i)}{GiW_{\theta_i,r_i}(\tilde{\theta}_i, \tilde{C}_i, \hat{r}_i, \nu_i)}.$$

Thus, we can draw samples from the *GiW* pdf having a stable AR model as the expected value and apply Monte Carlo evaluations to both expectations involved. For each $i \in i^*$, we take samples $\Theta_{ik} \sim GiW_{\Theta_i}\left(\tilde{\theta}_i, \tilde{C}_i, \hat{r}_i, \nu_i\right)$ and compute

$$\hat{\theta}_{is} \approx \frac{\sum_k \chi_{\theta_s^*}(\theta_{ik})\theta_{ik} q(\Theta_{ik})}{\sum_k \chi_{\Theta_s^*}(\theta_{ik}) q(\Theta_{ik})}. \tag{9.22}$$

This leads to the following modification of Algorithm 9.2.

Algorithm 9.3 (Point estimation of a stable AR component)

Initial mode

- *Perform the standard estimation, Chapter 8, giving the collection of statistics* $L_{i;\mathring{t}}, D_{i;\mathring{t}}, \nu_{i;\mathring{t}},\ i = 1, \ldots, \mathring{d}$.
- *Convert the collected statistics into their least-squares (LS) counterparts*

$$\begin{aligned}
\hat{\theta}_{i;\mathring{t}} &\equiv {}^{\lfloor \psi}L_{i;\mathring{t}}^{-1}\, {}^{\lfloor d\psi}L_{i;\mathring{t}} \equiv \textit{LS estimate of the regression coefficients}\\
\hat{r}_{i;\mathring{t}} &\equiv \frac{{}^{\lfloor d}D_{i;\mathring{t}}}{\nu_{i;\mathring{t}} - 2} \equiv \textit{LS estimate of the noise variance}\\
C_{i;\mathring{t}} &\equiv {}^{\lfloor \psi}L_{i;\mathring{t}}^{-1}\, {}^{\lfloor \psi}D_{i;\mathring{t}}^{-1} \left({}^{\lfloor \psi}L'_{i;\mathring{t}}\right)^{-1} \equiv \textit{LS covariance factor.}
\end{aligned}$$

- *Evaluate the* <u>*absolute value*</u> ρ *of the maximum eigenvalue of the state matrix (9.17) made of expected regression coefficients of all factors involved.*
- *Stop and take the unrestricted LS point estimates as the approximation of (9.16) if* $\rho < 1$ *as they define a stable AR model or equivalently belong to* Θ_s^* *(9.14).*
- *Select a number* $g_1 \in (0, 1)$*, say* $g_1 = 0.99$*, used for stabilization and set* $g = g_1/\rho$.

- *Define, for $i = 1, \ldots, \mathring{d}$, matrices T_i (9.19) and transform the statistics*

$$\tilde{\theta}_i = T_i \hat{\theta}_{i;\mathring{t}},\ \tilde{C}_i = C_{i;\mathring{t}} + \frac{\nu_{i;\mathring{t}} \left(\tilde{\theta}_i - \hat{\theta}_{i;\mathring{t}}\right) \left(\tilde{\theta}_i - \hat{\theta}_{i;\mathring{t}}\right)'}{(\nu_{i;\mathring{t}} - 2)\hat{r}_{i;\mathring{t}}}.$$

  Note that the standard rank-one updating can be used for evaluation of the $L'DL$ decomposition of the matrix $\tilde{C}_i$.
- *Select the upper bound $\mathring{n}$, say on the order of* $10^2 - 10^4$, *on the number n of generated samples and set $n = 0$.*
- *Select the bound $\mathring{m}$, say on the order of* $100 - - 1000$, *on the needed number m of stable samples and set $m = 0$.*
- *Initialize the vectors that will contain the generated stable point estimates $\hat{\theta}_{is} = 0,\ \hat{r}_{is} = 0$, and denominators $m_i = 0$, for $i = 1, \ldots, \mathring{d}$.*

Iterative mode

1. *Do while $n < \mathring{n}$ & $m < \mathring{m}$.*
2. *Set $n = n + 1$.*
3. *Generate samples $\theta_{in},\ r_{in} \sim GiW_{\theta_i, r_i}\left(\tilde{\theta}_i, \hat{r}_{i;\mathring{t}}, \tilde{C}_i, \nu_{i;\mathring{t}}\right)$, $i = 1, \ldots, \mathring{d}$,*

$$\begin{aligned} r_{in} &= \hat{r}_{i;\mathring{t}}(1 + e_r), \quad e_r \sim \mathcal{N}_{e_r}\left(0, \sqrt{\frac{2}{\nu_{i;\mathring{t}} - 4}}\right), \ cf.(8.23), \\ r_{in} &= \hat{r}_{i;\mathring{t}} \text{ if the sample } r_{in} \le 0, \\ \theta_{in} &= \tilde{\theta}_i + \sqrt{\hat{r}_i}\tilde{C}_i^{0.5} e, \quad e \sim \mathcal{N}_e(0, I_{\mathring{\theta}_i}). \end{aligned}$$

   The normal, moment-preserving, approximation of the marginal pdf $f(r_i|d(\mathring{t}))$ is adopted above.
4. *Test the component stability, i.e., inspect the maximum absolute values of eigenvalues of the state matrix (9.17) made of generated samples. Go to the beginning of* Iterative mode *if $\Theta_n \notin \Theta_s^*$, if the generated multivariate AR model is unstable.*
5. *Set, for $i = 1, \ldots, \mathring{d}$,*

$$\begin{aligned} \tilde{m}_i &= m_i + q\left(\Theta_{in}\right), \quad \tilde{\theta}_{is} = \hat{\theta}_{is} + \frac{q\left(\Theta_{in}\right)}{\tilde{m}_i}\left(\theta_{in} - \hat{\theta}_{is}\right) \\ \tilde{r}_{is} &= \hat{r}_{is} + \frac{q\left(\Theta_{in}\right)}{\tilde{m}_i}\left(r_{in} - \hat{r}_{is}\right), \ q(\cdot) \text{ is given by } (9.21). \end{aligned}$$

6. *Set $m = m + 1$, $\hat{\Theta}_{is} = \tilde{\Theta}_{is}$ and $m_i = \tilde{m}_i$, if $\tilde{\theta}_s$—made of $\tilde{\theta}_i$—is a stable AR model.*
7. *Go to the beginning of* Iterative mode.

Accept $\hat{\Theta}_s$ as the final estimate if $m = \mathring{m}$. Otherwise take the construction as unsuccessful.

9.1.3 Advices and their influence

The specialization of the model presented in Section 7.1.3, is implied by Agreement 9.1. Among other things, it implies that recognizable actions enter linearly regression vectors. Otherwise, the specialization to normal mixtures brings nothing new and it is elaborated in connection with the respective designs.

9.1.4 Practical presentation aspects

Multi-step-ahead predictors

Ideally, the estimated mixture should describe well both deterministic and stochastic relationships between observed history and future behavior. It is especially important for the discussed design part. The quality of the estimated mixture can be well judged according to its performance as a *multi-step-ahead* predictor. Its exact practical construction is difficult for the same reasons that make difficult the mixture estimation. Thus, an approximation is needed. The following approximation is based on our ability to evaluate the one-step-ahead predictor. Essentially, this predictor is used as a random generator of a next data item, which is inserted into the condition of the next predictor, etc.

This relatively short-horizon simulation is expected to reveal a significant information about the simulated model due to a large number of such simulation periods that coincides with the number of processed data records $\mathring{t}$. The algorithm is rather simple but surprisingly error prone. This has motivated the following detailed presentation of the algorithm. The simulated values are marked by the symbol ˜.

Algorithm 9.4 (Simulation-based multi-step predictor)

Initial mode

- *Specify the analyzed normal mixture, typically, estimate it as described in Chapter 8.*
- *Select the prediction horizon* $T \geq 1$.

Sequential mode, *running for* $t = 1, 2, \ldots,$

1. *Complement the state vector* ϕ_t *by* d_t.
2. *Specify recognizable actions* $u_{0;t+1}$ *and complement the regression vectors* $\psi_{\mathring{\Delta}c;t+1}$, $c \in c^*$.
3. *Generate* $c_{t+1} \sim [\alpha_1, \ldots, \alpha_{\mathring{c}}]$.

```
   Make copies $\tilde{\phi}_{c;t} = \phi_{c;t}$, $c \in c^*$ and set $\tilde{\psi}_{\mathring{\Delta}c_{t+1};t+1} \equiv \phi_{c_{t+1};t}$
        For $i = \mathring{\Delta}, \mathring{\Delta} - 1, \ldots, 1$
             Sample $\tilde{d}_{i;t+1} \sim \mathcal{N}_{d_{i;t+1}}\left(\theta'_{ic_{t+1}}\tilde{\psi}_{ic_{t+1};t+1}, r_{ic_{t+1}}\right)$.
             Complement $\tilde{\psi}_{(i-1)c_{t+1};t+1}$ by the sample $\tilde{d}_{i;t+1}$.
```

```
            end   of the cycle over $i$
      Take $\tilde{d}_{t+1}$ from $\mathring{d}$ initial entries of $\tilde{\Psi}_{c_{t+1};t+1} \equiv \tilde{\psi}_{0c_{t+1};t+1}$.
            For   $c = 1, \ldots, \mathring{c}$
                  Create $\tilde{\psi}_{\mathring{d}c;t+1} \equiv \tilde{\phi}_{c;t}$.
            end   of the cycle over $c$
          For   $\tau = t+1, \ldots, t+T$
                  Generate $c_\tau \sim [\alpha_1, \ldots, \alpha_{\mathring{c}}]$.
            For   $i = \mathring{d}, \mathring{d}-1, \ldots, 1$
                  Sample $\tilde{d}_{i;\tau} \sim \mathcal{N}_{d_{i;\tau}}\left(\theta'_{ic_\tau}\tilde{\psi}_{ic_\tau;\tau}, r_{ic_\tau}\right)$.
                  Complement $\tilde{\psi}_{(i-1)c_\tau;\tau}$ by the sample $\tilde{d}_{i;\tau}$.
            end   of the cycle over $i$
      Take the simulated $\tilde{d}_\tau$ from $\mathring{d}$ initial entries of $\tilde{\Psi}_{c_\tau;\tau} \equiv \tilde{\psi}_{0c_\tau;\tau}$.
            For   $c = 1, \ldots, \mathring{c}$
                  Create $\tilde{\psi}_{\mathring{d}c;\tau+1} = \tilde{\phi}_{c;\tau}$.
            end   of the cycle over $c$
          end   of the cycle over $\tau$
```

4. *Use the one-step-ahead predictor, Proposition 8.14, given by the state vectors $\tilde{\phi}_{c;t+T-1}$. For instance, evaluate moments of d_{t+T}.*

Remark(s) 9.4

1. *Except for the initial prediction time, when the recognizable actions are determined externally, the whole data record d_t should be simulated to get the real test of the overall model.*
2. *Component probabilities used for selecting the active mixture are either those estimated or their dynamic approximation; Section 7.1.2. The approximation can be restricted to the first step only or applied over the whole prediction horizon. The second variant is conjectured as the better one.*
3. *A sort of ergodic hypothesis is behind the construction of the algorithm: predictive relevance of samples is supposed. This aspect should be inspected more closely.*

Support of visible part of normal mixture

A graphic representation of the mixture predicting several data entries, say y, is shown to the operator. It requires evaluation of a multivariate interval

$[\underline{y}, \overline{y}]$ on which the values of the pdf

$$f(y|\mathcal{P}) = \sum_{c \in c^*} \alpha_c \mathcal{N}_y(\mu_c, r_c)$$

are high enough. We derive simple but useful algorithm for determining the interval $[\underline{y}, \overline{y}]$.

The component weights α_c, the vector-valued expected values μ_c and covariance matrices $r_c = L_c' D_c L_c$ are fixed in this task by the experience $\mathcal{P}$ in the condition. Normality implies that the highest value $(2\pi)^{-0.5\mathring{y}} \bar{\beta}$ of the involved weighted component is

$$(2\pi)^{-0.5\mathring{y}} \bar{\beta} \equiv \max_{c \in c^*, y \in (-\infty,\infty)^{\mathring{y}}} \alpha_c f(y|\mathcal{P}, c) = (2\pi)^{-0.5\mathring{y}} \max_{c \in c^*} \beta_c$$

$$\beta_c \equiv \frac{\alpha_c}{\sqrt{|r_c|}} = \frac{\alpha_c}{\sqrt{\prod_{i=1}^{\mathring{y}} D_{ic}}}.$$

Let us define the *visibility level* $K \in (0, 1)$, say $K = 0.01$. Then, the weighted pdf describing the weighted cth component is to be displayed on the appropriate interval $[\underline{y}, \overline{y}]$ on which $f(y|\mathcal{P}, c) \geq K(2\pi)^{-0.5\mathring{y}} \bar{\beta}$. The extreme values $\underline{y}_{ic} \equiv \mu_{ic} - x$, $\overline{y}_{ic} = \mu_{ic} + x$ for the cth component on the ith axis, $i = 1, \ldots, \mathring{y}$, are determined by the positive solution of the equation

$$\max_{x_\iota \in (-\infty,\infty),\, \iota \neq i} \exp\{-0.5\, [x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_{\mathring{y}}]\, r_c^{-1}$$
$$\times [x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_{\mathring{y}}]'\}$$
$$= \exp\left\{-0.5 \frac{x^2}{D_{ic}}\right\} = K \frac{\bar{\beta}}{\beta_c} \Leftrightarrow x = \sqrt{D_{ic} \zeta} \equiv \sqrt{D_{ic} 2 \ln\left(\frac{\beta_c}{K\bar{\beta}}\right)}.$$

The positive solution exists iff $\frac{\beta_c}{K\bar{\beta}} > 1$. The interval that is searched for has to cover the union of intervals $[\underline{y}_{ic}, \overline{y}_{ic}]$ for all components $c \in c^*$ and all axes $i \in \{1, \ldots, \mathring{y}\}$. These considerations lead to the overall algorithm.

Algorithm 9.5 (Practical support of a normal mixture)

Initial mode

- *Evaluate moments* $\mu_c, r_c \equiv L_c' D_c L_c$ *and weights* α_c *of the normal mixture of the interest.*
- *Specify the visibility level* $K \in (0, 1)$*, typically,* $K = 0.01$.
- *Define the initial values of the constructed interval* $\underline{y}_i = +\infty$, $\overline{y}_i = -\infty$, $i = 1, \ldots, \mathring{y}$.
- *Set* $\bar{\beta} = -\infty$.

Construction mode

```
   For   $c = 1, \ldots, \mathring{c}$
           $\beta_c = \frac{\alpha_c}{\sqrt{\prod_{i=1}^{\mathring{y}} D_{ic}}}, \quad \bar{\beta} = \max(\bar{\beta}, \beta_c).$
   end   of the cycle over $c$
 For   $c = 1, \ldots, \mathring{c}$
           $\beta_c \equiv \frac{\beta_c}{\bar{\beta}}, \quad \zeta_c = 2\ln\left(\frac{\beta_c}{K}\right).$
           Do if $\zeta_c > 0, \ \zeta_c = \sqrt{\zeta_c}$
   For   $i = 1, \ldots, \mathring{y}$
           $x_{ic} = \sqrt{D_{ic}}\zeta_c, \quad \underline{y}_i = \min(\underline{y}_i, \mu_{ic} - x_{ic}), \quad \overline{y}_i = \max(\overline{y}_i, \mu_{ic} + x_{ic})$
   end   of the cycle over $i$
           end of the condition $\zeta_c > 0$
 end   of the cycle over $c$
```

Rescaling

Both learning and design work on scaled data. Scaling is necessary both for numeric reasons and unification of various default values. The final presentation has to be done in the original user's units. It concerns predicted values only as we can always normalize data fed into the conditions, including those just contemplated. Thus, rescaling reduces to a simple, often affine, transformation of pdfs, cf. Proposition 2.5.

Projections to be presented

Operator can see low-dimensional pdfs only. Thus, a practical question arises — which of many possible projections is adequate?

The projected pdf is the ideal joint pdf resulting from the design. The optimized influence of the presentation action is based on the fact that the operator can willingly influence only the distribution of the shown quantities having indexes z_t. Thus, he can modify

$$f(d_t|d(t-1)) \to \frac{f(d_t|d(t-1))}{f(d_{z_t;t}|d(t-1))} \, {}^{\lfloor I}f(d_{z_t;t}|d(t-1)) \equiv {}^{\lfloor I}f(d_t|z_t, d(t-1)).$$

This relation models the influence of presentation actions z_t; Section 9.1.3. The strategy generating the actions z_t is optimized to get the modified pdf ${}^{\lfloor I}f(d_t|z_t, d(t-1))$ as close as possible to the user's ideal pdf. It leads to the clear conclusion: the marginal ${}^{\lfloor I}f(d_{z_t;t}|d(t-1))$ has to be presented, otherwise the optimal nature of the constructed advices is lost.

Point advice

Generally, advices to the operator are presented as marginal pdfs of a mixture resulting from an optimization with the common instruction *seek for a high-probability area.* Sometimes, a point advice is required. If the mixture is unimodal (if it contains single component), the maximizing argument is such an advice. It is harmonized with the above instruction. The question arises whether the same argument should be used in the multimodal case. The answer is positive. Such a choice can be interpreted as the maximum probable advice. It minimizes the distance of the ideal pdf to the user's ideal pdf while selecting the user's ideal pdf on advices as the optimized design knob; cf. Proposition 7.11.

On experimental evaluation of the KL divergence

During experimental verifications of the designed advisory system, the natural question arose how to evaluate experimentally the KL divergence of the objective pdf ${}^{\lfloor o}f$ of data (see Chapter 2) to the user's ideal pdf ${}^{\lfloor U}f$. The most straightforward approximation of the unknown ${}^{\lfloor o}f$ by a formal sample pdf fails. The sample pdf equals the average of the Dirac delta functions sitting on measured data and cannot be used in expressions of the type $\bullet\ln(\bullet)$. When approximating Dirac functions smoothly by Gaussian pdfs centered on measured data and having variance $R \to 0$, we also get infinite values. This makes us use the normal approximation but selecting R as the minimizer of the inspected KL divergence instead of considering $R \to 0$. It seems to be a proper solution of the problem. Let us describe it in detail.

Let ${}^{\lfloor o}f(d(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor o}f(d_t|\phi_{t-1})$, where the state vectors ϕ_t include both the state vector describing the objective dependence and the state vector in the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor U}f(d_t|\phi_{t-1})$. The normalized asymptotic KL divergence can be written as follows.

$$\overline{\lim}_{\mathring{t}\to\infty}\frac{1}{\mathring{t}}\mathcal{D}\left({}^{\lfloor o}f \,\middle|\middle|\, {}^{\lfloor U}f\right) = \overline{\lim}_{\mathring{t}\to\infty}\mathcal{E}\underbrace{\left[\int {}^{\lfloor o}f(d_t|\phi_{t-1})\ln\left(\frac{{}^{\lfloor o}f(d_t|\phi_{t-1})}{{}^{\lfloor U}f(d_t|\phi_{t-1})}\right) dd_t\right]}_{\omega(\phi_{t-1})}.$$

If we manage to evaluate at least approximately a finite sample version of $\omega(\phi_{t-1})$, we can use ordinary approximation

$$\mathcal{E}[\omega(\phi_{t-1})] \approx \frac{1}{\mathring{t}}\sum_{t=1}^{\mathring{t}}\omega(\tilde{\phi}_{t-1}),$$

hoping in ergodic behavior. The measured values are marked by the symbol ˜ in this paragraph.

As planned above, we approximate the objective pdf ${}^{\lfloor o}f(d_t|\phi_{t-1})$ at the measured point $\tilde{d}_t, \tilde{\phi}_{t-1}$ by the normal pdf ${}^{\lfloor o}f(d_t|\tilde{\phi}_{t-1}) \approx \mathcal{N}_{d_t}(\tilde{d}_t, R)$ with the optional covariance matrix R.

For the normal user's ideal pdf ${}^{\lfloor U}f(d_t|\tilde{\phi}_{t-1}) = \mathcal{N}_{d_t}\left({}^{\lfloor U}M(\tilde{\phi}_{t-1}),\ {}^{\lfloor U}R\right)$, Proposition 8.10 implies

$$\begin{aligned}\omega(\tilde{\phi}_{t-1}) = 0.5 \Big\{ & \ln \left| {}^{\lfloor U}RR^{-1} \right| - \mathring{d} + \mathrm{tr}\left[R\, {}^{\lfloor U}R^{-1}\right] \\ & + \left(\tilde{d}_t - {}^{\lfloor U}M(\tilde{\phi}_{t-1})\right)' \, {}^{\lfloor U}R^{-1} \left(\tilde{d}_t - {}^{\lfloor U}M(\tilde{\phi}_{t-1})\right) \Big\}.\end{aligned}$$

It is minimized by $R = {}^{\lfloor U}R$. This result and the sample-mean replacement of the expectation give the practical plausible indicator of the design quality

$$\frac{1}{2\mathring{t}} \sum_{t=1}^{\mathring{t}} \left(\tilde{d}_t - {}^{\lfloor U}M(\tilde{\phi}_{t-1})\right)' \, {}^{\lfloor U}R^{-1} \left(\tilde{d}_t - {}^{\lfloor U}M(\tilde{\phi}_{t-1})\right). \tag{9.23}$$

9.1.5 Quadratic forms in a fully probabilistic design

Section 7.1.4 expresses the KL divergence of the constructed ideal pdf on the p-data $d(\mathring{t})$ to the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$. It consists of the true user's ideal pdf $\prod_{t \in t^*} {}^{\lfloor U}f(d_{o;t}|d_o(t-1))$ and of the factor on the surplus data $d_{p+}(\mathring{t})$ of the p-system. According to the assumption (5.7), it coincides with the pdf $\prod_{t \in t^*} {}^{\lfloor I}f(d_{p+;t}|d(t-1))$ derived from the constructed ideal pdf ${}^{\lfloor I}f(d(\mathring{t}))$. The fully probabilistic design, Proposition 2.11, with this special target is described by Proposition 7.4. Its normal counterpart brings nothing new. The specific nature of the normal case can be exploited after expressing upper bounds on the KL divergence. Before progressing in this respect, the necessary manipulations with expected quadratic forms and the conditional KL divergence of normal pdfs are prepared. The formulas are expressed in the way suitable for an efficient evaluation that relies on the factorized form of normal components.

It is worthwhile stressing, that LDL' decomposition is used further on as it fits to the operations needed in the discussed design part of construction of the p-system.

Proposition 9.5 (Expected quadratic form) *Under Agreement 9.1, let us consider a normal parameterized component described by*

$$\begin{aligned} f(\mathit{\Delta}_t|u_{o;t}, \phi_{t-1}, \mathit{\Theta}) &\equiv \prod_{i \in i^*} \mathcal{N}_{\mathit{\Delta}_{i;t}}(\theta_i' \psi_{i;t}, r_i) \quad \textit{with} \\ \mathit{\Theta} &\equiv (\mathit{\Theta}_1, \ldots, \mathit{\Theta}_{\mathring{\mathit{\Delta}}}), \quad \mathit{\Theta}_i \equiv [\theta_i, r_i] \\ \psi_{i;t}' &\equiv [\mathit{\Delta}_{(i+1)\cdots\mathring{\mathit{\Delta}};t}', u_{o;t}', \phi_{t-1}'] \equiv [\mathit{\Delta}_{i+1;t}, \psi_{i+1;t}'] \\ & \qquad i = 0, 1, \ldots, \mathring{\mathit{\Delta}} - 1, \ \textit{recall } \psi_{0;t} \equiv \mathit{\Psi}_t, \\ \psi_{\mathring{\mathit{\Delta}};t}' &\equiv [u_{o;t}', \phi_{t-1}']. \end{aligned} \tag{9.24}$$

Let us consider a quadratic form in $\psi_{0;t} \equiv \mathit{\Psi}_t$ with the given kernel $L_0 D_0 L_0'$.

Then, the expected quadratic form, lifted by a constant k_0, reads

$$\begin{aligned}\mathcal{E}[k_0 + \Psi_t' L_0 D_0 L_0' \Psi_t | u_{o;t}, \phi_{t-1}] &\equiv \mathcal{E}[k_0 + \psi_{0;t}' L_0 D_0 L_0' \psi_{0;t} | \psi_{\mathring{\Delta};t}] \qquad (9.25)\\ &= k_{\mathring{\Delta}} + \psi_{\mathring{\Delta};t}' L_{\mathring{\Delta}} D_{\mathring{\Delta}} L_{\mathring{\Delta}}' \psi_{\mathring{\Delta};t}, \quad \textit{where}\end{aligned}$$

$$L_{i+1} D_{i+1} L_{i+1}' = {}^{\lfloor\psi}L_i \, {}^{\lfloor\psi}D_i \, {}^{\lfloor\psi}L_i' + \left(\theta_{i+1} + {}^{\lfloor\Delta\psi}L_i\right) {}^{\lfloor\Delta}D_i \left(\theta_{i+1} + {}^{\lfloor\Delta\psi}L_i\right)'$$

$$k_{i+1} = k_i + {}^{\lfloor\Delta}D_i r_{i+1}, \quad \textit{for } i = 0, \ldots, \mathring{\Delta} - 1, \quad \textit{with}$$

$$L_i \equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor\Delta\psi}L_i & {}^{\lfloor\psi}L_i \end{bmatrix}, \; D_i \equiv \mathrm{diag}\left[{}^{\lfloor\Delta}D_i, {}^{\lfloor\psi}D_i\right],$$

$$\textit{where } {}^{\lfloor\Delta}D_i \textit{ is scalar and } \mathring{D}_{i+1} = \mathring{D}_i - 1.$$

The updating the LDL' decomposition can be made using Algorithm 8.2.

Proof. The expectation is taken over entries of Δ_t since the remaining part of the data vector Ψ_t is fixed by the condition $\psi_{\mathring{\Delta};t} \equiv [u_{o;t}', \phi_{t-1}']'$. The the chain rule for expectations, Proposition 2.6, implies that we can evaluate conditional expectations of individual entries in the vector Δ_t one-by-one, starting from the first one. Taking a generic step and using the identity

$$\mathcal{E}\left[\Delta_{i+1}^2 | \psi_{i+1}\right] = r_{i+1} + \left\{\mathcal{E}[\Delta_{i+1} | \psi_{i+1}]\right\}^2, \quad \text{we have}$$

$$\begin{aligned}&\mathcal{E}\left[k_i + \psi_{i;t}' L_i D_i L_i' \psi_{i;t} | \psi_{i+1;t}\right] = \underbrace{k_i + {}^{\lfloor\Delta}D_i r_{i+1}}_{k_{i+1}}\\ &+ \psi_{i+1;t}' \begin{bmatrix} \theta_{i+1}' \\ I_{\mathring{\psi}_{i+1}} \end{bmatrix}' L_i D_i L_i' \begin{bmatrix} \theta_{i+1}' \\ I_{\mathring{\psi}_{i+1}} \end{bmatrix} \psi_{i+1;t} = k_{i+1}\\ &+ \psi_{i+1;t}' \underbrace{\left[{}^{\lfloor\psi}L_i \, {}^{\lfloor\psi}D_i \, {}^{\lfloor\psi}L_i' + \left(\theta_{i+1} + {}^{\lfloor\Delta\psi}L_i\right) {}^{\lfloor\Delta}D_i \left(\theta_{i+1} + {}^{\lfloor\Delta\psi}L_i\right)'\right]}_{L_{i+1} D_{i+1} L_{i+1}'} \psi_{i+1;t}.\end{aligned}$$

□

The treated quadratic forms do not contain linear term due to the inclusion of unit into regression vector. It can be numerically dangerous as it corresponds to the system description with a state at stability boundary. Consequently, the numerical noise causes divergence in dynamic programming. Moreover, when using this inclusion, the constant lift of the quadratic form is split into two parts and design algorithms need their sum. In software implementation, it is simply avoided by treating the *linear term of quadratic form* independently using the following proposition.

Proposition 9.6 (Expected linear form of factorized component) *Under Agreement 9.1, let us consider a normal parameterized component described by (9.24) and a linear form $l_0' \psi_{0;t}$ in $\psi_{0;t} \equiv \Psi_t$ given by a vector l_0.*

The expected linear form is evaluated recursively for $i = 0, \ldots, \mathring{\Delta} - 1$

$$\mathcal{E}[l_0' \Psi_t | u_{o;t}, \phi_{t-1}] \equiv \mathcal{E}[l_0' \psi_{0;t} | \psi_{\mathring{\Delta};t}] = l_{\mathring{\Delta}}' \psi_{\mathring{\Delta};t} \tag{9.26}$$
$$l_{i+1}' = {}^{\lfloor\Delta}l_i \theta_{i+1}' + {}^{\lfloor\psi}l_i' \ \ \textit{with}\ l_i \equiv \left[\ {}^{\lfloor\Delta}l_i \ \ {}^{\lfloor\psi}l_i'\right]',\ \ {}^{\lfloor\Delta}l_i \ \textit{is scalar},\ \mathring{l}_{i+1} = \mathring{l}_i - 1.$$

Proof. The expectation is taken over entries of Δ_t as the rest of the data vector Ψ_t is fixed by the condition $\psi_{\mathring{\Delta};t} \equiv [u_{o;t}', \phi_{t-1}']'$. The the chain rule for expectations, Proposition 2.6, implies that we can evaluate conditional expectations of individual entries in Δ_t one-by-one, starting from the first one

$$\mathcal{E}\left[l_i' \psi_{i;t} | \psi_{i+1;t}\right] = l_i' \begin{bmatrix} \theta_{i+1}' \\ I_{\mathring{\psi}_{i+1}} \end{bmatrix} \psi_{i+1;t} = \underbrace{\left[{}^{\lfloor\Delta}l_i \theta_{i+1}' + {}^{\lfloor\psi}l_i'\right]}_{l_{i+1}'} \psi_{i+1;t}.$$

□

A combination of Propositions 9.5, 9.6 gives a numerically safe evaluation of the expected value of the lifted quadratic form. The algorithm is based on a separate processing of the linear term in the treated quadratic form.

Algorithm 9.6 (Safe evaluation of the expected quadratic form)
Initial mode

- *Specify parameters* $\left(\bar{\theta}_i', r_i\right) \equiv \left([\theta_i', \mu_i], r_i\right)$ *of a component in the factorized form with the scalar offset* μ_i *at the last position.*
- *Specify the lift* $\bar{k}_0 \equiv k_0$ *and the kernel* $\bar{L}_0 \bar{D} \bar{L}_0'$ *of the lifted quadratic form in* $k_0 + \psi_{0;t}' \bar{L}_0 \bar{D} \bar{L}_0' \psi_{0;t}$ *whose expectation conditioned by* $\psi_{\mathring{\Delta};t}$ *is evaluated. The matrices* $\bar{L}_i, \bar{D}_i,\ i = 0, \ldots, \mathring{\Delta} - 1$, *are split*

$$\bar{L}_i = \begin{bmatrix} L_i & 0 \\ l_i' & 1 \end{bmatrix},\ \bar{D}_i = \begin{bmatrix} D_i & \\ & 0 \end{bmatrix}.$$

Recursive mode

For $\ i = 1, \ldots, \mathring{\Delta}$

$$L_{i+1} D_{i+1} L_{i+1}' = {}^{\lfloor\psi}L_i \, {}^{\lfloor\psi}D_i \, {}^{\lfloor\psi}L_i' + \left(\theta_{i+1} + {}^{\lfloor\Delta\psi}L_i\right) {}^{\lfloor\Delta}D_i \left(\theta_{i+1} + {}^{\lfloor\Delta\psi}L_i\right)'.$$

Updating of the LDL' *decomposition is made using Algorithm 8.2.*

$k_{i+1} = k_i + {}^{\lfloor\Delta}D_i r_{i+1}$, *with*

$$L_i \equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor\Delta\psi}L_i & {}^{\lfloor\psi}L_i \end{bmatrix},\ D_i \equiv \mathrm{diag}\left[{}^{\lfloor\Delta}D_i,\ {}^{\lfloor\psi}D_i\right],$$

where ${}^{\lfloor\Delta}D_i$ *is scalar and* $\mathring{D}_{i+1} = \mathring{D}_i - 1$

$l_{i+1}' = {}^{\lfloor\Delta}l_i \bar{\theta}_{i+1}' + {}^{\lfloor\psi}l_i'$ *with,* $\ l_i \equiv \left[\ {}^{\lfloor\Delta}l_i \ \ {}^{\lfloor\psi}l_i'\right]'$,

where ${}^{\lfloor\Delta}l_i$ *is scalar,* $\mathring{l}_{i+1} = \mathring{l}_i - 1$.

The operations $\ k_{i+1} \equiv k_i + 2 l_{\mathring{\psi}_{i+1}(i+1)},\ \ l_{\mathring{\psi}_{i+1}(i+1)} = 0$

```
    shift a constant from the linear term to the lift.
end of the cycle over i
```

The following auxiliary proposition is repeatedly used in approximate, fully probabilistic designs.

Proposition 9.7 (The conditional KL divergence) *Under Agreement 9.1, let* $\Delta_t = (\Delta_{o;t}, \Delta_{p+;t})$, $f(\Delta_t|u_{o;t}, d(t-1)) = \prod_{i=1}^{\mathring{\Delta}} \mathcal{N}_{\Delta_{i;t}}(\theta_i'\psi_{i;t}, r_i)$ *and*

$$^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d(t-1)) = \prod_{i=1}^{\mathring{\Delta}_o} \mathcal{N}_{\Delta_{i;t}}\left(^{\lfloor U}\theta_i'\psi_{i;t},\ ^{\lfloor U}r_i \right).$$

The regression coefficients $^{\lfloor U}\theta_i$ *of the user's ideal pdf* $^{\lfloor U}f(\cdot)$ *are complemented by zeros so that the regression vectors* $\psi_{i;t}$ *for the ith factor and the corresponding factor of the user's ideal pdf are common. Recall that, for* $i \in \{0, \ldots, \mathring{\Delta} - 1\}$,

$$\begin{aligned}\psi_{i;t} &= [\Delta'_{(i+1)\cdots\mathring{\Delta};t}, u'_{o;t}, \phi'_{t-1}]' \equiv [\Delta'_{(i+1)\cdots\mathring{\Delta};t}, \psi'_{\mathring{\Delta};t}]' \equiv [\Delta_{(i+1);t}, \psi'_{i+1;t}]' \\ \psi_{\mathring{\Delta};t} &= [u'_{o;t}, \phi'_{t-1}]'.\end{aligned}$$

Then,

$$\begin{aligned}\omega(u_{o;t}, \phi_{t-1}) &\equiv \omega(\psi_{\mathring{\Delta};t}) \\ &\equiv 2\int f(\Delta_t|u_{o;t}, \phi_{t-1}) \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, \phi_{t-1})}{^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, \phi_{t-1})} \right) d\Delta_t \\ &= k_{\mathring{\Delta}} + \psi'_{\mathring{\Delta};t} L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}} \psi_{\mathring{\Delta};t}.\end{aligned} \tag{9.27}$$

The lift $k_{\mathring{\Delta}}$ *and* kernel $L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}}$ *of the conditional KL divergence are found recursively*

```
For $i = 1, \ldots, \mathring{\Delta}$
    $L_i D_i L_i' = {}^{\lfloor \psi}L_{i-1}\, {}^{\lfloor \psi}D_{i-1}\, {}^{\lfloor \psi}L'_{i-1}$
        $+ \left(\theta_i + {}^{\lfloor \Delta\psi}L_{i-1)}\right) {}^{\lfloor \Delta}D_{i-1} \left(\theta_i + {}^{\lfloor \Delta\psi}L_{i-1}\right)'$
        $+\chi\left(i \le \mathring{\Delta}_o\right) \left(\theta_i - {}^{\lfloor U}\theta_i\right) {}^{\lfloor U}r_i^{-1} \left(\theta_i - {}^{\lfloor U}\theta_i\right)'$
    $k_i = k_{i-1} + {}^{\lfloor \Delta}D_{i-1} r_i + \chi\left(i \le \mathring{\Delta}_o\right) \left[\ln\left(\frac{^{\lfloor U}r_i}{r_i}\right) + \frac{r_i}{^{\lfloor U}r_i}\right]$,
    $k_0 = -\mathring{\Delta}_o,\ D_0 = 0_{\mathring{\psi},\mathring{\psi}},\ L_0 = I_{\mathring{\psi}}$
    $L_i \equiv \left[\begin{array}{cc} 1 & 0 \\ {}^{\lfloor \Delta\psi}L_i & {}^{\lfloor \psi}L_i \end{array}\right],\ D_i \equiv \mathrm{diag}\left[{}^{\lfloor \Delta}D_i,\ {}^{\lfloor \psi}D_i\right]$, where
    ${}^{\lfloor \Delta}D_i$ is scalar and $\mathring{D}_{i+1} = \mathring{D}_i - 1$.
end of the cycle over i
```

The updating of the LDL' (!) decomposition can be done by a double, if the set indicator $\chi(\cdot) = 1$, or single, if $\chi(\cdot) = 0$, use of Algorithm 8.2.

Proof. Let the innovations be split $\Delta = (\Delta_o, \Delta_{p+})$. Then,

$$(u_{o;t}, \phi_{t-1})$$

$$\begin{aligned}
&2\int f(\Delta_t|u_{o;t}, \phi_{t-1}) \ln\left(\frac{f(\Delta_t|u_{o;t}, d(t-1))}{f(\Delta_{p+;t}|u_{o;t}, d(t-1))\ {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d(t-1))}\right) d\Delta_t \\
&= \sum_{i=1}^{\mathring{\Delta}_o} \ln\left(\frac{{}^{\lfloor U}r_i}{r_i}\right) + \sum_{i=1}^{\mathring{\Delta}_o} \int f(\Delta_t|u_{o;t}, \phi_{t-1}) \\
&\times \left[-\frac{(\Delta_{i;t} - \theta_i'\psi_{i;t})^2}{r_i} + \frac{\left(\Delta_{i;t} - {}^{\lfloor U}\theta_i'\psi_{i;t}\right)^2}{{}^{\lfloor U}r_i}\right] d\Delta_t \qquad \underbrace{=}_{\substack{\textit{Proposition } 2.6 \\ \mathcal{E}[\bullet^2] = \mathcal{E}^2[\bullet] + \mathrm{cov}[\bullet]}} \\
&= \sum_{i=1}^{\mathring{\Delta}_o} \ln\left(\frac{{}^{\lfloor U}r_i}{r_i}\right) - \mathring{\Delta}_o + \sum_{i=1}^{\mathring{\Delta}_o} \int f(\Delta_t|u_{o;t}, \phi_{t-1}) \frac{(\Delta_{i;t} - {}^{\lfloor U}\theta_i'\psi_{i;t})^2}{{}^{\lfloor U}r_i} d\Delta_t \\
&\underbrace{=}_{\textit{Proposition 2.6}} \quad -\mathring{\Delta}_o + \sum_{i=1}^{\mathring{\Delta}_o} \left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_i}\right) + \frac{r_i}{{}^{\lfloor U}r_i}\right] \\
&+ \sum_{i=1}^{\mathring{\Delta}_o} \mathcal{E}\left[\psi_{i;t}'\left(\theta_i - {}^{\lfloor U}\theta_i\right)\ {}^{\lfloor U}r_i^{-1}\left(\theta_i - {}^{\lfloor U}\theta_i\right)'\psi_{i;t}\middle|\psi_{\mathring{\Delta};t}\right].
\end{aligned}$$

Let us define the kernel $L_{i-1}D_{i-1}L_{i-1}'$ of the lifted quadratic form $k_{i-1} + \psi_{i-1;t}'L_{i-1}D_{i-1}L_{i-1}'\psi_{i-1;t}$ for which we evaluate

$$\mathcal{E}[k_{i-1} + \psi_{i-1;t}'L_{i-1}D_{i-1}L_{i-1}'\psi_{i-1;t}|\psi_{\mathring{\Delta};t}].$$

Then, according to the equation (9.26) in the proof of Proposition 9.5, an intermediate lifted quadratic form arises $\tilde{k}_i + \psi_{i;t}'\tilde{L}_i\tilde{D}_i\tilde{L}_i'\psi_{i;t}$ with

$$\begin{aligned}
\tilde{L}_i\tilde{D}_i\tilde{L}_i' &= {}^{\lfloor\psi}L_{i-1}\ {}^{\lfloor\psi}D_{i-1}\ {}^{\lfloor\psi}L_{i-1}' + \left(\theta_i + {}^{\lfloor\Delta\psi}L_{i-1}\right)\ {}^{\lfloor\Delta}D_{i-1}\left(\theta_i + {}^{\lfloor\Delta\psi}L_{i-1}\right)' \\
\tilde{k}_i &= k_{i-1} + {}^{\lfloor\Delta}D_{i-1}r_i.
\end{aligned}$$

While $i \leq \mathring{\Delta}_o$, the expression

$$\ln\left(\frac{{}^{\lfloor U}r_i}{r_i}\right) + \frac{r_i}{{}^{\lfloor U}r_i} + \psi_{i;t}'\left(\theta_i - {}^{\lfloor U}\theta_i\right)\ {}^{\lfloor U}r_i^{-1}\left(\theta_i - {}^{\lfloor U}\theta_i\right)'\psi_{i;t}$$

has to be added to it.

$$L_i D_i L_i' = {}^{\lfloor\psi}L_{i-1}\,{}^{\lfloor\psi}D_{i-1}\,{}^{\lfloor\psi}L_{i-1}' + \left(\theta_i + {}^{\lfloor\Delta\psi}L_{i-1}\right)\,{}^{\lfloor\Delta}D_{i-1}\left(\theta_i + {}^{\lfloor\Delta\psi}L_{i-1}\right)'$$
$$+\,\chi\left(i \le \mathring{\Delta}_o\right)\frac{\left(\theta_i - {}^{\lfloor U}\theta_i\right)\left(\theta_i - {}^{\lfloor U}\theta_i\right)'}{{}^{\lfloor U}r_i}$$
$$k_i = k_{i-1} + {}^{\lfloor\Delta}D_{i-1}r_i + \chi\left(i \le \mathring{\Delta}_o\right)\left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_i}\right) + \frac{r_i}{{}^{\lfloor U}r_i}\right].$$

The updating can be made by a double or single use of Algorithm 8.2. The initial values are $k_0 = -\mathring{\Delta}_o$ and $D_0 = 0;\ L_0 = I =$ unit matrix. □

While performing the approximate fully probabilistic design, we work with a weighted version of the conditional KL divergence. The following slight extension of Proposition 9.7 supports the cases for which the function $-\ln(\gamma)$ determining the Bellman function is approximated by a lifted quadratic form.

Proposition 9.8 (The weighted conditional KL divergence) *Under Agreement 9.1, let*

$$\Delta_t = (\Delta_{o;t}, \Delta_{p+;t}) = \textit{(o-innovations, innovations in the surplus p-space)},$$
$$f(\Delta_t|u_{o;t}, d(t-1)) = \prod_{i=1}^{\mathring{\Delta}} \mathcal{N}_{\Delta_{i;t}}(\theta_i'\psi_{i;t}, r_i),$$
$${}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d(t-1)) = \prod_{i=1}^{\mathring{\Delta}_o} \mathcal{N}_{\Delta_{i;t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\ {}^{\lfloor U}r_i\right).$$

The regression coefficients ${}^{\lfloor U}\theta_i$ of the user's ideal pdf ${}^{\lfloor U}f(\cdot)$ are complemented by zeros so that regression vectors $\psi_{i;t}$ of the corresponding factors coincide. Recall, that $\psi_{\mathring{\Delta};t} = [u_{o;t}', \phi_{t-1}']'$ and for $i \in \{0, \ldots, \mathring{\Delta}-1\}$

$$\psi_{i;t} = [\Delta'_{(i+1)\cdots\mathring{\Delta};t}, u'_{o;t}, \phi'_{t-1}]' \equiv [\Delta'_{(i+1)\cdots\mathring{\Delta};t}, \psi'_{\mathring{\Delta};t}]' \equiv [\Delta_{(i+1);t}, \psi'_{i+1;t}]'.$$

Let, moreover,

$$\gamma(\phi_t) \equiv \exp\left[-0.5(k_\gamma + \phi_t' L_\gamma D_\gamma L_\gamma' \phi_t)\right], \quad \textit{where} \tag{9.28}$$
$$\phi_t \equiv [d'_{t\cdots(t-\partial+1)}, 1]' \Rightarrow \psi_{0;t} \equiv \Psi_t \equiv [\Delta_t', u_{o;t}', \phi_{t-1}']',$$
$$L_\gamma \equiv \ \textit{a lower triangular matrix with unit diagonal}$$
$$D_\gamma \equiv \ \textit{a diagonal matrix with nonnegative diagonal entries.}$$

Then,

$$\omega_\gamma(u_{o;t}, \phi_{t-1}) \equiv \tag{9.29}$$
$$\equiv 2\int f(\Delta_t|u_{o;t}, \phi_{t-1}) \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, \phi_{t-1})}{\gamma(\phi_t)\,{}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, \phi_{t-1})}\right) d\Delta_t$$
$$= k_{\mathring{\Delta}} + \psi'_{\mathring{\Delta};t} L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}} \psi_{\mathring{\Delta};t}.$$

Lift $k_{\mathring{\Delta}}$ and kernel $L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}}$ of the KL divergence are found recursively

For $\quad i = 1, \ldots, \mathring{\Delta}$

$$\begin{aligned} L_i D_i L_i' &= {}^{\lfloor\psi}L_{i-1}\, {}^{\lfloor\psi}D_{i-1}\, {}^{\lfloor\psi}L'_{i-1} \\ &+ \left(\theta_i + {}^{\lfloor\Delta\psi}L_{i-1}\right) {}^{\lfloor\Delta}D_{i-1} \left(\theta_i + {}^{\lfloor\Delta\psi}L_{i-1}\right)' \\ &+ \chi\left(i \le \mathring{\Delta}_o\right) \frac{\left(\theta_i - {}^{\lfloor U}\theta_i\right)\left(\theta_i - {}^{\lfloor U}\theta_i\right)'}{{}^{\lfloor U}r_i} \\ k_i &= k_{i-1} + {}^{\lfloor\Delta}D_{i-1} r_i + \chi\left(i \le \mathring{\Delta}_o\right)\left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_i}\right) + \frac{r_i}{{}^{\lfloor U}r_i}\right] \end{aligned}$$

end $\quad$ *of the cycle over i*

$$\begin{aligned} k_0 &= -\mathring{\Delta}_o + k_\gamma,\ L_0 D_0 L_0' = \mathcal{K} L_\gamma D_\gamma L_\gamma' \mathcal{K}' \\ \mathcal{K}' &\equiv \begin{bmatrix} I_{\mathring{d}(\partial-1)} & 0 & 0 \\ 0 & 0_{1,\mathring{d}} & 1 \end{bmatrix} \qquad (9.30) \\ L_i &\equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor\Delta\psi}L_i & {}^{\lfloor\psi}L_i \end{bmatrix},\ D_i \equiv \mathrm{diag}\left[{}^{\lfloor\Delta}D_i,\ {}^{\lfloor\psi}D_i\right], \end{aligned}$$

where ${}^{\lfloor\Delta}D_i$ is scalar and $\mathring{D}_{i+1} = \mathring{D}_i - 1$.

The updating of the LDL' (!) decomposition can be made by a double, if the set indicator $\chi(\cdot) = 1$, or single, if $\chi(\cdot) = 0$, use of Algorithm 8.2.

Proof. The function $\gamma(\phi_t)$ just adds nontrivial initial conditions. The matrix $\mathcal{K}$ extends the quadratic form in ϕ_t to the quadratic form in $\Psi_t \equiv \psi_{0;t}$. Otherwise, the proof of Proposition 9.7 can be copied. □

The role of the matrix $\mathcal{K}$ (9.30) that expresses the vector ϕ_t as a function of $\Psi_t \equiv \psi_{0;t}$ is more visible if we distinguish the entries corresponding to the position of the unit in the phase form of the state vector; Agreement 9.1.

Let $L_\gamma \equiv \begin{bmatrix} {}^{\lfloor\phi 0}L_\gamma & 0 \\ {}^{\lfloor\phi 1}L'_\gamma & 1 \end{bmatrix},\ D_\gamma = \mathrm{diag}\left[{}^{\lfloor\phi 0}D_\gamma,\ {}^{\lfloor 1}D_\gamma\right],\ {}^{\lfloor 1}D_\gamma$ is scalar. (9.31)

Then, $L'_\gamma \mathcal{K}' = \begin{bmatrix} {}^{\lfloor\phi 0}L'_\gamma & 0 & {}^{\lfloor\phi 1}L_\gamma \\ 0 & 0_{1,\mathring{d}} & 1 \end{bmatrix}$ and

$$\mathcal{K} L_\gamma D_\gamma L'_\gamma \mathcal{K}' = \begin{bmatrix} {}^{\lfloor\phi 0}L_\gamma\, {}^{\lfloor\phi 0}D_\gamma\, {}^{\lfloor\phi 0'}L_\gamma & 0 & {}^{\lfloor\phi 0}D_\gamma\, {}^{\lfloor\phi 1}L_\gamma \\ 0 & 0_{\mathring{d},\mathring{d}} & 0_{\mathring{d},1} \\ {}^{\lfloor\phi 1}L'_\gamma\, {}^{\lfloor\phi 0}L_\gamma & 0_{1,\mathring{d}} & {}^{\lfloor\phi 1}L'_\gamma\, {}^{\lfloor\phi 0}D_\gamma\, {}^{\lfloor\phi 1}L_\gamma + {}^{\lfloor 1}D_\gamma \end{bmatrix}.$$

It is straightforward to verify that the following algorithm provides LDL' decomposition of this kernel.

Algorithm 9.7 (Extension of kernel $\phi_t' L_\gamma D_\gamma L'_\gamma \phi_t \to \psi'_{0;t} L_0 D_0 L_0' \psi_{0;t}$)

$D_0 = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor\phi 0}D_\gamma\right], 0_{1,\mathring{d}},\ {}^{\lfloor 1}D_\gamma\right],\ L_0 = \begin{bmatrix} {}^{\lfloor\phi 0}L_\gamma & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor\phi 1}L'_\gamma & 0 & 1 \end{bmatrix},\ {}^{\lfloor 1}D_\gamma$ *is scalar.*

The following auxiliary proposition is useful in the industrial and simultaneous designs operating at the factor level.

Proposition 9.9 (Normal expectation of exp**(quadratic form))** *Let* $\psi'_{\mathring{\Delta}} = [u', \phi']$ *and* $f(u|\phi) = \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{i-\mathring{\Delta}}}(\theta_i' \psi_i, r_i) \equiv \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{\Delta_i}(\theta_i' \psi_i, r_i)$ *with known parameters and* $\psi_i' = [u'_{i+1-\mathring{\Delta}}, \psi'_{i+1}]$, $\mathring{\Delta} \le i < \mathring{d}$, $\psi_{\mathring{d}} = \phi$. *Let us consider the lifted quadratic positive semi-definite form* $k_{\mathring{\Delta}} + \psi'_{\mathring{\Delta}} L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}} \psi_{\mathring{\Delta}}$. *The lift* $k_{\mathring{\Delta}}$ *is nonnegative,* $L_{\mathring{\Delta}}$ *is a lower triangular matrix with unit diagonal and* $D_{\mathring{\Delta}}$ *is a diagonal matrix with nonnegative diagonal entries. Then,*

$$\begin{aligned}
&\mathcal{E}\left[\exp\left\{-0.5\left(k_{\mathring{\Delta}} + \psi'_{\mathring{\Delta}} L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}} \psi_{\mathring{\Delta}}\right)\right\} |\phi\right] \\
&\equiv \int f(u|\phi) \exp\left\{-0.5(k_{\mathring{\Delta}} + \psi'_{\mathring{\Delta}} L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}} \psi_{\mathring{\Delta}})\right\} du \\
&= \exp\left[-0.5(k_{\mathring{d}} + \phi' L_{\mathring{d}} D_{\mathring{d}} L'_{\mathring{d}} \phi)\right].
\end{aligned}$$

The lift and kernel are found recursively

For $i = \mathring{\Delta} + 1, \ldots, \mathring{d}$

$$\tilde{L}_i \tilde{D}_i \tilde{L}_i' = L_{i-1} D_{i-1} L'_{i-1} + [-1, \theta_i']' r_i^{-1} [-1, \theta_i'], \tag{9.32}$$

$$\tilde{L}_i = \begin{bmatrix} 1 & 0 \\ {}^{\lfloor 1\psi}\tilde{L}_i & L_i \end{bmatrix}, \ \tilde{D}_i = \operatorname{diag}\left[{}^{\lfloor 1}\tilde{D}_i, D_i\right], \ {}^{\lfloor 1}\tilde{D}_i \textit{ is scalar},$$

$$k_i = k_{i-1} + \ln\left(r_i \, {}^{\lfloor 1}\tilde{D}_i\right), \ \mathring{D}_i = \mathring{D}_{i-1} - 1.$$

end *of the cycle over* i

The updating of the LDL' *(!) decomposition can be done by Algorithm 8.2.*

Proof. The the chain rule for expectation, Proposition 2.6, implies the need to evaluate, for $i = \mathring{\Delta} + 1, \ldots, \mathring{d}$,

$$\begin{aligned}
&\int (2\pi r_i)^{-0.5} \\
&\quad \times \exp\left\{-\frac{k_{i-1} + [u, \psi_i'] \left([-1, \theta_i']' r_i^{-1} [-1, \theta_i'] + L_{i-1} D_{i-1} L'_{i-1}\right) [u, \psi_i']'}{2}\right\} du \\
&\equiv \int (2\pi r_i)^{-0.5} \exp\left\{-0.5\left(k_{i-1} + [u, \psi_i'] \tilde{L}_i \tilde{D}_i \tilde{L}_i' [u, \psi_i']'\right)\right\} du \\
&= \exp\left[-0.5\left(k_{i-1} + \ln\left(r_i \, {}^{\lfloor 1}\tilde{D}_i\right) + \psi_i' L_i D_i L_i' \psi_i\right)\right].
\end{aligned}$$

It remains to recall that $\psi_{\mathring{d}} \equiv \phi$. □

Within the considered framework, the evaluation of the KL divergences reduces to manipulations with expected quadratic forms. In the computationally advantageous factorized version, the correct way of computing the second moment for mixture models can easily be overlooked. This observation and applicability in special design versions makes us evaluate this moment on the component basis.

Proposition 9.10 (Expected quadratic form: matrix components) *Let L be a lower triangular matrix with unit diagonal and D be a diagonal matrix with nonnegative entries. Let the normal mixture with matrix components describe the evolution of the state vector in the phase form. Then, it holds*

$$\mathcal{E}[\phi_t' L D L' \phi_t | \psi_t, \Theta] = \sum_{c \in c^*} \left[\psi_t' A_c' L D L' A_c \psi_t + \alpha_c \mathrm{tr}\left({}^{\lfloor \Delta}L \, {}^{\lfloor \Delta}D \, {}^{\lfloor \Delta}L' r_c \right) \right]$$

$$A_c \equiv \sqrt{\alpha_c} \begin{bmatrix} \theta_c' \\ \Lambda \end{bmatrix}, \quad \Lambda \equiv \begin{bmatrix} I_{\mathring{\psi}-\mathring{\Delta}-1} & 0_{\mathring{\psi}-\mathring{\Delta}-1, \mathring{\Delta}+1} & 0 \\ 0_{1,\mathring{\psi}-\mathring{\Delta}-1} & 0_{1,\mathring{\Delta}+1} & 1 \end{bmatrix}, \quad \textit{where} \tag{9.33}$$

$$L = \begin{bmatrix} {}^{\lfloor \Delta}L & 0 \\ {}^{\lfloor \Delta \psi}L & {}^{\lfloor \psi}L \end{bmatrix}, \ D = \begin{bmatrix} {}^{\lfloor \Delta}D & 0 \\ 0 & {}^{\lfloor \psi}D \end{bmatrix}, \ {}^{\lfloor \Delta}L, \ {}^{\lfloor \Delta}D \ \textit{are } (\mathring{\Delta}, \mathring{\Delta})\textit{-matrices.}$$

Proof. It exploits the phase form of the state, known moments of components and identities: $\mathcal{E}[xx'] = \mathcal{E}[x]\mathcal{E}[x'] + \mathrm{cov}(x)$, $\mathcal{E}[x'Qx] = \mathrm{tr}[Q\mathrm{cov}(x)]$ valid for any vector and any symmetric $(\mathring{x}, \mathring{x})$-kernel Q. □

Now we are ready to focus on the fact that the individual predictors forming the optimized pdf ${}^{\lfloor I}f(d(\mathring{t}))$ are finite mixtures. This makes evaluation of the KL divergence $\mathcal{D}\left({}^{\lfloor I}f || {}^{\lfloor U}f \right)$ and consequently the application of Proposition 2.11 impossible. For this reason, we follow the direction outlined in Chapter 7 and use the Jensen inequality (2.14) for finding an upper bound on the KL divergence. It gives us a chance to optimize at least this upper bound. Its structure is demonstrated on the matrix form of the involved normal pdfs.

Proposition 9.11 (J divergence of a mixture to ${}^{\lfloor U}f$) *Let us consider the joint pdf on observed data $d(\mathring{t}) \in d^*(\mathring{t})$*

$$f(d(\mathring{t})) \equiv \prod_{t \in t^*} \sum_{c \in c^*} \alpha_{c;t} \mathcal{N}_{d_t}(\theta_c' \phi_{t-1}, r_c),$$

where the <u>matrix</u> parameters θ_c, r_c of normal components as well as their possibly past-data-dependent probabilistic weights $\alpha_{c;t}$ are known. Let ${}^{\lfloor U}f(d(\mathring{t})) = \prod_{t \in t^*} \mathcal{N}_{d_t}\left({}^{\lfloor U}\theta' \phi_{t-1}, {}^{\lfloor U}r \right)$ be another known normal pdf. If need be, the matrix regression coefficients θ_c in all components as well as in ${}^{\lfloor U}\theta$ are complemented by zeros so that the state vector ϕ_{t-1} is common to all of them. Then, the following inequality holds

$$\mathcal{D}\left(f \middle|\middle| {}^{\lfloor U}f \right) \le \mathcal{J}\left(f \middle|\middle| {}^{\lfloor U}f \right) \equiv 0.5 \mathcal{E}\left\{ \sum_{t \in t^*} \alpha_t' \omega(\phi_{t-1}) \right\} \quad \textit{with}$$

$$\omega(\phi_{t-1}) \equiv \left[\omega(1, \phi_{t-1}), \ldots, \omega(\mathring{c}, \phi_{t-1}) \right]'$$

$$\omega(c, \phi_{t-1}) \equiv 2 \int f(d_t | d(t-1), c) \ln\left(\frac{f(d_t | d(t-1), c)}{{}^{\lfloor U}f(d_t | d(t-1))} \right) dd_t$$

$$
\begin{aligned}
&= k_c + \phi_{t-1}' L_c D_c L_c' \phi_{t-1}, \ \textit{where} \\
k_c &\equiv \ln\left| {}^{\lfloor U}r r_c^{-1} \right| - \mathring{d} + \operatorname{tr}\left[r_c \, {}^{\lfloor U}r^{-1} \right] \\
L_c D_c L_c' &\equiv \left(\theta_c - {}^{\lfloor U}\theta\right) {}^{\lfloor U}r^{-1} \left(\theta_c - {}^{\lfloor U}\theta\right)'. \qquad (9.34)
\end{aligned}
$$

The J divergence $\mathcal{J}\left(f \,||\, {}^{\lfloor U}f\right)$ *is nonnegative and equal to zero iff* $r_c = {}^{\lfloor U}r$ *and* $\theta_c'\phi_{t-1} = {}^{\lfloor U}\theta'\phi_{t-1}$ $\forall t \in t^*$ *and* $c \in c^*$ *for which* $\alpha_{c;t} > 0$.

Proof. It is implied by a direct combination of Propositions 7.5, 8.10. □

9.2 Design of the advising strategy

Here, design variants of the academic, industrial and simultaneous advisory systems are discussed in the special case of normal components and the normal user's ideal pdf.

9.2.1 Academic design

The recommended pointers to components c_t are the actions of the academic p-system. They are described by the optimized academic strategy $\left\{ {}^{\lfloor I}f(c_t|d(t-1)) \right\}_{t\in t^*}$. The strategy determines, together with the estimated model of the o-system, the ideal pdfs

$$
{}^{\lfloor I}f(d_t, c_t|d(t-1)) \equiv \prod_{i=1}^{\mathring{d}} \mathcal{N}_{d_{ic_t;t}}\left(\theta_{ic_t}'\psi_{ic_t;t}, r_{ic_t}\right) {}^{\lfloor I}f(c_t|d(t-1)) \ \text{ with } d_t \equiv \Delta_t. \qquad (9.35)
$$

For the adopted fully probabilistic design, we have to specify the user's ideal pdf. We assume its normality on the o-data $d_{o;t}^*$. This true user's ideal pdf is extended on d_t^* in the way discussed in Section 5.1.5. It is extended further on c_t^* by specifying the target pf ${}^{\lfloor U}f(c_t|d(t-1))$ for the academic advices, i.e., for the recommended pointers c_t.

Initially, we always evaluate dangerous components (see Agreement 5.9 and Section 9.1.1), and reduce the support of ${}^{\lfloor U}f(c_t|d(t-1))$ to the nondangerous components. The optimal design of the academic p-system is solved by this choice if the reduced set contains just a single component. This component is offered to the operator as the designed ideal pdf. Otherwise, the normal version of the optimal academic advising strategy described by Proposition 7.10 has to be searched for.

Proposition 9.12 (Academic design with the γ-bound) *Let us consider the academic design for the o-system described by the mixture with normal components (9.1) having the state ϕ in the phase form; Agreement 9.1. Lack* *of recognizable actions implies that the innovation $\varDelta_t$ and the data record d_t coincide.*

Let $d_t = (d_{o;t}, d_{p+;t}) =$ (o-data, surplus p-data) and the user's ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$ on $d^*(\mathring{t})$ be defined by

$$
\begin{aligned}
{}^{\lfloor U}f(d_t|d(t-1)) &\equiv \prod_{i_o \in i_o^*} \mathcal{N}_{d_{i;t}}\left({}^{\lfloor U}\theta_i' \psi_{i;t},\ {}^{\lfloor U}r_i \right) {}^{\lfloor I}f(d_{p+;t}|d(t-1)) \\
i_o^* &\equiv \{1, \ldots, \mathring{d}_o\} \subset i^* \equiv \{1, \ldots, \mathring{d}\}.
\end{aligned} \tag{9.36}
$$

The parameters $\Theta_{ic} = [\theta_{ic}, r_{ic}]$, $i \in i^*$, $c \in c^*$, of the mixture model as well those determining the normal user's ideal pdf ${}^{\lfloor U}\Theta \equiv \left\{ {}^{\lfloor U}\theta_i,\ {}^{\lfloor U}r_i \right\}_{i \in i^*}$ are assumed to be known. Regression coefficients θ_{ic}, ${}^{\lfloor U}\theta_i$ are complemented by zeros so that the corresponding factors of the user's ideal pdf and the mixture model have the common regression vectors $\psi_{i;t} \equiv \left[d_{i+1;t}, \psi_{i+1;t}' \right]' = \left[d_{(i+1)\cdots\mathring{d};t}', \phi_{t-1}' \right]'$, $i < \mathring{d}$, $\psi_{\mathring{d};t} \equiv \phi_{t-1}$.

The recommended pointers c_t are allowed to have nonzero values at most for indexes in c^* that point to nondangerous components; Agreement 5.9.

Let us search for the causal advising strategy $\{d^*(t-1) \to c_t \in c^*\}_{t \in t^*}$, that, at least approximately, minimizes the KL divergence of ${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ to the user's ideal pdf

$$
{}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t})) = \prod_{t \in t^*} {}^{\lfloor U}f(d_t|d(t-1))\, {}^{\lfloor U}f(c_t|d(t-1)).
$$

In this description, the definition (9.36) is used and the user's ideal pf on c_{t+1} is considered in the form

$$
{}^{\lfloor U}f(c_{t+1}|d(t)) \propto {}^{\lfloor U}f(c_{t+1}) \exp\left[-0.5\, {}^{\lfloor U}\omega(c_{t+1}, \phi_t) \right] \tag{9.37}
$$

$$
\begin{aligned}
&{}^{\lfloor U}f(c_{t+1}) \text{ is a probabilistic vector} \\
{}^{\lfloor U}\omega(c_{t+1}, \phi_t) &\equiv {}^{\lfloor U}k_{c_{t+1};t} + \phi_t'\, {}^{\lfloor U}L_{c_{t+1};t}\, {}^{\lfloor U}D_{c_{t+1};t}\, {}^{\lfloor U}L_{c_{t+1};t}' \phi_t.
\end{aligned} \tag{9.38}
$$

The fixed, data independent, part ${}^{\lfloor U}f(c_{t+1})$ serves for the specification of support on nondangerous components. The lifts ${}^{\lfloor U}k_{c;t} \geq 0$ and kernels

$$
{}^{\lfloor U}L_{c_{t+1};t}\, {}^{\lfloor U}D_{c_{t+1};t}\, {}^{\lfloor U}L_{c_{t+1};t}',
$$

determined by the lower triangular matrices with unit diagonal ${}^{\lfloor U}L_{c_{t+1};t}$ and the diagonal matrices ${}^{\lfloor U}D_{c_{t+1};t}$, are assumed to be data independent. Mostly, they are even time-invariant. The lifted quadratic form determines a preferable user strategy of selecting recommended pointers $c_{t+1} \in c^*$.

The LDL' decomposition of kernels used below are split

$$
L \equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix}, \quad D \equiv \operatorname{diag}\left[{}^{\lfloor d}D,\ {}^{\lfloor \psi}D \right], \quad \text{where } {}^{\lfloor d}D \text{ is scalar.}
$$

Let us consider the following strategy, $c \in c^*$,

$$
\begin{aligned}
&{}^{\lfloor I}f(c_t|\phi_{t-1}) \propto {}^{\lfloor U}f(c_t)\exp[-0.5\omega_\gamma(c_t,\phi_{t-1})], \quad \textit{where} \qquad (9.39)\\
&\omega_\gamma(c_t,\phi_{t-1}) \equiv k_{c_t;t-1} + \phi'_{t-1}L_{c_t;t-1}D_{c_t;t-1}L'_{c_t;t-1}\phi_{t-1}.
\end{aligned}
$$

The lifts $k_{c_t;t-1}$ and kernels $L_{c_t;t-1}D_{c_t;t-1}L'_{c_t;t-1}$, $c \in c^*$, are evaluated against the time course $t = \mathring{t}, \mathring{t}-1, \ldots, 1$, with the initial values $k_{c;\mathring{t}} = 0$, $L_{c;\mathring{t}} = I_{\mathring{\phi}} =$ unit matrix, $D_{c;\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$.

```
For  $t = \mathring{t}, \ldots, 1$
  For  $c = 1, \ldots, \mathring{c}$
        Create the common initial conditions of recursions over factors
            $k_{0c} \equiv -\mathring{d}_o \equiv$ average lift.
            $L_{0c} = I_{\mathring{\Psi}},\ D_{0c} = 0_{\mathring{\Psi},\mathring{\Psi}} \equiv$ average kernel.
    For  $\tilde{c} = 1, \ldots, \mathring{c}$
            $\beta_{\tilde{c}} \propto {}^{\lfloor U}f(\tilde{c})\exp(-0.5k_{\tilde{c};t}),$    (9.40)
            $L_{0c}D_{0c}L'_{0c} \equiv L_{0c}D_{0c}L'_{0c} + \beta_{\tilde{c}}, \tilde{L}_{\tilde{c};t}\tilde{D}_{\tilde{c};t}\tilde{L}'_{\tilde{c};t}.$
            The kernel decomposition $\tilde{L}_{\tilde{c};t}$, $\tilde{D}_{\tilde{c};t}$ is specified using the split
            $L_{\tilde{c};t} \equiv \begin{bmatrix} {}^{\lfloor \psi 0}L_{\tilde{c};t} & 0 \\ {}^{\lfloor \psi 1}L'_{\tilde{c};t} & 1 \end{bmatrix}, \quad D_{\tilde{c};t} = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \psi 0}D_{\tilde{c};t}\right], {}^{\lfloor 1}D_{\tilde{c};t}\right],$
            ${}^{\lfloor 1}D_{\tilde{c};t}$ is scalar.
            $\tilde{L}_{\tilde{c};t} \equiv \begin{bmatrix} {}^{\lfloor \psi 0}L_{\tilde{c};t} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor \psi 1}L'_{\tilde{c};t} & 0 & 1 \end{bmatrix},$
            $\tilde{D}_{\tilde{c};t} = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \psi 0}D_{\tilde{c};t}\right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_{\tilde{c};t}\right].$
    end  of the cycle over $\tilde{c}$
    For  $i = 1, \ldots, \mathring{d}$
            $L_{ic}D_{ic}L'_{ic} = {}^{\lfloor \psi}L_{(i-1)c}\,{}^{\lfloor \psi}D_{(i-1)c}\,{}^{\lfloor \psi}L'_{(i-1)c}$
            $+ \left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right){}^{\lfloor d}D_{(i-1)c}\left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right)'$
            $+\chi\left(i \le \mathring{d}_o\right)\left(\theta_{ic} - {}^{\lfloor U}\theta_i\right){}^{\lfloor U}r_i^{-1}\left(\theta_{ic} - {}^{\lfloor U}\theta_i\right)',$
            $k_{ic} = k_{(i-1)c} + {}^{\lfloor d}D_{(i-1)c}r_{ic} + \chi\left(i \le \mathring{d}_o\right)\left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}}\right) + \frac{r_{ic}}{{}^{\lfloor U}r_i}\right].$
    end  of the cycle over $i$
            $L_{c;t-1}D_{c;t-1}L'_{c;t-1} \equiv L_{\mathring{d}c}D_{\mathring{d}c}L'_{\mathring{d}c} + {}^{\lfloor U}L_{c;t-1}\,{}^{\lfloor U}D_{c;t-1}\,{}^{\lfloor U}L'_{c;t-1}$
            $k_{c;t-1} \equiv k_{\mathring{d}c} + {}^{\lfloor U}k_{c;t-1}$
  end  of the cycle over $c$
end  of the cycle over $t$
```

Then, this strategy minimizes the KL divergence loss for the design horizon equal to one. For longer horizons, it minimizes the upper bound of the γ-type on the KL divergence (see Proposition 7.8) as it replaces the Bellman function

$$-\ln\left\{\sum_{c_{t+1}\in c^*} {}^{\lfloor U}f(c_{t+1})\exp[-0.5\omega_{\gamma}(c_{t+1},\phi_t)]\right\} \textit{ by the larger value} \tag{9.41}$$

$$\frac{1}{2}\sum_{c_{t+1}\in c^*}\beta_{c_{t+1}}\phi_t' L_{c_{t+1};t}D_{c_{t+1};t}L'_{c_{t+1};t}\phi_t + \textit{data and strategy independent term.}$$

Proof. Let us assume that the Bellman function is $-\ln(\gamma(\phi_t))$, with

$$\gamma(\phi_t) = \exp\left[-0.5\sum_{c_{t+1}\in c^*}\beta_{c_{t+1}}\phi_t' L_{c_{t+1};t}D_{c_{t+1};t}L'_{c_{t+1};t}\phi_t\right], \text{ where}$$
$$\beta_{c_{t+1}} \propto {}^{\lfloor U}f(c_{t+1})\exp[-0.5k_{c_{t+1};t}].$$

The kernels $L_{c_{t+1};t}D_{c_{t+1};t}L'_{c_{t+1};t}$ and the lifts $k_{c_{t+1};t}$ are conjectured to be independent of data.

This conjecture is true for $t=\mathring{t}$ with $k_{c_{\mathring{t}+1};\mathring{t}}=0,\ L_{c_{\mathring{t}+1};\mathring{t}}=I_{\mathring{\phi}},\ D_{c_{\mathring{t}+1};\mathring{t}}=0_{\mathring{\phi},\mathring{\phi}}$. We prove that this form is preserved during the backward induction for $t<\mathring{t}$, if the approximation (9.41) is used. It gives us also the algorithmic solution of the inspected problem.

First, we rewrite the assumed form of $\gamma(\phi_t)$ into the quadratic form in $\mathit{\Psi}_t \equiv \psi_{0;t}$, taking into account the phase form of the state $\phi_t = \mathcal{K}'\mathit{\Psi}_t$; see (9.30). This implies

$$\gamma(\phi_t) = \exp\left[-0.5\psi_{0;t}' L_0 D_0 L_0' \psi_{0;t}\right] \text{ with}$$
$$L_0 D_0 L_0' \equiv \sum_{c_{t+1}\in c^*}\beta_{c_{t+1}}\mathcal{K}L_{c_{t+1};t}D_{c_{t+1};t}L'_{c_{t+1};t}\mathcal{K}' \quad \textit{average kernel.}$$

Algorithm 9.7 implies the explicit expression of the kernel

$$L_0 D_0 L_0' \equiv \sum_{c_{t+1}\in c^*}\beta_{c_{t+1}}\tilde{L}_{c_{t+1};t}\tilde{D}_{c_{t+1};t}\tilde{L}'_{c_{t+1};t},$$
$$\tilde{L}_{c;t} \equiv \begin{bmatrix} {}^{\lfloor\psi 0}L_{c;t} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor\psi 1}L'_{c;t} & 0 & 1 \end{bmatrix}, \quad \tilde{D}_{c;t} = \operatorname{diag}\left[\operatorname{diag}\left[{}^{\lfloor\psi 0}D_{c;t}\right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_{c;t}\right].$$

The tth step of dynamic programming, with $\mathring{\mathit{\Delta}} \equiv \mathring{d}$, becomes

$$\min_{\left\{ {}^{\lfloor I}f(c_t|d(t-1))\right\}} \sum\nolimits_{c_t \in c^*} {}^{\lfloor I}f(c_t|d(t-1))$$

$$\times \left[0.5 \left(\tilde{\omega}_\gamma(c_t,\phi_{t-1}) + \underbrace{{}^{\lfloor U}\omega(c_t,\phi_{t-1})}_{(9.38)} \right) + \ln\left(\frac{{}^{\lfloor I}f(c_t|d(t-1))}{{}^{\lfloor U}f(c_t)} \right) \right]$$

$$\tilde{\omega}_\gamma(c_t,\phi_{t-1}) \quad \equiv \quad 2\int f(d_t|\phi_{t-1},c_t) \ln\left(\frac{f(d_{o;t}|d_{p+;t},\phi_{t-1},c_t)}{\gamma(\phi_t)\, {}^{\lfloor U}f(d_{o;t}|\phi_{t-1})} \right) dd_t.$$

The definition $\omega_\gamma(c_t,\phi_{t-1}) \equiv \tilde{\omega}_\gamma(c_t,\phi_{t-1}) + {}^{\lfloor U}\omega(c_t,\phi_{t-1})$ and use of the Proposition 9.8 imply

$$\begin{aligned}
\omega_\gamma(c_t,\phi_{t-1}) &= k_{\mathring{d}c_t;t-1} + {}^{\lfloor U}k_{c_t;t-1} + \phi'_{t-1} L_{c_t;t-1} D_{c_t;t-1} L'_{c_t;t-1}\phi_{t-1},\\
L_{c;t-1}D_{c;t-1}L'_{c;t-1} &\equiv L_{\mathring{d}c} D_{\mathring{d}c} L'_{\mathring{d}c} + {}^{\lfloor U}L_{c;t-1}\, {}^{\lfloor U}D_{c;t-1}\, {}^{\lfloor U}L'_{c;t-1},\\
L_{ic}D_{ic}L'_{ic} &= {}^{\lfloor \psi}L_{(i-1)c}\, {}^{\lfloor \psi}D_{(i-1)c}\, {}^{\lfloor \psi}L'_{(i-1)c}\\
&+ \left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right) {}^{\lfloor d}D_{(i-1)c} \left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right)'\\
&+ \chi\left(i \le \mathring{d}_o\right) \left(\theta_{ic} - {}^{\lfloor U}\theta_i\right) {}^{\lfloor U}r_i^{-1} \left(\theta_{ic} - {}^{\lfloor U}\theta_i\right)',\\
k_{ic} &= k_{(i-1)c} + {}^{\lfloor d}D_{(i-1)c} r_{ic} + \chi\left(i \le \mathring{d}_o\right) \left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}} \right) + \frac{r_{ic}}{{}^{\lfloor U}r_i} \right]
\end{aligned}$$

$i = 1,\ldots,\mathring{d}$, with the <u>common initial conditions</u>

$$L_{0c} = L_0,\ D_{0c} = D_0,\quad k_{0c} = -\mathring{d}_o + k_0.$$

The minimizing argument is

$$f(c_t|\phi_{t-1}) = {}^{\lfloor U}f(c_t) \exp[-0.5\omega_\gamma(c_t,\phi_{t-1})]/\tilde{\gamma}(\phi_{t-1}),$$

where $\tilde{\gamma}(\phi_{t-1})$ is normalizing factor, i.e.,

$$\begin{aligned}
\tilde{\gamma}(\phi_{t-1}) &= q \sum_{c_t \in c^*} \beta_{c_t} \exp\left[-0.5\phi_{t-1} L_{c_t;t-1} D_{c_t;t-1} L'_{c_t;t-1}\phi_{t-1}\right] \text{ with}\\
q &\equiv \sum_{c_t \in c^*} {}^{\lfloor U}f(c_t) \exp[-0.5 k_{c_t;t-1}].
\end{aligned}$$

The form of the reached minimum $-\ln(\tilde{\gamma}(\phi_{t-1}))$ differs from that assumed and cannot be practically used further on. The inequality between weighted arithmetic and geometric means implies that

$$\tilde{\gamma}(\phi_{t-1}) \ge q \exp\left[-0.5 \sum_{c_t \in c^*} \beta_{c_t} \phi'_{t-1} L_{c;t-1} D_{c;t-1} L'_{c;t-1} \phi_{t-1} \right] \equiv \gamma(\phi_{t-1}),$$

where the right-hand side has the same form as that assumed for $\gamma(\phi_t)$. The Bellman function $-\ln(\tilde{\gamma}(\phi))$ is decreasing function of the values $\tilde{\gamma}(\phi_t)$. Consequently, the adopted approximation (9.41), that replaces the function $\tilde{\gamma}(\phi_t)$

by the function $\gamma(\phi_t)$, bounds the minimized functional from above while preserving the advantageous analytical form. □

The proved proposition combined with the certainty-equivalence strategy gives the following algorithm.

Algorithm 9.8 (Fixed academic advising with the γ-bound)

Initial (offline) mode

- *Estimate the normal mixture model of the o-system with the state ϕ_t, Chapter 8, and use the point estimates of parameters in the definition of the individual components $\prod_{i=1}^{\mathring{d}} \mathcal{N}_{d_{ic}}(\theta'_{ic}\psi_{ic}, r_{ic})$, $c \in c^*$.*
- *Evaluate the steady-state behaviors of individual components; Section 9.1.1.*
- *Exclude dangerous components (Agreement 5.9) or stabilize them using Algorithm 9.3.*
- *Return to the learning phase if all components are dangerous.*
- *Take the nondangerous component as the ideal pdf offered to the operator and go to* Sequential mode *if $\mathring{c} = 1$.*
- *Specify the true user's ideal pdf on the response of the o-system*

$$^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) = \prod_{i=1}^{\mathring{d}_o} \mathcal{N}_{d_{i;t}}\left({}^{\lfloor U}\theta'_i\psi_{i;t},\ {}^{\lfloor U}r_i\right).$$

- *Specify <u>data and time invariant</u> part $^{\lfloor U}f(c_t)$ of the user's ideal pf on the recommended pointers. It is zero on dangerous components.*
- *Choose <u>data and time invariant lifts</u> $^{\lfloor U}k_{c_t}$ and <u>kernels</u> of the* user-specified KL divergence $^{\lfloor U}L_{c_t}$, $^{\lfloor U}D_{c_t}$. *It completes the definition of the user's ideal pf $^{\lfloor U}f(c_t|d(t-1))$; see (9.37).*
- *Select the length of the design horizon $\mathring{t} \geq 1$.*
- *Initialize the iterative mode by setting $L_{c;\mathring{t}} = I_{\mathring{\phi}}$, $D_{c;\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$, $k_{c;\mathring{t}} = 0$, $c \in c^*$.*

Iterative (offline) mode

- *Correct the quadratic forms, for $t = \mathring{t}, \mathring{t}-1, \ldots, 1$, according to the formulas (9.40).*
- *Take the terminal characteristics as those describing the approximate optimal* steady-state strategy*; Chapter 3,*

$$k_c \equiv k_{c;1},\ L_c \equiv L_{c;1},\ D_c \equiv D_{c;1},\ c \in c^*.$$

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Evaluate the values of the pf describing the optimal academic advising strategy*

$$^{\lfloor I}f(c_{t+1}|\phi_t) \propto {}^{\lfloor U}f(c_{t+1}) \exp\left[-0.5\left(k_{c_{t+1}} + \phi'_t L_{c_{t+1}} D_{c_{t+1}} L'_{c_{t+1}} \phi_t\right)\right].$$

3. *Present to the operator projections of the ideal pdf*

$$^{\lfloor I}f(d_{t+1}|\phi_t) = \sum_{c_{t+1}\in c^*} {}^{\lfloor I}f(c_{t+1}|\phi_t) \prod_{i=1}^{\mathring{d}} \mathcal{N}_{d_{ic_{t+1};t+1}} \left(\theta'_{ic_{t+1}} \psi_{ic_{t+1};t}, r_{ic_{t+1}}\right).$$

4. *Go to the beginning of* Sequential mode.

Remark(s) 9.5

1. *The solution is prepared for the factorized implementation, which is numerically safe.*
2. *The chosen form* $^{\lfloor U}f(c_t|d(t-1))$*, that coincides with that of the optimal advising strategy, allows us to stay within the computationally advantageous class while being able to*
   - *exclude a priori bad (dangerous) components; Agreement 5.9,*
   - *respect the "natural" requirement that advices should not be changed too quickly,*
   - *employ the IST strategy, Algorithm 7.5, in an adaptive implementation of the above strategy.*
3. *The redundant specification of probabilities* $^{\lfloor U}f(c)$ *and lifts* k_c *simplifies interpretation. The pf* $^{\lfloor U}f(c)$ *cares about the support of the desired recommended pointers. The lifts* k_c *distinguish data-invariant absolute values of individual probabilities. The* user-specified KL kernel*, given by* $^{\lfloor U}L_c$*,* $^{\lfloor U}D_c$*, prescribes data-dependent preferences among the recommended pointers.*
4. *All data-invariant parts in the conditional KL divergences have to be concentrated into the lift, including that occurring in the quadratic form written in terms of the state vectors with 1 as its entry. Ideally, the updating variant described by Algorithm 9.6 should be used.*
5. *The* grouped variant of advising*, Proposition 7.13, can be designed and used. Essentially, the initial conditions* k_{0c}, L_{0c}, D_{0c} *are constructed as described above only each nth step. In the intermediate recursions,* $k_{0c} = k_{c;t+1}$, $L_{0c} = L_{c;t+1}$, $D_{0c} = D_{c;t+1}$.
6. *The last two remarks apply for all advising algorithms discussed in this chapter.*

The online use of the fixed advisory system is extremely simple. It requires just evaluation of values of the normal-mixture predictor. Naturally, the quality of its advices heavily depends on the quality of the used mixture model. This makes us write down an adaptive version of the above algorithm. The increased complexity can be well counteracted by using the IST patch, Algorithm 7.5. The lifts $k_{c;t}$ and the kernels $L_{c;t}D_{c;t}L'_{c;t}$ of quadratic forms obtained at time t seem to be "natural" parameter ϑ needed: they are used as initial conditions for the design made at time $t+1$. For the IST strategy, the receding horizon $T = 1$ is expected to be mostly sufficient.

Let us write down the adaptive algorithm with receding-horizon, certainty-equivalence strategy complemented by the IST patch.

Algorithm 9.9 (Adaptive academic advising with the γ-bound)

Initial (offline) mode

- *Estimate the normal mixture model of the o-system with the state ϕ_t; Chapter 8.*
- *Evaluate the steady-state behaviors of individual components, Section 9.1.1.*
- *Exclude dangerous components, Agreement 5.9, or stabilize them using Algorithm 9.3.*
- *Return to the learning phase if all components are dangerous.*
- *Take the nondangerous component as the ideal pdf offered to the operator. Skip* Iterative mode *if $\mathring{c} = 1$.*
- *Specify the true user's ideal pdf on the response of the o-system*

$$ {}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) = \prod_{i=1}^{\mathring{d}_o} \mathcal{N}_{d_{i;t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\ {}^{\lfloor U}r_i\right). $$

- *Specify the data and time invariant part ${}^{\lfloor U}f(c_t)$ of the user's ideal pf on the recommended pointers. It is zero on dangerous components.*
- *Specify data invariant lifts ${}^{\lfloor U}k_{c_t;t-1}$ and kernels of the user-specified KL divergence ${}^{\lfloor U}L_{c_t;t-1}$, ${}^{\lfloor U}D_{c_t;t-1}$. It completes definition of the user's ideal pf (see 9.37)*

$$ {}^{\lfloor U}f(c_t|d(t-1)) \propto {}^{\lfloor U}f(c_t) \exp\left[-0.5\left(k_{c_t;t-1} + \phi_{t-1}' L_{c_t;t-1} D_{c_t;t-1} L_{c_t;t-1}' \phi_{t-1}\right)\right], $$

- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for t=1,2,...,*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update estimates of the model parameters; Section 8.5.*
3. *Initialize the iterative mode by setting $\tau = t + T$ and $L_{c;\tau} = I_{\mathring{\phi}}$, $D_{c;\tau} = 0_{\mathring{\phi},\mathring{\phi}}$, $k_{c;\tau} = 0$. The initialization of L, D, k is skipped for $t > 1$ if the IST strategy is used.*

Iterative mode

```
Correct the following quadratic forms using the current point estimates of
the factor parameters

For   $\tau = t + T, \ldots, t + 1$
   For   $c = 1, \ldots, \mathring{c}$
         Create the common initial conditions of recursions over factors
               $k_{0c} \equiv -\mathring{d}_o \equiv$ average lift,
               $L_{0c} = I_{\mathring{\Psi}}$,  $D_{0c} = 0_{\mathring{\Psi},\mathring{\Psi}} \equiv$ average kernel.
```

```
        For $\tilde{c} = 1, \ldots, \mathring{c}$
            $\beta_{\tilde{c}} \propto {}^{\lfloor U}f(\tilde{c}) \exp(-0.5 k_{\tilde{c};\tau})$,
            $L_{0c} D_{0c} L'_{0c} \equiv L_{0c} D_{0c} L'_{0c} + \beta_{\tilde{c}} \tilde{L}_{\tilde{c};\tau} \tilde{D}_{\tilde{c};\tau} \tilde{L}'_{\tilde{c};\tau}$.
            Terms $\tilde{L}_{\tilde{c};\tau}$, $\tilde{D}_{\tilde{c};\tau}$ are specified using the split
            $L_{\tilde{c};\tau} \equiv \begin{bmatrix} {}^{\lfloor \psi 0}L_{\tilde{c};\tau} & 0 \\ {}^{\lfloor \psi 1}L'_{\tilde{c};\tau} & 1 \end{bmatrix}$,  $D_{\tilde{c};\tau} = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \psi 0}D_{\tilde{c};\tau}\right], {}^{\lfloor 1}D_{\tilde{c};\tau}\right]$,
            ${}^{\lfloor 1}D_{\tilde{c};\tau}$ is scalar.
            $\tilde{L}_{\tilde{c};\tau} \equiv \begin{bmatrix} {}^{\lfloor \psi 0}L_{\tilde{c};\tau} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor \psi 1}L'_{\tilde{c};\tau} & 0 & 1 \end{bmatrix}$,
            $\tilde{D}_{\tilde{c};\tau} = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \psi 0}D_{\tilde{c};\tau}\right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_{\tilde{c};\tau}\right]$.
        end  of the cycle over $\tilde{c}$
        For $i = 1, \ldots, \mathring{d}$
            $L_{ic} D_{ic} L'_{ic} = {}^{\lfloor \psi}L_{(i-1)c} \, {}^{\lfloor \psi}D_{(i-1)c} \, {}^{\lfloor \psi}L'_{(i-1)c}$
            $+ \left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right) {}^{\lfloor d}D_{(i-1)c} \left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right)'$
            $+ \chi\left(i \le \mathring{d}_o\right) \left(\theta_{ic} - {}^{\lfloor U}\theta_i\right) {}^{\lfloor U}r_i^{-1} \left(\theta_{ic} - {}^{\lfloor U}\theta_i\right)'$
            $k_{ic} = k_{(i-1)c} + {}^{\lfloor d}D_{(i-1)c} r_{ic} + \chi\left(i \le \mathring{d}_o\right) \left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}}\right) + \frac{r_{ic}}{{}^{\lfloor U}r_i}\right]$.
        end  of the cycle over $i$
            $L_{c;\tau-1} D_{c;\tau-1} L'_{c;\tau-1} \equiv L_{\mathring{d}c} D_{\mathring{d}c} L'_{\mathring{d}c} + {}^{\lfloor U}L_{c;\tau-1} \, {}^{\lfloor U}D_{c;\tau-1} \, {}^{\lfloor U}L'_{c;\tau-1}$
            $k_{c;\tau-1} \equiv k_{\mathring{d}c} + {}^{\lfloor U}k_{c;\tau-1}$.
    end  of the cycle over $c$
end  of the cycle over $\tau$
```

4. *Evaluate the pf describing the optimal academic advising strategy*

$$ {}^{\lfloor I}f(c_{t+1}|\phi_t) \propto {}^{\lfloor U}f(c_{t+1}) \exp\left[-0.5(k_{c_{t+1};t} + \phi'_t L_{c_{t+1};t} D_{c_{t+1};t} L'_{c_{t+1};t} \phi_t)\right]. $$

5. *Present to the operator projections of the ideal pdf*

$$ {}^{\lfloor I}f(d_{t+1}|\phi_t) = \sum_{c_{t+1} \in c^*} {}^{\lfloor I}f(c_{t+1}|\phi_t) \prod_{i \in i^*} \mathcal{N}_{d_{ic_{t+1};t+1}} \left(\theta'_{ic_{t+1}} \psi_{ic_{t+1};t}, r_{ic_{t+1}}\right). $$

6. *Go to the beginning of* Sequential mode*.*

For completeness, let us write down a version of this algorithm that selects the most probable recommended pointer. Within the iterative evaluations, the maximizer ${}^{\lfloor I}c_{\tau+1}$ of ${}^{\lfloor I}f(c_{\tau+1}|\phi_\tau)$ is selected under simplifying assumption that $\phi_\tau = \phi_t$. This assumption allows us to preserve computational feasibility. It brings no additional approximation when using the IST strategy with the horizon $T = 1$.

Algorithm 9.10 (Adaptive, most probable, academic advising)

Initial (offline) mode

- *Estimate the normal mixture model of the o-system with the state ϕ_t; Chapter 8.*
- *Evaluate the steady-state behaviors of individual components; Section 9.1.1.*
- *Exclude dangerous components, Agreement 5.9, or stabilize them using Algorithm 9.3.*
- *Return to the learning phase if all components are dangerous.*
- *Take the nondangerous component as the ideal pdf offered to the operator. Skip* Iterative mode *if $\mathring{c} = 1$.*
- *Specify the true user's ideal pdf on the response of the o-system*

$$
{}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) = \prod_{i=1}^{\mathring{d}_o} \mathcal{N}_{d_{i;t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\ {}^{\lfloor U}r_i\right).
$$

- *Select the length of the receding horizon $T \geq 1$.*
- *Select randomly the pointer ${}^{\lfloor I}c_{T+1} \in c^*$.*

Sequential (online) mode*, running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update estimates of model parameters; Section 8.5.*
3. *Initialize the iterative mode by setting $\tau = t + T$ and $L_{c;\tau} = I_{\mathring{\phi}}$, $D_{c;\tau} = 0_{\mathring{\phi},\mathring{\phi}}$, $k_{c;\tau} = 0$. The initialization of L, D, k is skipped for $t > 1$ if the IST strategy is used.*

Iterative mode

Correct the following quadratic forms using the current point estimates of the factor parameters.

```
For   $\tau = t + T, \ldots, t + 1$
   For   $c = 1, \ldots, \mathring{c}$
         Create the common initial conditions of recursions over factors
            $L_{0c}D_{0c}L_{0c}' \equiv \tilde{L}_{\lfloor I_{c_{\tau+1};\tau}}\tilde{D}_{\lfloor I_{c_{\tau};\tau}}\tilde{L}'_{\lfloor I_{c_{\tau+1};\tau}},$
            $k_{0c} \equiv -\mathring{d}_o + k_{\lfloor I_{c_{\tau+1};\tau}},$   where
            $\tilde{L}_{\lfloor I_{c_{\tau+1};\tau}},\ \tilde{D}_{\lfloor I_{c_{\tau+1};\tau}}$ are specified using the split
            $L_{\lfloor I_{c_{\tau+1};\tau}} \equiv \begin{bmatrix} {}^{\lfloor \psi 0}L_{\lfloor I_{c_{\tau+1};\tau}} & 0 \\ {}^{\lfloor \psi 1}L'_{\lfloor I_{c_{\tau+1};\tau}} & 1 \end{bmatrix},$
```

```
        $D_{{}^{\lfloor I}c_{\tau+1};\tau} = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \psi 0}D_{{}^{\lfloor I}c_{\tau+1};\tau}\right], {}^{\lfloor 1}D_{{}^{\lfloor I}c_{\tau+1};\tau}\right]$,
        ${}^{\lfloor 1}D_{{}^{\lfloor I}c_{\tau+1};\tau}$ is scalar.
        $\tilde{L}_{{}^{\lfloor I}c_{\tau+1};\tau} \equiv \begin{bmatrix} {}^{\lfloor \psi 0}L_{{}^{\lfloor I}c_{\tau+1};\tau} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor \psi 1}L'_{{}^{\lfloor I}c_{\tau+1};\tau} & 0 & 1 \end{bmatrix}$,
        $\tilde{D}_{{}^{\lfloor I}c_{\tau+1};\tau} = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \psi 0}D_{{}^{\lfloor I}c_{\tau+1};\tau}\right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_{{}^{\lfloor I}c_{\tau+1};\tau}\right]$.
    For $i = 1, \ldots, \mathring{d}$
        $L_{ic}D_{ic}L'_{ic} = {}^{\lfloor \psi}L_{(i-1)c}\, {}^{\lfloor \psi}D_{(i-1)c}\, {}^{\lfloor \psi}L'_{(i-1)c}$
        $+ \left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right) {}^{\lfloor d}D_{(i-1)c} \left(\theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c}\right)'$
        $+\chi\left(i \leq \mathring{d}_o\right)\left(\theta_{ic} - {}^{\lfloor U}\theta_i\right) {}^{\lfloor U}r_i^{-1} \left(\theta_{ic} - {}^{\lfloor U}\theta_i\right)'$,
        $k_{ic} = k_{(i-1)c} + {}^{\lfloor d}D_{(i-1)c} r_{ic} + \chi\left(i \leq \mathring{d}_o\right)\left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}}\right) + \frac{r_{ic}}{{}^{\lfloor U}r_i}\right]$.
    end of the cycle over $i$
        $L_{c;\tau-1}D_{c;\tau-1}L'_{c;\tau-1} \equiv L_{\mathring{d}c}D_{\mathring{d}c}L'_{\mathring{d}c} + {}^{\lfloor U}L_{c;\tau-1}\, {}^{\lfloor U}D_{c;\tau-1}\, {}^{\lfloor U}L'_{c;\tau-1}$
        $k_{c;\tau-1} \equiv k_{\mathring{d}c} + {}^{\lfloor U}k_{c;\tau-1}$.
  end of the cycle over $c$
        Select ${}^{\lfloor I}c_\tau \equiv \mathrm{Arg}\min_{c \in c^*}\left(k_{c;\tau} + \phi'_t L_{c;\tau}D_{c;\tau}L'_{c;\tau}\phi_t\right)$
        using the simplifying assumption $\phi_\tau = \phi_t$.
end of the cycle over $\tau$
```

4. *Present to the operator projections of the ideal pdf*

$$
{}^{\lfloor I}f(d_{t+1}|\phi_t) = \prod_{i=1}^{\mathring{d}} \mathcal{N}_{d_{i\,{}^{\lfloor I}c_{t+1};t+1}}\left(\theta'_{i\,{}^{\lfloor I}c_{t+1}}\psi_{i\,{}^{\lfloor I}c_{t+1};t+1}, r_{i\,{}^{\lfloor I}c_{t+1}}\right).
$$

5. *Go to the beginning of* Sequential mode*.*

Remark(s) 9.6

1. *In a former formulation tried, the probabilities α, determining component weights, Agreement 5.4, were taken as academic actions. The corresponding design was much more complex. Experiments indicated, however, reasonable properties of such a strategy. Thus, it might be useful to inspect this direction in future.*
2. *It is worth stressing that the selected strategy tries to minimize not only the Jensen upper bound on the KL divergence as it was needed for the previous variant, but also its γ-bound.*
3. *The choice of the user-specified KL divergence determining ${}^{\lfloor U}f(c_{t+1}|\phi_t)$ influences substantially properties of the designed strategy. Mostly, we rely*

on the discussion presented in Section 7.2.2. Specifically, grouped versions of algorithms are implemented; see Proposition 7.13 and Remark 5.

9.2.2 Industrial design

The industrial design has to be used whenever component weights have objective meaning and cannot be influenced by the operator. We address it using ideas of Proposition 7.15 proved in Section 7.2.3. It should be stressed that the more effective and simpler simultaneous design should be used (see Section 9.2.3) whenever the component weights are under the operator control.

The normal components with explicitly marked recognizable actions

$$\left\{ f(\mathit{\Delta}_t|u_{o;t},d(t-1),c)\, f(u_{o;t}|d(t-1),c) \equiv \prod_{i=1}^{\mathring{\mathit{\Delta}}} \mathcal{N}_{\mathit{\Delta}_{i;t}}\left(\theta'_{ic}\psi_{ic;t},r_{ic}\right) \times \prod_{i=\mathring{\mathit{\Delta}}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\mathit{\Delta}});t}}\left(\theta'_{ic}\psi_{ic;t},r_{ic}\right) \right\}_{t\in t^*,c\in c^*} \tag{9.42}$$

and their weights $\{\alpha_c\}_{c\in c^*}$ are assumed to be known (well-estimated). Here, we use the convention that the data record is ordered as in previous sections $d_t = (\mathit{\Delta}_{o;t}, \mathit{\Delta}_{p+;t}, u_{o;t})$.

The following extension (see Section 5.1.5) of the true user's ideal pdf is considered ${}^{\lfloor U}f(d(\mathring{t})) \equiv$

$$\begin{aligned}
{}^{\lfloor U}f(d(\mathring{t})) &\equiv \\
&\prod_{t\in t^*} {}^{\lfloor U}f(\mathit{\Delta}_{o;t}|u_{o;t},d_o(t-1))\, {}^{\lfloor I}f(\mathit{\Delta}_{p+;t}|u_{o;t},d(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1))
\end{aligned}$$

$$\begin{aligned}
{}^{\lfloor U}f(\mathit{\Delta}_{o;t}|u_{o;t},d_o(t-1)) &\equiv \prod_{i=1}^{\mathring{\mathit{\Delta}}_o} \mathcal{N}_{\mathit{\Delta}_{i;t}}\left({}^{\lfloor U}\theta'_i\psi_{i;t},\ {}^{\lfloor U}r_i\right),\ \mathit{\Delta}_{io;t} = \mathit{\Delta}_{i;t} \text{ for } i \le \mathring{\mathit{\Delta}}_o, \\
{}^{\lfloor U}f(u_{o;t}|d_o(t-1)) &\equiv \prod_{i=\mathring{\mathit{\Delta}}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\mathit{\Delta}});t}}\left({}^{\lfloor U}\theta'_i\psi_{i;t},\ {}^{\lfloor U}r_i\right).
\end{aligned} \tag{9.43}$$

The upper bound on the KL divergence, serving as the loss whose expectation is optimized, reads

$$\begin{aligned}
Z(d(\mathring{t})) &\equiv \sum_{t\in t^*} \int {}^{\lfloor I}f(u_{o;t}|d(t-1)) \\
&\quad \times \left[\ln\left(\frac{{}^{\lfloor I}f(u_{o;t}|d(t-1))}{{}^{\lfloor U}f(u_{o;t}|d(t-1))}\right) + \omega(u_{o;t},d(t-1))\right] du_{o;t} \\
\omega(u_{o;t},d(t-1)) &\equiv \sum_{c\in c^*} f(c|u_{o;t},d(t-1))\omega(c,u_{o;t},d(t-1))
\end{aligned}$$

$$\omega(c, u_{o;t}, d(t-1)) \equiv \ln\left(\frac{f(c|u_{o;t}, d(t-1))}{\alpha_c}\right) \tag{9.44}$$
$$+ \int f(\Delta_t|u_{o;t}, d(t-1), c) \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, d(t-1), c)}{{}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1))}\right) d\Delta_t.$$

Its optimization calls for the evaluation of the weights

$$f(c|u_{o;t}, d(t-1)) \equiv \frac{\alpha_c f(u_{o;t}|d(t-1), c)}{\sum_{c\in c^*} \alpha_c f(u_{o;t}|d(t-1), c)}. \tag{9.45}$$

For the considered normal mixture, they have to be approximated. It seems to be reasonable to approximate them by α_c within the design horizon and to use the values (9.45) in the final step only. The following proposition applies Proposition 7.15 in conjunction with this approximation.

Proposition 9.13 (Industrial design with the bound (7.22)) *Let the joint pdf*

$$\begin{aligned}&{}^{\lfloor I}f(\Delta(\mathring{t}), u_o(\mathring{t}))\\ &\equiv \prod_{t\in t^*} \frac{\sum_{c\in c^*} \alpha_c f(\Delta_t|u_{o;t}, d(t-1), c) f(u_{o;t}|d(t-1), c)}{\sum_{c\in c^*} \alpha_c f(u_{o;t}|d(t-1), c)}\, {}^{\lfloor I}f(u_{o;t}|d(t-1)),\end{aligned}$$

with components (9.43), be determined by the optional industrial advising strategy described by pdfs $\left\{ {}^{\lfloor I}f(u_{o;t}|d(t-1))\right\}_{t\in t^*}$.

Let us search for the strategy approximately minimizing the expected value of the loss $Z(d(\mathring{t}))$ *(9.44). It is the upper bound on the KL divergence* $\mathcal{D}\left({}^{\lfloor I}f|| {}^{\lfloor U}f\right)$ *to the user's ideal pdf (9.43). Moreover, let us assume that* $f(c|u_{o;t}, d(t-1)) \approx \alpha_c$*, cf. (9.45), for* $t > 1$.

In description of the strategy, we use the definition (9.30)

$$\mathcal{K}' \equiv \begin{bmatrix} I_{\mathring{d}(\partial-1)} & 0 & 0 \\ 0 & 0_{1,\mathring{d}} & 1 \end{bmatrix},\quad L_t = \begin{bmatrix} {}^{\lfloor \phi 0}L_t & 0 \\ {}^{\lfloor \phi 1}L_t & 1 \end{bmatrix},\quad D_t = \mathrm{diag}\left[{}^{\lfloor \phi 0}D_t,\ {}^{\lfloor 1}D_t\right].$$

${}^{\lfloor 1}D_t$ *is scalar. We also use the following split of matrices defining LDL′ decompositions of involved kernels*

$$L_{ic} \equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor \Delta\psi}L_{ic} & {}^{\lfloor \psi}L_{ic} \end{bmatrix},\ D_{ic} \equiv \mathrm{diag}\left[{}^{\lfloor \Delta}D_{ic},\ {}^{\lfloor \psi}D_{ic}\right],\quad \textit{where}$$
$${}^{\lfloor \Delta}D_{ic} \textit{ is scalar and } \mathring{D}_{(i+1)c} = \mathring{D}_{ic} - 1.$$

Then, with the introduced notation, the specialization of the approximately optimal strategy described in Proposition 7.15 becomes

$${}^{\lfloor I}f(u_{o;t}|d(t-1)) = \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta});t}}\left({}^{\lfloor I}\theta'_{i;t-1}\psi_{i;t},\ {}^{\lfloor I}r_{i;t-1}\right). \tag{9.46}$$

The regression coefficients ${}^{\lfloor I}\theta_{i;t-1}$ *and variance* ${}^{\lfloor I}r_{i;t-1}$ *are generated by the following algorithm.*

```
        Set $L_{\mathring{t}} = I_{\mathring{\phi}},\ D_{\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$.        (9.47)
For   $t = \mathring{t},\ldots,1$
        $L_{\mathring{\Delta}} = I_{\mathring{\psi}}, D_{\mathring{\Delta}} = 0_{\mathring{\psi},\mathring{\psi}}$
  For   $c = 1,\ldots,\mathring{c}$
        $L_{0c} \equiv \begin{bmatrix} {}^{\lfloor \phi 0}L_t & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor \phi 1}L_t' & 0 & 1 \end{bmatrix}$
        $D_{0c} \equiv \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \phi 0}D_t\right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_t\right],$  ${}^{\lfloor 1}D_t$ is scalar.
    For   $i = 1,\ldots,\mathring{\Delta}$
        $L_{ic}D_{ic}L_{ic}' = {}^{\lfloor \psi}L_{(i-1)c}\, {}^{\lfloor \psi}D_{(i-1)c}\, {}^{\lfloor \psi}L_{(i-1)c}'$
        $+ \left(\theta_{ic} + {}^{\lfloor \Delta\psi}L_{(i-1)c}\right) {}^{\lfloor \Delta}D_{(i-1)c} \left(\theta_{ic} + {}^{\lfloor \Delta\psi}L_{(i-1)c}\right)'$
        $+\chi\left(i \le \mathring{\Delta}_o\right)\left(\theta_{ic} - {}^{\lfloor U}\theta_i\right) {}^{\lfloor U}r_i^{-1} \left(\theta_{ic} - {}^{\lfloor U}\theta_i\right)'.$
    end   of the cycle over $i$
        $L_{\mathring{\Delta}}D_{\mathring{\Delta}}L_{\mathring{\Delta}}' = L_{\mathring{\Delta}}D_{\mathring{\Delta}}L_{\mathring{\Delta}}' + \alpha_c L_{\mathring{\Delta}c}D_{\mathring{\Delta}c}L_{\mathring{\Delta}c}'$
  end   of the cycle over $c$
  For   $i = \mathring{\Delta}+1,\ldots,\mathring{d}$
        $\tilde{L}_i\tilde{D}_i\tilde{L}_i' = L_{i-1}D_{i-1}L_{i-1}' + \left[-1, {}^{\lfloor U}\theta_i'\right]' {}^{\lfloor U}r_i^{-1} \left[-1, {}^{\lfloor U}\theta_i'\right]$
        $\tilde{L}_i \equiv \begin{bmatrix} 1 & 0 \\ -\,{}^{\lfloor I}\theta_{i;t-1} & L_i \end{bmatrix}$
        $\tilde{D}_i \equiv \mathrm{diag}\left[{}^{\lfloor I}r_{i;t-1}^{-1}, \mathrm{diag}[D_i]\right],$   ${}^{\lfloor I}r_{i;t-1}$ is scalar.
  end   of the cycle over $i$
        $L_{t-1} \equiv L_{\mathring{d}},\ D_{t-1} \equiv D_{\mathring{d}}.$
end   of the cycle over $t$
```

The ideal pdf shown to the operator is

$$
{}^{\lfloor I}f(d_{o;t}|d(t-1)) \equiv \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{(i-\mathring{\Delta})o;t}}\left({}^{\lfloor I}\theta_{ic;t-1}'\psi_{i;t};\ {}^{\lfloor I}r_{ic;t-1}\right)
$$

$$
\times \sum_{c\in c^*} f(c|u_{o;t}, d(t-1)) \prod_{i=1}^{\mathring{\Delta}_o} \mathcal{N}_{\Delta_{i;t}}(\theta_{ic}'\psi_{i;t}; r_{ic}) \quad \textit{with}
$$

$$f(c|u_{o;t}, d(t-1)) \propto \alpha_c \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta});t}}(\theta'_{ic}\psi_{i;t}, r_{ic}). \tag{9.48}$$

Proof. The proof is based on Proposition 9.13. Let us assume that the Bellman function has the form $-\ln(\gamma(d(t))) = 0.5\phi'_t L_t D_t L'_t \phi_t$ with data independent kernel $L_t D_t L'_t$. It is true for $t = \mathring{t}$ for the claimed terminal conditions.

Specialization of (7.40) should confirm this assumption and provide the corresponding algorithm.

We assume that $f(c|u_{o;t}, d(t-1)) \approx \alpha_c$ so that the first term in the definition (7.40) of $\omega(c, u_{o;t}, d(t-1))$ is zero and

$$\begin{aligned}
&2\omega_\gamma(u_{o;t}, d(t-1)) \\
&\equiv \sum_{c\in c^*} \alpha_c \int f(\Delta_t|u_{o;t}, d(t-1), c) \times \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, d(t-1), c)}{\gamma(d(t))\, {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1))}\right) d\Delta_t \\
&= \sum_{c\in c^*} \alpha_c \left[k_{\mathring{\Delta}c} + \psi'_{\mathring{\Delta};t} L_{\mathring{\Delta}c} D_{\mathring{\Delta}c} L'_{\mathring{\Delta}c} \psi_{\mathring{\Delta};t}\right] \equiv k_{\mathring{\Delta}} + \psi'_{\mathring{\Delta};t} L_{\mathring{\Delta}} D_{\mathring{\Delta}} L'_{\mathring{\Delta}} \psi_{\mathring{\Delta};t}.
\end{aligned}$$

The individual lifts $k_{\mathring{\Delta}c}$ and kernels $L_{\mathring{\Delta}c}, D_{\mathring{\Delta}c}$ in the above sum are evaluated using Proposition 9.8 that provides the values of the weighted conditional KL divergence. The respective recursions start from the common initial condition $k_{0c} = k_\gamma = 0$, $L_{0c}D_{0c}L'_{0c} = \mathcal{K}L_t D_t L'_t \mathcal{K}'$. Results are summed into a single shifted quadratic form. The resulting strategy is proportional to the product ${}^{\lfloor U}f(u_{o;t}|d(t-1))\exp[-0.5\omega(u_{o;t}, \phi_{t-1})]$ Thus, the quadratic form is increased by the quadratic form in the exponent of the normal user's ideal pdf on recognizable actions. The integration over $u_{o;t}$ can be performed recursively according to Proposition 9.9. It shows that $\gamma(d(t-1))$ preserves the assumed form. The average lift cancels in normalization and can be fixed at zero. Consequently, individual lifts need not be evaluated at all. It completes the full design step. The algorithm (9.47) puts these evaluations together. □

This proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm that adds just the learning part to the algorithm described.

Algorithm 9.11 (Industrial advising: (7.22), $f(c|u_{o;t}, d(t-1)) \approx \alpha_c$**)**

Initial (offline) mode

- *Estimate the mixture model of the o-system with the state ϕ_t, Chapter 8.*
- *Specify the true user's ideal pdf (9.43) on the response of the o-system.*
- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update estimates of the model parameters, Section 8.5, if you deal with the adaptive advisory system.*

3. *Initialize the iterative mode by setting $\tau = t + T$ and $L_\tau = I$, $D_\tau = 0$. Omit the initialization of L_τ, D_τ if $t > 0$ and the IST strategy is used.*
   Iterative mode
   - *Apply the algorithm given in Proposition 9.13 while replacing t by τ and stopping at $\tau = t + 1$.*
4. *Present to the operator projections of the ideal pdf (9.48) with pdfs*

$$f(\varDelta_{o;t}|u_{o;t+1}, \phi_t, c), \;\; f(u_{o;t+1}|\phi_t, c)$$

derived from the cth learned mixture component.

5. *Go to the beginning of* Sequential mode*.*

Remark(s) 9.7
A grouped version of the industrial design is implemented; see Proposition 7.13 and Remark 5.

Problem 9.1 (Improvements of industrial design) *The specific normal form of the adopted models could and should be used for constructing improved approximations both of the upper bound on the KL divergence and on $f(c|u_{o;t}, d(t-1))$. Preliminary inspections have confirmed that this research direction is a promising one.*

An application of the simultaneous design with ${}^{\lfloor U}f(c_t|d(t-1))$ manipulated so that ${}^{\lfloor I}f(c_t|d(t-1)) = \alpha_{c_t}$ is worth considering as well.

At the end of this section, we attach a version of industrial design with a quadratic criterion. It has a tight connection with the multiple-model direction studied currently in control theory [112]. As such, it is of independent interest. This makes us to present it. The optimization is made via dynamic programming written for a quadratic additive loss, Proposition 2.9.

Proposition 9.14 (Quadratic design) *Let us consider the data-driven design, Agreement 2.8, and search for the optimal admissible strategy*

$$\left\{d^*(t-1) \to u^*_{o;t}\right\}_{t\in t^*}$$

acting on an extending experience formed by $d(t)$. Let the system be described by a known normal mixture. Then, the optimal strategy minimizing expected value of the quadratic loss

$$\sum_{t\in t^*} \left(\phi'_t \, {}^{\lfloor\phi}L \, {}^{\lfloor\phi}D \, {}^{\lfloor\phi}L'\phi_t + u'_{o;t} \, {}^{\lfloor u}L \, {}^{\lfloor u}D \, {}^{\lfloor u}L' u_{o;t}\right), \qquad (9.49)$$

determined by the constant decomposed penalization kernels ${}^{\lfloor\phi}L \, {}^{\lfloor\phi}D \, {}^{\lfloor\phi}L'$ and ${}^{\lfloor u}L \, {}^{\lfloor u}D \, {}^{\lfloor u}L'$, is generated by the following algorithm.

Algorithm 9.12 (Linear quadratic advising for mixture models)

Initial mode

- *Construct the matrices* $A_c,\ c \in c^*$ *according to (9.33).*
- *Choose penalization kernels* ${}^{\lfloor\phi}L\, {}^{\lfloor\phi}D\, {}^{\lfloor\phi}L'$ *and* ${}^{\lfloor u}L\, {}^{\lfloor u}D\, {}^{\lfloor u}L'$*.*
- *Set* $L = I_{\mathring{\phi}},\ D = 0_{\mathring{\phi},\mathring{\phi}}$*.*

Iterative mode

1. *Transform* $D,\ L$ *to* $\bar{D},\ \bar{L}$ *of the same type so that*
$$\bar{L}\bar{D}\bar{L}' = LDL' + {}^{\lfloor\phi}L\, {}^{\lfloor\phi}D\, {}^{\lfloor\phi}L'. \tag{9.50}$$
2. *Compute* $\bar{A}_c = \bar{L}A_c,\ c \in c^*$*, while exploiting the special form of the shifting matrix* A*; see (9.33).*
3. *Transform* $\bar{A}_c,\ c \in c^*$*, and* $\bar{D}$*, to* $\tilde{L}, \tilde{D}$ *so that the following equality holds*
$$\sum_{c \in c^*} \bar{A}_c' \bar{D} \bar{A}_c + \left[{}^{\lfloor u}L',\, 0\right]'\, {}^{\lfloor u}D \left[{}^{\lfloor u}L',\, 0\right] = \tilde{L}\tilde{D}\tilde{L}'. \tag{9.51}$$
4. *Split* $\tilde{L},\ \tilde{D}$
$$\tilde{L} \equiv \begin{bmatrix} {}^{\lfloor u}\tilde{L} & 0 \\ {}^{\lfloor u\phi}\tilde{L} & {}^{\lfloor\phi}\tilde{L} \end{bmatrix},\quad \tilde{D} = \mathrm{diag}\left[{}^{\lfloor u}\tilde{D},\, {}^{\lfloor\phi}\tilde{D}\right],\quad \textit{where} \tag{9.52}$$
${}^{\lfloor u}\tilde{L},\ {}^{\lfloor u}\tilde{D}$ *are of* $(\mathring{u}, \mathring{u})$*-type.*
5. *Set* $L = {}^{\lfloor\phi}\tilde{L}$ *and* $D = {}^{\lfloor\phi}\tilde{D}$*.*
6. *Go to the beginning of* Iterative mode *until convergence is observed.*

Sequential (online) mode, *running for* $t = 1, 2, \ldots,$
Generate the optimal recognizable actions by the linear feedback
$$u_{o;t} = -\left({}^{\lfloor u}\tilde{L}'\right)^{-1}\, {}^{\lfloor u\phi}\tilde{L}'\phi_{t-1}. \tag{9.53}$$

Proof. For the considered finite horizon $\mathring{t}$ and additive loss, the optimal strategy can be constructed valuewise. The optimal recognizable inputs are minimizing arguments in the functional Bellman equation (2.22)

$$\mathcal{V}(d(t-1)) = \min_{u_{o;t} \in u^*_{o;t}} \mathcal{E}\left[\phi_t'\, {}^{\lfloor\phi}L\, {}^{\lfloor\phi}D\, {}^{\lfloor\phi}L'\phi_t + u_{o;t}'\, {}^{\lfloor u}L\, {}^{\lfloor u}D\, {}^{\lfloor u}L' u_{o;t} + \mathcal{V}(d(t)) | u_{o;t}, d(t-1)\right],$$

for $t = \mathring{t}, \mathring{t}-1, \ldots, 1$, starting with $\mathcal{V}(d(\mathring{t})) = 0$. In the considered case of the normal mixture with the common state in the phase form, the Bellman function depends on the state only. It is conjectured to be of the lifted quadratic form
$$\mathcal{V}(d(t)) = \mathcal{V}(\phi_t) = k_t + \phi_t' L_t D_t L_t' \phi_t \tag{9.54}$$

with the nonnegative scalar lift k_t and the decomposed kernel $L_tD_tL_t'$ independent of data.

For $\mathring{t}$, the conjecture is valid with $L_{\mathring{t}} = I_{\mathring{\phi}}$, $D_{\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$ and $k_{\mathring{t}} = 0$. According to the inductive assumption for a generic t, the right-hand side of the Bellman equation is

$$\begin{aligned}
&\min_{u_{o;t}} \mathcal{E}\left[\phi_t'\left({}^{\lfloor\phi}L\,{}^{\lfloor\phi}D\,{}^{\lfloor\phi}L' + L_tD_tL_t'\right)\phi_t + u_{o;t}'\,{}^{\lfloor u}L\,{}^{\lfloor u}D\,{}^{\lfloor u}L'u_{o;t}\middle|\,\psi_t\right] + k_t \\
&\underbrace{=}_{(9.50)} \min_{u_{o;t}} \mathcal{E}\left[\phi_t'\,{}^{\lfloor\phi}\bar{L}\,{}^{\lfloor\phi}\bar{D}\,{}^{\lfloor\phi}\bar{L}'\phi_t + u_{o;t}'\,{}^{\lfloor u}L\,{}^{\lfloor u}D\,{}^{\lfloor u}L'u_{o;t}\middle|\,\psi_t\right] + k_t \qquad (9.55)\\
&\underbrace{=}_{(9.33)} \min_{u_{o;t}} \psi_t'\left(\sum_{c\in c^*} A_c'\,{}^{\lfloor\phi}\bar{L}\,{}^{\lfloor\phi}\bar{D}\,{}^{\lfloor\phi}\bar{L}'A_c\right)\psi_t + u_{o;t}'\,{}^{\lfloor u}L\,{}^{\lfloor u}D\,{}^{\lfloor u}L'u_{o;t}\\
&+ \underbrace{\operatorname{tr}\left[[I_{\mathring{\Delta}},0]\bar{L}\bar{D}\bar{L}'[I_{\mathring{\Delta}},0]'\sum_{c\in c^*}\alpha_c r_c\right] + k_t}_{k_{t-1}}\\
&\underbrace{=}_{(9.51)} \min_{u_{o;t}} \psi_t'\tilde{L}\tilde{D}\tilde{L}'\psi_t + k_{t-1} \underbrace{=}_{(9.52)} \phi_{t-1}'\,{}^{\lfloor\phi}\tilde{L}\,{}^{\lfloor\phi}\tilde{D}\,{}^{\lfloor\phi}\tilde{L}'\phi_{t-1} + k_{t-1}
\end{aligned}$$

with the minimizing argument $u_{o;t}' = -\phi_{t-1}'\,{}^{\lfloor u\phi}\tilde{L}\,{}^{\lfloor u}\tilde{L}^{-1}$.

These evaluations prove both the claimed form and provide the algorithm giving the optimal feedback. □

Remark(s) 9.8

1. *The algorithm is described directly in the way suitable for a numerically efficient, factorized solution based on Algorithm* dydr *8.3.*
2. *For simplicity, the regulation problem is considered. An extension to the combined regulation and tracking is straightforward and standard one.*
3. *The optimization is made for the known model. As usual inlinear-quadratic design problems, the normality assumption serves us for learning. In the design, the knowledge of the second moment (9.33) is only needed.*
4. *The key message is that the state matrix corresponding to the mixture is not the weighted mean of state matrices of individual components. The combination has to be done in the mean square sense as reflected in Algorithm 9.12.*
5. *Knowledge on linear-quadratic control theory can be simply exploited. For instance, if there is a controller that makes the expected loss — divided by $\mathring{t}$ — finite for $\mathring{t} \to \infty$, the designed strategy makes this loss finite, too. Also, a sufficient rank of penalizing matrices and existence of a stabilizing controller imply that the designed one is stabilizing, etc.*
6. *The optimization is equivalent to a solution of the discrete-time Riccati equation [2, 79]. Thus, the presented evaluations can be viewed as its extension to mixture models.*

7. *The IST strategy, Section 4.2.1, should be used in the adaptive context.*
8. *Reducing the derived algorithm to the case with empty $u_{o;t}$, we get the second noncentral moment assigned to the whole mixture. This evaluation can serve us for practical test of the* stability of the mixture *as a whole.*

Problem 9.2 (Extension of linear-quadratic optimization art) *It is obvious that the well developed linear-quadratic optimization art [2] can be extended to mixture models. It is worth to do it in detail.*

9.2.3 Simultaneous academic and industrial design

Whenever applicable, the simultaneous academic and industrial design provides the best problem formulation and solution. The considered actions of the academic part of the simultaneous p-system $c_t \in c^* \equiv \{1,\ldots,\mathring{c}\}$ are to be generated by a causal strategy $d^*(t-1) \to c^*$. The industrial part generates the recommended recognizable actions $d^*(t-1) \to u^*_{o;t}$. The strategy determines the potential ideal pdfs, among which the best one is searched for,

$$\begin{aligned}
&{}^{\lfloor I}f(d_t,c_t|d(t-1)) = f(\Delta_{o;t}|\Delta_{p+;t},u_{o;t},d(t-1),c_t)\\
&\times f(\Delta_{p+;t}|u_{o;t},d(t-1),c_t)\,{}^{\lfloor I}f(c_t|u_{o;t},d(t-1))\,{}^{\lfloor I}f(u_{o;t}|d(t-1)).
\end{aligned}\tag{9.56}$$

The normal user's ideal pdf is used

$$\begin{aligned}
&{}^{\lfloor U}f(d_t,c_t|d(t-1)) = {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t},d_o(t-1))\\
&\times f(\Delta_{p+;t}|u_{o;t},d(t-1),c_t)\,{}^{\lfloor U}f(u_{o;t}|d_o(t-1))\,{}^{\lfloor U}f(c_t|u_{o;t},d(t-1))\\
&\propto \prod_{i=1}^{\mathring{\Delta}_o}\mathcal{N}_{\Delta_{i;t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\,{}^{\lfloor U}r_i\right)\prod_{i=\mathring{\Delta}_o+1}^{\mathring{\Delta}}\mathcal{N}_{\Delta_{i;t}}\left(\theta'_{ic_t}\psi_{ic_t;t},r_{ic_t}\right)\\
&\times\prod_{i=\mathring{\Delta}+1}^{\mathring{d}}\mathcal{N}_{u_{o(i-\mathring{\Delta});t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\,{}^{\lfloor U}r_i\right)\\
&\times\,{}^{\lfloor U}f(c_t)\exp\left[-0.5\left({}^{\lfloor U}k_{c_t;t-1}+\psi_t'\,{}^{\lfloor U}L_{c_t;t-1}\,{}^{\lfloor U}D_{c_t;t-1}\,{}^{\lfloor U}L'_{c_t;t-1}\psi_t\right)\right].
\end{aligned}$$

The following elements are involved.

$f(\Delta_{o;t}|\Delta_{p+;t},u_{o;t},d(t-1),c_t) \equiv \prod_{i=1}^{\mathring{\Delta}_o}\mathcal{N}_{\Delta_{ic_t;t}}\left(\theta'_{ic_t}\psi_{ic_t;t},r_{ic_t}\right)$ is the pdf derived from the c_t-th learned component describing the o-innovations;

$f(\Delta_{p+;t}|u_{o;t},d(t-1),c_t) \equiv \prod_{i=\mathring{\Delta}_o+1}^{\mathring{\Delta}}\mathcal{N}_{\Delta_{i;t}}\left(\theta'_{ic_t}\psi_{ic_t;t},r_{ic_t}\right)$ is the pdf derived from the c_t-th learned component describing the surplus innovations of the p-system;

$\left\{{}^{\lfloor I}f(c_t,u_{o;t}|d(t-1)) \equiv {}^{\lfloor I}f(c_t|u_{o;t},d(t-1))\,{}^{\lfloor I}f(u_{o;t}|d(t-1))\right\}_{t\in t^*}$ is the optimized simultaneous strategy;

${}^{\lfloor U}f(\Delta_{o;t}|u_{o;t},d_o(t-1)) = \prod_{i=1}^{\mathring{\Delta}_o}\mathcal{N}_{\Delta_{i;t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\,{}^{\lfloor U}r_i\right)$ is the true user's ideal pdf on the o-innovations; see Section 5.1.5;

${}^{\lfloor U}f(u_{o;t}|d_o(t-1)) = \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta});t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\ {}^{\lfloor U}r_i\right)$ is the true user's ideal pdf on the recognizable actions $u_{o;t}$; see Section 5.1.5;

${}^{\lfloor U}f(c_t|u_{o;t}, d(t-1)) \propto$

$$\propto {}^{\lfloor U}f(c_t)\exp\left[-0.5\left({}^{\lfloor U}k_{c_t;t-1} + \psi_t'\,{}^{\lfloor U}L_{c_t;t-1}\,{}^{\lfloor U}D_{c_{t-1};t}\,{}^{\lfloor U}L'_{c_t;t-1}\psi_t\right)\right]$$

is the pf representing the optional knob of the p-system that can respect special requirements like stability of advices. It is determined by the time invariant pf ${}^{\lfloor U}f(c_t)$ and by the lifted and decomposed quadratic forms in the exponent. Notice that the quadratic forms in the regression vector ψ_t are considered, i.e., the desirable c_t can be made dependent on $u_{o;t}$. The function ${}^{\lfloor U}\omega(c_t,\psi_t) \equiv {}^{\lfloor U}k_{c_t;t-1} + \psi_t'\,{}^{\lfloor U}L_{c_t;t-1}\,{}^{\lfloor U}D_{c_t;t-1}\,{}^{\lfloor U}L'_{c_t;t-1}\psi_t$ can be interpreted as a version of the user-specified KL divergence.

Proposition 9.15 (Simultaneous fully probabilistic design) *Let us consider the simultaneous academic and industrial design for the o-system described by the normal mixture with the state ϕ in the phase form. The data record d_t contains both innovations $\Delta_t = (\Delta_{o;t}, \Delta_{p+;t})$ =(innovations in d_o^*, innovations in d_{p+}^*) and the unrestricted recognizable actions $u_{o;t}$. The o-data $d_{o;t}$ consist of $(\Delta_{o;t}, u_{o;t})$ = (o-innovations, recognizable actions). The data record is ordered $d = (\Delta_o, \Delta_{p+}, u_o)$.*

The assumed influence of advices and the user's ideal pdf are described by (9.56). The parameters $\Theta_{ic} = [\theta_{ic}, r_{ic}]$ of the mixture model, as well those determining the true user's ideal pdf,

$${}^{\lfloor U}\Theta_i \equiv \left[{}^{\lfloor U}\theta_i,\ {}^{\lfloor U}r_i\right],\ i \in i^* \equiv \left\{1,\ldots,\mathring{\Delta}_o\right\} \cup \left\{\mathring{\Delta}+1,\ldots,\mathring{d}\right\}$$

are known and fixed. The regression coefficients are complemented by zeros so that, for each index i in the set i^*, the corresponding factors in the user's ideal pdf and mixture model have the common regression vectors $\psi_{i;t} \equiv [d_{i+1;t}, \psi'_{i+1;t}]' = [d'_{(i+1)\cdots\mathring{d};t}, \phi'_{t-1}]'$, $i < \mathring{d}$, $\psi_{\mathring{d};t} \equiv \phi_{t-1}$.

Let us search for the advising strategy $\left\{{}^{\lfloor I}f(c_t, u_{o;t}|d(t-1))\right\}_{t\in t^*}$ selecting both the recommended pointers c_t and the recognizable actions $u_{o;t}$ that, at least approximately, minimize the KL divergence of

$${}^{\lfloor I}f(d(\mathring{t}), c(\mathring{t})) = \prod_{t\in t^*} f(d_t, c_t|d(t-1)) \quad \textit{to the user's ideal pdf}$$

$${}^{\lfloor U}f(d(\mathring{t}), c(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor U}f(d_t, c_t|d(t-1)) \ \textit{with its factors given by (9.56).}$$

The definition of the KL divergence, the chain rule and marginalization imply that the minimized KL divergence can be interpreted as the expected value of the loss function,

$$\sum_{t\in t^*,c_t\in c^*}\int f(\varDelta_t|u_{o;t},d(t-1),c_t)\,{}^{\lfloor I}f(c_t|u_{o;t},d(t-1))\,{}^{\lfloor I}f(u_{o;t}|d(t-1)) \tag{9.57}$$
$$\times\ln\left(\frac{f(\varDelta_{o;t}|\varDelta_{p+;t},u_{o;t},d(t-1),c_t)\,{}^{\lfloor I}f(c_t|u_{o;t},d(t-1))\,{}^{\lfloor I}f(u_{o;t}|d(t-1))}{{}^{\lfloor U}f(\varDelta_{o;t}|u_{o;t},d_o(t-1))\,{}^{\lfloor U}f(c_t|u_{o;t},d_o(t-1))\,{}^{\lfloor U}f(u_{o;t}|d_o(t-1))}\right)dd_t.$$

Let us consider the following strategy

$$\begin{aligned}
{}^{\lfloor I}f(c_t,u_{o;t}|\phi_{t-1}) &\propto {}^{\lfloor U}f(c_t)\exp\left[-0.5\omega_\gamma(c_t,\phi_{t-1})\right]\,{}^{\lfloor I}f(u_{o;t}|\phi_{t-1},c_t)\\
\omega_\gamma(c_t,\phi_{t-1}) &\equiv k_{c_t;t-1}+\phi'_{t-1}L_{c_t;t-1}D_{c_t;t-1}L'_{c_t;t-1}\phi_{t-1} \qquad (9.58)\\
{}^{\lfloor I}f(u_{o;t}|\phi_{t-1},c_t) &\equiv \prod_{i=\mathring{\varDelta}+1}^{\mathring{d}}\mathcal{N}_{u_{o(i-\mathring{\varDelta})};t}\left({}^{\lfloor I}\theta'_{ic_t;t-1}\psi_{i;t},\ {}^{\lfloor I}r_{ic_t;t-1}\right).
\end{aligned}$$

The lifts $k_{c_t;t-1}$ and kernels $L_{c_t;t-1}D_{c_t;t-1}L'_{c_t;t-1}$, defining $\omega_\gamma(c_t,\phi_{t-1})$, and parameters ${}^{\lfloor I}\varTheta_{ic_t;t-1}\equiv\left[{}^{\lfloor I}\theta'_{ic_t;t-1},{}^{\lfloor I}r_{ic_t;t-1}\right]$, defining pdfs ${}^{\lfloor I}f(u_{o;t}|\phi_{t-1},c_t)$, are generated recursively. In the description of recursions, the following split of LDL' kernels is used

$$L\equiv\begin{bmatrix}1 & 0\\ {}^{\lfloor d\psi}L & {}^{\lfloor\psi}L\end{bmatrix},\quad D\equiv\mathrm{diag}\left[{}^{\lfloor d}D,\ {}^{\lfloor\psi}D\right],\quad \textit{where } {}^{\lfloor d}D \textit{ is scalar.}$$

An average quadratic form *$\phi'_{t-1}L_{\gamma;t-1}D_{\gamma;t-1}L'_{\gamma;t-1}\phi_{t-1}$ is used in the specification of the initial conditions of recursions defining the strategy. Its following split is used*

$$L_{\gamma;t}\equiv\begin{bmatrix}{}^{\lfloor\phi0}L_{\gamma;t} & 0\\ {}^{\lfloor\phi1}L'_{\gamma;t} & 1\end{bmatrix},\ D_{\gamma;t}=\mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor\phi0}D_{\gamma;t}\right],\ {}^{\lfloor1}D_{\gamma;t}\right],\ \ {}^{\lfloor1}D_{\gamma;t}\ \textit{is scalar.}$$

The recursions are described by the following formulas.

```
    Set $L_{\gamma;\mathring{t}}=I_{\mathring{\phi}}=\mathring{\phi}$-unit matrix, $D_{\gamma;\mathring{t}}=0_{\mathring{\phi},\mathring{\phi}}\equiv(\mathring{\phi},\mathring{\phi})$-zero matrix.
For  $t=\mathring{t},\dots,1$
          $q=0$
  For  $c=1,\dots,\mathring{c}$
          $L_{0c}\equiv\begin{bmatrix}{}^{\lfloor\phi0}L_{\gamma;t} & 0 & 0\\ 0 & I_{\mathring{d}} & 0\\ {}^{\lfloor\phi1}L'_{\gamma;t} & 0 & 1\end{bmatrix},\ D_{0c}=\mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor\phi0}D_{\gamma;t}\right],0_{1,\mathring{d}},\ {}^{\lfloor1}D_{\gamma;t}\right]$
          $k_{0c}\equiv-\mathring{d}_o$
    For  $i=1,\dots,\mathring{\varDelta}$
          $L_{ic}D_{ic}L'_{ic}={}^{\lfloor\psi}L_{(i-1)c}\,{}^{\lfloor\psi}D_{(i-1)c}\,{}^{\lfloor\psi}L'_{(i-1)c}$
          $+\left(\theta_{ic}+{}^{\lfloor d\psi}L_{(i-1)c}\right){}^{\lfloor d}D_{(i-1)c}\left(\theta_{ic}+{}^{\lfloor d\psi}L_{(i-1)c}\right)'$
          $+\chi\left(i\le\mathring{\varDelta}_o\right)\left(\theta_{ic}-{}^{\lfloor U}\theta_i\right){}^{\lfloor U}r_i^{-1}\left(\theta_{ic}-{}^{\lfloor U}\theta_i\right)'$,
      $k_{ic} \equiv k_{(i-1)c} + {}^{\lfloor d}D_{(i-1)c} r_{ic} + \chi\left(i \leq \mathring{\Delta}_o\right)\left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}}\right) + \frac{r_{ic}}{{}^{\lfloor U}r_i}\right]$
      $\mathring{D}_{ic} = \mathring{D}_{(i-1)c} - 1$
    end of the cycle over i
      $L_{\mathring{\Delta}c} D_{\mathring{\Delta}c} L'_{\mathring{\Delta}c} \equiv L_{\mathring{\Delta}c} D_{\mathring{\Delta}c} L'_{\mathring{\Delta}c} + {}^{\lfloor U}L_{c;t-1}\, {}^{\lfloor U}D_{c;t-1}\, {}^{\lfloor U}L'_{c;t-1}$
      $k_{\mathring{\Delta}c} \equiv k_{\mathring{\Delta}c} + {}^{\lfloor U}k_{c;t-1}$
  end of the cycle over c
      $L_{\gamma;t-1} = I_{\mathring{\phi}},\ D_{\gamma;t-1} = 0_{\mathring{\phi},\mathring{\phi}}$
  For $c = 1, \ldots, \mathring{c}$
    For $i = \mathring{\Delta} + 1, \ldots, \mathring{d}$
      $\tilde{L}_{ic}\tilde{D}_{ic}\tilde{L}'_{ic} = L_{(i-1)c} D_{(i-1)c} L'_{(i-1)c} + \left[-1,\ {}^{\lfloor U}\theta'_i\right]'\ {}^{\lfloor U}r_i^{-1}\left[-1,\ {}^{\lfloor U}\theta'_i\right]$
      $\tilde{L}_{ic} = \begin{bmatrix} 1 & 0 \\ -\,{}^{\lfloor I}\theta_{ic;t-1} & L_{ic} \end{bmatrix}$
      $\tilde{D}_{ic} = \mathrm{diag}\left[{}^{\lfloor I}r^{-1}_{ic;t-1}, D_{ic}\right]$, ${}^{\lfloor I}r_{ic;t-1}$ is scalar,
      $k_{ic} = k_{(i-1)c} + \ln\left({}^{\lfloor U}r_i\, {}^{\lfloor I}r^{-1}_{ic;t-1}\right)$
    end of the cycle over i
      $L_{c;t-1} \equiv L_{\mathring{d}c},\ D_{c;t-1} \equiv D_{\mathring{d}c},\ k_{c;t-1} \equiv k_{\mathring{d}c}$
      $\beta_c \equiv {}^{\lfloor U}f(c)\exp(-0.5 k_{\mathring{d}c})$
      $q = q + \beta_c$
  end of the cycle over c
  For $c = 1, \ldots, \mathring{c}$
      $\beta_c = \frac{\beta_c}{q}$
      $L_{\gamma;t-1} D_{\gamma;t-1} L'_{\gamma;t-1} = L_{\gamma;t-1} D_{\gamma;t-1} L'_{\gamma;t-1} + \beta_c L_{c;t-1} D_{c;t-1} L'_{c;t-1}.$
  end of the cycle over c
end of the cycle over t
```

The updating of the LDL' (!) decomposition can be done by Algorithm 8.2. Then, this strategy minimizes the KL divergence for the horizon equal one. For other horizons, it minimizes the upper bound on the expected loss as it replaces the correct Bellman function, cf. (9.58),

$$-\ln\left\{\sum_{c_t \in c^*} {}^{\lfloor U}f(c_t)\exp[-0.5\omega_\gamma(c_t, \phi_{t-1})]\right\} \text{ by the larger value}$$

$$0.5\phi'_{t-1}\left(\sum_{c_t \in c^*} \beta_{c_t} L_{c_t;t-1} D_{c_c;t-1} L'_{c_t;t-1}\right)\phi_{t-1}. \tag{9.59}$$

Proof. Let us assume that the Bellman function is $-\ln(\gamma(\phi_t))$, with

$$\begin{aligned}\gamma(\phi_t) &= \exp\left[-0.5 \sum_{c_{t+1}\in c^*} \beta_{c_{t+1}} \underline{\omega}_{\gamma}(c_{t+1},\phi_t)\right] \equiv \exp\left[-0.5\omega_{\gamma}(\phi_t)\right]\\ \beta_{c_{t+1}} &\propto {}^{\lfloor U}f(c_{t+1})\exp[-0.5k_{c_{t+1};t}]\end{aligned}$$

$$\underline{\omega}_{\gamma}(c_{t+1},\phi_t) \equiv \phi_t' L_{c_{t+1};t} D_{c_{t+1};t} L_{c_{t+1};t}' \phi_t, \text{ i.e. } \omega_{\gamma}(\phi_t) = \phi_t' L_{\gamma;t} D_{\gamma;t} L_{\gamma;t}' \phi_t \text{ with}$$

$$L_{\gamma;t} D_{\gamma;t} L_{\gamma;t}' \equiv \sum_{c_{t+1}\in c^*} \beta_{c_{t+1}} L_{c_{t+1};t} D_{c_{t+1};t} L_{c_{t+1};t}'.$$

The kernels $L_{c_{t+1};t} D_{c_{t+1};t} L_{c_{t+1};t}'$ and the scalar lifts $k_{c_{t+1};t}$ are conjectured to be independent of data. This conjecture is correct for $t = \mathring{t}$ with $k_{c_{\mathring{t}+1};\mathring{t}} = 0$, $L_{c_{\mathring{t}+1};\mathring{t}} = I_{\mathring{\phi}}$, $D_{c_{\mathring{t}+1};\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$. We prove that this form is preserved during the backward induction for $t < \mathring{t}$ if the approximation (9.59) is used. It gives us also the algorithmic solution of the inspected problem.
Under the inductive assumption, the weighted conditional KL divergence is

$$\begin{aligned}2\tilde{\omega}_{\gamma}(c_t,\psi_t) &\equiv 2\int f(\Delta_t|\psi_t,c_t)\ln\left[\frac{f(\Delta_{o;t}|\Delta_{p+;t},\psi_t,c_t)}{\gamma(\phi_t)\,{}^{\lfloor U}f(\Delta_{o;t}|\psi_t)}\right] d\Delta_t\\ &= k_{\mathring{\Delta}c_t} + \psi_t' L_{\mathring{\Delta}c_t} D_{\mathring{\Delta}c_t} L_{\mathring{\Delta}c_t}' \psi_t,\end{aligned}$$

where, according to Proposition 9.8, the lifts and the kernels are found recursively for $i = 1,\ldots,\mathring{\Delta}$

$$\begin{aligned}L_{ic}D_{ic}L_{ic}' &= {}^{\lfloor\psi}L_{(i-1)c}\,{}^{\lfloor\psi}D_{(i-1)c}\,{}^{\lfloor\psi}L_{(i-1)c}'\\ &+ \left(\theta_{ic} + {}^{\lfloor\Delta\psi}L_{(i-1)c}\right){}^{\lfloor\Delta}D_{(i-1)c}\left(\theta_{ic} + {}^{\lfloor\Delta\psi}L_{(i-1)c}\right)'\\ &+ \chi\left(i \le \mathring{\Delta}_o\right)\left(\theta_{ic} - {}^{\lfloor U}\theta_i\right){}^{\lfloor U}r_i^{-1}\left(\theta_{ic} - {}^{\lfloor U}\theta_i\right)'\\ k_{ic} &= k_{(i-1)c} + {}^{\lfloor\Delta}D_{(i-1)c}r_{ic} + \chi\left(i \le \mathring{\Delta}_o\right)\left[\ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}}\right) + \frac{r_{ic}}{{}^{\lfloor U}r_i}\right].\end{aligned}$$

The initial conditions, implied by the general condition $\gamma(d(\mathring{t})) = 1$, are

$$\begin{aligned}k_{0c} &= -\mathring{\Delta}_o, \quad L_{0c}D_{0c}L_{0c}' = \mathcal{K}L_{\gamma;t}D_{\gamma;t}L_{\gamma;t}'\mathcal{K}'\\ \mathcal{K}' &\equiv \begin{bmatrix} I_{\mathring{d}(\partial-1)} & 0 & 0\\ 0 & 0_{1,\mathring{d}} & 1\end{bmatrix}\\ L_{ic} &\equiv \begin{bmatrix} 1 & 0\\ {}^{\lfloor\Delta\psi}L_{ic} & {}^{\lfloor\psi}L_{ic}\end{bmatrix},\ D_{ic} \equiv \operatorname{diag}\left[{}^{\lfloor\Delta}D_{ic},\ {}^{\lfloor\psi}D_{ic}\right], \quad \text{where}\\ &{}^{\lfloor\Delta}D_{ic} \text{ is scalar and } \mathring{D}_{(i+1)c} = \mathring{D}_{ic} - 1.\end{aligned}$$

The optimal pdf ${}^{\lfloor I}f(c_t, u_{o;t}|d(t-1))$ is the minimizer of

$$\sum_{c_t \in c^*} \int {}^{\lfloor I}f(c_t, u_{o;t}|d(t-1))$$
$$\times \left[\omega_\gamma(c_t, \psi_t) + \ln\left(\frac{{}^{\lfloor I}f(c_t, u_{o;t}|d(t-1))}{{}^{\lfloor U}f(c_t|u_{o;t}, d_o(t-1))\, {}^{\lfloor U}f(u_{o;t}|\phi_{t-1})}\right)\right] du_{o;t}, \text{ where}$$
$$\omega_\gamma(c_t, \psi_t) \equiv \tilde{\omega}_\gamma(c_t, \psi_t) + {}^{\lfloor U}\omega(c_t, \psi_t).$$

Thus, ω_γ is defined by the kernel $L_{\mathring{\Delta}c} D_{\mathring{\Delta}c} L'_{\mathring{\Delta}c}$ of $\tilde{\omega}_\gamma$ increased by ${}^{\lfloor U}L_{c;t-1}\, {}^{\lfloor U}D_{c;t-1}\, {}^{\lfloor U}L'_{c;t-1}$. Also, the lift ${}^{\lfloor U}k_{c;t-1}$ is added to $k_{\mathring{\Delta}c}$.

The minimizer is

$$ {}^{\lfloor I}f(c_t, u_{o;t}|\phi_{t-1}) = \frac{{}^{\lfloor U}f(c_t)\, {}^{\lfloor U}f(u_{o;t}|\phi_{t-1}) \exp[-0.5\omega_\gamma(c_t, \psi_t)]}{\tilde{\gamma}(\phi_{t-1})}$$
$$\tilde{\gamma}(\phi_{t-1}) \equiv \sum_{c_t \in c^*} {}^{\lfloor U}f(c_t) \int {}^{\lfloor U}f(u_{o;t}|\phi_{t-1}) \exp[-0.5\omega_\gamma(c_t, \psi_t)]\, du_{o;t}.$$

We modify this expression so that the marginal pf ${}^{\lfloor I}f(c_t|\phi_{t-1})$ become visible. For it, we use the explicit form of the user's ideal pdf on the recognizable actions when expressing the product

$$ {}^{\lfloor U}f(u_{o;t}|\phi_{t-1}) \exp[-0.5\omega_\gamma(c_t, \psi_t)]$$
$$= \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \left(2\pi\, {}^{\lfloor U}r_i\right)^{-0.5} \exp\left\{-\frac{1}{2}\left(k_{\mathring{\Delta}c_t} + \psi_t' L_{\mathring{\Delta}c_t} D_{\mathring{\Delta}c_t} L'_{\mathring{\Delta}c_t} \psi_t\right)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\left(\sum_{i=\mathring{\Delta}+1}^{\mathring{d}} [u_{o(i-\mathring{\Delta});t}, \psi'_{i;t}] \frac{\left[-1, {}^{\lfloor U}\theta_i'\right]' \left[-1, {}^{\lfloor U}\theta_i'\right]}{{}^{\lfloor U}r_i} [u_{o(i-\mathring{\Delta});t}, \psi'_{i;t}]'\right)\right\}.$$

Completion of the squares in individual entries of $u_{o(i-\mathring{\Delta});t}$ defines sequence of kernels for $i = \mathring{\Delta}+1, \ldots, \mathring{d}$, $c \in c^*$,

$$\tilde{L}_{ic}\tilde{D}_{ic}\tilde{L}'_{ic} \equiv L_{(i-1)c} D_{(i-1)c} L'_{(i-1)c} + \frac{\left[-1, {}^{\lfloor U}\theta_i'\right]' \left[-1, {}^{\lfloor U}\theta_i'\right]}{{}^{\lfloor U}r_i}$$
$$\tilde{L}_{ic} = \begin{bmatrix} 1 & 0 \\ -\, {}^{\lfloor I}\theta_{ic;t-1} & L_{ic} \end{bmatrix}, \ \tilde{D}_{ic} \equiv \operatorname{diag}\left[{}^{\lfloor I}r^{-1}_{ic;t-1}, \operatorname{diag}[D_{ic}]\right], \ {}^{\lfloor I}r^{-1}_{ic;t-1} \text{ is scalar.}$$

Thus,

$$ {}^{\lfloor I}f(c_t, u_{o;t}|\phi_{t-1}) = \frac{{}^{\lfloor U}f(c_t)}{\tilde{\gamma}(\phi_{t-1})}$$
$$\times \exp\left\{-0.5\left[k_{c_t;t-1} + \phi'_{t-1} L_{c_t;t-1} D_{c_t;t-1} L'_{c_t;t-1} \phi_{t-1}\right]\right\}$$
$$\times \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta});t}}\left({}^{\lfloor I}\theta'_{ic;t-1}\psi_{i;t}, {}^{\lfloor I}r_{ic;t-1}\right).$$

The integration over u_{oi}, performed while normalizing, increases lifts to

$$k_{c_t;t-1} \equiv k_{\mathring{\Delta} c_t} + \sum_{i=\mathring{\Delta}+1}^{\mathring{d}} \ln\left(\frac{{}^{\lfloor U}r_i}{{}^{\lfloor I}r_{ic;t-1}}\right) \text{ and } L_{c_t;t-1} \equiv L_{\mathring{d}c_t},\ \ D_{c_t;t-1} \equiv D_{\mathring{d}c_t}.$$

$$\begin{aligned}
\tilde{\gamma}(\phi_{t-1}) &\equiv \sum_{c_t \in c^*} {}^{\lfloor U}f(c_t) \exp\left\{-\frac{1}{2}\left[k_{c_t;t-1} + \phi'_{t-1} L_{c_t;t-1} D_{c_t;t-1} L'_{c_t;t-1} \phi_{t-1}\right]\right\} \\
&= q \sum_{c_t \in c^*} \beta_{c_t} \exp\left[-0.5 \phi'_{t-1} L_{c_t;t-1} D_{c_t;t-1} L'_{c_t;t-1} \phi_{t-1}\right] \\
&= q \sum_{c_t \in c^*} \beta_{c_t} \exp\left[-0.5 \underline{\omega}(c_t;\phi_{t-1})\right] \quad \text{with} \\
\beta_{c_t} &= \frac{{}^{\lfloor U}f(c_t) \exp\left[-0.5 k_{c_t;t-1}\right]}{q}, \quad q \equiv \sum_{c_t \in c^*} {}^{\lfloor U}f(c_t) \exp\left[-0.5 k_{c_t;t-1}\right].
\end{aligned}$$

The obtained minimum $-\ln(\tilde{\gamma}(\phi_{t-1}))$ is approximated from above by replacing weighted arithmetic mean by weighted geometric mean. It reproduces

$$\begin{aligned}
\gamma(\phi_{t-1}) &= \exp\left[-0.5 \phi'_{t-1} L_{\gamma;t-1} D_{\gamma;t-1} L'_{\gamma;t-1} \phi_{t-1}\right] \quad \text{with} \\
L_{\gamma;t-1} D_{\gamma;t-1} L'_{\gamma;t-1} &\equiv \sum_{c_t \in c^*} \beta_{c_t} L_{c_t;t-1} D_{c_t;t-1} L'_{c_t;t-1}.
\end{aligned}$$

□

The proposition, combined with the certainty-equivalence strategy, justifies the algorithm designing the fixed simultaneous advisory system.

Algorithm 9.13 (Fixed simultaneous advising with the γ-bound)

Initial (offline) mode

- *Estimate normal mixture model of the o-system with the state ϕ_t in the phase form; Chapter 8.*
- *Specify the true user's ideal pdf on the response of the o-system*

$$\begin{aligned}
{}^{\lfloor U}f(d_{o;t}|d_o(t-1)) &\equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) \\
&= \prod_{i=1}^{\mathring{\Delta}_o} \mathcal{N}_{\Delta_{i;t}}\left({}^{\lfloor U}\theta'_i \psi_{i;t},\ {}^{\lfloor U}r_i\right) \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta});t}}\left({}^{\lfloor U}\theta'_i \psi_{i;t},\ {}^{\lfloor U}r_i\right).
\end{aligned}$$

- *Specify the data and time invariant part ${}^{\lfloor U}f(c_t)$ of the user's ideal pf on the recommended pointers. It is zero on dangerous components.*
- *Specify data and time invariant lifts ${}^{\lfloor U}k_{c_t}$ and data and time invariant kernels of the user-specified KL divergence ${}^{\lfloor U}L_{c_t}$, ${}^{\lfloor U}D_{c_t}$. It completes the definition of the user's ideal pf $f(c_t|d(t-1))$; see (9.56). Notice that the quadratic form in the regression vector (not only in the state ϕ) is to be considered.*

- *Select the length of the design horizon* $\mathring{t} \geq 1$.
- *Initialize the iterative mode by setting* $L_{\gamma;\mathring{t}} = I_{\mathring{\phi}},\ D_{\gamma;\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$.

Iterative (offline) mode

1. *Correct the conditional KL divergences, for* $t = \mathring{t}, \ldots, 1$, *as given in (9.58).*
2. *Denote the final characteristics of the approximate optimal* steady-state strategy, *cf. Chapter 3,* $i = \mathring{\Delta} + 1, \ldots, \mathring{d},\ c \in c^*$,

$$k_c \equiv k_{c;1},\ L_c \equiv L_{c;1},\ D_c \equiv D_{c;1},\ \theta_{ic} \equiv {}^{\lfloor I}\theta_{ic;1},\ r_{ic} \equiv {}^{\lfloor I}r_{ic;1}.$$

Sequential (online) mode, *running for* $t = 1, 2, \ldots,$

1. *Acquire the data record* d_t *and determine the state vector* ϕ_t.
2. *Evaluate the ideal pf on pointers*

$${}^{\lfloor I}f(c_{t+1}|\phi_t) \propto {}^{\lfloor U}f(c_{t+1}) \exp\left[-0.5\left(k_{c_{t+1}} + \phi'_{t-1} L_{c_{t+1}} D_{c_{t+1}} L'_{c_{t+1}} \phi_{t-1}\right)\right].$$

3. *Present to the operator selected projections of the ideal pdf*

$${}^{\lfloor I}f(d_{t+1}|\phi_t) = \sum_{c_{t+1} \in c^*} {}^{\lfloor I}f(c_{t+1}|\phi_t) \prod_{i=1}^{\mathring{d}} \mathcal{N}_{d_{ic_{t+1};t+1}}(\theta'_{ic_{t+1}} \psi_{ic_{t+1};t}, r_{ic_{t+1}}).$$

4. *Go to the beginning of* Sequential mode.

Proposition 9.15, combined with the receding-horizon, certainty-equivalence strategy, justifies the following adaptive design algorithm.

Algorithm 9.14 (Adaptive simultaneous advising with γ-bound)
Initial (offline) mode

- *Estimate normal mixture model of the o-system with the state* ϕ_t *in the phase form; Chapter 8.*
- *Specify the true user's ideal pdf on the response of the o-system*

$$\begin{aligned}{}^{\lfloor U}f(d_{o;t}|d_o(t-1)) &\equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) \\ &= \prod_{i=1}^{\mathring{\Delta}_o} \mathcal{N}_{\Delta_{i;t}}\left({}^{\lfloor U}\theta'_i \psi_{i;t},\ {}^{\lfloor U}r_i\right) \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{i-\mathring{\Delta};t}}\left({}^{\lfloor U}\theta'_i \psi_{i;t},\ {}^{\lfloor U}r_i\right).\end{aligned}$$

- *Specify* <u>*data invariant*</u> *part* ${}^{\lfloor U}f(c_t)$ *of the user ideal on the recommended pointers.*
- *Specify* <u>*data invariant*</u> *lifts* ${}^{\lfloor U}k_{c;t}$ *and the kernels* ${}^{\lfloor U}L_{c;t},\ {}^{\lfloor U}D_{c;t}$ *that define the user's ideal pf on recommended pointers*

$$\begin{aligned}&f(c_t|u_{o;t}, d(t-1)) \\ &\quad \propto {}^{\lfloor U}f(c_t) \exp\left[-0.5\left({}^{\lfloor U}k_{c_t;t-1} + \psi'_t\, {}^{\lfloor U}L_{c_t;t-1}\, {}^{\lfloor U}D_{c_t;t-1}\, {}^{\lfloor U}L'_{c_t;t-1} \psi_t\right)\right].\end{aligned}$$

- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update estimates of the model parameters; Section 8.5.*
3. *Initialize the iterative mode by setting $\tau = t + T$, $L_{\gamma;\tau} = I_{\mathring{\phi}}$, $D_{\gamma;\tau} = 0_{\mathring{\phi},\mathring{\phi}}$.*

$$L_{\gamma;\tau} \equiv \begin{bmatrix} {}^{\lfloor \phi 0}L_{\gamma;\tau} & 0 \\ {}^{\lfloor \phi 1}L'_{\gamma;\tau} & 1 \end{bmatrix}, \; D_{\gamma;\tau} \equiv \left[\operatorname{diag} \left[{}^{\lfloor \phi 0}D_{\gamma;\tau} \right], {}^{\lfloor 1}D_{\gamma;\tau} \right], \; {}^{\lfloor 1}D_{\gamma;\tau} \text{ is scalar.}$$

Omit the initialization of $L_{\gamma;\tau}$, $D_{\gamma;\tau}$ if $t > 1$ and the IST strategy is adopted.

Iterative mode

Correct the lifted quadratic forms defining the KL divergences using the current point estimates of parameters.

```
Set $L_{\gamma;t+T} = I_{\mathring{\phi}} = \mathring{\phi}$-unit matrix, $D_{\gamma;t+T} = 0_{\mathring{\phi},\mathring{\phi}} \equiv (\mathring{\phi}, \mathring{\phi})$-zero.
    For $\tau = t + T, \ldots, t + 1$
              $q = 0$
      For $c = 1, \ldots, \mathring{c}$
              $L_{0c} \equiv \begin{bmatrix} {}^{\lfloor \phi 0}L_{\gamma;\tau} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor \phi 1}L'_{\gamma;\tau} & 0 & 1 \end{bmatrix}$
              $D_{0c} = \operatorname{diag} \left[ \operatorname{diag} \left[ {}^{\lfloor \phi 0}D_{\gamma;\tau} \right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_{\gamma;\tau} \right], \quad k_{0c} \equiv -\mathring{d}_o$
        For $i = 1, \ldots, \mathring{\Delta}$
          $L_{ic}D_{ic}L'_{ic} = {}^{\lfloor \psi}L_{(i-1)c} \, {}^{\lfloor \psi}D_{(i-1)c} \, {}^{\lfloor \psi}L'_{(i-1)c}$
          $+ \left( \theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c} \right) {}^{\lfloor d}D_{(i-1)c} \left( \theta_{ic} + {}^{\lfloor d\psi}L_{(i-1)c} \right)'$
          $+\chi \left( i \leq \mathring{\Delta}_o \right) \left( \theta_{ic} - {}^{\lfloor U}\theta_i \right) {}^{\lfloor U}r_i^{-1} \left( \theta_{ic} - {}^{\lfloor U}\theta_i \right)',$
          $k_{ic} \equiv k_{(i-1)c} + {}^{\lfloor d}D_{(i-1)c} r_{ic} + \chi \left( i \leq \mathring{\Delta}_o \right) \left[ \ln \left( \frac{{}^{\lfloor U}r_i}{r_{ic}} \right) + \frac{r_{ic}}{{}^{\lfloor U}r_i} \right]$
              $\mathring{D}_{ic} = \mathring{D}_{(i-1)c} - 1$
        end of the cycle over $i$
              $L_{\mathring{\Delta}c}D_{\mathring{\Delta}c}L'_{\mathring{\Delta}c} \equiv L_{\mathring{\Delta}c}D_{\mathring{\Delta}c}L'_{\mathring{\Delta}c} + {}^{\lfloor U}L_{c;t-1} \, {}^{\lfloor U}D_{c;\tau-1} \, {}^{\lfloor U}L'_{c;\tau-1}$
              $k_{\mathring{\Delta}c} \equiv k_{\mathring{\Delta}c} + {}^{\lfloor U}k_{c;\tau-1}$
      end of the cycle over $c$
              $L_{\gamma;\tau-1} = I_{\mathring{\phi}}, \; D_{\gamma;\tau-1} = 0_{\mathring{\phi},\mathring{\phi}}$
      For $c = 1, \ldots, \mathring{c}$
```

```
        For   $i = \mathring{\Delta} + 1, \ldots, \mathring{d}$
            $\tilde{L}_{ic}\tilde{D}_{ic}\tilde{L}'_{ic} = L_{(i-1)c}D_{(i-1)c}L'_{(i-1)c} + \left[-1,\ {}^{\lfloor U}\theta'_i\right]'\ {}^{\lfloor U}r_i^{-1}\left[-1,\ {}^{\lfloor U}\theta'_i\right]$
            $\tilde{L}_{ic} = \begin{bmatrix} 1 & 0 \\ -\ {}^{\lfloor I}\theta_{ic;\tau-1} & L_{ic} \end{bmatrix}$
            $\tilde{D}_{ic} = \mathrm{diag}\left[{}^{\lfloor I}r^{-1}_{ic;\tau-1}, \mathrm{diag}[D_{ic}]\right]$,  ${}^{\lfloor I}r_{ic;\tau-1}$ is scalar,
            $k_{ic} = k_{(i-1)c} + \ln\left({}^{\lfloor U}r_i\ {}^{\lfloor I}r^{-1}_{ic;\tau-1}\right)$
        end   of the cycle over i
                $L_{c;\tau-1} \equiv L_{\mathring{d}c},\ D_{c;\tau-1} \equiv D_{\mathring{d}c},\ k_{c;\tau-1} \equiv k_{\mathring{d}c}$
                $\beta_c \equiv {}^{\lfloor U}f(c)\exp(-0.5k_{\mathring{d}c})$
                $q = q + \beta_c$
    end   of the cycle over c
    For   $c = 1, \ldots, \mathring{c}$
        $\beta_c = \frac{\beta_c}{q}$
        $L_{\gamma;\tau-1}D_{\gamma;\tau-1}L'_{\gamma;\tau-1} = L_{\gamma;\tau-1}D_{\gamma;\tau-1}L'_{\gamma;\tau-1} + \beta_c L_{c;\tau-1}D_{c;\tau-1}L'_{c;\tau-1}.$
    end   of the cycle over c
end   of the cycle over $\tau$
```

The updating of the LDL′ (!) decomposition can be done by Algorithm 8.2.

4. *Evaluate the ideal pf on pointers*

$$ {}^{\lfloor I}f(c_{t+1}|\phi_t) \propto {}^{\lfloor U}f(c_{t+1})\exp[-0.5(k_{c_{t+1};t} + \phi'_t L_{c_{t+1};t}D_{c_{t+1};t}L'_{c_{t+1};t}\phi_t)]. $$

5. *Present to the operator projections of the ideal pdf*

$$ {}^{\lfloor I}f(d_{t+1}|\phi_t) = \sum_{c_{t+1}\in c^*} {}^{\lfloor I}f(c_{t+1}|\phi_t)\prod_{i=1}^{\mathring{\Delta}} \mathcal{N}_{d_{ic_{t+1};t+1}}(\theta'_{ic_{t+1}}\psi_{ic_{t+1};t}, r_{ic_{t+1}}) $$
$$ \times \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta})c_{t+1};t+1}}\left({}^{\lfloor I}\theta'_{ic_{t+1};t}\psi_{ic_{t+1};t},\ {}^{\lfloor I}r_{ic_{t+1};t}\right). $$

6. *Go to the beginning of* Sequential mode*.*

Remark(s) 9.9

1. *It is worth stressing that the user's ideal pf on c_t is conditioned on ψ_t. Thus, it can modify dependence of c_t and $u_{o;t}$. In algorithmic terms, the constructed coefficients ${}^{\lfloor I}\theta_{ic;t}$ are influenced by the optional ${}^{\lfloor U}L_c$, which is the $(\mathring{\psi}, \mathring{\psi})$-dimensional matrix.*

2. *Other design variants, like selection of the most probable advices or the grouped version, can be constructed directly according to Chapter 7, using the above evaluations.*

9.3 Interaction with an operator

Here, we elaborate the design of strategies generating presentation and signaling actions, using the model presented in Section 7.1.3, in the case of normal mixtures and normal user's ideal pdfs.

9.3.1 Assigning priorities

We present a normal version of the choice of presentation priorities, i.e., the choice of the quantities that should be shown to the operator most urgently. It is based on Proposition 7.19, on the restriction to $\mathring{z}_t = 1$ and on the simplifying assumption that the probabilities of pointers $c \in c^*$ to components used in the design are close to the estimated weights α_c

$$f(c|d_{z_t;t}, d(t-1)) \approx \alpha_c. \tag{9.60}$$

We also assume that the optimized component weights are approximated by data and time invariant values

$${}^{\lfloor I}f(c|d(t-1)) \approx {}^{\lfloor I}\alpha_c. \tag{9.61}$$

Proposition 9.16 (Presentation design; the bound (7.47), $\mathring{z}_t = 1$)
Let us consider that the academic, industrial or simultaneous design has provided the optimal advising mixture

$${}^{\lfloor I}f(d_t|d(t-1)) = \sum_{c_t \in c^*} {}^{\lfloor I}f(d_t|d(t-1), c_t)\, {}^{\lfloor I}f(c_t|d(t-1)).$$

Let the signaling strategy make the operator fully alert, i.e., signaling actions $s(\mathring{t}) \equiv 1$. Let us assume that $\mathring{z}_t = 1$ and (9.60), (9.61) hold in addition to the assumptions of Proposition 7.19. Let us specify the user data invariant ideal pf ${}^{\lfloor U}f(z_t)$ on the set of possible presentation actions $z^* \equiv \{1, \ldots, \mathring{d}_o\}$.

Let us define a presentation strategy that assigns the higher priority to the entries of $d_{z_t;t}$ the higher are the values of the following pf

$$f(z_t|\phi_{t-1}) \propto {}^{\lfloor U}f(z_t) \exp\left[-0.5\left(k_{z_t} + \phi'_{t-1} L_{z_t;t} D_{z_t;t} L'_{z_t;t} \phi'_{t-1}\right)\right]. \tag{9.62}$$

The involved lifts and kernels are generated by the following algorithm.

```
Set $L_{\mathring{t}} = I_{\mathring{\phi}},\ D_{\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$.    (9.63)
For $t = \mathring{t}, \ldots, 1$
          $q = 0$
  For $z_t = 1, \ldots, \mathring{d}_o$
        Permute entry $d_{z_t;t}$ at the tail position in the used ideal pdf
          $^{\lfloor U}f(d_t|d(t-1)) \underbrace{\rightarrow}_{Proposition\ 9.1} \prod_{i\in i^*} \mathcal{N}_{d_{i;t}}\left({}^{\lfloor U}\theta'_{iz_t}\psi_{i;t},\ {}^{\lfloor U}r_{iz_t}\right)$.
        Reflect this permutation into kernel of the quadratic form $L_t, D_t$
        $L_t,\ D_t \underbrace{\rightarrow}_{Proposition\ 8.4} L_{z_t;t},\ D_{z_t;t}$
          Set $L_{z_t} = I_{\mathring{\phi}+1},\ D_{z_t} = 0_{\mathring{\phi}+1,\mathring{\phi}+1},\ k_{z_t} = 0$
    For $c = 1, \ldots, \mathring{c}$
      Permute entry $d_{z_t;t}$ to the tail position in estimated and ideal pdfs
      $^{\lfloor I}f(d_t|d(t-1),c) \underbrace{\rightarrow}_{Proposition\ 9.1} \prod_{i\in i^*} \mathcal{N}_{d_{i;t}}\left({}^{\lfloor I}\theta'_{icz_t}\psi_{i;t},\ {}^{\lfloor I}r_{icz_t}\right)$,
      $f(d_t|d(t-1),c) \underbrace{\rightarrow}_{Proposition\ 9.1} \prod_{i\in i^*} \mathcal{N}_{d_{i;t}}\left(\theta'_{icz_t}\psi_{i;t}, r_{icz_t}\right)$.
      Assign $k_{0cz_t} \equiv 0,\ L_{0cz_t} \equiv \begin{bmatrix} {}^{\lfloor \phi 0}L_{z_t;t} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor \phi 1}L'_{z_t;t} & 0 & 1 \end{bmatrix}$,
      $D_{0cz_t} \equiv \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \phi 0}D_{z_t;t}\right], 0_{1,\mathring{d}},\ {}^{\lfloor 1}D_{z_t;t}\right]$, scalar ${}^{\lfloor 1}D_{z_t;t}$.
      For $i = 1, \ldots, \mathring{d}-1$
          $L_{icz_t}D_{icz_t}L'_{icz_t} = {}^{\lfloor \psi}L_{(i-1)cz_t}\, {}^{\lfloor \psi}D_{(i-1)cz_t}\, {}^{\lfloor \psi}L'_{(i-1)cz_t}$
          $+ \left(\theta_{icz_t} + {}^{\lfloor d\psi}L_{(i-1)cz_t}\right) {}^{\lfloor d}D_{(i-1)cz_t} \left(\theta_{icz_t} + {}^{\lfloor d\psi}L_{(i-1)cz_t}\right)'$
          $k_{icz_t} = k_{(i-1)cz_t} + {}^{\lfloor d}D_{(i-1)cz_t} r_{icz_t}$, where
          $L_{(i-1)cz_t} = \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L_{(i-1)cz_t} & {}^{\lfloor \psi}L_{(i-1)cz_t} \end{bmatrix}$
          $D_{(i-1)cz_t} = \mathrm{diag}\left[{}^{\lfloor d}D_{(i-1)cz_t}, \mathrm{diag}\left[{}^{\lfloor \psi}D_{(i-1)cz_t}\right]\right]$
          ${}^{\lfloor d}D_{(i-1)cz_t}$ is scalar.
      end of the cycle over $i$
          $L_{z_t}D_{z_t}L'_{z_t} = L_{z_t}D_{z_t}L'_{z_t} + \alpha_c L_{(\mathring{d}-1)cz_t} D_{(\mathring{d}-1)cz_t} L'_{(\mathring{d}-1)cz_t}$,
          $k_{z_t} = k_{z_t} + \alpha_c k_{(\mathring{d}-1)cz_t}$.
    end of the cycle over $c$
          $l_{z_t} = {}^{\lfloor d\psi}L_{z_t},\ \delta_{z_t} = {}^{\lfloor d}D_{z_t},\ L_{z_t} = {}^{\lfloor \psi}L_{z_t},\ D_{z_t} = {}^{\lfloor \psi}D_{z_t}$.
        For   $c = 1, \ldots, \mathring{c}$
                $L_{z_t} D_{z_t} L'_{z_t} = L_{z_t} D_{z_t} L'_{z_t}$
                $+ {}^{\lfloor I}\alpha_c \left\{ \left( {}^{\lfloor I}\theta_{\mathring{d}cz_t} + l_{z_t} \right) \delta_{z_t} \left( {}^{\lfloor I}\theta_{\mathring{d}cz_t} + l_{z_t} \right)' \right.$
                $\left. + \left( {}^{\lfloor I}\theta_{\mathring{d}cz_t} - {}^{\lfloor U}\theta_{\mathring{d}z_t} \right) {}^{\lfloor U}r^{-1}_{\mathring{d}z_t} \left( {}^{\lfloor I}\theta_{\mathring{d}cz_t} - {}^{\lfloor U}\theta_{\mathring{d}z_t} \right)' \right\},$
                $k_{z_t} = k_{z_t} + {}^{\lfloor I}\alpha_c \left\{ {}^{\lfloor d}D_{cz_t} \, {}^{\lfloor I}r_{\mathring{d}cz_t} + \ln\left( \frac{{}^{\lfloor U}r_{\mathring{d}z_t}}{{}^{\lfloor I}r_{\mathring{d}cz_t}} \right) + \frac{{}^{\lfloor I}r_{\mathring{d}cz_t}}{{}^{\lfloor U}r_{\mathring{d}z_t}} \right\}.$
        end   of the cycle over $c$
                $\beta_{z_t} = {}^{\lfloor U}f(z_t) \exp(-0.5 k_{z_t}), \; q = q + \beta_{z_t}$
    end   of the cycle over $z_t$
                $L_{t-1} = I_{\mathring{\phi}}, \; D_{t-1} = 0_{\mathring{\phi},\mathring{\phi}}$
    For   $z_t = 1, \ldots, \mathring{d}_o$
                $\beta_{z_t} = \frac{\beta_{z_t}}{q}$
                $L_{t-1} D_{t-1} L'_{t-1} = L_{t-1} D_{t-1} L'_{t-1} + \beta_{z_t} L_{z_t} D_{z_t} L'_{z_t}.$
    end   of the cycle over $z_t$
end   of the cycle over $t$
```

Then, this presentation strategy minimizes the upper bound on the KL divergence, implied by the inequality described in Proposition 7.7 for horizon equal 1 and the γ-bound (see Proposition 7.8) on it when $\mathring{t} > 1$.

Proof. Let us assume that

$$-\ln(\gamma(d(t)) = 0.5 \sum_{z=1}^{\mathring{d}_o} \beta_z \phi'_t L_{z;t} D_{z;t} L'_{z;t} \phi_t = 0.5 \phi'_t L_t D_t L'_t \phi_t.$$

It is true for $t = \mathring{t}$ with $L_{\mathring{t}} = I_{\mathring{\phi}}$, $D_{\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$. As usual, the backward induction serves us for verifying this assumption and constructing the resulting algorithms.

For a generic t, the inductive assumption and the assumption (9.60) imply that the function, cf. (7.50),

$$\omega(c, d_{z_t;t}, d(t-1)) = \int f(d_{\bar{z}_t;t} | d_{z_t;t}, d(t-1), c) \times \ln\left(\frac{f(d_{\bar{z}_t;t} | d_{z_t;t}, d(t-1), c)}{\gamma(d(t)) \, {}^{\lfloor U}f(d_{\bar{z}_t;t} | d_{z_t;t}, d(t-1))} \right) dd_{\bar{z}_t;t}$$

has to be evaluated. Note that the first term in its definition is zero due to the assumption (9.60). Proposition 9.8 describes evaluation of the weighted conditional KL divergence. In order to make it applicable, we permute the considered entry $d_{z_t;t}$ on the last position in d_t. It is achieved by a sequential use of Proposition 9.1. At the same time, the corresponding permutations of the kernel determining $\gamma(d(t))$ have to be applied; see Proposition 8.4. These permutation have to be also made both for the normal components ${}^{\lfloor I}f(d_t|d(t-1),c)$ resulting from a previous design and for the user's ideal pdf ${}^{\lfloor U}f(d_t|d(t-1))$.

After all these permutations, the application of Proposition 9.8 gives recursions, applied for the dimension $\mathring{d}-1$ for computing $\omega(c, d_{z_t;t}, d(t-1)) = k_{c,z_t} + [d_{z_t;t}, \phi'_{t-1}]L_{cz_t;t-1}D_{cz_t;t-1}L'_{cz_t;t-1}[d_{z_t;t}, \phi'_{t-1}]'$.

The assumption (9.60) also implies that the second term in the definition of $\omega(z_t, d(t-1))$ (7.50) becomes

$$
{}^{\lfloor I}\mathcal{E}\left\{\sum_{c\in c^*}\alpha_c\left(k_{c,z_t} + [d_{z_t;t}, \phi'_{t-1}]L_{cz_t;t-1}D_{cz_t;t-1}L'_{cz_t;t-1}[d_{z_t;t}, \phi'_{t-1}]'\right)\middle| z_t, \phi_{t-1}, c_t\right\}
$$
$$
\equiv k_{c_t,z_t;t-1} + \phi'_{t-1}L_{c_tz_t;t-1}D_{c_tz_t;t-1}L'_{c_tz_t;t-1}\phi_{t-1},
$$

where ${}^{\lfloor I}\mathcal{E}\{\cdot\}$ stresses that the expectation is taken with respect to the optimal mixture ${}^{\lfloor I}f(d_{z_t;t}|d(t-1),c_t)$ obtained in the previous design of the advisory system.

The result is obtained in two steps. First, the computationally demanding mixing of factorized lifts k_{cz_t} and kernels $L_{cz_t;t-1}D_{cz_t;t-1}L'_{cz_t;t-1}$ is made. Then, the integration over $d_{z_t;t}$ is made according to Proposition 9.5. In the same cycle over $c \in c^*$, the conditional KL divergence is evaluated

$$
\int {}^{\lfloor I}f(d_{z_t;t}|d(t-1),c_t)\ln\left(\frac{{}^{\lfloor I}f(d_{z_t;t}|d_{p+;t}, d(t-1), c_t)}{{}^{\lfloor U}f(d_{z_t;t}|d(t-1),c_t)}\right) dd_{z_t;t}
$$
$$
= {}^{\lfloor I}k_{c_tz_t;t-1} + \phi'_{t-1}\,{}^{\lfloor I}L_{c_tz_t;t-1}\,{}^{\lfloor I}D_{c_tz_t;t-1}\,{}^{\lfloor I}L'_{c_tz_t;t-1}\phi_{t-1}
$$

using Proposition 9.7. Adopting the assumption (9.61), we mix intermediate results, see (7.50), and get

$$
\omega_\gamma(z_t, d(t-1)) = \sum_{c_t\in c^*}{}^{\lfloor I}\alpha_{c_t}\Big[\,{}^{\lfloor I}k_{c_tz_t;t-1} + k_{c_tz_t;t-1}
$$
$$
+\,\phi'_{t-1}\left(L_{c_tz_t;t-1}D_{c_tz_t;t-1}L'_{c_tz_t;t-1} + {}^{\lfloor I}L_{c_tz_t;t-1}\,{}^{\lfloor I}D_{c_tz_t;t-1}\,{}^{\lfloor I}L'_{c_tz_t;t-1}\right)\phi_{t-1}\Big]
$$
$$
\equiv {}^{\lfloor I}k_{z_t;t-1} + \phi'_{t-1}\,{}^{\lfloor I}L_{z_t;t-1}\,{}^{\lfloor I}D_{z_t;t-1}\,{}^{\lfloor I}L'_{z_t;t-1}\phi_{t-1}.
$$

It defines the pf on the presentation priorities. Its normalizing factor used in the next optimization steps is approximated by replacing the weighted arithmetic mean by its smaller geometric counterpart. It guarantees that an upper bound is minimized and the assumed form of γ is reproduced with

$$
\omega(d(t-1)) = \phi'_{t-1}\left[\sum_{z_t\in z^*}\beta_{z_t}\,{}^{\lfloor I}L_{z_t;t-1}\,{}^{\lfloor I}D_{z_t;t-1}\,{}^{\lfloor I}L'_{z_t;t-1}\right]\phi_{t-1}
$$

$$
\text{with } \beta_{z_t} \propto {}^{\lfloor U}f(z_t)\exp\left(-0.5\,{}^{\lfloor I}k_{z_t;t-1}\right).
$$

□

Algorithm 9.15 (Presentation with the bound (7.47) and $\mathring{\hat{z}}_t = 1$)

Initial (offline) mode

- *Estimate the normal mixture model of the o-system with the state ϕ_t in the phase form; Chapter 8.*
- *Specify the user's ideal pdf on d_t, c_t in the form*

$$\begin{aligned}
&{}^{\lfloor U}f(d_t, c_t|d(t-1)) = {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d_o(t-1)) \\
&\times f(\Delta_{p+;t}|u_{o;t}, d(t-1), c_t)\, {}^{\lfloor U}f(c_t|u_{o;t}, d(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1)) \\
&\propto \prod_{i=1}^{\mathring{\Delta}_o} \mathcal{N}_{\Delta_{i;t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\ {}^{\lfloor U}r_i\right) \prod_{i=\mathring{\Delta}_o+1}^{\mathring{\Delta}} \mathcal{N}_{\Delta_{i;t}}\left(\theta_i'\psi_{i;t}, r_i\right) \\
&\times \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta});t}}\left({}^{\lfloor U}\theta_i'\psi_{i;t},\ {}^{\lfloor U}r_i\right) \\
&\times\ {}^{\lfloor U}f(c_t)\exp\left[-0.5\left({}^{\lfloor U}k_{c_t} + \psi_t'\, {}^{\lfloor U}L_{c_t}\, {}^{\lfloor U}D_{c_t}\, {}^{\lfloor U}L_{c_t}'\psi_t\right)\right].
\end{aligned}$$

- *Select the user's ideal pf ${}^{\lfloor U}f(z_t|d(t-1))$ on the scalar priority actions $z_t \in z^* \equiv \{1, \ldots, \mathring{d}_o\}$.*
- *Select the number $\mathring{\hat{z}}$ of quantities to be presented to the operator. Typically, $\mathring{\hat{z}} < 10$.*
- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update estimates of the model parameters, Section 8.5, if you deal with the adaptive advisory system.*
3. *Design the academic, industrial or simultaneous strategy generating the ideal pdf*

$$\begin{aligned}
&{}^{\lfloor I}f(d_t|\phi_{t-1}) = \sum_{c_t\in c^*} {}^{\lfloor I}f(c_t|\phi_{t-1})\, {}^{\lfloor I}f(d_t|\phi_{t-1}, c_t) = \sum_{c_t\in c^*} \\
&\prod_{i=1}^{\mathring{\Delta}} \mathcal{N}_{d_{i;t}}(\theta_{ic}'\psi_{i;t}, r_{ic}) \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \mathcal{N}_{u_{o(i-\mathring{\Delta});t}}\left({}^{\lfloor I}\theta_{ic;t-1}'\psi_{i;t},\ {}^{\lfloor I}r_{ic;t-1}\right) {}^{\lfloor I}f(c_t|\phi_{t-1}), \\
&{}^{\lfloor I}f(c_{t+1}|\phi_t) \propto {}^{\lfloor U}f(c_{t+1})\exp\left[-0.5\left(k_{c_{t+1}} + \phi_{t-1}'L_{c_{t+1}}D_{c_{t+1}}L_{c_{t+1}}'\phi_{t-1}\right)\right].
\end{aligned}$$

4. *Initialize the iterative mode by setting $\tau = t+T$ and $L_\tau = I_{\mathring{\phi}}$, $D_\tau = 0_{\mathring{\phi},\mathring{\phi}}$. The initialization of L_τ, D_τ is skipped if $t > 1$ and the IST strategy is used.*

5. *Apply the iterative part of the algorithm described in Proposition 9.16 on the range $[t+1, t+T]$. It gives the pf $f(z_{t+1}|d(t))$.*
6. *Order the values $^{\lfloor I}f(z_{t+1}|\phi_t) \propto {}^{\lfloor U}f(z_{t+1}|\phi_t)\exp[-\omega_\gamma(z_{t+1},\phi_t)]$.*
7. *Present projections of the ideal pdf $^{\lfloor I}f(d_{\hat{z}_{t+1};t+1}|\phi_t)$ to the operator. The entries of $\mathring{\hat{z}}$-vector $\hat{z}_{t+1}$ are indexes $z_t \in \{1,\ldots,\mathring{d}_o\}$ chosen so that they have the highest values of $^{\lfloor I}f(z_{t+1}|\phi_t)$.*
8. *Go to the beginning of* Sequential mode*.*

Problem 9.3 (Improvements of presentation strategies) *The presented solution is far from being satisfactory. In spite of significant approximations made, the result is cumbersome and highly computationally demanding. Probably, a direct use of these approximations before estimating the KL divergence could simplify the solution.*

Alternatively, the academic, industrial or simultaneous designs with $d_{o;t} = d_{z;t}$ with several different zth in $\{1,\ldots,\mathring{d}_o\}$ could be performed and those options resulting to the smallest KL divergences presented. The logic behind this is obvious: the operator can optimize only those quantities of which he is aware. This justifies the reduction of his observable data space to the presented data space.

Of course, other directions have to be thought of, too.

9.3.2 Stimulating the operator

The signaling strategy makes the operator alert. It asks him to follow the advised actions, when the ideal pdf, resulting from the academic, industrial or simultaneous design, gives a significantly smaller KL divergence to the user's ideal pdf than the KL divergence of the estimated model to this user's ideal pdf.

The model relating the signaling action s_t to the response of the optimized guided o-system is

$$
\begin{aligned}
&^{\lfloor I}f(d_t,s_t|d(t-1)) \equiv {}^{\lfloor I}f(d_t|s_t,d(t-1))\,{}^{\lfloor I}f(s_t|d(t-1)),\ \ s_t \in s^* \equiv \{0,1\}\\
&^{\lfloor I}f(d_t|s_t=0,d(t-1)) \equiv f(d_t|d(t-1)) \equiv \underbrace{\sum_{c\in c^*}\alpha_c\prod_{i=1}^{\mathring{d}}\mathcal{N}_{d_{i;t}}\left(\theta'_{ic}\psi_{i;t},r_{ic}\right)}_{\text{learned mixture}}\\
&^{\lfloor I}f(d_t|s_t=1,d(t-1)) \equiv {}^{\lfloor I}f(d_t|d(t-1)) \qquad (9.64)\\
&\qquad\qquad \equiv \underbrace{\sum_{c_t\in c^*}{}^{\lfloor I}f(c_t|\phi_{t-1})\prod_{i=1}^{\mathring{d}}\mathcal{N}_{d_{i;t}}\left({}^{\lfloor I}\theta'_{ic}\psi_{i;t},\ {}^{\lfloor I}r_{ic}\right)}_{\text{designed mixture}}\\
&^{\lfloor I}f(c_t|\phi_{t-1}) \propto {}^{\lfloor U}f(c_t)\exp\left\{-0.5\left[k_{c_t;t-1}+\phi'_{t-1}L_{c_t;t-1}D_{c_t;t-1}L'_{c_t;t-1}\phi_{t-1}\right]\right\}.
\end{aligned}
$$

Note that $^{\lfloor I}\theta_{ic} = \theta_{ic},\ {}^{\lfloor I}r_{ic} = r_{ic}$ for $i \le \mathring{\Delta}$ as the ideal pdfs use the estimated factors predicting innovations without any change.

The model (9.64) is a special version of the model (9.35) used in the academic design. Thus, the design of the signaling strategy reduces to it. It is reasonable to assume that periods of operator activity or nonactivity have to be relatively long. Thus, the grouped version of the design (Proposition 7.13) is to be used.

Proposition 9.17 (Grouped signaling design with the γ-bound) *Let us consider that the academic, industrial or simultaneous design of the advisory system has provided the ideal pdf ${}^{\lfloor I}f(d_t|\phi_{t-1})$ that is built into the used model (9.64). Then, the following signaling strategy minimizes the γ-type bound on the KL divergence under the grouping constraint $f(s_\tau|d(tn)) = f(s_{tn+1}|d(tn))$ for $\tau \in \{tn+1, \ldots, t(n+1)\}$.*

$$f(s_{nt+1}|\phi_{nt}) \propto {}^{\lfloor U}f(s_{nt+1}|\phi_{nt})$$
$$\times \exp\left[-0.5\left(k_{s_{nt+1};nt+1} + \phi'_{nt} L_{s_{nt+1};nt+1} D_{s_{nt+1};nt+1} L'_{s_{nt+1};nt+1} \phi_{nt}\right)\right].$$

It is defined for $s_{nt+1} \in \{0,1\}$ recursively starting with the average kernel

$$L_{\gamma;\mathring{t}} = I_{\mathring{\phi}},\ D_{\gamma;\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}.$$

```
For  $t = \mathring{t}, \ldots, \mathring{t} - i \times n + 1, \ldots, 1$
          $q = 0$
  For  $s = 0, \ldots, 1$
          $L_{s;t-n} = I_{\mathring{\phi}},\ D_{s;t-n} = 0_{\mathring{\phi},\mathring{\phi}},\ k_s = 0$
    For  $c = 1, \ldots, \mathring{c}$
          $L_{0sc} \equiv L_{\gamma;t},\ D_{0sc} \equiv D_{\gamma;t},\ k_{0sc} = 0.$
      For  $\tilde{n} = 1, \ldots, n$
          $L_{0\tilde{n}sc} = \begin{bmatrix} {}^{\lfloor \phi 0}L_{\mathring{d}(\tilde{n}-1)sc} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor \phi 1}L'_{\mathring{d}(\tilde{n}-1)sc} & 0 & 1 \end{bmatrix}$
          $D_{00sc} = \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \phi 0}D_{\mathring{d}(\tilde{n}-1)sc}\right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_{\mathring{d}(\tilde{n}-1)sc}\right].$
          $k_{0\tilde{n}sc} = k_{\mathring{d}(\tilde{n}-1)sc}.$     It uses the split
          $L_{0(\tilde{n}-1)sc} \equiv \begin{bmatrix} {}^{\lfloor \phi 0}L_{\mathring{d}(\tilde{n}-1)sc} & 0 \\ {}^{\lfloor \phi 1}L'_{\mathring{d}(\tilde{n}-1)sc} & 1 \end{bmatrix}$
          $D_{0(\tilde{n}-1)sc} \equiv \mathrm{diag}\left[\mathrm{diag}\left[{}^{\lfloor \phi 0}D_{\mathring{d}(\tilde{n}-1)sc}\right], {}^{\lfloor 1}D_{\mathring{d}(\tilde{n}-1)sc}\right]$
          ${}^{\lfloor 1}D_{\mathring{d}(\tilde{n}-1)sc}$ is scalar.
        For  $i = 1, \ldots, \mathring{d}$
          $L_{(i-1)\tilde{n}sc} \equiv \begin{bmatrix} 1 & 0 \\ {}^{\lfloor \Delta\psi}L_{(i-1)\tilde{n}sc} & {}^{\lfloor \psi}L_{(i-1)\tilde{n}sc} \end{bmatrix}$
          $D_{(i-1)\tilde{n}sc} \equiv \mathrm{diag}\left[{}^{\lfloor \Delta}D_{(i-1)\tilde{n}sc}, \mathrm{diag}\left[{}^{\lfloor \psi}D_{(i-1)\tilde{n}sc}\right]\right]$
          ${}^{\lfloor\Delta}D_{(i-1)\tilde{n}sc}$ is scalar.
          $L_{i\tilde{n}sc}D_{i\tilde{n}sc}L'_{i\tilde{n}sc} = {}^{\lfloor\psi}L_{(i-1)\tilde{n}sc}\,{}^{\lfloor\psi}D_{(i-1)\tilde{n}sc}\,{}^{\lfloor\psi}L'_{(i-1)\tilde{n}sc}$
        $+\chi\left(i \le \mathring{\Delta}_o\right)$
        $\times\left[\left(\theta_{ic} + {}^{\lfloor\Delta\psi}L_{(i-1)\tilde{n}sc}\right){}^{\lfloor\Delta}D_{(i-1)\tilde{n}sc}\left(\theta_{ic} + {}^{\lfloor\Delta\psi}L_{(i-1)\tilde{n}sc}\right)'\right.$
        $\left.+\left(\theta_{ic} - {}^{\lfloor U}\theta_{ic}\right){}^{\lfloor U}r_i^{-1}\left(\theta_{ic} - {}^{\lfloor U}\theta_{ic}\right)'\right] + \chi\left(\mathring{\Delta}_o < i \le \mathring{\Delta}\right)$
        $\times\left(\theta_{isc} + {}^{\lfloor\Delta\psi}L_{(i-1)\tilde{n}sc}\right){}^{\lfloor\Delta}D_{(i-1)sc}\left(\theta_{isc} + {}^{\lfloor\Delta\psi}L_{(i-1)\tilde{n}sc}\right)'$
        $+\chi\left(\mathring{\Delta} < i\right)$
        $\times\left[\left(\theta_{isc} + {}^{\lfloor\Delta\psi}L_{(i-1)\tilde{n}sc}\right){}^{\lfloor\Delta}D_{(i-1)sc}\left(\theta_{isc} + {}^{\lfloor\Delta\psi}L_{(i-1)\tilde{n}sc}\right)'\right.$
        $\left.+\left(\theta_{isc} - {}^{\lfloor U}\theta_{ic}\right){}^{\lfloor U}r_i^{-1}\left(\theta_{isc} - {}^{\lfloor U}\theta_{ic}\right)'\right]$
        $k_{i\tilde{n}sc} = k_{i\tilde{n}sc} + \chi\left(i \le \mathring{\Delta}_o\right)$
        $\times\left(k_{(i-1)\tilde{n}sc} + {}^{\lfloor\Delta}D_{(i-1)\tilde{n}sc}r_{ic}\ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}}\right) + \frac{r_{ic}}{{}^{\lfloor U}r_i}\right)$
        $+\chi\left(\mathring{\Delta}_o < i \le \mathring{\Delta}\right){}^{\lfloor\Delta}D_{(i-1)\tilde{n}sc}r_{isc}$
        $+\chi\left(\mathring{\Delta} < i\right)\left({}^{\lfloor\Delta}D_{(i-1)\tilde{n}sc}r_{isc} + \ln\left(\frac{{}^{\lfloor U}r_i}{r_{isc}}\right) + \frac{r_{isc}}{{}^{\lfloor U}r_i}\right)$
      end    of the cycle over $i$
    end    of the cycle over $\tilde{n}$
  end    of the cycle over $c$
  For    $c = 1, \ldots, \mathring{c}$
          $L_{s;t-n}D_{s;t-n}L'_{s;t-n} = L_{s;t-n}D_{s;t-n}L'_{s;t-n}$
          $+\chi(s=0)\left\{\frac{\alpha_c}{n}L_{\mathring{d}nsc}D_{\mathring{d}nsc}L'_{\mathring{d}nsc}\right\}$
          $+\chi(s=1)\left\{{}^{\lfloor U}f(c)\left(\frac{1}{n}L_{\mathring{d}nsc}D_{\mathring{d}nsc}L'_{\mathring{d}nsc} + {}^{\lfloor I}L_c\,{}^{\lfloor I}D_c\,{}^{\lfloor I}L'_c\right)\right\}$
          $k_{s;t-n} = k_{s;t-n} + \chi(s=0)\frac{\alpha_c}{n}k_{\mathring{d}nsc}$
          $+\chi(s=1)\left\{{}^{\lfloor U}f(c)\left(\frac{1}{n}k_{\mathring{d}n1c} + {}^{\lfloor I}k_c\right)\right\}$
  end    of the cycle over $c$
          $\beta_s \equiv {}^{\lfloor U}f(s)\exp\left(-0.5k_{s;t-n}\right),\ q = q + \beta_s$
end    of the cycle over $s$
```

$$L_{\gamma;t-n}D_{\gamma;t-n}L'_{\gamma;t-n} \equiv \frac{\beta_0}{q}L_{0;t-n}D_{0;t-n}L'_{0;t-n} + \frac{\beta_1}{q}L_{1;t-n}D_{1;t-n}L'_{1;t-n}.$$

end of the cycle over t

Proof. Let us assume that $-2\ln(\gamma(d(n(t+1))))$

$$= \sum_{s_{n(t+1)}\in s^*} \beta_{s_{n(t+1)};t}\phi'_{n(t+1)}L_{s_{n(t+1)};n(t+1)}D_{s_{n(t+1)};n(t+1)}L'_{s_{n(t+1)};n(t+1)}\phi_{n(t+1)}$$

$$\equiv \phi'_{n(t+1)}L_{\gamma;n(t+1)}D_{\gamma;n(t+1)}L'_{\gamma;n(t+1)}\phi_{n(t+1)},$$

$$\beta_{s_{n(t+1)};n(t+1)} \propto {}^{\lfloor U}f(s_{n(t+1)})\exp[-0.5k_{s_{n(t+1)};n(t+1)}].$$

The lifts $k_{s_t;t}$ as well as the kernels are assumed to be independent data. It is true for $n(t+1) = \mathring{t}$ with $k_{s_{\mathring{t}};\mathring{t}} = 0,\ L_{\gamma;\mathring{t}} = I_{\mathring{\phi}},\ D_{\gamma;\mathring{t}} = 0_{\mathring{\phi},\mathring{\phi}}$. We use the backward induction to verify this assumption and to derive the design algorithm.

Let us adopt the inductive assumption, approximate the KL divergence from above using Jensen inequality (2.14) and evaluate, cf. Proposition 7.13,

$$\omega_\gamma(s_{nt+1} = s, c, \phi_{nt}) \equiv \mathcal{E}\left[\frac{1}{n}\sum_{\tau=nt+1}^{n(t+1)}\int f(d_\tau|d(\tau-1), s_{nt+1}, c)\right.$$

$$\left.\times \ln\left(\frac{f(d_{o;\tau}|d_{p+;\tau}, \phi_{\tau-1}, s_{nt+1}, c)}{\gamma(\phi_{n(t+1)})\,{}^{\lfloor U}f(d_{o;\tau}|\phi_{\tau-1})}\right) dd_\tau|\phi_{nt}\right]$$

$$\underbrace{=}_{\text{Proposition 9.7}} k_{n\mathring{d}sc} + \phi'_{nt}L_{n\mathring{d}sc}D_{n\mathring{d}sc}L'_{n\mathring{d}sc}\phi'_{nt}$$

With kernels and lifts found recursively starting from

$$L_{0sc} \equiv L_{\gamma;nt},\ D_{0sc} \equiv D_{\gamma;nt},\ k_{0sc} = 0.$$

For $\tilde{n} = 1, \ldots, n$

$$L_{0\tilde{n}sc} = \begin{bmatrix} {}^{\lfloor\phi 0}L_{\mathring{d}(\tilde{n}-1)sc} & 0 & 0 \\ 0 & I_{\mathring{d}} & 0 \\ {}^{\lfloor\phi 1}L'_{\mathring{d}(\tilde{n}-1)sc} & 0 & 1 \end{bmatrix}$$

$$D_{00sc} = \operatorname{diag}\left[\operatorname{diag}\left[{}^{\lfloor\phi 0}D_{\mathring{d}(\tilde{n}-1)sc}\right], 0_{1,\mathring{d}}, {}^{\lfloor 1}D_{\mathring{d}(\tilde{n}-1)sc}\right]$$

$$k_{0\tilde{n}sc} = k_{\mathring{d}(\tilde{n}-1)sc}$$

For $i = 1, \ldots, \mathring{d}$

$$L_{i\tilde{n}sc}D_{i\tilde{n}sc}L'_{i\tilde{n}sc} = {}^{\lfloor\psi}L_{(i-1)\tilde{n}sc}\,{}^{\lfloor\psi}D_{(i-1)\tilde{n}sc}\,{}^{\lfloor\psi}L'_{(i-1)\tilde{n}sc} + \chi\left(i \le \mathring{\varDelta}_o\right)$$

$$\times\left[\left(\theta_{ic} + {}^{\lfloor\varDelta\psi}L_{(i-1)\tilde{n}sc}\right){}^{\lfloor\varDelta}D_{(i-1)\tilde{n}sc}\left(\theta_{ic} + {}^{\lfloor\varDelta\psi}L_{(i-1)\tilde{n}sc}\right)'\right.$$

$$\left.+\left(\theta_{ic} - {}^{\lfloor U}\theta_{ic}\right){}^{\lfloor U}r_i^{-1}\left(\theta_{ic} - {}^{\lfloor U}\theta_{ic}\right)'\right] + \chi\left(\mathring{\varDelta}_o < i \le \mathring{\varDelta}\right)$$

```
        $\times \left(\theta_{isc} + {}^{\lfloor \Delta\psi}L_{(i-1)\tilde{n}sc}\right) {}^{\lfloor \Delta}D_{(i-1)sc} \left(\theta_{isc} + {}^{\lfloor \Delta\psi}L_{(i-1)\tilde{n}sc}\right)'$
        $+\chi\left(\mathring{\Delta} < i\right) \left[\left(\theta_{isc} + {}^{\lfloor \Delta\psi}L_{(i-1)\tilde{n}sc}\right) {}^{\lfloor \Delta}D_{(i-1)sc} \left(\theta_{isc} + {}^{\lfloor \Delta\psi}L_{(i-1)\tilde{n}sc}\right)'\right.$
        $\left. + \left(\theta_{isc} - {}^{\lfloor U}\theta_{ic}\right) {}^{\lfloor U}r_i^{-1} \left(\theta_{isc} - {}^{\lfloor U}\theta_{ic}\right)'\right], \quad k_{i\tilde{n}sc} = k_{i\tilde{n}sc}$
        $+\chi\left(i \le \mathring{\Delta}_o\right) \left(k_{(i-1)\tilde{n}sc} + {}^{\lfloor \Delta}D_{(i-1)\tilde{n}sc} r_{ic} \ln\left(\frac{{}^{\lfloor U}r_i}{r_{ic}}\right) + \frac{r_{ic}}{{}^{\lfloor U}r_i}\right)$
        $+\chi\left(\mathring{\Delta}_o < i \le \mathring{\Delta}\right) {}^{\lfloor \Delta}D_{(i-1)\tilde{n}sc} r_{isc}$
        $+\chi\left(\mathring{\Delta} < i\right) \left({}^{\lfloor \Delta}D_{(i-1)\tilde{n}sc} r_{isc} + \ln\left(\frac{{}^{\lfloor U}r_i}{r_{isc}}\right) + \frac{r_{isc}}{{}^{\lfloor U}r_i}\right)$
    end    of the cycle over $i$
end    of the cycle over $\tilde{n}$
```

Note that the splitting to subsets respects both the distinction of the o- and p-data and the fact that parameters of factors predicting innovations are independent of the signaling action s_t. For specification of the optimal strategy, it remains to mix contributions of distances of individual components. We take into account the form of the component weights for the estimated and designed models. For it we set, $L_0 = L_1 = I_{\mathring{\phi}}$, $D_0 = D_1 = 0_{\mathring{\phi},\mathring{\phi}}$, $k_0 = k_1 = 0$.

```
For    $c = 1, \ldots, \mathring{c}$
        $L_0 D_0 L_0' = L_0 D_0 L_0' + \frac{\alpha_c}{n} L_{\mathring{d}n0c} D_{\mathring{d}n0c} L'_{\mathring{d}n0c}, \; k_0 = k_0 + \frac{\alpha_c}{n} k_{\mathring{d}n0c}$
        $L_1 D_1 L_1' = L_1 D_1 L_1' + {}^{\lfloor U}f(c) \left(\frac{1}{n} L_{\mathring{d}n1c} D_{\mathring{d}n1c} L'_{\mathring{d}n1c} + {}^{\lfloor I}L_c \, {}^{\lfloor I}D_c \, {}^{\lfloor I}L_c'\right)$
        $k_1 = k_1 + {}^{\lfloor U}f(c) \left(\frac{1}{n} k_{\mathring{d}n1c} + {}^{\lfloor I}k_c\right).$
end    of the cycle over $c$
```

The standard minimization implies that the optimal strategy is

$$ {}^{\lfloor I}f(s_{nt+1}|d(nt)) \propto {}^{\lfloor U}f(s_t) \exp\left[-0.5\left(k_{s_{nt+1}} + \phi'_{nt} L_{s_{nt+1}} D_{s_{nt+1}} L'_{s_{nt+1}} \phi'_{nt}\right)\right]. $$

Its normalizing factor $\gamma(\phi_{nt})$ defines the exact Bellman function to be transferred to the next optimization step. Use of the inequality between the arithmetic and geometric means preserves the assumed form of $-\ln(\gamma(d(nt))$ while guaranteeing that the upper bound is minimized. □

Remark(s) 9.10

1. *The probability* ${}^{\lfloor I}f(s_{tn+1} = 0|\phi_{tn})$ *is to be mapped on traffic lights. Psychologically, it seems to be reasonable to use a nonlinear mapping given by a pair of thresholds* $0 < w < W < 1$

$$\begin{aligned}&\textit{Call for action if } {}^{\lfloor I}f(s_{tn+1}=0|\phi_{tn}) \leq w. \\ &\textit{Allow nonaction if } {}^{\lfloor I}f(s_{tn+1}=0|\phi_{tn}) \geq W > w. \\ &\textit{Otherwise, do not change current signalling action.}\end{aligned} \tag{9.65}$$

2. *The common factors of both models can be exploited when the grouping rate $n = 1$. Otherwise, this property brings no advantage.*
3. *Attempts to use the ω-bound were done but with no significant influence on the algorithm. It is worth checking possible alternatives.*

The result is rather complex. Thus, a fixed version of signaling will be probably used. Let us write down the corresponding fixed signaling algorithm.

Algorithm 9.16 (Fixed signaling based on Proposition 9.17)
Initial (offline) mode

- *Estimate normal mixture model of the o-system with the state ϕ_t; see Chapter 8.*
- *Specify the true user's ideal pdf on $d_{o;t}, c_t$.*
- *Specify the user "alarm probability" pf ${}^{\lfloor U}f(s_t = 0)$.*
- *Perform a fixed academic, industrial or simultaneous design.*
- *Compute steady-state lifts k_s and kernels $L_s, D_s,\ s \in \{0, 1\}$ according to Proposition 9.17.*
- *Specify thresholds w, W determining the mapping (9.65).*

Sequential (online) mode, *running for $t = 1, 2, \ldots$,*

1. *Acquire d_t and evaluate the state ϕ_t.*
2. *Evaluate signaling probabilities*

$$f(s_{t+1}|\phi_t) \propto {}^{\lfloor U}f(s_{t+1}) \exp\left[-0.5\left(k_{s_{t+1}} + \phi_t' L_{s_{t+1}} D_{s_{t+1}} L_{s_{t+1}}' \phi_t\right)\right].$$

3. *Convert the value $f(s_{t+1} = 0|\phi_t)$ into color using the nonlinearity (9.65). The call action is the stronger the closer is $f(s_{t+1} = 0|\phi_t)$ to zero.*
4. *Go to the beginning of* Sequential mode*.*

Problem 9.4 (Completion and development of advising design) *The described algorithms cover all formulated tasks. The overall state is, however, far from being satisfactory. The following problems have to be at least addressed*

- *completion and comparison of all variants,*
- *design of new variants,*
- *simplification,*
- *creation of a guide on choosing of tuning knobs,*
- *inclusion of stopping rules,*
- *analysis of theoretical properties known from control area (like controllability),*
- *robustness analysis, . . .*

10

Learning with Markov-chain factors and components

Markov-chain factors allow us to incorporate logical quantities and to work with mixed real and discrete components. Mixtures consisting solely of Markov-chain components form a Markov chain. In spite of this, the use of mixtures makes sense as they have the potential to provide parsimonious parameterization. A detailed description of all learning aspects related to Markov-chain factors, components and their mixtures forms the content of this chapter.

We deal with Markov-chain factors predicting scalar discrete-valued $d_t \in d^* \equiv \{1, \ldots, \mathring{d}\}$. The factor is described by the pf

$$f(d_t|d(t-1), \Theta) \equiv f(d_t|\psi_t, \Theta) = \prod_{d \in d^*} \prod_{\psi \in \psi^*} \left[{}^{\lfloor d|\psi}\Theta \right]^{\delta_{[d,\psi']', \Psi_t}}, \text{ where } \quad (10.1)$$

$\psi_t \in \psi^* \equiv \{1, \ldots, \mathring{\psi} < \infty\}$ is a finite-dimensional regression vector with a finite number of different values that can be constructed in a recursive way from the observed data;

$\Psi_t \equiv [d_t, \psi_t']'$ is a finite-dimensional, finite-valued data vector;

$\Theta \in \Theta^* \equiv \left\{ {}^{\lfloor d|\psi}\Theta \geq 0,\ \sum_{d \in d^*} {}^{\lfloor d|\psi}\Theta = 1,\ \forall \psi \in \psi^* \right\}$ are unknown transition probabilities;

$\delta_{\cdot,\cdot}$ is the Kronecker symbol that equals 1 for identical arguments and it is zero otherwise.

This chapter layout copies that of the general Chapter 6. Common tools in Section 10.1 concern mainly properties of the Dirichlet pdf that is the conjugate one to the Markov-chain model. Data preprocessing, Section 10.2, adds to common operations those specific to Markov chains, namely, quantization of continuous-valued signals and coding of discrete regression vectors. After discussion of prior-knowledge elicitation, Section 10.3, the key construction of the prior pdf is specialized to Markov chains in Section 10.4. Adequate versions of approximate estimation, Section 10.5, structure estimation, Section 10.6, and model validation, Section 10.7, form the rest of this chapter.

It is fair to forewarn that the presented algorithms are elaborated much less than their normal counterpart. Applicability of the general approaches of

Chapter 6 and presentation of a wide range of open research problems form the main messages of this chapter.

10.1 Common tools

10.1.1 Dirichlet pdf as a conjugate prior

Markov-chain factors belong to the exponential family (see Section 3.2) so that they possess a conjugate prior. The following correspondence to (3.6) holds

$$
\begin{aligned}
f(d_t|d(t-1),\Theta) &= \exp\left[\sum_{[d,\psi']'=\Psi\in\Psi^*} \delta_{[d,\psi']',\Psi_t} \ln\left({}^{\lfloor d|\psi}\Theta\right)\right]\\
&\equiv A(\Theta)\exp\left[\langle B(\Psi_t), C(\Theta)\rangle + D(\Psi)\right]
\end{aligned}
$$

$$
A(\Theta)\equiv 1,\quad B_{\tilde{\Psi}}(\Psi)\equiv\delta_{\tilde{\Psi},\Psi},\ C(\Theta_{\tilde{\Psi}}) = \ln\left({}^{\lfloor\tilde{d}|\tilde{\psi}}\Theta\right),\ \text{with } \tilde{\Psi}\equiv\left[\tilde{d},\tilde{\psi}'\right]'\in\Psi^*,
$$

$$
D(\Psi)=0,\quad \langle B,C\rangle \equiv \sum_{\tilde{\Psi}\in\Psi^*} B_{\tilde{\Psi}}(\Psi)C_{\tilde{\Psi}}(\Theta).
$$

This correspondence determines the conjugate prior (3.13) in the form known as the *Dirichlet pdf*

$$
Di_\Theta(V)\equiv f(\Theta|V) = \frac{\prod_{\Psi=[d,\psi']'\in\Psi^*}\left[{}^{\lfloor d|\psi}\Theta\right]^{{}^{\lfloor d|\psi}V-1}\chi_{\Theta^*}(\Theta)}{\mathcal{I}(V)}, \tag{10.2}
$$

where $\chi_{\Theta^*}(\Theta)$ is indicator of Θ^*.

The value of the normalizing integral $\mathcal{I}(V)$ and constraint on the statistic $V = [\,{}^{\lfloor d|\psi}V]_{d\in d^*,\psi\in\psi^*}$ that guarantees finiteness of $\mathcal{I}(V)$ are described below together with other properties of this important pdf.

Within this chapter the indexes at Θ (similarly as for statistic V) corresponding to the treated data vector $\Psi = [d',\psi']'$ are placed as the left upper index ${}^{\lfloor d|\psi}$. It makes iterative formulas with a lot of other indexes a bit more readable.

Proposition 10.1 (Basic properties and moments of *Di* pdf) *The normalization integral of the Dirichlet pdf $Di_\Theta(V)$ is*

$$
\begin{aligned}
\mathcal{I}(V) &= \prod_{\psi\in\psi^*}\mathcal{B}\left({}^{\lfloor\cdot|\psi}V\right) \equiv \prod_{\psi\in\psi^*}\frac{\prod_{d\in d^*}\Gamma\left({}^{\lfloor d|\psi}V\right)}{\Gamma\left(\sum_{d\in d^*}{}^{\lfloor d|\psi}V\right)} \\
\Gamma(x) &\equiv \int_0^\infty z^{x-1}\exp(-z)\,dz < \infty \textit{ for } x>0.
\end{aligned}
\tag{10.3}
$$

Thus, the normalization factor is finite iff ${}^{\lfloor d|\psi}V>0,\ \forall\Psi\equiv[d,\psi']'\in\Psi^*$. *This condition is met for all posterior statistics* ${}^{\lfloor d|\psi}V_t$ *if* ${}^{\lfloor d|\psi}V_0>0$.

The Dirichlet pdf has the following marginal pdfs and moments $Di_{{}^{\lfloor d|\psi}\Theta}(V)$

$$\equiv f\left({}^{\lfloor d|\psi}\Theta|V\right) = \frac{\left[{}^{\lfloor d|\psi}\Theta\right]^{{}^{\lfloor d|\psi}V-1}\left[1-{}^{\lfloor d|\psi}\Theta\right]^{{}^{\lfloor\psi}\nu-{}^{\lfloor d|\psi}V-1}\chi_{[0,1]}\left({}^{\lfloor d|\psi}\Theta\right)}{\mathcal{I}\left({}^{\lfloor d|\psi}V\right)}$$

$${}^{\lfloor\psi}\nu \equiv \sum_{d\in d^*} {}^{\lfloor d|\psi}V,\quad \mathcal{I}\left({}^{\lfloor d|\psi}V\right) \equiv \frac{\Gamma\left({}^{\lfloor d|\psi}V\right)\Gamma\left({}^{\lfloor\psi}\nu-{}^{\lfloor d|\psi}V\right)}{\Gamma\left({}^{\lfloor\psi}\nu\right)} \tag{10.4}$$

$$\mathcal{E}\left[{}^{\lfloor d|\psi}\Theta\middle| V,\Psi\right] = \frac{{}^{\lfloor d|\psi}V}{{}^{\lfloor\psi}\nu}$$

$$\mathcal{E}\left[{}^{\lfloor d|\psi}\Theta\,{}^{\lfloor\tilde{d}|\tilde{\psi}}\Theta\middle| V,\Psi,\tilde{\Psi}\right] = (1-\delta_{\Psi,\tilde{\Psi}})\mathcal{E}\left[{}^{\lfloor d|\psi}\Theta\middle| V,\Psi\right]\mathcal{E}\left[{}^{\lfloor\tilde{d}|\tilde{\psi}}\Theta\middle| V,\tilde{\Psi}\right]$$

$$+\delta_{\Psi,\tilde{\Psi}}\left\{\mathcal{E}\left[{}^{\lfloor d|\psi}\Theta\middle| V,\Psi\right]\right\}^2\frac{1+{}^{\lfloor d|\psi}V^{-1}}{1+{}^{\lfloor\psi}\nu^{-1}}$$

$$\operatorname{cov}\left[{}^{\lfloor d|\psi}\Theta,\ {}^{\lfloor\tilde{d}|\tilde{\psi}}\Theta\middle| V,\Psi,\tilde{\Psi}\right] = \delta_{\Psi,\tilde{\Psi}}\left\{\mathcal{E}\left[{}^{\lfloor d|\psi}\Theta\middle| V,\Psi\right]\right\}^2\frac{{}^{\lfloor d|\psi}V^{-1}-{}^{\lfloor\psi}\nu^{-1}}{1+{}^{\lfloor\psi}\nu^{-1}}$$

$$\mathcal{E}\left[\ln\left({}^{\lfloor d|\psi}\Theta\right)\middle| V,\Psi\right] = \frac{\partial}{\partial\,{}^{\lfloor d|\psi}V}\ln\left(\Gamma\left({}^{\lfloor d|\psi}V\right)\right) - \frac{\partial}{\partial\,{}^{\lfloor\psi}\nu}\ln\left(\Gamma\left({}^{\lfloor\psi}\nu\right)\right). \tag{10.5}$$

The predictive pdf is $f(d|\psi,V) = \dfrac{{}^{\lfloor d|\psi}V}{{}^{\lfloor\psi}\nu} \equiv \mathcal{E}\left[{}^{\lfloor d|\psi}\Theta\middle| V,\Psi\right].$ (10.6)

Proof. Evaluation of the normalization integral is the basic step. It can be done inductively starting with $\mathring{d} = 2$, using the definition of the Euler beta function, coinciding with $\mathcal{B}$ for $\mathring{d} = 2$ and its relationship to the gamma function $\Gamma(\cdot)$ defined by (10.3). The rest relies mostly on the known recursion $\Gamma(x+1) = x\Gamma(x)$, [156].

The evaluation $\mathcal{E}\left[\ln\left({}^{\lfloor d|\psi}\Theta\right)\middle| V,\Psi\right]$ is the only involved step. To demonstrate it, let us assume that $\Theta \sim Di_{\Theta}(m,n) = \mathcal{B}^{-1}(m,n)\Theta^{m-1}(1-\Theta)^{n-1}$. Then,

$$\begin{aligned}\mathcal{E}[\ln(\Theta)] &= \mathcal{B}^{-1}(m,n)\int_0^1 \ln(\Theta)\Theta^{m-1}(1-\Theta)^{n-1}\,d\Theta\\ &= \mathcal{B}^{-1}(m,n)\int_0^1 \frac{\partial}{\partial m}\Theta^{m-1}(1-\Theta)^{n-1}\,d\Theta\\ &= \mathcal{B}^{-1}(m,n)\frac{\partial}{\partial m}\mathcal{B}(m,n) = \frac{\partial}{\partial m}\ln(\mathcal{B}(m,n))\\ &= \frac{\partial}{\partial m}\ln(\Gamma(m)) - \frac{\partial}{\partial(m+n)}\ln(\Gamma(m+n)).\end{aligned}$$

Substitution $m = {}^{\lfloor d|\psi}V$ and $m+n = {}^{\lfloor\psi}\nu$ gives the result (10.5). □

We need to know the KL divergence of a pair of Dirichlet pdfs. Parameters corresponding to different regression vectors are independent. Consequently, we can consider a fixed, and for notation simplicity, empty regression vector.

Proposition 10.2 (KL divergence of Di pdfs) *Let $f(\Theta) = Di_{\Theta}(V)$, $\tilde{f}(\Theta) = Di_{\Theta}\left(\tilde{V}\right)$ be a pair of Dirichlet pdfs describing parameters*

$$\Theta \equiv \left({}^{\lfloor 1}\Theta, \ldots, {}^{\lfloor \mathring{d}}\Theta\right) \in \Theta^* = \left\{ {}^{\lfloor d}\Theta > 0,\ \sum_{d \in d^*} {}^{\lfloor d}\Theta = 1 \right\},\ d^* \equiv \left\{1, \ldots, \mathring{d}\right\}.$$

Then, their KL divergence is given by the formula

$$\begin{aligned}\mathcal{D}(f\|\tilde{f}) = &\sum_{d=1}^{\mathring{d}} \left[\left({}^{\lfloor d}V - {}^{\lfloor d}\tilde{V}\right) \frac{\partial}{\partial\, {}^{\lfloor d}V} \ln\left(\Gamma\left({}^{\lfloor d}V\right)\right) + \ln\left(\frac{\Gamma\left({}^{\lfloor d}\tilde{V}\right)}{\Gamma\left({}^{\lfloor d}V\right)}\right)\right] \\ &- (\nu - \tilde{\nu})\frac{\partial}{\partial \nu}\ln(\Gamma(\nu)) + \ln\left(\frac{\Gamma(\nu)}{\Gamma(\tilde{\nu})}\right) \\ &\nu \equiv \sum_{d=1}^{\mathring{d}} {}^{\lfloor d}V,\quad \tilde{\nu} \equiv \sum_{d=1}^{\mathring{d}} {}^{\lfloor d}\tilde{V}. \end{aligned} \tag{10.7}$$

Proof. The majority of evaluations is prepared by Proposition 10.1. For the considered pdfs, the KL divergence (2.25) gets the form (use (10.3), (10.4))

$$\begin{aligned}\mathcal{D} &= \mathcal{E}\left[\sum_{d=1}^{\mathring{d}} \left({}^{\lfloor d}V - {}^{\lfloor d}\tilde{V}\right) \ln\left({}^{\lfloor d}\Theta\right) | V\right] - \ln\left(\frac{\mathcal{B}(V)}{\mathcal{B}(\tilde{V})}\right) \\ &= \sum_{d=1}^{\mathring{d}} \left({}^{\lfloor d}V - {}^{\lfloor d}\tilde{V}\right) \mathcal{E}\left[\ln\left({}^{\lfloor d}\Theta\right) | \, {}^{\lfloor d}V\right] - \ln\left(\frac{\mathcal{B}(V)}{\mathcal{B}(\tilde{V})}\right) \\ &\underbrace{=}_{(10.5)} \sum_{d=1}^{\mathring{d}} \left[\left({}^{\lfloor d}V - {}^{\lfloor d}\tilde{V}\right)\left(\frac{\partial}{\partial\, {}^{\lfloor d}V}\ln\left(\Gamma\left({}^{\lfloor d}V\right)\right) - \frac{\partial}{\partial\nu}\ln(\Gamma(\nu))\right)\right] - \ln\left(\frac{\mathcal{B}(V)}{\mathcal{B}\left(\tilde{V}\right)}\right) \\ &= \sum_{d=1}^{\mathring{d}} \left({}^{\lfloor d}V - {}^{\lfloor d}\tilde{V}\right) \frac{\partial}{\partial\, {}^{\lfloor d}V}\ln\left(\Gamma\left({}^{\lfloor d}V\right)\right) - (\nu - \tilde{\nu})\frac{\partial}{\partial\nu}\ln(\Gamma(\nu)) - \ln\left(\frac{\mathcal{B}(V)}{\mathcal{B}(\tilde{V})}\right) \\ &= \sum_{d=1}^{\mathring{d}} \left[\left({}^{\lfloor d}V - {}^{\lfloor d}\tilde{V}\right) \frac{\partial}{\partial\, {}^{\lfloor d}V}\ln\left(\Gamma\left({}^{\lfloor d}V\right)\right) + \ln\left(\frac{\Gamma\left({}^{\lfloor d}\tilde{V}\right)}{\Gamma\left({}^{\lfloor d}V\right)}\right)\right] \\ &-(\nu - \tilde{\nu})\frac{\partial}{\partial\nu}\ln(\Gamma(\nu)) + \ln\left(\frac{\Gamma(\nu)}{\Gamma(\tilde{\nu})}\right).\end{aligned}$$

There, we have used linearity of expectation, the definitions of the beta function $\mathcal{B}$ and of the statistic $\nu \equiv \sum_{d \in d^*} {}^{\lfloor d}V$, $\tilde{\nu} \equiv \sum_{d \in d^*} {}^{\lfloor d}\tilde{V}$. □

10.1.2 Estimation and prediction with Markov-chain factors

The Markov-chain factors predicting discrete-valued d_t belong to the exponential family. Thus, their estimation with the conjugate prior pdf and subsequent prediction reduce to algebraic operations; see Proposition 3.2.

Proposition 10.3 (Estimation and prediction with Markov chains) *Let us consider a fixed discrete value ψ of the regression vector. Let natural conditions of decision making, Requirement 2.5, hold, the treated factor (10.1) is Markov-chain and a Dirichlet conjugate prior $Di_{\Theta}(V_0)$ (3.13), $V_0 \equiv \left[{}^{\lfloor 1|\psi}V_0, \ldots, {}^{\lfloor \mathring{d}|\psi}V_0 \right]$ as well as conjugate alternatives $Di_{\Theta}({}^{\lfloor A}V_t)$, ${}^{\lfloor A}V_t \equiv \left[{}^{\lfloor 1|\psi\, A}V_t, \ldots, {}^{\lfloor \mathring{d}|\psi\, A}V_t \right]$ with the forgetting factor $\lambda \in [0,1]$, are used in stabilized forgetting; see Section 3.1. Then, the posterior pdf is also the Dirichlet pdf $Di_{\Theta}(V_t)$ with*

$$V_t \equiv \left[{}^{\lfloor 1|\psi}V_t, \ldots, {}^{\lfloor \mathring{d}|\psi}V_t \right].$$

The sufficient statistic V_t evolves according to the recursion

$$\begin{aligned} {}^{\lfloor d|\psi}V_t &= \lambda \left({}^{\lfloor d|\psi}V_{t-1} + \delta_{\Psi,\Psi_t} \right) + (1-\lambda)\, {}^{\lfloor d|\psi\, A}V_t \\ & {}^{\lfloor d|\psi}V_0 \text{ given, } [d, \psi']' \equiv \Psi \in \Psi^*. \end{aligned} \tag{10.8}$$

The predictive pdf is

$$f(d|\psi, V_t) = \frac{{}^{\lfloor d|\psi}V_t}{\sum_{\tilde{d} \in d^*} {}^{\lfloor \tilde{d}|\psi}V_t}. \tag{10.9}$$

Proof. It is a direct consequence of the Bayes rule applied to the member of the exponential family. □

Remark(s) 10.1

1. *The array $V = \left[{}^{\lfloor d|\psi}V \right]_{d \in d^*, \psi \in \psi^*}$ is also known as the* occurrence table.
2. *The obtained prediction corresponds with understanding of the probability as a relative frequency of events. Here, the data are just split in subsets corresponding to different values of regression vectors ψ. Numerically, the Bayesian set up adds nontrivial initial conditions to the occurrence table.*
3. *The predictive pdf is the key element needed for approximate estimation; see Section 10.5.*

Problem 10.1 (Reduction of dimensionality) *The dimensionality of the occurrence table might be easily excessive, mostly because of the high cardinality $\mathring{\psi}$. Then, a reduced parameterization is necessary. Promising techniques exploiting a sparse nature of this table are described in [157, 165]. They could be a reasonable start for solving this practically important problem.*

Problem 10.2 (Relation to nonparametric Bayesian estimation) *Markov chains have a tight connection to* nonparametric Bayesian estimation, *[166]. Inspection whether this connection can be used in our context is worth considering. It could extend applicability of the achieved results significantly.*

10.1.3 Likelihood on variants

Predictors based on Markov chains serve for comparing competitive descriptions of the same discrete valued data. Proposition 10.1 provides immediately a Markov-chain variant of Proposition 6.1.

Proposition 10.4 (Markov-chain mixture *one-step-ahead predictor*)

Use of online estimation of a Markov-chain mixture within the framework of the adaptive advisory system *provides a one-step-ahead predictor in the form*

$$\begin{aligned} f(d_{t+1}|d(t)) &= \sum_{c\in c^*} \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c};t}} \prod_{i\in i^*} f(d_{ic;t+1}|d(t),c) \qquad (10.10)\\ &\equiv \sum_{c\in c^*} \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c};t}} \prod_{i\in i^*} \frac{\mathcal{I}(d(t+1)|ic)}{\mathcal{I}(d(t)|ic)}\\ &\equiv \sum_{c\in c^*} \hat{\alpha}_{c;t} \prod_{i\in i^*} f(d_{ic;t+1}|\psi_{ic;t+1},d(t),c), \quad \hat{\alpha}_{c;t} \equiv \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c};t}}. \end{aligned}$$

$$\mathcal{I}(d(t)|ic) \equiv \mathcal{I}(V_{ic;t}) = \mathcal{B}(V_{ic;t}) \equiv \frac{\prod_{d_{ic}\in d_{ic}^*} \Gamma\left({}^{\lfloor d_{ic}|\psi_{ic}}V_{ic;t}\right)}{\Gamma\left(\sum_{d_{ic}\in d_{ic}^*} {}^{\lfloor d_{ic}|\psi_{ic}}V_{ic;t}\right)},$$

$$\Gamma(x) = \int_0^\infty z^{x-1}\exp(-z)\,dz < \infty \text{ for } x>0, \text{ where}$$

$${}^{\lfloor d_{ic}|\psi_{ic}}V_{ic;t} = \text{the occurrence table of the factor } ic \text{ collected up to time } t.$$

For the fixed advisory system, the predictor is given by the last equality in (10.10) *with* $f(d_{ic;t+1}|\psi_{ic;t+1},d(t),c)$, $\hat{\alpha}_{c;t}$ *replaced by* $f(d_{ic;t+1}|\psi_{ic;t+1},d(0),c)$, $\hat{\alpha}_{c;0}$. *The used factor predictors can be expressed as follows.*

$$f(d_{ic;t+1}|\psi_{ic;t+1},d(t),c) = \frac{{}^{\lfloor d_{ic;t+1}|\psi_{ic;t+1}}V_{ic;t}}{\sum_{d_{ic}\in d_{ic}^*} {}^{\lfloor d_{ic}|\psi_{ic;t+1}}V_{ic;t}}, \text{ for an adaptive advising,}$$

$$f(d_{ic;t+1}|\psi_{ic;t+1},d(0),c) = \frac{{}^{\lfloor d_{ic;t+1}|\psi_{ic;t+1}}V_{ic;0}}{\sum_{d_{ic}\in d_{ic}^*} {}^{\lfloor d_{ic}|\psi_{ic;t+1}}V_{ic;0}}, \text{ for a fixed advising.}$$

Proof. Omitted. □

10.1.4 Branch-and-bound techniques

For Markov-chain parameterized factors with conjugate prior pdfs, various $Di_\Theta(V)$ pdfs are alternative descriptions inspected and compared. Thus, alternative sets of sufficient statistics V are branched and bounded as outlined in Section 6.1.3. The maximized functional F then becomes the function, namely, the v-likelihood $f(d(\mathring{t})|V)$ of data $d(\mathring{t})$ conditioned on the sufficient statistics V determining the $Di_{\Theta}(V)$ pdf that are used for eliminating unknown parameters

$$f(d(\mathring{t})|V) = \int f(d(\mathring{t})|\Theta) Di_{\Theta}(V)\, d\Theta \underbrace{=}_{(10.10)} \prod_{t \in t^*} f(d_t|d(t-1)).$$

Evaluation of approximate v-likelihood values for Markov-chain mixtures is implied by Proposition 10.4. Conceptually, general procedures for function maximization can be used. Generic high dimensions of occurrence matrices V make us stay within the considered framework of the branch-and-bound techniques. Specific versions are described in Section 10.4.

10.2 Data preprocessing

Except for the handling of missing data, no general preprocessing is expected when the modelled data are "naturally" discrete-valued. Often, however, the discrete-valued data arise by quantizing underlying continuous-valued signals. Then, the complete preprocessing arsenal described in Chapter 6 is needed. Here, the aspects specific to the intended quantization are briefly discussed.

10.2.1 Use of physical boundaries

The ability to respect hard physical bounds is a key advantage of Markov-chain models applied to continuous-valued signals. The uneven occurrence of the underlying scalar signal values within the inspected range is more the rule than the exception. This calls for an uneven specification of quantization levels. It allows us either "aggregate" signals so that differences between the signal distribution and target are well characterized or to exploit a full range of the inspected information channel. The following simple proposition gives guidelines for the latter case.

Proposition 10.5 (Signal quantization) *Let $f(x) > 0$ be pdf of a scalar continuous-valued quantity $x \in x^* \equiv [\underline{x}, \overline{x}]$ with known finite boundaries $\underline{x}$, $\overline{x}$. Let us consider a grid x_ι, $\iota = 0, \ldots, \mathring{\iota}$,*

$$\underline{x} = x_0 < x_1 < \cdots < x_{\mathring{\iota}-1} < x_{\mathring{\iota}} = \overline{x}$$

with a given finite number of grid points $\mathring{\iota} + 1$.

Let $\hat{f} \in \hat{f}^* \equiv \left\{ \text{pdfs on } x^* \text{ with constant values } \hat{f}_\iota \text{ on } (x_{\iota-1}, x_\iota) \right\}$. The best approximation $\hat{f}(x)$ of the pdf $f(x)$ taken from $\hat{f}^*$ such that

$$\mathcal{D}\left(f \middle|\middle| \hat{f}\right) \to \min_{\{\hat{f}_\iota\}_{\iota \in \iota^*}} \quad \text{and} \quad \mathcal{D}\left(\hat{f} \middle|\middle| \text{uniform pdf}\right) \to \min_{\{x_\iota\}_{\iota \in \iota^*}}, \ \text{is}$$

$$\hat{f}(x) = \frac{x_\iota - x_{\iota-1}}{\mathring{\iota}} \chi_{[x_{\iota-1}, x_\iota]}(x)$$

on the grid specified by the requirement

$$\frac{\bar{x} - \underline{x}}{\mathring{\iota}} = \int_{x_{\iota-1}}^{x_\iota} f(x)\, dx. \tag{10.11}$$

Proof. First a fixed grid is considered, then the part of $\mathcal{D}\left(f\middle|\middle|\hat{f}\right)$ influenced by values $\hat{f}_\iota$ has the form

$$\sum_{\iota=1}^{\mathring{\iota}} \int_{x_{\iota-1}}^{x_\iota} f(x) \ln\left(\frac{1}{\hat{f}_\iota}\right) dx = \sum_{\iota=1}^{\mathring{\iota}} \left(\int_{x_{\iota-1}}^{x_\iota} f(x)\, dx\right) \ln\left(\frac{1}{\hat{f}_\iota}\right)$$
$$= \sum_{\iota=1}^{\mathring{\iota}} \left(\int_{x_{\iota-1}}^{x_\iota} f(x)\, dx\right) \ln\left(\frac{\int_{x_{\iota-1}}^{x_\iota} f(x)\, dx}{\hat{f}_\iota}\right)$$
$$- \sum_{\iota=1}^{\mathring{\iota}} \left(\int_{x_{\iota-1}}^{x_\iota} f(x)\, dx\right) \ln\left(\int_{x_{\iota-1}}^{x_\iota} f(x)\, dx\right).$$

In this expression, the last term is independent of the optimized values and the previous term is the KL divergence minimized by $\hat{f}_\iota = \int_{x_{\iota-1}}^{x_\iota} f(x)\, dx$.

The second KL divergence measures uncertainty of the resulting approximation and it reaches the smallest zero value for uniform pdf $\hat{f}$. □

This simple Proposition hints how to quantize scalar signal $d(\mathring{t})$.

Algorithm 10.1 (Signal quantization)

1. *Select the number* $\mathring{\iota}$ *of quantization levels.*
2. *Fit a rich normal static mixture to the learning data* $d(\mathring{t})$.
3. *Evaluate the corresponding distribution function as the integral of the pdf estimated in the previous step.*
4. *Split the range [0,1] of values of the distribution function into* $\mathring{\iota}$*-intervals of the same length.*
5. *Project these values on the axis of arguments of the distribution function. It gives the grid searched for.*

Remark(s) 10.2

1. *The recommended simple quantization rule touches on an extensive field reviewed well in the paper [167].*
2. *The choice of the grid so that the quantized version is as close as possible to a uniform pdf tries to avoid the situation that the subsequently learned models have to deal with rare events.*
3. *Any mixture can be fitted to experimental data for the discussed purpose. Even an ordinary histogram can be used. Mixture estimation serves here as a smoothed version of a histogram.*

4. *Dependencies are neglected in the construction. Essentially, a sort of ergodic behavior is implicitly assumed and the estimated pdf is the limiting distribution. This approach seems to be sufficient for the considered quantization purpose.*
5. *Proposition 10.5 can be and should be applied to increments of the signal, too. For smooth signals, that are expected in the physical world, the probability that the signal change is greater than some level is practically zero. This fact is revealed by the described algorithm that "discovers" a band of nonzero transition probabilities. It helps to decrease significantly the number of estimated parameters and makes treatment of such signals practicable.*

Problem 10.3 (Algorithmic solution of quantization) *The basic ideas presented above have to be converted into a complete set of algorithms.*

10.2.2 Removal of high-frequency noise

Quantization of signal values suppresses small (within the quantization step) changes in signal values. High-frequency noise usually has relatively small amplitudes. Consequently, quantization suppresses "naturally" high-frequency noise. This brief argument describes more of a tendency than a general rule. Thus, a high-frequency-noise removal should generally be considered before quantization.

10.2.3 Suppression of outliers

In the discussed context, outliers are dangerous as they extend the range of the signal to be quantized. They increase the number of levels to be considered for achieving the target precision. Proposition 10.5 helps in this respect, too. By definition, outliers occur rarely so that their, even wide, range can be labelled by a few discrete values only.

10.2.4 Coding of regression vectors

Both predicted data item d and regression vector $\tilde{\psi}$ are discrete valued. The sufficient statistics V describing individual factors are arrays indexed by data vectors $\Psi = \left[d, \tilde{\psi}'\right]'$. For an easy software manipulation, it is reasonable to keep d as an independent index and to map the regression vector $\tilde{\psi}$ to a scalar quantity ψ. The chosen mapping, called *coding*, influences the formal structure of the statistics $V = \left[{}^{\lfloor d|\tilde{\psi}}V\right] = \left[{}^{\lfloor d|\psi}V\right]$, $d \in d^* \equiv \{1, \ldots, \mathring{d}\}$, $\psi \in \psi^* \equiv \{1, \ldots, \mathring{\psi}\}$. Treatment of the potentially sparse matrix $[V_{d|\psi}]$ is simpler if its nontrivial entries are clustered in a few contiguous areas. Coding may either enhance or destroy this property.

It seems that delayed items of a single quantity, say $\tilde{d}_t, \ldots, \tilde{d}_{t-\partial}$, each with $\mathring{d}$ possible values, should be coded to scalar d_t by the "thermometer" code

$$d_t \equiv \sum_{i=0}^{\partial} \mathring{d}^{\partial - i} \tilde{d}_{t-i} = \mathring{d}^{\partial} \tilde{d}_t + \mathring{d}^{-1} \left(d_{t-1} - \tilde{d}_{t-\partial-1} \right). \tag{10.12}$$

This coding reflects presumption that the older values have a smaller influence on the predicted value. Universality of this option and combinations of regressors of a different nature are questionable and the problem should be studied in detail. Vast amount of experience available in signal processing can definitely be exploited.

Description of graph-type dependencies is another aspect of the same problem. Experience from the field known as *Bayesian networks* [168, 169] can help.

10.2.5 Coding of signal values

The mixture consisting exclusively of Markov-chain components is again a Markov chain. It stimulates the natural question of whether it makes sense to use them at all. The affirmative answer follows from the possibility to spare estimated parameters. The coding of signal values discussed here serves as an example of this property.

Let us have a continuous-valued signal x_t whose evolution is described by the pdf $f(x_t|x(t-1)) \equiv f(x_t|x_{t-1})$ with a finite support covered say by $[0,1]^2$. Binary expansion of $x_t \approx \sum_{i=1}^{\mathring{d}} d_{i;t} 2^{-i}$ makes x_t (approximately) equivalent to discrete-valued multivariate data record d_t with entries $d_{i;t} \in \{0,1\}$. The data d_t forms the Markov chain that can be approximated by a specific mixture. For describing it, it is sufficient to consider $\mathring{d} = 2$. Then, the following approximation

$$\begin{aligned} f(d_t|d_{t-1}) &= f(d_{1;t}|d_{2;t}, d_{1;t-1}, d_{2;t-1}) f(d_{2;t}|d_{1;t-1}, d_{2;t-1}) \\ &\approx f(d_{1;t}|d_{1;t-1}) \left[\alpha f(d_{2;t}|d_{2;t-1}, d_{1;t-1} = 0) \right. \\ &\left. + \; (1-\alpha) f(d_{2;t}|d_{2;t-1}, d_{1;t-1} = 1) \right] \\ &\equiv f(d_{1;t}|d_{1;t-1}) \left[\alpha f(d_{2;t}|d_{2;t-1}, c = 1) + (1-\alpha) f(d_{2;t}|d_{2;t-1}, c = 2) \right] \end{aligned}$$

can be justified by an expected dominant influence of $d_{1;t-1}$ on $d_{1;t}$ and by introducing the "averaged" influence of $d_{1;t-1}$ on the evolution of $d_{2;t-1} \to d_{2;t}$. The original Markov chain on the left-hand side has 12 free parameters, the approximate right-hand side has 7 free parameters, including the probability α quantifying the average influence of $d_{1;t-1}$. This number increases to 9 if we do not assume that $f(d_{1;t}|d_{1;t-1})$ is a common factor. The difference is much more significant for higher $\mathring{d}$s and longer memory of the approximated model.

10.3 Use of prior knowledge at the factor level

Here, general ideas of quantification of prior knowledge at the factor level, Section 6.3, are specialized to Markov-chain factors.

10.3.1 Internally consistent fictitious data blocks

Processing of internally consistent data blocks coincides with learning Markov-chain factors; see Section 10.1.2. Thus, we have to quantify individual knowledge items $K_k,\ k = 1, \ldots, \mathring{k}$ and then to specialize merging of individual pdfs $f(\Theta|K_k)$ expressing them.

10.3.2 Translation of input-output characteristics into data

Quantification of the knowledge of the initial moments (6.26) of the predictive pdf $f(d|\psi)$ given by a fixed regression vector ψ is assumed

$$\hat{d} = \sum_{d \in d^*} d f(d|\psi),\ \hat{r} = \sum_{d \in d^*} \left(d - \hat{d}\right)^2 f(d|\psi). \tag{10.13}$$

The general Proposition 6.4 gives the constructed prior out of the conjugate class Dirichlet pdfs. We need to stay within it. It makes us derive the following specialized proposition.

Proposition 10.6 (Knowledge of input-output characteristics) *Let us fix a regression vector ψ. Consider Dirichlet pdfs $Di_\theta(V)$ of the unknown parameter θ with entries $\theta_d \equiv {}^{\lfloor d|\psi}\Theta$ and determined by the vector statistics V with entries ${}^{\lfloor d}V \equiv {}^{\lfloor d|\psi}V,\ d \in d^*$. The symbol ${}^{\lfloor d|\psi}V$ denotes entry of a positive occurrence array; see Section 10.1.1.*

We search for the $Di_\theta(V)$ pdf that fulfills constraints (10.13) and minimizes the KL divergence to the flat pre-prior $Di_\theta(\varepsilon \mathbf{1})$ given by $\varepsilon > 0,\ \varepsilon \to 0$ that multiplies vector of units $\mathbf{1} = [1, \ldots, 1]'$ *of the appropriate length.*

Such a pdf is determined by the statistics

$${}^{\lfloor d}V = \frac{1}{ad^2 + bd + c},\ d \in d^* = \{1, \ldots, \mathring{d}\} \tag{10.14}$$

with constants a, b, c solving equations

$$\hat{d} = \frac{\sum_{d \in d^*} \frac{d}{ad^2+bd+c}}{\sum_{d \in d^*} \frac{1}{ad^2+bd+c}},\ \hat{r} = \frac{\sum_{d \in d^*} \frac{\left(d-\hat{d}\right)^2}{ad^2+bd+c}}{\sum_{d \in d^*} \frac{1}{ad^2+bd+c}}, \tag{10.15}$$

while minimizing the resulting distance.

Proof. For $\nu \equiv \sum_{d\in d^*} {}^{\lfloor d}V$ and $\bar{\nu} \equiv \varepsilon \mathring{d}$, the optimized functional $\mathcal{D}\left(f||\bar{f}\right)$ is the function of statistics V, $\bar{V} = \varepsilon \mathbf{1}$ (see Proposition 10.2)

$$\mathcal{D}\left(f||\bar{f}\right) = \sum_{d=1}^{\mathring{d}} \left[\left({}^{\lfloor d}V - {}^{\lfloor d}\bar{V} \right) \frac{\partial}{\partial\, {}^{\lfloor d}V} \ln\left(\Gamma\left({}^{\lfloor d}V \right)\right) + \ln\left(\frac{\Gamma\left({}^{\lfloor d}\bar{V} \right)}{\Gamma\left({}^{\lfloor d}V \right)} \right) \right]$$
$$-(\nu - \bar{\nu}) \frac{\partial}{\partial \nu} \ln(\Gamma(\nu)) + \ln\left(\frac{\Gamma(\nu)}{\Gamma(\bar{\nu})} \right).$$

Its derivative with respect to ${}^{\lfloor d}V$ is

$$\frac{\partial \mathcal{D}\left(f||\bar{f}\right)}{\partial\, {}^{\lfloor d}V} = \frac{\partial \ln\left(\Gamma\left({}^{\lfloor d}V \right)\right)}{\partial\, {}^{\lfloor d}V} + \left({}^{\lfloor d}V - {}^{\lfloor d}\bar{V} \right) \frac{\partial^2 \ln\left(\Gamma\left({}^{\lfloor d}V \right)\right)}{\partial\, {}^{\lfloor d}V^2} - \frac{\partial \ln\left(\Gamma\left({}^{\lfloor d}V \right)\right)}{\partial\, {}^{\lfloor d}V}$$
$$-\frac{\partial \ln(\Gamma(\nu))}{\partial \nu} + \frac{\partial \ln(\Gamma(\nu))}{\partial \nu} - (\nu - \bar{\nu}) \frac{\partial^2 \ln(\Gamma(\nu))}{\partial \nu^2}$$
$$= \left({}^{\lfloor d}V - {}^{\lfloor d}\bar{V} \right) \frac{\partial^2 \ln\left(\Gamma\left({}^{\lfloor d}V \right)\right)}{\partial\, {}^{\lfloor d}V^2} - (\nu - \bar{\nu}) \frac{\partial^2 \ln(\Gamma(\nu))}{\partial \nu^2}.$$

Similarly, $\dfrac{\partial \hat{d}}{\partial\, {}^{\lfloor d}V} = \nu^{-1}(d - \hat{d}),\ \dfrac{\partial \hat{r}}{\partial\, {}^{\lfloor d}V} = \nu^{-1} \left[\left(d - \hat{d} \right)^2 - \hat{r} \right].$

This gives derivatives of the Lagrangian function that, for $\varepsilon \to 0$, can be rewritten into the form

$$0 = {}^{\lfloor d}V \frac{\partial^2 \ln\left(\Gamma\left({}^{\lfloor d}V \right)\right)}{\partial V_d^2} + 0.5ad^2 + 0.5bd + 0.5(c-1),$$

where the constants a, b, c are made of Lagrangian multipliers and terms independent of d. Using expansion of $\ln\left(\Gamma\left({}^{\lfloor d}V \right)\right)$ [156], the involved second derivative can be approximated by

$$\frac{\partial^2 \ln\left(\Gamma\left({}^{\lfloor d}V \right)\right)}{\partial\, {}^{\lfloor d}V^2} \approx 0.5\, {}^{\lfloor d}V^{-1} \left(1 - {}^{\lfloor d}V^{-1} \right).$$

Putting this approximation into the condition for extreme, we get the linear equation for ${}^{\lfloor d}V$ that determines its dependence on d. The optional constants a, b, c are selected so that constraints are fulfilled. The selection is made unique by selecting c so that the resulting distance is the smallest one. □

Remark(s) 10.3

1. *For a given a, b, the value c determines the sum $\nu = \sum_{d\in d^*} (ad^2 + bd + c)^{-1}$. The minimized functional is its increasing function while $\nu > \bar{\nu}$. It hints at the choice of c as the solution of the equation $\bar{\nu} = \sum_{d\in d^*} (ad^2 + bd + c)^{-1}$. This statement is not included into Proposition, as it is not proved that such a triple (a, b, c) exist while guaranteeing that ${}^{\lfloor d}V > 0$. A nontrivial solution can be expected iff meaningful options $\hat{d} \in \left(1, \mathring{d}\right)$ and $\hat{r} < \left(0.5\mathring{d}\right)^2$ are made.*
2. *Other constraints can be and should be considered, respecting the discrete nature of data.*

10.3.3 Merging of knowledge pieces

We have got a collection of pdfs $f(\Theta|K_k) = Di_{\Theta}(V_k)$ $k \in k^* \equiv \{1,\ldots,\mathring{k}\}$ after processing individual internally consistent data blocks and individual knowledge items. We use them for construction of a single posterior pdf $\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k})) = Di_{\Theta}(V_{\mathring{t}})$ that merges them all together with available real data $d(\mathring{t})$. For this, the general Proposition 6.5 applies. It leads directly to the Di_{Θ} counterpart of Algorithm 6.3.

Algorithm 10.2 (Merging Dirichlet knowledge sources)

1. *Construct prior $Di_{\Theta}(V_k)$ pdfs reflecting internally consistent data blocks by applying ordinary Bayesian estimation and (or) reflecting individual knowledge pieces according to Proposition 10.6.*
2. *Evaluate the likelihood function $\mathcal{L}(\Theta, d(\mathring{t})) = Di_{\Theta}(V_{0;\mathring{t}})$ by applying Proposition 10.3 with zero initial conditions for recursive evaluation of the occurrence matrix V.*
3. *Evaluate posterior pdfs $f(\Theta|d(t), K_k) = Di_{\Theta}(V_{0;\mathring{t}} + V_k)$ and v-likelihood*
$$f(d(\mathring{t})|K_k) = \frac{\mathcal{I}(V_{0;\mathring{t}} + V_k)}{\mathcal{I}(V_k)}$$
corresponding to prior pdfs $f(\Theta|K_k) = Di_{\Theta}(V_k)$ $k \in k^*$. The normalization integral $\mathcal{I}(V)$ is given by the formula (10.3).
Regularization by a flat pre-prior pdf $Di_{\Theta}(\varepsilon\mathbf{1})$, given by small $\varepsilon > 0$ has to be considered if some V_k contains zero values.
4. *Evaluate the weights expressing the posterior confidence to individual knowledge items*
$$\beta_{k|d(\mathring{t})} = \frac{f(d(\mathring{t})|K_k)}{\sum_{\tilde{k}\in k^*} f(d(\mathring{t})|K_{\tilde{k}})} = \frac{\frac{\mathcal{I}(V_{0;\mathring{t}}+V_k)}{\mathcal{I}(V_k)}}{\sum_{\tilde{k}\in k^*} \frac{\mathcal{I}(V_{0;\mathring{t}}+V_{\tilde{k}})}{\mathcal{I}(V_{\tilde{k}})}}, \quad k \in k^*.$$
5. *Determine the merger as the posterior pdf to be used*
$$\begin{aligned}\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k})) &= Di_{\Theta}\left(V_{0;\mathring{t}} + \sum_{k\in k^*} \beta_{k|d(\mathring{t})} V_k\right)\\ &= Di_{\Theta}(V_{0;\mathring{t}}) Di_{\Theta}\left(\sum_{k\in k^*} \beta_{k|d(\mathring{t})} V_k\right).\end{aligned}$$
Notice that the used “prior” pdf is seen in the last formula.

10.4 Construction of the prior estimate

Here, Markov-chain counterparts of Section 6.4 are presented.

10.4.1 Iterative construction of prior pdf

The danger of a plain repetitive use of the Bayes rule, discussed at the general level in Section 6.4.1, gets the following form within the conjugate class of Dirichlet pdfs. After $\mathring{n}$th repetitive application of the (approximate) Bayes rule we get the posterior pdf

$$Di_{\Theta}\left(\sum_{n=1}^{\mathring{n}} V_{n;\mathring{t}}\right), \text{ where } V_n \text{ is the sufficient statistic evaluated at } n\text{th step.}$$

Taking into account the basic properties of the Dirichlet pdf, Proposition 10.1, it can be seen that point estimates $\hat{\Theta}$ of unknown Θ would be around the point estimate equal to nonweighted average of point estimates obtained in respective iterations. It can be a reasonable value. Its apparent uncertainty would be, however, too small as it is inversely proportional to $\mathring{n}\mathring{t}$. This observation clarifies why the flattening may help; see Section 6.4.3. It acts as forgetting that prefers newer, hopefully better, values of $V_{n;\mathring{t}}$ and prevents the unrealistically fast growth of the estimated precision.

10.4.2 Common bounding mapping

The common bounding mapping (see Section 6.4.2) just cancels the "least fit candidates". For Markov-chain mixtures, the posterior pdfs are given by the vector statistic κ, describing component weights, and by the collection of the occurrence matrices V_{ic}. These are finite-dimensional objects. They may contain, however, a large number of entries. This fact makes us search for a compromise between two contradictory requirements on the bounding mapping

- preserve as much information as possible from previous iteration steps,
- store as little as possible alternatives.

As said in Chapters 6, 8, we have no general guideline how to reach the adequate compromise. At present, we stress the second item and bound the number of alternatives to two, at most to three.

10.4.3 Flattening mapping

We specialize the flattening operation to Markov-chain mixtures.

Proposition 10.7 (Optimal flattening mapping for Di **factors)** *Let* $\tilde{f} = Di_{\Theta}\left(\tilde{V}\right)$ *and* $\bar{f} = Di_{\Theta}\left(\bar{V}\right)$ *be a pair of Di pdfs defined on the common support* Θ^**. Then, the pdf* $\hat{f}$ *defined on* Θ^* *and minimizing the functional* $\mathcal{D}\left(\hat{f}||\tilde{f}\right) + q\mathcal{D}\left(\hat{f}||\bar{f}\right)$, $q > 0$ *is* $Di_{\Theta}\left(\hat{V}\right)$ *with*

$$\hat{V} = \Lambda\tilde{V} + (1-\Lambda)\bar{V}, \text{ where } \Lambda \equiv 1/(1+q) \in (0,1). \tag{10.16}$$

Proof. Omitted. □

The application of this result to whole mixtures is straightforward as the posterior pdf of its parameters is a product of the posterior pdfs of Dirichlet form related to both weights α and parameters of individual factors.

Proposition 10.8 (Optimal flattening for Markov-chain mixtures) *Let parameters of a pair of Markov-chain mixtures be described by factors* $\tilde{f}_{ic}(\Theta_{ic}) = Di_{\Theta_{ic}}\left(\tilde{V}_{ic}\right)$, *weights* $\tilde{f}(\alpha) = Di_\alpha(\tilde{\kappa})$, *and similarly* $\bar{f}_{ic}(\Theta_{ic}) = Di_{\Theta_{ic}}\left(\bar{V}_{ic}\right)$, $\bar{f}(\alpha) = Di_\alpha(\bar{\kappa})$, $i \in i^*, c \in c^*$. *The factors with the same indexes ic are defined on the common support* Θ^*_{ic}. *Then, the pdf* $\hat{f}$ *minimizing the functional*

$$\mathcal{D}\left(\hat{f}||\tilde{f}\right) + q\mathcal{D}\left(\hat{f}||\bar{f}\right),\ q > 0$$

preserves the Dirichlet functional form. The resulting pdfs are given by the statistics, $i \in \{1, \ldots, \mathring{d}\},\ c \in c^*$,

$$\hat{V}_{ic} = \Lambda\tilde{V}_{ic} + (1-\Lambda)\bar{V}_{ic},\ \hat{\kappa}_c = \Lambda\tilde{\kappa}_c + (1-\Lambda)\bar{\kappa}_c,\ \Lambda \equiv 1/(1+q) \in (0,1). \quad (10.17)$$

Proof. Omitted. □

Flattening of the component-weights estimates is universal. It preserves the Dirichlet form with statistics being a convex combination of statistics of estimates among which the compromise is searched for. The same applies to estimates of Markov-chain factors that are of Dirichlet type, too. The choice of the flattening rate is fully described by Propositions 6.7, 6.8, 6.11 and 6.13.

10.4.4 Geometric mean as branching mapping

Geometric-mean branching mapping $\mathcal{A}$ (6.15) maps a pair of arguments $\hat{f}_n(\Theta)$, $n = 1, 2$, of the maximized functional (6.38) $\lfloor^h\mathcal{L}\left(f_n(d(\mathring{t})|\Theta)\hat{f}_n(\Theta), d(\mathring{t})\right) \equiv \hat{f}_n(d(\mathring{t})) \equiv \int f_n(d(\mathring{t})|\Theta)\hat{f}_n(\Theta)\, d\Theta$ on a new, hopefully better, candidate $\hat{f}_n(\Theta)$.

Recall that $f_n(d(\mathring{t})|\Theta)\hat{f}_n(\Theta) \propto Di_\Theta(V_{n;\mathring{t}})$ is an approximate posterior likelihood computed by approximate estimation when conjugate $\hat{f}_n(\Theta) = Di_\Theta(V_{n;0})$, $n = 1, 2$, are used as prior pdfs.

The new candidate is defined

$$\begin{aligned}\hat{f}_3(\Theta|d(\mathring{t})) &\propto \hat{f}_1^\lambda(\Theta|d(\mathring{t}))\hat{f}_2^{1-\lambda}(\Theta|d(\mathring{t}))\\ &\equiv Di_\Theta(V_{3;\mathring{t}}) \equiv Di_\Theta\left(\lambda V_{1;\mathring{t}} + (1-\lambda)V_{2;\mathring{t}}\right) \qquad (10.18)\end{aligned}$$

$$\lambda = \frac{\hat{f}_1(d(\mathring{t}))}{\hat{f}_1(d(\mathring{t})) + \hat{f}_2(d(\mathring{t}))},\ \hat{f}_n(d(\mathring{t})) = \int \hat{f}_n(d(\mathring{t})|\Theta)\hat{f}_n(\Theta)\, d\Theta,\quad n = 1, 2. \qquad (10.19)$$

Use of (10.18), (10.19) and Proposition 10.4 describes the new candidate for the better estimate.

Proposition 10.9 (Geometric-mean branching of Dirichlet pdfs)
Let

$$\hat{f}_n(\Theta) \equiv Di_\alpha(\kappa_{n;0}) \prod_{c\in c^*} \prod_{i \in i^*} Di_{\Theta_{ic}}(V_{icn;0}),$$

$n = 1, 2,$ *be prior pdfs employed in approximate estimation (see Section 10.5) of the Markov-chain mixture*

$$f(d_t|d(t-1), \Theta) \equiv \sum_{c\in c^*} \alpha_c \prod_{i\in i^*} {}^{\lfloor d_{ic;t}|\psi_{ic;t}}\Theta_{ic}.$$

The posterior pdfs preserve the functional forms of the prior pdfs and they are described by statistics $V_{icn;\mathring{t}}$ *and* $\kappa_{cn;\mathring{t}}$.

Their geometric mean $f_3(\Theta|V_{3;\mathring{t}})$ *keeps this form and its statistics are*

$$\begin{aligned} V_{3ic;\mathring{t}} &= \lambda_{1;\mathring{t}} V_{ic1;\mathring{t}} + (1-\lambda_{1;\mathring{t}}) V_{ic2;\mathring{t}} \\ \kappa_{3ic;\mathring{t}} &= \lambda_{1;\mathring{t}} \kappa_{ic1;\mathring{t}} + (1-\lambda_{1;\mathring{t}}) \kappa_{ic2;\mathring{t}} \\ \lambda_{1;\mathring{t}} &= \left[1 + \prod_{t\in t^*} \frac{\sum_{c\in c^*} \prod_{i\in i^*} \frac{\mathcal{I}(V_{ic2;t})}{\mathcal{I}(V_{ic2;t-1})} \frac{\kappa_{c2;t-1}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c}2;t-1}}}{\sum_{c\in c^*} \prod_{i\in i^*} \frac{\mathcal{I}(V_{ic1;t})}{\mathcal{I}(V_{ic1;t-1})} \frac{\kappa_{c1;t-1}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c}1;t-1}}} \right]^{-1}. \end{aligned} \tag{10.20}$$

The normalization integral $\mathcal{I}(V)$ *is given by the formula (10.3).*

Proof. Omitted. □

10.4.5 Random branching of statistics

The deterministic search algorithms are often stuck at a local optimum. This makes us complement the deterministic search by a random one. Here, we discuss this possibility in connection with Markov-chain factors.

Approximate estimation, Section 10.5, provides the posterior pdf $f(\Theta|d(\mathring{t}))$ as a product of the posterior Di_Θ pdfs corresponding to the individual, Markov-chain parameterized, factors and to the posterior Dirichlet pdf on component weights. Thus, the following considerations and computations may deal with individual factors only and the random search step may be applied to their subselection.

The $Di_\Theta(V)$ factor is determined by the finite-dimensional (almost) sufficient statistics $V \in V^* \equiv \{$ collection of occurrence tables $V > 0$ with dimensions compatible with estimated $\Theta\,\}$. Knowing this, we generate a random sample $\tilde{V}$ in V^* describing the modified factor. It gives us a new guess $\tilde{f}(\Theta|d(\mathring{t})) \equiv Di_\Theta(\tilde{V})$. By flattening it towards a flat pdf $\bar{f} = Di_\Theta(\bar{V})$, given by a small positive $\bar{V}$ (see Proposition 6.6), we get a new guess of the prior pdf $\hat{f}_{new}(\Theta) = Di_\Theta(\Lambda\tilde{V} + (1-\Lambda)\bar{V})$. Even for a single factor, the computational burden induced by the random choice is usually prohibitive. Thus, it is reasonable to exploit interpretation of parts of statistics and change them partially only. At present, it seems wise

- to estimate the factor structure before generating new statistics;
- to change the point estimate of parameters using normal approximation, based on coincidence of moments, cf. Proposition 10.1, of Dirichlet pdf

$$
\begin{aligned}
{}^{\lfloor d|\psi}\tilde{\hat{\Theta}} &= \max\left[{}^{\lfloor d|\psi}\bar{\hat{\Theta}},\ {}^{\lfloor d|\psi}\hat{\Theta}\left(1+e\rho\sqrt{\frac{{}^{\lfloor d|\psi}\hat{\Theta}^{-1}-1}{{}^{\lfloor\psi}\nu+1}}\right)\right] \qquad (10.21)\\
& e\sim\mathcal{N}_e(0,1), \quad \text{where}\\
{}^{\lfloor d|\psi}\bar{\hat{\Theta}}>0 \quad & \text{is a preselected lower bound}\\
& \rho\in\rho^*\equiv(1,5) \text{ scales the generating noise}\\
{}^{\lfloor d|\psi}\tilde{\hat{\Theta}} &\equiv {}^{\lfloor d|\psi}\tilde{\hat{\Theta}}\frac{\nu}{\nu-{}^{\lfloor d|\psi}\hat{\Theta}+{}^{\lfloor d|\psi}\tilde{\hat{\Theta}}}\\
\tilde{\nu} &\equiv \nu\left(1-{}^{\lfloor d|\psi}\hat{\Theta}+{}^{\lfloor d|\psi}\tilde{\hat{\Theta}}\right) \text{ and define the new occurrence table}\\
{}^{\lfloor d|\psi}V &\equiv \tilde{\nu}\,{}^{\lfloor d|\psi}\tilde{\hat{\Theta}}.
\end{aligned}
$$

Statistics κ determining the Dirichlet pdf of component weights are modified in the same ways as in (8.52). It gives a chance to be modified even to the components with $\hat{\alpha}_c \approx 0$. The normal pdf cut so that the smallest κ_c does not decrease is approximately used. It gives equivalent of (8.52)

$$
\tilde{\kappa}_c=\max\left[\kappa_c\left(1+\frac{\rho}{\sqrt{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c}}}}e_c\right),\min_{\tilde{c}\in c^*}\kappa_{\tilde{c}}\right],\ \ \rho\in(1,5),\ \ f(e_c)=\mathcal{N}_{e_c}(0,1).
$$

Remark(s) 10.4

1. *The chosen scaling of new point estimates to the occurrence table preserves the value of* ${}^{\lfloor\psi}\nu=\sum_{d\in d^*}{}^{\lfloor d|\psi}V$ *that determines uncertainty of the inspected Dirichlet pdf.*
2. *Random generating of statistics can be used as a part of various compound algorithms constructing the prior pdf.*
3. *The optional value* ρ *is a tuning knob of generators. Its recommended range stems from standard properties of the normal pdf used instead of the correct Dirichlet pdf. At present, values of* ρ *in a more narrow range* $\rho^*\equiv(1,2)$ *seem to be preferable.*

10.4.6 Prior-posterior branching

The prior-posterior branching (see Section 6.4.6) starts at some prior pdf, performs an approximate estimation, usually quasi-Bayes estimation (see Algorithm 6.13), and flattens the resulting posterior pdf so that it provides a new alternative to the prior pdf used. The specialization of the general results on flattening gives directly the following iterative Bayesian learning of Markov-chain mixtures.

Algorithm 10.3 (Prior-posterior branching with geometric mean)

Initial mode

- *Select the upper bound $\mathring{n}$ on the number of iterations n and set $n = 1$.*
- *Select a sufficiently rich structure of the mixture, i.e., specify the number of components $\mathring{c}$ and the ordered lists of factors allocated to considered components. The factor structure for each ic is determined by the structure of the corresponding data vector Ψ_{ic}.*
- *Select a flat pre-prior pdf $\bar{f}(\Theta)$ in the form (6.3) with Dirichlet pdfs describing individual factors. It means, select pre-prior statistics $\bar{V}_{ic}$, $\bar{\kappa}_c$ determining Dirichlet pdfs on the transition probabilities ${}^{\lfloor d|\psi}\Theta$ and on component weights α with a high variance; cf. Proposition 10.1. Typically, $\bar{V}_{ic} = \varepsilon$ (entrywise) and $\bar{\kappa}_c = \varepsilon > 0$, ε small. They serve as alternatives in flattening.*
- *Select a prior pdf $f(\Theta)$ in the form (6.3) with Dirichlet pdfs describing individual factors. It means, select prior statistics V_{ic}, κ_c determining the involved Dirichlet pdfs; cf. Proposition 10.1. Generally, $\bar{f}(\Theta) \neq f(\Theta)$.*
- *Set $\hat{f}_{1n}(\Theta) \equiv f(\Theta)$, i.e., set $V_{ic1n;0} = V_{ic}$, $\kappa_{c1n;0} = \kappa_c$ giving $\hat{f}_{1n}(\Theta_{ic}) = \prod_{\Psi\in\Psi^*} Di_{\,{}^{\lfloor d|\psi}\Theta_{ic}}\left({}^{\lfloor d|\psi}V_{ic1n;0}\right)$ and $\hat{f}_{1n}(\alpha) = Di_{\alpha}(\kappa_{1n;0})$.*
- *Compute the posterior pdf*

$$\tilde{f}_{1n}(\Theta|d(\mathring{t})) = Di_{\alpha}(\kappa_{1n;\mathring{t}}) \prod_{i\in i^*,\Psi\in\Psi^*} Di_{\,{}^{\lfloor d|\psi}\Theta_{ic}}\left({}^{\lfloor d|\psi}V_{ic1n;\mathring{t}}\right)$$

  using an approximate Bayesian estimation that starts at $\hat{f}_{1n}(\Theta)$; see Section 10.5.
- *Evaluate v-likelihood l_{1n} resulting from the use of $\hat{f}_{1n}(\Theta)$.*
- *Apply flattening operation to $\tilde{f}_{1n}(\Theta|d(\mathring{t}))$ according to Proposition 6.7. Call the resulting pdf $\hat{f}_{2n}(\Theta)$. For the component weights α, the flattening rate is*

$$\Lambda_D \equiv \frac{\sum_{c\in c^*}(\kappa_c - \bar{\kappa}_c)}{\sum_{\tilde{c}\in c^*}(\kappa_{\tilde{c}1n;\mathring{t}} - \bar{\kappa}_{\tilde{c}})}.$$

  For Dirichlet estimates of Markov-chain factors, the flattening is determined by

$$\Lambda_{ic|\psi} \equiv \frac{\sum_{d\in d^*}\left({}^{\lfloor d|\psi}V_{ic} - {}^{\lfloor d|\psi}\bar{V}_{ic}\right)}{\sum_{\tilde{d}\in d^*}\left({}^{\lfloor \tilde{d}|\psi}V_{ic1n;\mathring{t}} - {}^{\lfloor \tilde{d}|\psi}\bar{V}_{ic}\right)}.$$

  It provides a new guess of the prior pdf of the form (6.3) given by

$$\begin{aligned}\kappa_{c2n;0} &= \Lambda_D \kappa_{c1n;\mathring{t}} + (1-\Lambda_D)\bar{\kappa}_c\\ {}^{\lfloor d|\psi}V_{ic2n;0} &= \Lambda_{ic|\psi}\, {}^{\lfloor d|\psi}V_{ic1n;\mathring{t}} + (1-\Lambda_{ic|\psi})\, {}^{\lfloor d|\psi}\bar{V}_{ic}.\end{aligned}$$

- *Compute the posterior pdf*

$$\tilde{f}_{2n}(\Theta|d(\mathring{t})) = Di_{\alpha}(\kappa_{2n;\mathring{t}}) \prod_{i\in i^*,c\in c^*,d\in d^*,\psi\in\psi^*} Di_{\,{}^{\lfloor d|\psi}\Theta_{ic}}\left({}^{\lfloor d|\psi}V_{icn;\mathring{t}}\right)$$

using an approximate Bayesian estimation that starts at $\hat{f}_{2n}(\Theta)$; see Section 10.5.

- *Evaluate v-likelihood l_{2n} resulting from the use of $\hat{f}_{2n}(\Theta)$.*
- *Set $\bar{l}_n = \max(l_{1n}, l_{2n})$.*

Iterative mode

1. *Apply the geometric branching to the pair $\tilde{f}_{1n}(\Theta|d(\mathring{t}))$, $\tilde{f}_{2n}(\Theta|d(\mathring{t}))$ with v-likelihood l_{1n}, l_{2n}, respectively. For $\lambda \equiv \frac{l_{1n}}{l_{1n}+l_{2n}}$, it gives $\tilde{f}_{3n}(\Theta|d(\mathring{t}))$ of the form (6.3) with Dirichlet descriptions of factors. It is determined by the statistics*
$$\kappa_{c3n;\mathring{t}} = \lambda \kappa_{c1n;\mathring{t}} + (1-\lambda)\kappa_{c2n;\mathring{t}}, \quad V_{ic3n;\mathring{t}} = \lambda V_{ic1n;\mathring{t}} + (1-\lambda) V_{ic2n;\mathring{t}}.$$
2. *Apply the flattening operation to $\tilde{f}_{3n}(\Theta|d(\mathring{t}))$ according to Proposition 6.7. Call the resulting pdf $\hat{f}_{3n}(\Theta)$. It preserves the form (6.3). The flattening rates are*
$$\Lambda_D \equiv \frac{\sum_{c\in c^*}\left(\kappa_{c3n} - \bar{\kappa}_c\right)}{\sum_{\tilde{c}\in c^*}\left(\kappa_{\tilde{c}3n;\mathring{t}} - \bar{\kappa}_{\tilde{c}}\right)}, \quad \Lambda_{ic|\psi} \equiv \frac{\sum_{d\in d^*}\left({}^{\lfloor d|\psi}V_{ic3n} - {}^{\lfloor d|\psi}\bar{V}_{ic}\right)}{\sum_{\tilde{d}\in d^*}\left({}^{\lfloor \tilde{d}|\psi}V_{ic3n;\mathring{t}} - {}^{\lfloor \tilde{d}|\psi}\bar{V}_{ic}\right)}.$$
They provide the new statistics
$$\begin{aligned}\kappa_{c3n;0} &= \Lambda_D \kappa_{c3n;\mathring{t}} + (1-\Lambda_D)\bar{\kappa}_c\\ {}^{\lfloor d|\psi}V_{ic3n;0} &= \Lambda_{ic|\psi}\, {}^{\lfloor d|\psi}V_{ic3n;\mathring{t}} + (1-\Lambda_{ic|\psi})\, {}^{\lfloor d|\psi}\bar{V}_{ic}.\end{aligned}$$
3. *Evaluate v-likelihood l_{3n} resulting from the use of $\hat{f}_{3n}(\Theta)$ for determining of $\tilde{f}_{3n}(\Theta|d(\mathring{t}))$. The approximation prepared for the fixed advisory system has to be used; see Section 6.1.2.*
4. *Choose among the triple $\tilde{f}_{in}(\Theta|d(\mathring{t}))$, $i \in \{1,2,3\}$, the pair with the highest v-likelihood values and call these pdfs $\tilde{f}_{1(n+1)}(\Theta|d(\mathring{t}))$, $\tilde{f}_{2(n+1)}(\Theta|d(\mathring{t}))$ with v-likelihood $l_{1(n+1)}$, $l_{2(n+1)}$.*
5. *Go to the beginning of* Iterative mode *with $n = n+1$ if*
$$\bar{l}_{n+1} \equiv \max(l_{1(n+1)}, l_{2(n+1)}) > \bar{l}_n$$
or if $\bar{l}_{n+1}, \bar{l}_n$ are the same according to Proposition 6.2 and $n < \mathring{n}$.
6. *Stop and select among $\hat{f}_{in}(\Theta)$, $i = 1, 2$ that leading to the higher value of l_{in} and take it as the prior pdf constructed.*

Remark(s) 10.5

1. *The structure of the estimated mixture does not change during iterations. Thus, it has to be sufficiently reflected in the prior pdf $f(\Theta)$.*
2. *Prediction of the v-likelihood of connected with the prior pdf obtained flattened geometric mean is possible; see Section 10.6.3.*

10.4.7 Branching by forgetting

Branching by forgetting (see Section 6.4.7) makes parallel recursive estimation without forgetting and with a fixed forgetting with forgetting factor smaller than one. The alternative pdfs are compared according to their v-likelihood values. At the time moment, when one of them is a sure winner they are bounded into a single pdf and whole process is repeated. Here, the algorithm is specialized to Markov-chain mixtures.

Algorithm 10.4 (Online branching with forgetting)

Initial mode

- *Select a sufficiently rich structure of the mixture, i.e., specify the number of components $\mathring{c}$ and the ordered lists of factors allocated to the considered components. The factor structure for each ic is determined by the structure of the corresponding data vector Ψ_{ic}.*
- *Select a constant $\rho \approx 3-5$ defining the significant difference of v-log-likelihood values.*
- *Set the record counter $t = 0$ and select the statistics $V_{ic;0}$ determining the prior pdf*

$$\hat{f}(\Theta_{ic}) = \prod_{\Psi\in\Psi^*} Di_{{}^{\lfloor d|\psi}\Theta_{ic}}\left({}^{\lfloor d|\psi}V_{ic;0}\right),\ i\in i^*,\ c\in c^*,\ \Psi=[d,\psi']',$$

  for factors as well as for the component weights $\hat{f}(\alpha) = Di_{\alpha}(\kappa_0)$.
- *Choose a fixed relatively small forgetting factor $\lambda < 1$.*
- *Select a fixed pre-prior alternative pdf used in the stabilized forgetting; see Proposition 3.1. The alternative is usually taken as a flat pre-prior pdf of the Dirichlet version of (6.3) given by small ${}^{\lfloor A}\kappa = {}^{\lfloor Ad|\psi}V_{ic} = \varepsilon > 0$, ε small.*
- *Set $\hat{f}_1(\Theta|d(t)) = \hat{f}_\lambda(\Theta|d(t)) = \hat{f}(\Theta|d(t))$.*
- *Initialize v-log-likelihood $l_{1;t} = 0$, $l_{\lambda;t} = 0$ assigned to the respective forgetting alternatives.*

Data processing mode*, running for $t = 1, 2, \ldots,$*

1. *Collect new data d_t and construct the data vector Ψ_t.*
2. *Update $\hat{f}_1(\Theta|d(t-1))$ to $\hat{f}_1(\Theta|d(t))$ using approximate estimation, Section 10.5, with the forgetting factor 1.*
3. *Re-compute the v-log-likelihood $l_{1;t-1}$ to $l_{1;t}$ by adding the logarithm of the mixture prediction $\ln(f(d(t)|d(t-1)))$ obtained for the "prior" pdf $\hat{f}_1(\Theta|d(t-1))$.*
4. *Update $\hat{f}_\lambda(\Theta|d(t-1))$ to $\hat{f}_\lambda(\Theta|d(t))$ using approximate estimation with the forgetting factor λ and the chosen alternative. Thus, after obtaining statistics $\kappa_{\lambda 1;t}$, $V_{ic\lambda 1;t}$ through updating of $\kappa_{\lambda;t-1}$, $V_{ic\lambda;t-1}$, with forgetting 1, forget $\kappa_{\lambda;t} = \lambda\kappa_{\lambda 1;t} + (1-\lambda)\,{}^{\lfloor A}\kappa$, $V_{ic\lambda;t} = \lambda V_{ic\lambda 1;t} + (1-\lambda)\,{}^{\lfloor A}V_{ic}$.*

5. *Recompute the v-log-likelihood* $l_{\lambda;t-1}$ *to* $l_{\lambda;t}$ *by adding the logarithm of the mixture prediction* $\ln(f(d(t)|d(t-1)))$ *obtained for the "prior" pdf* $\hat{f}_\lambda(\Theta|d(t-1))$.
6. *Go to Step 1 with* $t = t+1$ *if* $|l_{1;t} - l_{\lambda;t-1}| < \rho$.
7. *Set* $\hat{f}(\Theta|d(t)) = \hat{f}_1(\Theta|d(t))$ *if* $l_{1;t} > l_{\lambda;t}$ *otherwise set* $\hat{f}(\Theta|d(t)) = \hat{f}_\lambda(\Theta|d(t))$.
   This setting means assignment of the appropriate sufficient statistics.
8. *Go to the beginning of* Data processing mode *if* $t \leq \mathring{t}$. *Otherwise stop and take* $\hat{f}_1(\Theta|d(\mathring{t}))$ *as the final estimate.*

Remark(s) 10.6

1. *Speeding up of the learning is the main expectation connected with this algorithm. The model with no forgetting is expected to be long-run winner. The estimation with forgetting is to be switched off when the estimation without forgetting is better the majority of the time.*
2. *The technique can be directly combined with the prior-posterior branching.*
3. *It will be necessary to get experience with the choice of the fixed forgetting factor* λ *in the case of Markov chains. The option* $\lambda \approx 0.6$ *seems to be satisfactory for normal mixtures. A similar value represents a reasonable start for a closer inspection in the Markov-chain case.*

10.4.8 Branching by factor splitting

This section specializes the algorithm described in Section 6.4.8. It covers important cases when we have no clue about the structure of the mixture. The algorithm simply splits factors suspicious for hiding more modes and makes a new learning attempt. Selection of the structure of the mixture and its constituents is harder than in the case of normal mixtures as the most parsimonious structure is not unique. However, it causes no harm: the predictive properties are decisive in the advising mode.

New factors should remain in the class of Dirichlet pdfs. The flattened and sharpened factors found according to Proposition 6.14 have this property.

Proposition 10.10 (Flattened/sharpened Markov-chain factors)
Let us consider pdfs $f = Di_\Theta(V)$, $\bar{f} = Di_\Theta\left(\bar{V}\right)$. *The pdf minimizing the KL divergence* $\mathcal{D}\left(\hat{f}\middle|\middle| f\right)$, *and having the KL divergence*

$$\mathcal{D}\left(\hat{f}\middle|\middle| \bar{f}\right) = \omega\mathcal{D}\left(f\,||\bar{f}\right)\ , \ \omega \neq 1 \text{ has the form} \tag{10.22}$$
$$\hat{f} = Di_\Theta(\lambda V + (1-\lambda)\bar{V})$$

with the scalar λ *chosen so that condition (10.22) is met.*

Proof. Proposition 6.14 implies the form of $\hat{f}$ and condition (10.22) determines the values of λ. □

Problem 10.4 (Shifted Di factor) *The explicit shift in expected value of some function $g(\Theta)$ of parameters Θ is intuitively preferable. The choice $g(\Theta) = [\ln(\Theta_1), \ldots, \ln(\Theta_{\mathring{d}})]'$ guarantees that the factor found according to Proposition 6.15 stays within the desirable Dirichlet class. The statistics of the resulting pdf solve complex algebraic equations resulting from necessary conditions of the extreme. This is not elaborated further but is worth considering.*

10.4.9 Hierarchical selection of split factors

The *hierarchical splitting* described here relies on our ability to estimate three factors instead of a single one.

During approximate estimation, Section 10.5, the ith factor within the cth component is assigned a weight $w_{ic;t} \in [0, 1]$ that expresses the expectation that the data item $d_{ic;t}$ is generated by this factor. Then, the modified parameterized factor

$$f_w(d_{ic;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1)) \equiv f_w(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}) \propto [f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic})]^{w_{ic;t}}$$

is used in the "ordinary" Bayes rule when updating the parameter estimates of this factor. Thus, for estimation purposes, we can inspect this factor independently of others and drop (temporarily) the subscripts w, c.

We want to check whether the discussed factor explains data that should be modelled by more factors while assuming that Requirement 6.2 is met, i.e., the split factor is conditioned by a regression vector that includes the true one. We formulate hypothesis H_0, that the objective pdf has two components

$$^{\lfloor o}f(d_t|\psi_t) = \beta\, ^{\lfloor d_t|\psi_t}\Theta_1 + (1-\beta)\, ^{\lfloor d_t|\psi_t}\Theta_2,\ \beta \in (0,1), \tag{10.23}$$

where Θ_1, Θ_2 have structures differing from that of Θ in the split factor as they contain more zero entries.

The alternative hypothesis H_1 assumes that $^{\lfloor o}f(d_t|\psi_t) = {}^{\lfloor d_t|\psi_t}\Theta$, i.e., denies the need for splitting.

We estimate parameters of the mixture (6.63) together with the factor in question in order to decide on the need to split. With $f(H_0) = f(H_1)$, modelling no prejudice, the Bayes rule gives

$$\text{Probability}\left(\text{split is needed}|d(\mathring{t})\right) \equiv f(H_0|d(\mathring{t})) = \frac{f(d(\mathring{t})|H_0)}{f(d(\mathring{t})|H_0) + f(d(\mathring{t})|H_1)}. \tag{10.24}$$

The v-likelihood values $f(d(\mathring{t})|H_0)$, $f(d(\mathring{t})|H_1)$ are obtained as a byproduct of the approximate estimation; see Section 10.5.

The factor is split if the probability (10.24) is high enough. The estimated factors of the mixture (10.23) are natural candidates for its replacement. We have to define initial estimates of the "small mixture" (10.23). Its components should be dissimilar and their mixture should be close to the mixture split. We shall reach this property by splitting (possibly in a random way)

$$\Psi^* = \Psi_1^* \cup \Psi_2^*, \ \Psi_1^* \cap \Psi_2^* = \emptyset \tag{10.25}$$

so that cardinalities of both sets are approximately equal. For the given occurrence matrix ${}^{\lfloor d|\psi}V_0$ defining the prior pdf on the considered factor, we specify initial occurrence matrices, $\forall \psi \in \psi^*$ and $\varepsilon \in \left(0, \min_{d\in d^*, \psi\in\psi^*} {}^{\lfloor d|\psi}V_0\right)$,

$${}^{\lfloor d|\psi}V_{1;0} = \begin{cases} {}^{\lfloor d|\psi}V_0 - \varepsilon, \forall \ \Psi \in \Psi_1^* \\ \varepsilon, \qquad\qquad \forall \ \Psi \in \Psi_2^* \end{cases}, \quad {}^{\lfloor d|\psi}V_{2;0} = \begin{cases} {}^{\lfloor d|\psi}V_0 - \varepsilon, \forall \ \Psi \in \Psi_2^* \\ \varepsilon, \qquad\qquad \forall \ d \in \Psi_1^* \end{cases}. \tag{10.26}$$

We also set the initial statistics $\kappa_{1;0} = \kappa_{2;0} = \kappa_0$, where κ_0 is the initial statistics determining estimates of component weights α. With these options, we can describe the Markov-chain version of Algorithm 6.10

Algorithm 10.5 (Hierarchical split of Markov-chain factors)

Initial mode

- *Construct the initial estimate of the mixture parameters; see Algorithm 6.8,*

$$f(\Theta) \propto Di_\alpha(\kappa_0) \prod_{c\in c^*} \prod_{d=1}^{\mathring{d}} \prod_{\psi\in\psi^*} \left[{}^{\lfloor d|\psi}\Theta\right]^{{}^{\lfloor d|\psi}V_0 - 1}.$$

- *Select significance level $\bar{P} \in (0,1)$, $\bar{P} \approx 1$ controlling the need to split.*
- *Select $\varepsilon \in \left(0, \min_{d\in d^*, \psi\in\psi^*} {}^{\lfloor d|\psi}V_0\right)$.*
- *Define sets Ψ_1^*, Ψ_2^* according to (10.25).*
- *Assign to each factor $i = 1, \ldots, \mathring{d}$, $c \in c^*$, a prior pdf*

$$f(\beta_{ic}) = Di_{\beta_{ic}}\left(0.5\kappa_{c;0}\,[1,1]\right),$$

and, for $j = 1, 2$, cf. (10.26),

$$f(\Theta_{jic}|d(0)) = \prod_{\psi\in\psi^*} Di_{\{{}^{\lfloor d|\psi}\Theta_{jic}\}_{d\in d^*}}\left({}^{\lfloor d|\psi}V_{jic;0}, \, d \in d^*\right)$$

on parameters of the "small" mixture (10.23). Indices ic stress "membership" of this mixture with the factor ic.
- *Initialize likelihood $l_{ic;0|H_0}$, $l_{ic;0|H_1}$, $i \in i^*$, $c \in c^*$.*

Sequential mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire data d_t and complement data vectors $\Psi_{i;t} = [d_{i;t}, \psi_{it}']'$.*
2. *Perform a single step of an approximate estimation, i.e., update*

$$f(\Theta|d(t-1)) \rightarrow f(\Theta|d(t)),$$

Section 10.5. It generates weights $w_{ic;t}$ measuring the degree with which data are assigned to respective factors.

Update at the same time respective values of the v-likelihood $l_{ic;t|H_1} \equiv f(d_{ic;t}|\psi_{ic;t}, d(t-1))l_{ic;t-1|H_1}$. *Here,* $d(t-1)$ *in condition stresses that the predictors for all possible discrete values* $d_{ic;t}|\psi_{ic;t}$ *are computed using the observed data* $d(t-1)$.

3. *Weight new data by respective* $w_{ic;t}$ *and update estimates of parameters of "small" mixtures (10.23)*

$$f(\beta_{ic}, \Theta_{1ic}, \Theta_{2ic}|d(t-1)) \rightarrow f(\beta_{ic}, \Theta_{1ic}, \Theta_{2ic}|d(t)).$$

Update also values of the v-likelihood $l_{ic;t|H_0} = f(d_{ic;t}|\psi_{ic;t}, d(t-1))l_{ic;t-1|H_0}$ *using predictors* $f(d_{ic;t}|\psi_{ic;t}, d(t-1))$ *constructed from the estimated two-component mixtures.*

4. *Go to the beginning of* Sequential mode *while* $t \leq \mathring{t}$.

Selection mode

1. *For* $i = 1, \ldots, \mathring{d}, \; c \in c^*$.
2. *Compare* $P_{ic} \equiv \frac{l_{ic;\mathring{t}|H_0}}{l_{ic;\mathring{t}|H_0} + l_{ic;\mathring{t}|H_1}}$ *with* $\bar{P}$.
3. *Split* $f(\Theta_{ic}|d(\mathring{t})) \rightarrow \left(f(\Theta_{1ic}|d(\mathring{t})), f(\Theta_{2ic}|d(\mathring{t}))\right)$ *if* $P_{ic} \geq \bar{P}$.

10.4.10 Techniques applicable to static mixtures

The specific Markov-chain form of the considered model brings nothing specific to this counterpart of Chapter 6. The validity of the assumptions of Proposition 6.17 is worth mentioning only.

10.5 Approximate parameter estimation

10.5.1 Quasi-Bayes estimation

A direct application of Algorithm 6.13 gives the specialized algorithm.

Algorithm 10.6 (Quasi-Bayes algorithm with common factors)

Initial (offline) mode

1. *Select statistics* $V_{ic;0}$ *determining prior distributions of the individual factors* $Di_{\Theta_{ic}}(V_{ic;0})$.
2. *Select initial values* $\kappa_{c;0} > 0$ *determining Dirichlet distribution* $Di_{\alpha}(\kappa_0)$ *of weights* α, *say,* $\sum_{c \in c^*} \kappa_{c;0} \approx$ *10% of the length* $\mathring{t}$ *of the processed data.*
3. *Select forgetting factor* λ *and alternative statistics* ${}^{\lfloor A}V$ *needed for stabilized forgetting; see (10.8).*

Sequential (online) mode, *running for* $t = 1, 2, \ldots,$

1. *Acquire the data record* d_t.

2. *Evaluate values of the predictive pdfs (2.47) for respective factors*

$$f(d_{ic;t}|\psi_{ic;t}, d(t-1), c) = \frac{\mathcal{I}(d(t)|ic)}{\mathcal{I}(d(t-1)|ic)} = \frac{{}^{\lfloor d_{ic;t}|\psi_{ic;t}}V_{t-1}}{\sum_{\tilde{d}_{ic}=1}^{\mathring{d}_i} {}^{\lfloor \tilde{d}_{ic}|\psi_{ic;t}}V_{t-1}}.$$

3. *Compute the values of predictive pdfs for respective components*

$$f(d_t|d(t-1), c) = \prod_{i \in i^*} f(d_{ic;t}|\psi_{ic;t}, d(t-1), c), \ c \in c^*.$$

4. *Compute the probabilities $w_{c;t}$, using the formula (6.86)*

$$w_{c;t} \equiv f(c_t = c|d(t)) = \frac{\frac{\kappa_{c;t-1}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c};t-1}} f(d_t|d(t-1), c)}{\sum_{\tilde{c}\in c^*} \frac{\kappa_{\tilde{c};t-1}}{\sum_{\tilde{c}\in c^*} \kappa_{\tilde{c};t-1}} f(d_t|d(t-1), \tilde{c})}$$
$$= \frac{\kappa_{c;t-1} f(d_t|d(t-1), c)}{\kappa_{\tilde{c};t-1} \sum_{\tilde{c}\in c^*} f(d_t|d(t-1), \tilde{c})}.$$

5. *Update scalars $\kappa_{c;t} = \kappa_{c;t-1} + w_{c;t}$; cf. (6.88).*
6. *Update Bayesian parameter estimates of different factors, i.e., update the corresponding statistics ${}^{\lfloor d|\psi}V_{ic;t-1} \rightarrow {}^{\lfloor d\psi}V_{ic;t}$ equivalent to the weighted version of (10.8)*

$$\begin{aligned} {}^{\lfloor d|\psi}V_{ic;t} &= \lambda \left({}^{\lfloor d|\psi}V_{ic;t-1} + w_{ic;t} \delta_{\Psi_{ic;t}, \Psi} \right) + (1-\lambda)\, {}^{\lfloor Ad|\psi}V_{ic} \\ w_{ic;t} &= \sum_{\tilde{c} \in c_i^*} w_{\tilde{c};t}. \end{aligned} \tag{10.27}$$

   The set c_i^* consists of pointers to components that contain the ith factor.
7. *Evaluate, if need be, characteristics of the Bayesian estimate of the component weights and factor parameters.*
8. *Go to the beginning of* Sequential mode *while data are available.*

10.5.2 EM estimation

A direct application of EM Algorithm 6.15 to Markov-chain factors gives the adequate specialization.

Algorithm 10.7 (EM mixture estimation with common factors)
Initial mode

- *Select the upper bound $\mathring{n}$ on the number n of iterations and set $n = 0$.*
- *Select point estimates $\hat{\Theta}_{icn} \equiv \left[{}^{\lfloor d|\psi}\hat{\Theta}_{icn} \right]$ of parameters $\left[{}^{\lfloor d|\psi}\Theta_{ic} \right] \equiv$ probabilities of $d_{ic} = d$ when conditioned by the regression vector $\psi_{ic} = \psi$ within the cth component.*
- *Select point estimates $\hat{\alpha}_{cn} = 1/\mathring{c}$ of components weights α_c.*

Iterative mode

1. *Fill the occurrence matrices* $V_{ic;0} = \varepsilon$ *(entrywise) and the statistic* $\kappa_{c;0} = \varepsilon$*, where* $\varepsilon > 0,\ \varepsilon \approx 0$.

   Sequential mode*, running for* $t = 1, 2, \ldots,$

   a) *Acquire data record* d_t *and construct the data vectors* $\Psi_{ic;t}$.

   b) *Compute values of the predictive pdfs*

   $$f\left(d_{ic;t}|\psi_{ic;t}, \hat{\Theta}_{icn}, c\right) \equiv {}^{\lfloor d_{ic;t}|\psi_{ic;t}}\hat{\Theta}_{icn}$$

   for each individual factor $i \in i^* = \left\{1, \ldots, \mathring{d}\right\}$ *in all components* $c \in c^*$ *using the parameter estimates* $\hat{\Theta}_{icn}$ *that are constant during time cycle of the sequential mode.*

   c) *Compute the values of the predictive pdfs*

   $$f(d_t|\phi_{c;t-1}, \hat{\Theta}_{cn}, c) \equiv \prod_{i \in i^*} {}^{\lfloor d_{ic;t}|\psi_{ic;t}}\hat{\Theta}_{icn}$$

   for each individual component $c \in c^*$.

   d) *Compute the probabilities* $w_{c;t}$ *approximating* δ_{c,c_t}

   $$w_{c;t} = \frac{f\left(d_t|\phi_{c;t-1}, \hat{\Theta}_{cn}, c\right) \hat{\alpha}_{cn}}{\sum_{\tilde{c} \in c^*} f(d_t|\phi_{\tilde{c};t-1}, \hat{\Theta}_{\tilde{c}n}, \tilde{c})\hat{\alpha}_{\tilde{c}n}}.$$

   e) *Update the statistics* $\kappa_{c;t} = \kappa_{c;t-1} + w_{c;t},\ c \in c^*$.

   f) *Update the statistics determining log-likelihood values describing different factors*

   $$\begin{aligned} {}^{\lfloor d|\psi}V_{ic;t} &= {}^{\lfloor d|\psi}V_{ic;t-1} + w_{ic;t}\delta_{\Psi_{ic;t},\Psi}, \text{ where} \\ w_{ic;t} &= \sum_{\tilde{c} \in c_i^*} w_{\tilde{c};t}. \end{aligned} \qquad (10.28)$$

   The set $c_i^* \equiv$ *consists of pointers to components that contain the ith factor.*

   g) *Go to the beginning of* Sequential mode *if* $t \leq \mathring{t}$. *Otherwise continue.*

2. *Find the new point estimates*

   $${}^{\lfloor d_i|\psi}\hat{\Theta}_{ic(n+1)} = \frac{{}^{\lfloor d_i|\psi}V_{ic;\mathring{t}}}{\sum_{d_i \in d_i^*} {}^{\lfloor d_i|\psi}V_{ic;\mathring{t}}} \quad \text{and} \quad \hat{\alpha}_{c(n+1)} = \frac{\kappa_{c;\mathring{t}}}{\sum_{\tilde{c} \in c^*} \kappa_{\tilde{c};\mathring{t}}}.$$

   These values maximize the nth approximate likelihood.

3. *Stop if the attained likelihood values*

   $$\sum_{c \in c^*} \kappa_{c;\mathring{t}} \left[\ln(\hat{\alpha}_{c(n+1)}) + \sum_{i \in i^*, \Psi \in \Psi^*} {}^{\lfloor d|\psi}V_{ic;\mathring{t}} \ln\left({}^{\lfloor d|\psi}\hat{\Theta}_{ic(n+1)}\right)\right] \qquad (10.29)$$

   do not increase further. Otherwise set $n = n + 1$ *and go to the beginning of* Iterative mode.

10.5.3 Batch quasi-Bayes estimation

Here, Algorithm 6.16 is specialized to Markov chains. In this way, processing-order-independent, batch quasi-Bayes estimation of Markov chains is gained.

Algorithm 10.8 (BQB estimation with common factors)

Initial mode

- *Select the upper bound $\mathring{n}$ on the number n of iterations and set $n = 0$.*
- *Select the statistics $\bar{V}_{ic}$, $\bar{\kappa}_c$ determining the alternative pdf used in flattening operation.*
- *Select the statistics V_{ic} determining prior pdfs of individual Markov-chain factors $Di_{\Theta_{ic}}(V_{ic})$.*
- *Select initial values $\kappa_c > 0$ describing estimates of component weights.*

Iterative mode

1. *Make copies $V_{ic;0} = V_{ic}$ and $\kappa_{c;0} = \kappa_c$ and set v-log-likelihood l_0 of the mixture to zero.*

   Sequential mode*, running for $t = 1, 2, \ldots,$*

   a) *Acquire the data record d_t and construct the data vectors $\Psi_{ic;t}$.*

   b) *Compute values of the predictive pdfs*

   $$f(d_{ic;t}|\psi_{ic;t}, c) = \int f(d_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c) f(\Theta_{ic}|V, \kappa)\, d\Theta_{ic}$$
   $$= \frac{{}^{\lfloor d_{ic;t}|\psi_{ic;t}}V_{ic}}{\sum_{\tilde{d}\in\tilde{d}^*_{ic}} {}^{\lfloor \tilde{d}|\psi_{ic;t}}V_{ic}}.$$

   for each individual factor $i \in i^* = \{1, \ldots, \mathring{d}\}$ in all components $c \in c^*$ using the statistics V that are constant during the time cycle.

   c) *Compute the values of predictive pdfs*

   $$f(d_t|\phi_{c;t-1}, c) \equiv \prod_{i\in i^*} f_n(d_{ic;t}|\psi_{ic;t}, c)$$

   for each individual component $c \in c^*$.

   d) *Compute the probabilities $w_{c;t}$ approximating δ_{c,c_t}*

   $$w_{c;t} \propto f(d_t|\phi_{c;t-1}, c)\kappa_c$$

   using statistics κ_c that are constant during the time cycle.

   e) *Update statistics determining posterior pdfs evolving from copies of prior statistics $V_{ic;0}$ and $\kappa_{c;0}$.*

   $$\begin{aligned} {}^{\lfloor d|\psi}V_{ic;t} &= {}^{\lfloor d|\psi}V_{ic;t-1} + w_{ic;t}\delta_{\Psi_{ic;t},\Psi_{ic}} \qquad (10.30)\\ \kappa_{c;t} &= \kappa_{c;t-1} + w_{c;t}, \text{ where}\\ w_{ic;t} &= \sum_{\tilde{c}\in c^*_i} w_{\tilde{c};t}. \end{aligned}$$

   The set c^*_i consists of pointers to components that contain ith factor.

f) Update the v-log-likelihood of the mixture

$$l_t = l_{t-1} + \ln\left[\sum_{c\in c^*} \frac{\kappa_{c;t}}{\sum_{\tilde{c}\in c^*}\kappa_{\tilde{c};t}} f(d_t|\psi_{c;t}, c)\right].$$

g) Go to the beginning of Sequential mode *if $t \le \mathring{t}$. Otherwise continue.*

2. *Stop if the v-likelihood $l_{\mathring{t}}$ of the mixture does not increase among iterations. Otherwise apply flattening operation to $f(\Theta_{ic}|d(\mathring{t}))$ and $Di_\alpha(\kappa_{c;\mathring{t}})$*

$$V_{ic} = \Lambda_n V_{ic|\mathring{t}} + (1-\Lambda_n)\bar{V}_{ic},\ \kappa_{ic} = \Lambda_n \kappa_{ic;\mathring{t}} + \Lambda_n \bar{\kappa}_c,\ \Lambda_n \to 0.5.$$

3. *Increase the counter to $n = n + 1$.*
4. *Stop if $n > \mathring{n}$; otherwise go to the beginning of* Iterative mode*.*

10.6 Structure estimation

10.6.1 Estimation of factor structure

Estimation of a factor structure can be and will be based on Bayesian structure estimation using the fact that the Markov chain belongs to the nested exponential family. The technique is applicable to a Markov chain with a low-dimensional regression vector ψ_r corresponding to the richest structure.

Proposition 10.11 (Structure estimation of nested Markov chains) *Let us consider Markov chain $f(d|\psi_r, \Theta_r) = {}^{\lfloor d|\psi_r}\Theta_r$ with the richest regression vector ψ_r and another model $f(d|\psi_s, \Theta) = {}^{\lfloor d|\psi_s}\Theta$ describing the same data $d(\mathring{t})$.*

Let $\psi_s = N\psi_r$, where N is time invariant matrix selecting entries of ψ_s from ψ_r. Let us consider Dirichlet pdfs $\prod_{\psi_S\in\psi_S^*} Di_{\,{}^{\lfloor\cdot|\psi_S,S}\Theta}\left({}^{\lfloor\cdot|\psi_S,S}V_0\right)$ as conjugate prior pdfs for both models $S \in \{s, r\}$.

Let the initial V-statistics be related by the nesting mapping

$${}^{\lfloor d|\psi_s,s}V_t = \sum_{\psi_r\in\{\psi_s = N\psi_r\}} {}^{\lfloor d|\psi_r,r}V_t, \quad \text{for } t = 0. \tag{10.31}$$

Then, the same property holds for all $t \in t^*$ if we use Bayes estimation or quasi-Bayes estimation with common data weights for both models.

The v-likelihood for the regression vector having structure $S \in \{s, r\}$ is

$$f(d(\mathring{t})|S) = \prod_{\psi_S\in\psi_S^*} \frac{\mathcal{B}\left({}^{\lfloor\cdot|\psi_S,S}V_{\mathring{t}}\right)}{\mathcal{B}\left({}^{\lfloor\cdot|\psi_S,S}V_0\right)}. \tag{10.32}$$

The multivariate beta function

$$\mathcal{B}(v) = \frac{\prod_{i=1}^{\mathring{v}} \mathit{\Gamma}(v_i)}{\mathit{\Gamma}\left(\sum_{i=1}^{\mathring{v}} v_i\right)}$$

is well defined for vectors v with positive entries v_i for which Euler gamma function $\mathit{\Gamma}$ is defined.

Let $f(s)$ and $f(r)$ be prior probabilities of s, r. Then,

$$f(s|d(\mathring{t})) = \left(1 + \frac{f(d(\mathring{t})|r)f(r)}{f(d(\mathring{t})|s)f(s)}\right)^{-1}. \tag{10.33}$$

Proof. It is a specialized version of Proposition 3.3 and extended by the use of the Bayes rule on competitive structures $S \in \{s, r\}$. □

The following algorithm uses this proposition for computing posterior probabilities of all nested structures within a given richest one.

Algorithm 10.9 (Estimate of Markov-chain factor structure)

Initial mode

- *Select the structure of a richest regression vector.*
- *Specify the flat prior statistics ${}^{\lfloor d|\psi_r,r}V_0 > 0$, ${}^{\lfloor d|\psi_r,r}V_0 \approx 0$.*
- *Collect the posterior statistics ${}^{\lfloor d|\psi_r,r}V_{\mathring{t}}$ using Bayes, batch quasi-Bayes or quasi-Bayes estimation.*
- *Specify prior probabilities $f(s)$ of competitive structures, typically a uniform one on a priori possible structures.*
- *Define the initial values of v-log-likelihood of competitive a priori possible structures $l_s = \ln(f(s))$.*

Structure estimation

1. *Do for all $\psi_r \in \psi_r^*$.*
   a) *Select the structure s determining ψ_s nested in ψ_r and having $f(s) > 0$; ψ_r has to be included among selections made.*
   b) *Evaluate for $d \in d^*$ and $t \in \{0, \mathring{t}\}$*
   $$ {}^{\lfloor d|\psi_s}V_{s;t} = \sum_{\psi_r \in \{\psi_s = N\psi_r\}} {}^{\lfloor d|\psi_r,r}V_t.$$
   c) *Compute increments of the v-log-likelihood*
   $$l_{\psi_s} = \sum_{d \in d^*} \ln\left(\frac{\mathit{\Gamma}\left({}^{\lfloor d|\psi_s,s}V_{\mathring{t}}\right)}{\mathit{\Gamma}\left({}^{\lfloor d|\psi_s,s}V_0\right)}\right) - \ln\left[\frac{\mathit{\Gamma}\left(\sum_{\tilde{d}\in d^*} {}^{\lfloor \tilde{d}|\psi_s,s}V_{\mathring{t}}\right)}{\mathit{\Gamma}\left(\sum_{\tilde{d}\in d^*} {}^{\lfloor \tilde{d}|\psi_s,s}V_0\right)}\right].$$
   d) *Update v-log-likelihood $l_s = l_s + l_{\psi_s}$.*
2. *Shift the v-log-likelihood $l_s = l_s - \max_{\{s:f(s)>0\}} l_s$*
3. *Compute posterior probabilities of individual structures $f(s|d(\mathring{t})) \propto \exp(l_s)$.*
4. *Select an estimate of a structure, typically, the MAP estimate.*

Remark(s) 10.7

1. *The prior probabilities are an efficient tool for reducing the space in which the best structure is searched for. For instance, it can reflect additional information that arises from smoothness of the originally continuous-valued regression vector; cf. [170].*
2. *Flat prior pdf is chosen in order to violate the assumption on nested prior pdfs only slightly. This condition can be met by a more careful choice of V_0. Generally, this problem should be studied but negligible practical consequences are expected.*
3. *The assumption that updating of various structures is performed with a common weight is the most serious approximation we use. It is enforced by the wish to reduce the need for updating of statistics for all structures that can be generated from ψ_r.*
4. *The initial and terminal values of statistics are sufficient for structure estimation when no forgetting is used. Otherwise, the overall predictor is a product of one-stage-ahead predictors that do not cancel in the product. This simple observation is often overlooked.*

Problem 10.5 (Connection with Bayesian networks) *Prior information resulting from the underlying dependence knowledge as studied in Bayesian-networks techniques can be used here [168, 169]. Its use is, however, infrequent for the assumed applications in which a detailed inspection is a priori prevented by the problem complexity. It may be important in specific cases.*

10.6.2 Estimation of component structure

Similar to the normal case, this task is still unsolved in an efficient algorithmic way. Bayesian-networks technology [168, 169] seems to be the proper direction to be inspected. In the mixed discrete-continuous case of the primary interest, the simple rule "place the Markov factors at the tail of the component" seems to be sufficient.

10.6.3 Merging and cancelling of components

Initiation by factor splitting is almost always employed for determining the number of components. The resulting, usually excessive, number of components has to be reduced. Thus, algorithms reducing the *number of components* form an important part of mixture estimation. A reduced and tailored set of general algorithms given in Section 6.6.4 is described here. General discussion is skipped and algorithms predicting the values of the v-log-likelihood are considered only. Factor-based merging of components and cancelling of components are described.

The formula (6.115) predicts v-likelihood after merging. Algorithm 6.24 uses it for a systematic merging of factors. It is specialized here.

Algorithm 10.10 (Systematic merging of Markov-chain factors)

Initial mode

- *Estimate a mixture with Markov-chain factors and conjugate Dirichlet prior pdf. The estimate is described by the collection of occurrence matrices*
$$\{V_{ic;t}\}_{c\in c^*,i=1,\ldots,\mathring{d},t\in\{0,\mathring{t}\}}\,.$$
The factors with the common i are supposed to describe the same entry of $d_{i;t}$ irrespective of the component number.
- *Initialize the list with rows $\rho = (i, c, \tilde{c}) \equiv i$th factor is common for components $c, \tilde{c}$. Usually, ρ is initialized as the empty one.*
- *Evaluate the individual normalization factors (see Proposition 10.1)*
$$\ln(\mathcal{I}(V_{ic;t})),\ \forall c \in c^*, i = 1, \ldots, \mathring{d}, t \in \{0, \mathring{t}\}.$$

Evaluation mode

```
For   $i = 1, \ldots, \mathring{d}$
        Set pointers $c = 1, \tilde{c} = 2$ to trial components to be merged.
  Test of the common structure
        Set the indicator of the common structure $cs = 0$.
        Set $cs = -1$ if the structures of $\Theta_{ic}$ and $\Theta_{i\tilde{c}}$ differ.
      Do if $cs = 0$
        Create the trial merger
        $\tilde{V}_{i;\mathring{t}} = V_{ic;\mathring{t}} + V_{i\tilde{c};\mathring{t}},\ \tilde{V}_{i;0} = V_{ic;0} + V_{i\tilde{c};0}$.
        Evaluate increment $\tilde{l}$ of the log-v-likelihood
        $\tilde{l} = +\left\{\ln(\mathcal{I}(\tilde{V}_{i;\mathring{t}})) - \ln(\mathcal{I}(V_{ic;\mathring{t}})) - \ln(\mathcal{I}(V_{i\tilde{c};\mathring{t}}))\right\}$
            $- \left\{\ln(\mathcal{I}(\tilde{V}_{i;0})) - \ln(\mathcal{I}(V_{ic;0})) - \ln(\mathcal{I}(V_{i\tilde{c};0}))\right\}$
      end of the test on $cs = 0$
      Do if $\tilde{l} \le 0$ or $cs < 0$
        Set $\tilde{c} = \tilde{c} + 1$.
        Go to the Test of the common structure  if $\tilde{c} \le \mathring{c}$.
        Otherwise continue.
        Set $c = c + 1$ and $\tilde{c} = c + 1$.
        Go to the beginning of Test of the common structure  if $c < \mathring{c}$.
        Otherwise go to the end of cycle over $i$.
      else replace prior and posterior estimates of factors with indexes
        $ic$ and $i\tilde{c}$ by the trial merger.
        Extend the list of common factors by $\rho = [\rho; (i, c, \tilde{c})]$.
```

```
            end of the test on improvement of v-likelihood and of $cs < 0$
    end   of the cycle over $i$
          Merging of components
  For   $c = 1, \ldots, \mathring{c} - 1$
    For   $\tilde{c} = c + 1, \ldots, \mathring{c}$
          Set $\kappa_{\tilde{c};\mathring{t}} = \kappa_{\tilde{c};\mathring{t}} + \kappa_{c;\mathring{t}}$,  $\kappa_{\tilde{c};0} = \kappa_{\tilde{c};0} + \kappa_{c;0}$ and cancel the component $c$
          if the components consist of common factors only.
    end   of the cycle over $\tilde{c}$
  end   of the cycle over $c$
```

Component cancelling

Cancelling of spurious components is also based on predicting induced change of the v-likelihood. The general approach of Chapters 6 and 8 is just specialized here. Selection of parameter values for which $f(d_t|d(t-1), \Theta_c, c) = g(d(t))$ makes this case specific. Obviously, it is sufficient to take ${}^{\lfloor d|\psi}\Theta = \frac{1}{\mathring{d}}$.

Algorithm 10.11 (Systematic cancelling of Markov components)

Initial mode

- *Estimate a Markov-chain mixture described by the collection of occurrence matrices*
$$\{V_{ic;t}\}_{c\in c^*, i=1,\ldots,\mathring{d}, t\in\{0,\mathring{t}\}}\,.$$
The factors with the common i *are supposed to describe the same entry of* $d_{i;t}$ *irrespective of the component number.*
- *Evaluate the individual normalization factors* $\ln(\mathcal{I}(V_{ic;t}))$, $\forall c \in c^*, i = 1, \ldots, \mathring{d}, t \in \{0, \mathring{t}\}$; *see Proposition 10.1.*
- *Set* $c = 1$ *and* $K = \ln(\mathring{d})$.

Evaluation mode

```
    Do while $c \le \mathring{c}$ and $\mathring{c} > 1$
                Set  $l = 0$.
      For   $i = 1, \ldots, \mathring{d}$
                $l = l - (V_{ic:\mathring{t}} - V_{ic;0})K + \ln(\mathcal{I}(V_{ic;0})) - \ln\left(\mathcal{I}(V_{ic;\mathring{t}})\right)$.
      end   of the cycle over $i$
    If $l > 0$
       Swap $c$ with $\mathring{c}$ and set $\mathring{c} = \mathring{c} - 1$,  i.e., cancel the component
       Stop if $\mathring{c} = 1$
    else
    Set $c = c + 1$
  end of the test on v-log-likelihood increase
```

10.7 Model validation with Markov-chain components

This part applies the general model validation given in Section 6.7 to Markov-chain components. The presentation is appropriately reduced.

10.7.1 Test of data homogeneity

This part is a reduced counterpart of Section 6.7.1. We deal with the situation when quality of observed behavior is classified ex post and the following hypotheses are formulated:

$H_0 \equiv$ The difference in observed consequences is due to the inseparable influence of external conditions and management. Under this hypothesis, a single mixture describes the standard $d(\mathring{t}_s)$ as well as the labelled $d(\mathring{t}_e)$ data, i.e., for $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$

$$f(d(\mathring{t})|H_0) = \int f(d(\mathring{t})|\Theta, H_0) f(\Theta|H_0)\, d\Theta. \tag{10.34}$$

$H_1 \equiv$ The difference in observed consequences is caused by differences in management. Different mixtures should be used for the standard $d(\mathring{t}_s)$ and the labelled $d(\mathring{t}_e)$ data, i.e., for $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$

$$f(d(\mathring{t})|H_1) = \int f(d(\mathring{t}_s)|\Theta_s, H_1) f(\Theta_s|H_1)\, d\Theta_s \int f(d(\mathring{t}_e)|\Theta_e, H_1) f(\Theta_e|H_1)\, d\Theta_e \tag{10.35}$$

with possibly different structures of both mixtures.

With these elements and no prejudice, $f(H_0) = f(H_1)$, the Bayes rule provides the posterior pdf $f(H_0|d(\mathring{t}))$. The common model can be accepted as a good one if this probability is high enough. The test of these hypotheses is described by the following algorithm.

Algorithm 10.12 (Test of Markov-chain data homogeneity)

1. *Run the complete model estimations on the standard $d(\mathring{t}_s)$, labelled $d(\mathring{t}_e)$ and concatenated $d(\mathring{t}) \equiv (d(\mathring{t}_s), d(\mathring{t}_e))$ data. This provides*

$$f(\Theta|d(\mathring{t}_\iota)) = Di_\alpha(\kappa_{\mathring{t}_\iota}) \prod_{c \in c^*} \prod_{d \in d^*, \psi \in \psi^*} Di_{\,\lfloor d|\psi}\Theta_c\left({}^{\lfloor d|\psi}V_{c;\mathring{t}_\iota}\right),\ \iota \in \iota^* \equiv \{s, e, \emptyset\}.$$

2. *The corresponding v-likelihood values indexed by $\iota \in \iota^*$ are obtained as by product of approximate estimations; see (10.10).*

3. *Determine the probability that a single standard model should be used*

$$f(standard|d(\mathring{t})) \equiv f(H_0|d(\mathring{t})) = \frac{f(d(\mathring{t})|H_0)}{f(d(\mathring{t})|H_0) + f(d(\mathring{t}_s)|H_1)f(d(\mathring{t}_e)|H_1)}. \tag{10.36}$$

4. *Use the single model further on if $f(standard|d(\mathring{t}))$ is close to 1. The factors that were active on $f(d(\mathring{t}_e))$ are potentially dangerous.*
5. *Use both mixtures independently if $f(standard|d(\mathring{t}))$ is close to 0. The danger of causing the situation labelled by e should be signaled whenever the model fitted to $d(\mathring{t}_e)$ makes better predictions than the model fitted to the standard data.*

10.7.2 Learning results and forgetting-based validation

The test on homogeneous data (see Algorithm 10.12) can be directly used for model validation if we take $d(\mathring{t}_s)$ as learning data and $d(\mathring{t}_v)$ as validation data unused in learning. Of course, the results of the test have to be re-interpreted appropriately. We expect that the Markov-chain counterpart of Algorithm 8.18 with varying cutting moments and forgetting-based validation, Section 6.7.3, will be mostly used.

10.7.3 Other indicators of model validity

Other techniques mentioned in Chapter 6 such as human inspection of low-dimensional projections, Section 6.7.4, or checking of operating modes, Section 6.7.5, have to be used for model validation. It cannot be over-stressed that the list of validation tests is still very much open.

Problem 10.6 (Completion of the Markov-chain suite) *In the studied context, much less experience is available with Markov chains than with normal mixtures. This chapter indicates clearly that algorithms proposed in Chapter 6 are applicable to the Markov-chain case. It also shows that a lot of aspects still have to be elaborated in order to get a complete suite of algorithms covering this important class of models.*

11

Design with Markov-chain mixtures

Markov-chain factors allow us to incorporate logical quantities and to work with mixed real- and discrete- valued data. The description of design aspects related to mixtures consisting completely of Markov chains forms the content of the chapter. The mixed case is addressed in this Chapter 13.

Formally, the majority of approximations used in connection with the normal case are unnecessary in the case of Markov-chain mixtures as we operate algebraically on finite-dimensional tables. The design is complex due the dimensionality of these tables.

Similarly, as in normal case, Chapter 9, we restrict the presentation to the case with the state in the phase form; see Agreement 5.4. This fits the considered applications and decreases the computational load caused by dimensionality. The restriction to the phase form allows us to exploit, without repetitive references, the relevant notation and relationships introduced in Agreement 9.1.

The chapter starts with a brief summary of common tools, Section 11.1. Then, academic, industrial and simultaneous designs are described in Section 11.2. The concluding Section 11.3 provides strategies supporting interactions with an operator.

11.1 Common tools

11.1.1 Model projections in design

Steady-state distribution

Evaluation of the steady-state distribution of the observed data helps us to recognize good and bad factors and components. The large number of data available justifies the assumption that the uncertainty of parameters is negligible after estimation. It implies that we can evaluate the steady-state distribution of the state vector assuming that parameters Θ are given. For the adopted phase form, the stationary pf $f(\phi_t) \equiv f([d_t', \ldots, d_{t-\partial+1}']')$ has to fulfill the identity

$$f(d_t, \ldots, d_{t-\partial+1}) = \sum_{d_{t-\partial}} \Theta_{d_t|\phi_{t-1}} f(d_{t-1}, \ldots, d_{t-\partial}) \tag{11.1}$$

$$= \sum_{d_{t-\partial}} \prod_{i=1}^{\mathring{d}} \Theta_{d_{i;t}|\psi_{i;t}} f(d_{t-1}, \ldots, d_{t-\partial}).$$

Note that the indexes of Θ related to data vector are written again right lower subscript and that the inclusion of unit in ϕ_t is superfluous in the Markov-chain case.

The full solution of (11.1) is feasible in low-dimensional cases only. Often, it is, however, sufficient to get some stationary characteristics, typically, moments. Then, it is reasonable to simulate this Markov chain and to estimate the steady-state moments from the observed realization. We rely on ergodic behavior and, ideally, apply an adapted recursive estimation of steady-state moments with a sequential stopping rule [94].

Problem 11.1 (Analysis of the Markov chain) *Classification of states of the classical Markov chain is well elaborated and gives a well-understood picture of the behavior of such dynamic system; see e.g., [79, 171]. A similar picture for higher-order Markov chains with the state in the phase form would be useful. It does not belong to the standard textbook content and should be collected.*

Marginal and conditional distributions

Low-dimensional marginal and conditional pfs provide the main technical tool for presenting results of the design of the p-system. The algorithmic description of their evaluation is given here for a known (well-estimated) Markov-chain mixture.

Essentially, Proposition 7.2 is specialized here. The specialization is split in Proposition 11.2 dealing with marginal pfs and Proposition 11.3 addressing the conditioning. Both rely on a specific structure of components, Agreement 5.4. Thus, we need a tool for changing it.

Proposition 11.1 (Permutation of an adjacent pair of factors) *Let us consider a pair of adjacent Markov-chain factors with known parameters*

$$f(\Delta_1, \Delta_2|\psi_1, \psi_2) = \Theta_{\Delta_1|\Delta_2,\psi_1} \Theta_{\Delta_2|\psi_2} = \tilde{\Theta}_{\Delta_1|\psi} \tilde{\Theta}_{\Delta_2|\Delta_1,\psi}, \quad \textit{where} \tag{11.2}$$

Θ *are known transition probabilities,*
$\tilde{\Theta}$ *are transition probabilities after permutation,*
ψ *contains the union of entries in* ψ_1, ψ_2*; the entry, which is at least in one of them, is put on the corresponding position of* ψ.

Then, the parameters $\tilde{\Theta}$, describing the permuted Δ_1, Δ_2, are generated by the following algorithm.

```
For $\psi \in \psi_1^* \cup \psi_2^*$                                                    (11.3)
  For $\Delta_1 \in \Delta_1^*$
        Set $\tilde{\Theta}_{\Delta_1|\psi} = 0$
    For   $\Delta_2 \in \Delta_2^*$
        Evaluate the joint probability $\Theta_{\Delta_1,\Delta_2|\psi} \equiv \Theta_{\Delta_1|\Delta_2,\psi}\Theta_{\Delta_2|\psi}$
        $\Theta_{\Delta_1|\Delta_2,\psi} = \Theta_{\Delta_1|\Delta_2,\psi_1}$ for $\psi_1$ included in $\psi$ and
        $\Theta_{\Delta_2|\psi} = \Theta_{\Delta_2|\psi_2}$ for $\psi_2$ in $\psi$.
        Add $\tilde{\Theta}_{\Delta_1|\psi} = \tilde{\Theta}_{\Delta_1|\psi} + \Theta_{\Delta_1,\Delta_2|\psi}$.
    end of the cycle over $\Delta_2$
    For   $\Delta_2 \in \Delta_2^*$
        Set $\tilde{\Theta}_{\Delta_2|\Delta_1,\psi} = \frac{\Theta_{\Delta_1,\Delta_2|\psi}}{\tilde{\Theta}_{\Delta_1|\psi}}$
    end of the cycle over $\Delta_2$
  end of the cycle over $\Delta_1$
end of the cycle over $\psi$
```

Proof. The result is implied by the alternative expressions for the joint probability $f(x,y) = f(x|y)f(y) = f(y|x)f(x)$ and by marginalization. □

Any desired change of the component structure can be performed by such pairwise permutations.

Proposition 11.2 (Marginal predictors for known mixtures) *Let us consider the known Markov-chain mixture in the factorized form*

$$f(\Delta_t|u_{o;t}, d(t-1)) = \sum_{c\in c^*} \alpha_c \prod_{i\in i^*} \Theta_{\Delta_{i;t}|\psi_{ic;t},c}, \quad \textit{where}$$

$\psi_{ic;t} = \left[\Delta'_{(i+1)\cdots\mathring{\Delta}c;t}, u'_{o;t}, \phi'_{c;t-1}\right]' \equiv \left[\Delta'_{(i+1)\cdots\mathring{\Delta}c;t}, \psi'_{\mathring{\Delta}c;t}\right]'$ *are regression vectors,* $i \in i^* \equiv \{1,\ldots,\mathring{\Delta}-1\}$, $c \in c^*$,

$\psi'_{\mathring{\Delta}c;t}$ *is the common part of regression vectors forming the component* c,

$\Delta_{i;t}$ *are entries of innovations* Δ_t *and* $u_{o;t}$ *are recognizable o-actions,*

$\Theta_{\Delta_{i;t}|\psi_{ic;t},c} =$ *are known transition probabilities of the cth component.*

Then, the marginal pf of $\Delta_{\underline{i}\cdots\mathring{\Delta}}$, $\underline{i} \in \{1,\ldots,\mathring{\Delta}\}$, *conditioned on the common part of the regression vector* $\psi_{\mathring{\Delta}} = \left[u'_{o;t}, \phi'_{t-1}\right]'$ *is obtained by omitting leading factors in the above product, i.e.,*

$$f(\Delta_{\underline{i}\cdots\mathring{\Delta};t}|u_{o;t}, d(t-1)) = \sum_{c\in c^*} \alpha_c \prod_{i=\underline{i}}^{\mathring{\Delta}} \Theta_{\Delta_{i;t}|\psi_{ic;t},c}.$$

Proof. Omitted. □

For reference purposes, evaluation of conditional pfs is described.

Proposition 11.3 (Conditional pfs of Markov-chain mixture) *Let us consider the known Markov-chain mixture in the factorized form*

$$f(\Delta_t|u_{o;t}, d(t-1)) = \sum_{c\in c^*} \alpha_c \prod_{i\in i^*} \Theta_{\Delta_{i;t}|\psi_{ic;t},c}, \quad \textit{where}$$

$\psi_{ic;t} = \left[\Delta'_{(i+1)\cdots\mathring{\Delta}c;t}, u'_{o;t}, \phi'_{c;t-1}\right]' \equiv \left[\Delta_{(i+1)c;t}, \psi'_{(i+1)c;t}\right]'$ *are regression vectors, $i \in \{0,1,\ldots,\mathring{\Delta}-1\}$, with the common part $\psi_{\mathring{\Delta}c;t} \equiv \left[u'_{oc;t}, \phi'_{c;t-1}\right]'$,*
$\Delta_{ic;t}$ *are entries of innovations $\Delta_{c;t}$, $u_{oc;t}$ are recognizable o-actions and $\phi_{c;t-1}$ are observable states of individual components,*
$\Theta_{\Delta_{i;t}|\psi_{i;t},c_t}$ *are known transition probabilities of the cth component.*

For simplicity, let the order of factors for all components be common, i.e., $\Delta_{c;t} = \Delta_t$. Let $j \in j^* \equiv \{\underline{j},\ldots,\mathring{i}\}$, $\underline{j} \leq \mathring{i}$ point to selected entries of Δ_t. Let $\underline{k} \in (\underline{j},\mathring{i})$ point to a subset of the selected entries in Δ_t. Then, the predictor of $\Delta_{\underline{j}\cdots\underline{k}-1;t}$ conditioned on $u_{o;t}, d(t-1)$ and $\Delta_{\underline{k}\cdots\mathring{i};t}$ is the ratio of Markov-chain mixtures

$$f(\Delta_{\underline{j}\cdots\underline{k}-1;t}|\Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1)) = \frac{\sum_{c\in c^*}\alpha_c \prod_{j=\underline{j}}^{\mathring{i}} \Theta_{\Delta_{j;t}|\psi_{j;t},c}}{\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c}} \prod_{\tilde{j}=\underline{k}}^{\mathring{i}} \Theta_{\Delta_{\tilde{j};t}|\psi_{\tilde{j};t},\tilde{c}}}. \quad (11.4)$$

For a given regression vector, the resulting predictor can be interpreted as the Markov-chain mixture

$$\begin{aligned} &f(\Delta_{\underline{j}\cdots\underline{k}-1;t}|\Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1)) \\ &\quad= \sum_{c\in c^*} A_c(\Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1)) \prod_{j=\underline{j}}^{\underline{k}-1} \Theta_{\Delta_{j;t}|\psi_{j;t},c} \\ &A_c(\Delta_{\underline{k}\cdots\mathring{i};t}, u_{o;t}, d(t-1)) = \frac{\alpha_c \prod_{j=\underline{k}}^{\mathring{i}} \Theta_{\Delta_{j;t}|\psi_{j;t},c}}{\sum_{\tilde{c}\in c^*}\alpha_{\tilde{c}} \prod_{j=\underline{k}}^{\mathring{i}} \Theta_{\Delta_{j;t}|\psi_{j;t},\tilde{c}}}. \end{aligned} \quad (11.5)$$

Proof. Conditioning $f(\beta|\gamma) = \frac{f(\beta,\gamma)}{f(\gamma)} = \frac{f(\beta,\gamma)}{\int f(\beta,\gamma)\, d\beta}$ implies the result. □

Remark(s) 11.1

1. *The formula (11.5) shows that the model obtained through the estimation of the full predictor followed by marginalization is richer than the directly estimated low-dimensional predictor: the mixture obtained by conditioning has data-dependent component weights.*

2. *Permutations of factors according to Proposition 11.1 have to be made if the considered entries are not at the assumed positions.*

Problem 11.2 (Graphical presentation of advices) *The general advice "try to reach data configurations having a high probability" applies in the Markov-chain case, too. The adequate and informative graphical representation has to be chosen or developed for the presentation of the optimized ideal low-dimensional pfs.*

11.1.2 Basic operations for fully probabilistic design

This section prepares basic evaluations needed in connection with the fully probabilistic design.

Proposition 11.4 (Expected loss for a factorized Markov chain) *Let us consider a Markov-chain component described by*

$$f(\mathit{\Delta}_t|u_{o;t},\phi_{t-1},\mathit{\Theta}) = \prod_{i\in i^*} \mathit{\Theta}_{\mathit{\Delta}_{i;t}|\psi_{i;t}} \quad \textit{with} \tag{11.6}$$

$$\mathit{\Theta} \equiv \left\{\mathit{\Theta}_{\mathit{\Delta}_i|\psi_i} \geq 0\right\}_{\mathit{\Delta}_i\in\mathit{\Delta}_i,\psi_i\in\psi_i^*,i=1,\ldots,\mathring{\mathit{\Delta}}}$$

with finite sets of all involved indexes

$$\psi'_{i;t} \equiv \left[\mathit{\Delta}'_{(i+1)\cdots\mathring{\mathit{\Delta}};t}, u'_{o;t}, \phi'_{t-1}\right] \equiv \left[\mathit{\Delta}_{i+1;t}, \psi'_{i+1;t}\right]$$

$$i = 0,1,\ldots,\mathring{\mathit{\Delta}}-1, \quad \psi'_{\mathring{\mathit{\Delta}};t} \equiv \left[u'_{o;t},\phi'_{t-1}\right].$$

Let $\omega_0(\mathit{\Psi}_t)$ be a nonnegative array indexed by $\mathit{\Psi}_t = \psi_{0;t}$. Then, the expected value

$$\mathcal{E}[\omega_0(\mathit{\Psi}_t)|u_{o;t},\phi_{t-1}] \equiv \omega_{\mathring{\mathit{\Delta}}}(\psi_{\mathring{\mathit{\Delta}};t}), \quad \textit{where } \omega_{\mathring{\mathit{\Delta}}}(\psi_{\mathring{\mathit{\Delta}};t}) \textit{ is obtained recursively}$$

$$\omega_i(\psi_{i;t}) = \sum_{\mathit{\Delta}_{i;t}\in\mathit{\Delta}_i^*} \omega_{i-1}([\mathit{\Delta}_{i;t},\psi_{i;t}])\mathit{\Theta}_{\mathit{\Delta}_{i;t}|\psi_{i;t}} \textit{ starting from } \omega_0(\psi_{0;t}). \tag{11.7}$$

Proof. The expectation is taken over entries of $\mathit{\Delta}_t$ as the remaining part of the data vector $\mathit{\Psi}_t$ is fixed by the condition $\psi_{\mathring{\mathit{\Delta}};t} \equiv \left[u'_{o;t},\phi'_{t-1}\right]'$. The the chain rule for expectations, Proposition 2.6, implies that we can evaluate conditional expectations over individual entries in $\mathit{\Delta}_t$ one-by-one starting from the first one. □

When performing fully probabilistic designs, we use the following auxiliary proposition.

Proposition 11.5 (The conditional KL divergence) *Let innovations $\Delta_t = (\Delta_{o;t}, \Delta_{p+;t})$ and*

$$f(\Delta_t|u_{o;t}, d(t-1)) = \prod_{i=1}^{\mathring{\Delta}} \Theta_{\Delta_{i;t}|\, {}^{\lfloor l}\psi_{i;t}}$$

$${}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d(t-1)) = \prod_{i=1}^{\mathring{\Delta}_o} {}^{\lfloor U}\Theta_{\Delta_{i;t}|\, {}^{\lfloor U}\psi_{i;t}}.$$

The transition probabilities $\Theta_{\Delta_{i;t}|\, {}^{\lfloor l}\psi_{i;t}}$ and user's ideal pf ${}^{\lfloor U}\Theta_{\Delta_{i;t}|\, {}^{\lfloor U}\psi_{i;t}}$ are extended as follows. $\Theta_{\Delta_{i;t}|\, {}^{\lfloor l}\psi_{i;t}} = \Theta_{\Delta_{i;t}|\psi_{i;t}}$, ${}^{\lfloor U}\Theta_{\Delta_{i;t}|\, {}^{\lfloor U}\psi_{i;t}} = {}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}$, where $\psi_{i;t}$ contains the union of entries from respective regression vectors ${}^{\lfloor l}\psi_{i;t}$, ${}^{\lfloor U}\psi_{i;t}$. In other words, the regression vectors $\psi_{i;t}$ for the ith learned factor and the corresponding factor of the user's ideal pf are common. Recall that $\psi_{\mathring{\Delta};t} = \left[u'_{o;t}, \phi'_{t-1}\right]'$ and

$$\psi_{i;t} = \left[\Delta'_{(i+1)\cdots\mathring{\Delta};t}, u'_{o;t}, \phi'_{t-1}\right]' \equiv \left[\Delta_{(i+1);t}, \psi'_{i+1;t}\right]' \text{ for } i \in \{0, \ldots, \mathring{\Delta} - 1\}.$$

Then, $$\omega(u_{o;t}, \phi_{t-1}) \equiv \sum_{\Delta_t \in \Delta^*} f(\Delta_t|u_{o;t}, \phi_{t-1}) \ln\left(\frac{f(\Delta_{o;t}|\Delta_{p+;t}, u_{o;t}, \phi_{t-1})}{{}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, \phi_{t-1})}\right)$$
$$= \omega_{\mathring{\Delta}}(\psi_{\mathring{\Delta};t}). \tag{11.8}$$

The conditional KL divergence $\omega_{\mathring{\Delta}}(\psi_{\mathring{\Delta};t})$ is found recursively

$$\omega_i(\psi_{i;t}) = {}^{\lfloor \psi}\omega_{i-1}(\psi_{i;t})$$
$$+ \sum_{\Delta_{i;t} \in \Delta_i^*} \Theta_{\Delta_{i;t}|\psi_{i;t}} \left[{}^{\lfloor \Delta\Psi}\omega_{i-1}([\Delta_{i;t}, \psi_{i;t}]) + \chi\left(i \le \mathring{\Delta}_o\right) \ln\left(\frac{\Theta_{\Delta_{i;t}|\psi_{i;t}}}{{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}}\right)\right]$$

In the recursion, running for $i = 1, \ldots, \mathring{\Delta}$, we set $\omega_0(\psi_{0;t}) = 0$,

${}^{\lfloor \psi}\omega_{i-1}(\psi_{i;t})$ *= the part of $\omega_{i-1}(\psi_{i-1;t})$ independent of entries $\Delta_{i;t}$,*

${}^{\lfloor \Delta\Psi}\omega_{i-1}([\Delta_{i;t}, \psi_{i;t}])$ *= the part of $\omega_{i-1}(\psi_{i-1;t})$ that depends also on $\Delta_{i;t}$.*

Proof. Let the innovations be split $\Delta = (\Delta_o, \Delta_{p+})$. Then, $\omega(u_{o;t}, \phi_{t-1})$

$$\equiv \sum_{\Delta_t \in \Delta^*} f(\Delta_t|u_{o;t}, \phi_{t-1}) \ln\left(\frac{f(\Delta_t|u_{o;t}, d(t-1))}{f(\Delta_{p+;t}|u_{o;t}, d(t-1))\, {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}, d(t-1))}\right)$$

$$= \sum_{\Delta_t \in \Delta^*} \Theta_{\Delta_t|u_{o;t},\phi_t} \ln\left(\prod_{i=1}^{\mathring{\Delta}_o} \frac{\Theta_{\Delta_{i;t}|\psi_{i;t}}}{{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}}\right)$$

$$\underbrace{=}_{\text{Proposition 2.6}} \sum_{\Delta_{\mathring{\Delta};t} \in \Delta^*_{\mathring{\Delta};t}} \Theta_{\Delta_{\mathring{\Delta};t}|\psi_{\mathring{\Delta};t}} \left\{ \chi(\mathring{\Delta} \le \mathring{\Delta}_o) \ln\left(\frac{\Theta_{\Delta_{\mathring{\Delta};t}|\psi_{\mathring{\Delta};t}}}{{}^{\lfloor U}\Theta_{\Delta_{\mathring{\Delta};t}|\psi_{\mathring{\Delta};t}}}\right)\right.$$

$$\underbrace{+\cdots+\sum_{\Delta_{2;t}\in\Delta^*_{2;t}}\Theta_{\Delta_{2;t}|\psi_{2;t}}\chi(2\leq\mathring{\Delta}_o)\ln\left(\frac{\Theta_{\Delta_{2;t}|\psi_{2;t}}}{{}^{\lfloor U}\Theta_{\Delta_{2;t}|\psi_{2;t}}}\right)+\omega_1(\psi_{1;t})}_{\omega_2(\psi_{2;t})}\Bigg\}.$$

$$\omega_1(\psi_{1;t})\equiv\sum_{\Delta_{1;t}\in\Delta^*_{1;t}}\Theta_{\Delta_{1;t}|\psi_{1;t}}\chi(1\leq\mathring{\Delta}_o)\ln\left(\frac{\Theta_{\Delta_{1;t}|\psi_{1;t}}}{{}^{\lfloor U}\Theta_{\Delta_{1;t}|\psi_{1;t}}}\right)$$

The part of $\omega_{i-1}(\psi_{i-1;t})$ independent of $\Delta_{i;t}$, called ${}^{\lfloor\psi}\omega_i(\psi_{i;t})$, simply adds to the part whose expectation is taken. It consists of the sum part called ${}^{\lfloor\Delta\Psi}\omega_i(\psi_{i-1;t})$ and of the "penalization" term $\chi(i\leq\mathring{\Delta}_o)\ln\left(\frac{\Theta_{\Delta_{i;t}|\psi_{i;t}}}{{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}}\right)$. □

While performing the fully probabilistic design, we work with a shifted version of the conditional KL divergence. The following slight extension of Proposition 11.5 supports the cases for which the Bellman function is given by a table $\omega_\gamma(\phi_t)$.

Proposition 11.6 (The shifted conditional KL divergence) *Let innovations $\Delta_t=(\Delta_{o;t},\Delta_{p+;t})$ and*

$$f(\Delta_t|u_{o;t},d(t-1))=\prod_{i=1}^{\mathring{\Delta}}\Theta_{\Delta_{i;t}|\psi_{i;t}},\quad {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t},d(t-1))=\prod_{i=1}^{\mathring{\Delta}_o}{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}.$$

The transition probabilities $\Theta_{\Delta_{i;t}|\psi_{i;t}}$ and the pf ${}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}$ describing the user's ideal pf are extended (see the formulation of Proposition 11.5) so that the regression vectors $\psi_{i;t}$ for the ith factor and the corresponding factor of the user's ideal pf are common. Recall that $\psi_{\mathring{\Delta};t}=\left[u'_{o;t},\phi'_{t-1}\right]'$ and

$$\psi_{i;t}=\left[\Delta'_{(i+1)\cdots\mathring{\Delta};t},u'_{o;t},\phi'_{t-1}\right]'\equiv\left[\Delta_{(i+1);t},\psi'_{i+1;t}\right]'\ \textit{for}\ i\in\{0,\ldots,\mathring{\Delta}-1\}.$$

Let, moreover, the Bellman function be the table $\omega_\gamma(\phi_t)$ with

$$\phi_t\equiv\left[d'_{t\cdots(t-\partial+1)}\right]'\Rightarrow\psi_{0;t}\equiv\Psi_t\equiv\left[\Delta'_t,u'_{o;t},\phi'_{t-1}\right]'.$$

Then, the lifted KL divergence

$$\omega_\gamma(u_{o;t},\phi_{t-1})\equiv\sum_{\Delta_t\in\Delta^*}\Theta_{\Delta_t|u_{o;t},\phi_{t-1}}\left[\ln\left(\frac{\Theta_{\Delta_{o;t}|\Delta_{p+;t},u_{o;t},\phi_{t-1}}}{{}^{\lfloor U}\Theta_{\Delta_{o;t}|u_{o;t},\phi_{t-1}}}\right)+\omega_\gamma(\phi_t)\right]$$

equals $\omega_{\mathring{\Delta}}(\psi_{\mathring{\Delta};t})$. The value $\omega_{\mathring{\Delta}}(\psi_{\mathring{\Delta};t})$ is found recursively

$$\omega_i(\psi_{i;t})={}^{\lfloor\psi}\omega_{i-1}(\psi_{i;t})$$
$$+\sum_{\Delta_{i;t}\in\Delta^*_i}\Theta_{\Delta_{i;t}|\psi_{i;t}}\left\{{}^{\lfloor\Delta\Psi}\omega_{i-1}([\Delta_{i;t},\psi_{i;t}])+\chi\left(i\leq\mathring{\Delta}_o\right)\ln\left(\frac{\Theta_{\Delta_{i;t}|\psi_{i;t}}}{{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}}\right)\right\}.$$

In the recursion, running for $i=1,\ldots,\mathring{\Delta}$, we denoted

$$^{\lfloor\psi}\omega_{i-1}(\psi_{i;t}) = \textit{the part of } \omega_{i-1}(\psi_{i-1;t}) \textit{ independent of } \Delta_{i;t}$$

$$^{\lfloor\Delta\Psi}\omega_{i-1}([\Delta_{i;t}, \psi_{i;t}]) = \textit{the part of } \omega_{i-1}(\psi_{i-1;t}) \textit{ that depends also on } \Delta_{i;t}.$$

The recursions start from the following initial condition

$$\omega_0(\psi_{0;t}) = \omega_\gamma(\phi_t) \ \ \textit{for any } d_{t-\partial} \textit{ extending } \phi_t \to \psi_{0;t} = \Psi_t.$$

Notice that during the evaluations the array ω_i "loses" entries with indexes $\Delta_i, [\Delta_i, \psi_i]$. On the other hand, ω_0 extends ω_γ so that the dimensionality loss is compensated.

Proof. The table $\omega(\phi_t)$ just adds nontrivial initial conditions to the recursion described in Proposition 11.5. It arises by extending $\omega(\phi_t)$ so that it can be taken as the function Ψ_t. □

11.1.3 Dangerous components

The discrete and finite nature of the data space modelled by Markov chains implies that there is no such universally dangerous component like the unstable one for the normal components. Thus, we have to define as dangerous those components that, when permanently active, lead to a much higher distance to the user's ideal pf than the other components; cf. Section 7.2.2.

For a quantitative expression of this statement, we write down the Markov-chain version of Proposition 7.3. It evaluates recursively the KL divergence of a pair of multivariate Markov chains $^{\lfloor I}f(d(\mathring{t}))$, $^{\lfloor U}f(d(\mathring{t}))$. The transition matrices $^{\lfloor I}\Theta_{d|\psi}$, $^{\lfloor U}\Theta_{d|\psi}$ generating them

$$^{\lfloor I}f(d(\mathring{t})) = \prod_{t\in t^*}\prod_{i=1}^{\mathring{d}} {}^{\lfloor I}\Theta_{d_{i;t}|\psi_{i;t}}, \quad ^{\lfloor U}f(d(\mathring{t})) = \prod_{t\in t^*}\prod_{i=1}^{\mathring{d}} {}^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}}, \tag{11.9}$$

describe a fixed component and the user's ideal pf, respectively. We assume that for surplus data of the p-system it holds that

$$^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}} = {}^{\lfloor I}\Theta_{\Delta_{i;t}|\psi_{i;t}} \text{ for } i = \mathring{\Delta}_o + 1, \ldots, \mathring{\Delta}. \tag{11.10}$$

The transition probabilities $^{\lfloor I}\Theta_{\Delta_{i;t}|\psi_{i;t}}$ may result from optimization, for instance, of the recognizable actions. For the considered purpose, we need not distinguish innovations and recognizable actions explicitly. Thus, we can assume that $\Delta_t = d_t$.

Proposition 11.7 (Recursive evaluation of the KL divergence) *Let $\Delta_t = (\Delta_{o;t}, \Delta_{p+;t}) = d_t$. Let pfs $^{\lfloor I}f(d(\mathring{t}))$, $^{\lfloor U}f(d(\mathring{t}))$ be generated according to (11.9) and fulfill (11.10). Then, the KL divergence of $^{\lfloor I}f(d(\mathring{t}))$ and $^{\lfloor U}f(d(\mathring{t}))$*

$$\mathcal{D}\left(^{\lfloor I}f(d(\mathring{t}))\middle|\middle| \, ^{\lfloor U}f(d(\mathring{t}))\right) \equiv \omega_\gamma(\phi_0) \textit{ is obtained recursively for} \tag{11.11}$$

$$t = \mathring{t}, \mathring{t}-1, \ldots, 1, \quad \textit{starting at } \omega_\gamma(\phi_{\mathring{t}}) = 0$$

$$\omega_\gamma(\phi_{t-1}) \equiv \sum_{\Delta_t \in \Delta^*} \prod_{i=1}^{\mathring{\Delta}} {}^{\lfloor I}\Theta_{\Delta_{i;t}|\psi_{i;t}} \left[\ln \left(\prod_{j=1}^{\mathring{\Delta}} \frac{{}^{\lfloor I}\Theta_{\Delta_{j;t}|\psi_{j;t}}}{{}^{\lfloor U}\Theta_{\Delta_{j;t}|\psi_{j;t}}} \right) + \omega_\gamma(\phi_t) \right].$$

The table $\omega_{\mathring{\Delta}}(\psi_{\mathring{\Delta};t}) \equiv \omega_\gamma(\phi_{t-1})$ *is found recursively*

$$\omega_i(\psi_{i;t}) = {}^{\lfloor \psi}\omega_{i-1}(\psi_{i;t}) \tag{11.12}$$
$$+ \sum_{\Delta_{i;t} \in \Delta_i^*} \Theta_{\Delta_{i;t}|\psi_{i;t}} \left\{ {}^{\lfloor \Delta\Psi}\omega_{i-1}([\Delta_{i;t}, \psi_{i;t}]) + \chi\left(i \le \mathring{\Delta}_o\right) \ln\left(\frac{\Theta_{\Delta_{i;t}|\psi_{i;t}}}{{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}} \right) \right\}.$$

In the recursion, running for $i = 1, \ldots, \mathring{\Delta}$, *we denoted*

${}^{\lfloor \psi}\omega_{i-1}(\psi_{i;t}) =$ *the part of* $\omega_{i-1}(\psi_{i-1;t})$ *independent of* $\Delta_{i;t}$,

${}^{\lfloor \Delta\Psi}\omega_{i-1}([\Delta_{i;t}, \psi_{i;t}]) =$ *the part of* $\omega_{i-1}(\psi_{i-1;t})$ *that depends also on* $\Delta_{i;t}$.

The recursions start from the following initial condition

$\omega_0(\psi_{0;t}) = \omega_\gamma(\phi_t)$ *for any* $d_{t-\partial}$ *extending* $\phi_t \to \psi_{0;t} = \Psi_t$.

Proof. It combines directly Propositions 7.3 and 11.6. □

Remark(s) 11.2

1. *The index* γ *is used in order to keep continuity with the notation used in Chapters 7 and 9.*
2. *The recursion (11.12) is linear with respect to the array* $\omega_\gamma(\cdot)$.
3. *It makes sense to consider the stationary version of (11.12), i.e., the case with* $\mathring{t} \to \infty$. *It is closely related to the average KL divergence per step, i.e.* $\lim_{\mathring{t}\to\infty} \frac{1}{\mathring{t}} \mathcal{D}\left({}^{\lfloor I}f(d(\mathring{t})) || {}^{\lfloor U}f(d(\mathring{t}))\right)$. *The full art the Markov-chain decision processes can be used for its analysis [79, 171]. Even numerical techniques elaborated in this area, e.g., [172], should be used for its evaluation and optimization.*

11.2 Design of the advising strategy

Particular design versions are based on a straightforward application of the fully probabilistic design with the special target resulting from the difference between sets d_o^* and d_p^*; see Proposition 7.4. Influence of advices on the interconnection of the p- and o-systems is modelled as presented in Section 7.1.3.

11.2.1 Academic design

The recommended pointers c_t to components are the only actions of the academic p-system. They are generated by the optimized academic strategy $\left\{ {}^{\lfloor I}f(c_t|d(t-1)) \right\}_{t \in t^*}$. It determines, together with the estimated Markov-chain mixture of the o-system, the ideal pfs to be presented to the operator

$$^{\lfloor I}f(d_t, c_t|d(t-1)) \equiv \prod_{i=1}^{\mathring{d}} \Theta_{d_{i;t}|\psi_{ic_t;t},c_t} \, ^{\lfloor I}f(c_t|d(t-1)) \text{ with } d_t \equiv \Delta_t, \; \mathring{d} = \mathring{\Delta}. \tag{11.13}$$

The user's ideal pf needed for the fully probabilistic design is selected as a Markov chain on the o-data $d^*_{o;t}$. This true user's ideal pf is extended on d^*_t in the way discussed in Section 5.1.5. It is extended further on c^*_t by specifying the target pf $^{\lfloor U}\Theta_{c_t|\phi_{t-1}}$ for the academic advices. Altogether,

$$^{\lfloor U}f(d_t, c_t|d(t-1)) = \prod_{i=1}^{\mathring{d}_o} \, ^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}} \, ^{\lfloor I}f(d_{p+;t}|c_t, d(t-1)) \, ^{\lfloor U}\Theta_{c_t|\phi_{t-1}}.$$

In evaluations, we use the freedom of choice of the ideal pf $^{\lfloor U}\Theta_{c_t|\phi_{t-1}}$ for the actions c_t of the constructed academic p-system. We always evaluate initially dangerous components (see Section 11.1.3) and reduce the support of $^{\lfloor U}\Theta_{c_t|\phi_{t-1}}$ to nondangerous ones. If the reduced set contains just a single component, the optimal design of the academic p-system is solved by this choice and the component is presented to the operator as the designed ideal pf. Otherwise, the Markov-chain version of the optimal academic advising strategy described by Proposition 7.10 has to be searched for.

Proposition 11.8 (Optimal fully probabilistic academic design) *Let us consider the design of the academic advisory system for the o-system described by the mixture with Markov-chain components having the state ϕ_t in the phase form. The innovation Δ_t and data record d_t coincide since no recognizable actions are available. Let $d_t = (d_{o;t}, d_{p+;t})$ =(o-data, surplus p-data) and the user's ideal pf $^{\lfloor U}f(d(\mathring{t}))$ on $d^*(\mathring{t})$ be generated by*

$$^{\lfloor U}f(d_t, c_t|d(t-1)) \equiv \prod_{i_o \in i_o^*} \, ^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}} \, ^{\lfloor I}f(d_{p+;t}|d(t-1)) \, ^{\lfloor U}\Theta_{c_t|\phi_{t-1}} \tag{11.14}$$
$$i_o^* \equiv \left\{1, \ldots, \mathring{d}_o\right\} \subset i^* \equiv \left\{1, \ldots, \mathring{d}\right\}.$$

The parameters $\Theta_{d_{i;t}|\psi_{i;t},c} \geq 0$, $\sum_{d_i \in d_i^*} \Theta_{d_i|\psi_{ic}} = 1$ of the Markov-chain mixture as well those determining the Markov-chain user's ideal pf are assumed to be known. The probabilities $\Theta_{d_{i;t}|\psi_{i;t},c}$, $^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}}$ are extended (cf. Proposition 11.5) so that the corresponding factors of the user ideal and the Markov-chain mixture have common regression vectors $\psi_{i;t} \equiv \left[d_{i+1;t}, \psi'_{i+1;t}\right]' = \left[d'_{(i+1)\cdots\mathring{d};t}, \phi'_{t-1}\right]'$, $i < \mathring{d}$, $\psi_{\mathring{d};t} \equiv \phi_{t-1}$.

The recommended pointers c_t are allowed to have nonzero values at most for those indexes in c^* that point to nondangerous components; Section 11.1.3. Then, the optimal causal academic advisory strategy, minimizing the KL divergence of $^{\lfloor I}f(d(\mathring{t}), c(\mathring{t}))$ to the user ideal

$$^{\lfloor U}f(d(\mathring{t}), c(\mathring{t})) = \prod_{t \in t^*} \, ^{\lfloor U}f(d_t, c_t|\phi_{t-1}),$$

is described by the following formulas initialized by $\omega_{\gamma}(\phi_{\mathring{t}}) = 0, \quad \forall \phi_{\mathring{t}} \in \phi^*$.

```
For $t = \mathring{t}, \ldots, 1$
    For $\phi_{t-1} \in \phi^*$
            Set $\omega_{\gamma}(\phi_{t-1}) = 0$.
      For $c_t = 1, \ldots, \mathring{c}$
        For $d_t = 1, \ldots, \mathring{d}$
            Set $\Psi_t \equiv \psi_{0;t} = [d_t', \phi_{t-1}']' \equiv [\phi_t', d_{t-\partial}']'$        (11.15)
            $\omega_0(c_t, \psi_{0;t}) \equiv \omega_0(\psi_{0;t}) = \omega_{\gamma}(\phi_t)$ for any $d_{t-\partial}$
            extending $\phi_t \rightarrow \psi_{0;t} = \Psi_t$.
        end of the cycle over $d_t$
        For $i = 1, \ldots, \mathring{d}$
            $\omega_i(c_t, \psi_{i;t}) = {}^{\lfloor \psi}\omega_{i-1}(c_t, \psi_{i;t})$
          For $d_{i;t} = 1, \ldots, \mathring{d}_i$
            $\omega_i(c_t, \psi_{i;t}) = \omega_i(c_t, \psi_{i;t}) + \Theta_{\Delta_{i;t}|\psi_{i;t}, c_t}$
            $\times \left[ {}^{\lfloor d\Psi}\omega_{i-1}(c_t, [d_{i;t}, \psi_{i;t}]) + \chi\left(i \leq \mathring{d}_o\right) \ln\left( \frac{\Theta_{d_{i;t}|\psi_{i;t}, c_t}}{{}^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}}} \right) \right]$.
          end of the cycle over $d_{i;t}$
         In the above recursion, we use
           ${}^{\lfloor \psi}\omega_{i-1}(c_t, \psi_{i;t})$ = the part of $\omega_{i-1}(c_t, \psi_{i-1;t})$ independent of $d_{i;t}$,
           ${}^{\lfloor d\Psi}\omega_{i-1}(c_t, [d_{i;t}, \psi_{i;t}])$ = the part of $\omega_{i-1}(c_t, \psi_{i-1;t})$
           that depends also on $d_{i;t}$.
        end of the cycle over $i$
           Set ${}^{\lfloor I}f(c_t|\phi_{t-1}) = {}^{\lfloor U}\Theta_{c_t|\phi_{t-1}} \exp[-\omega_{\mathring{d}}(c_t, \phi_{t-1})]$,
           $\omega_{\gamma}(\phi_{t-1}) = \omega_{\gamma}(\phi_{t-1}) + {}^{\lfloor I}f(c_t|\phi_{t-1})$.
      end of the cycle over $c_t$
      For $c_t = 1, \ldots, \mathring{c}$
            ${}^{\lfloor I}f(c_t|\phi_{t-1}) = \frac{{}^{\lfloor I}f(c_t|\phi_{t-1})}{\omega_{\gamma}(\phi_{t-1})}, \quad \omega_{\gamma}(\phi_{t-1}) = -\ln(\omega_{\gamma}(\phi_{t-1}))$
      end of the cycle over $c_t$
    end of the cycle over $\phi_{t-1}$
end of the cycle over $t$
```

Proof. We show that the Bellman function is the table $\omega_{\gamma}(\phi_t)$. For $t = \mathring{t}$, it holds with $\omega(\phi_{\mathring{t}}) = 0$. Backward induction implies that for a generic t

$$\omega_{\gamma}(\phi_{t-1}) = \sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|\phi_{t-1}) \left[\ln\left(\frac{{}^{\lfloor I}f(c_t|\phi_{t-1})}{{}^{\lfloor U}f(c_t|\phi_{t-1})} \right) + \omega_{\gamma}(c_t, \phi_{t-1}) \right], \quad \text{with}$$

$$\omega_\gamma(c_t, \phi_{t-1}) \equiv \sum_{d_t \in d^*} f(d_t|\phi_{t-1}, c_t) \left[\ln \left(\frac{f(d_t|\phi_{t-1}, c_t)}{{}^{\lfloor U} f(d_t|\phi_{t-1})} \right) + \omega_\gamma(\phi_t) \right].$$

It implies the form of the minimizing strategy with the achieved minimum being the table

$$\omega_\gamma(\phi_{t-1}) = -\ln \left[\sum_{c_t \in c^*} {}^{\lfloor U} f(c_t|\phi_{t-1}) \exp[-\omega_\gamma(c_t, \phi_{t-1})] \right].$$

Proposition 11.6 provides evaluations of the shifted KL divergence. □

The proved proposition combined with the certainty-equivalence strategy gives the following algorithm.

Algorithm 11.1 (Optimal fixed academic advising)

Initial (offline) mode

- *Estimate the Markov-chain mixture describing the o-system with the state ϕ_t in the phase form; Chapter 10.*
- *Exclude dangerous components; Sections 11.1.3, 11.1.1.*
- *Return to the learning phase if all components are dangerous.*
- *Take the nondangerous component as the ideal pf offered to the operator and stop if $\mathring{c} = 1$.*
- *Specify the true user's ideal pf on the response of the o-system*

$$^{\lfloor U} f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U} f(d_{o;t}|\phi_{t-1}) = \prod_{i=1}^{\mathring{d}_o} {}^{\lfloor U} \Theta_{d_{i;t}|\psi_{i;t}}.$$

- *Specify* <u>time invariant</u> *user's ideal pf ${}^{\lfloor U} f(c_t|\phi_{t-1}) = {}^{\lfloor U} \Theta_{c_t|\phi_{t-1}}$ on the recommended pointers. It is zero on dangerous components.*
- *Select the length of the design horizon $\mathring{t} \geq 1$.*
- *Initialize the iterative mode by setting $\omega_\gamma(\phi_{\mathring{t}}) = 0$.*

Iterative (offline) mode

- *Correct the arrays, for $t = \mathring{t}, \ldots, 1$, as given in (11.15).*
- *Store the terminal characteristics $\omega_\gamma(c, \phi)$, $c \in c^*$, $\phi \in \phi^*$, determining the optimal* steady-state strategy*; Chapter 3.*

Sequential (online) mode*, running for $t = 1, 2, \ldots$,*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Evaluate the values of the pf describing the optimal academic advising strategy*

$$^{\lfloor I} \Theta_{c_{t+1}|\phi_t} \propto {}^{\lfloor U} \Theta_{c_{t+1}|\phi_t} \exp[-\omega_\gamma(c_{t+1}, \phi_t)].$$

3. *Present to the operator projections of the ideal pf*

$$ {}^{\lfloor I}\Theta_{d_{t+1}|\phi_t} = \sum_{c_{t+1}\in c^*} {}^{\lfloor I}\Theta_{c_{t+1}|\phi_t} \prod_{i=1}^{\mathring{d}} \Theta_{d_{ic_{t+1};t+1}|\psi_{i;t+1}}. $$

4. *Go to the beginning of* Sequential mode*.*

The adaptive version of the academic design with certainty-equivalence strategy and IST patch is described by the following algorithm.

Algorithm 11.2 (Optimal adaptive academic advising)

Initial (offline) mode

- *Estimate the Markov-chain mixture describing the o-system with the state ϕ_t in the phase form; Chapter 10.*
- *Specify the true user's ideal pf on the response of the o-system*

$$ {}^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) = \prod_{i=1}^{\mathring{d}_o} {}^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}}. $$

- *Specify a* <u>*time invariant*</u> *user's ideal pf ${}^{\lfloor U}f(c_t|\phi_{t-1}) = {}^{\lfloor U}\Theta_{c_t|\phi_{t-1}}$ on the recommended pointers.*
- *Select the length of the design horizon $T \geq 1$.*

Sequential (online) mode*, running for $t = 1, 2, \ldots,$*

1. *Complement the state vector ϕ_t by newly acquired data d_t and construct the data vector $\Psi_t = [d_t', \phi_{t-1}']'$.*
2. *Update the estimates of the Markov mixture using quasi-Bayes estimation, Algorithm 6.13.*
3. *Initialize the iterative mode by setting $\omega_\gamma(\phi_{\mathring{t}}) = 0$. Skip this step if $t > 1$ and IST patch is used.*
4. *Correct iteratively the arrays ω_γ as given in (11.15) with τ in the role of t, $\tau = t + T, \ldots, t + 1$.*
5. *Evaluate the values of the pf describing the optimal academic advising strategy*
$$ {}^{\lfloor I}\Theta_{c_{t+1}|\phi_t} \propto {}^{\lfloor U}\Theta_{c_{t+1}|\phi_t} \exp[-\omega_\gamma(c_{t+1}, \phi_t)]. $$
6. *Present to the operator projections of the ideal pf*

$$ {}^{\lfloor I}\Theta_{d_{t+1}|\phi_t} = \sum_{c_{t+1}\in c^*} {}^{\lfloor I}\Theta_{c_{t+1}|\phi_t} \prod_{i=1}^{\mathring{d}} \Theta_{d_{ic_{t+1};t+1}|\psi_{i;t+1}}. $$

7. *Go to the beginning of* Sequential mode*.*

Remark(s) 11.3

1. *The strategy with maximum probable advices can simply be derived by copying general considerations of Chapter 7.*
2. *The strategy with grouped advices can also be simply derived by copying results given in Chapter 7.*
3. *Design and application of the academic advising is simple. Manipulations with high-dimensional arrays are the most complex and restrictive operations. Current technology allows, however, handle rather extensive memories. It is important, especially in online mode, when read-only operations dominate.*

11.2.2 Industrial design

The industrial design is used whenever component weights have objective meaning and the operator has no influence on them. It is addressed here.

Similarly, as in previous chapters, the following order of entries in the data record d_t is supposed $d_t = (\varDelta_{o;t}, \varDelta_{p+;t}, u_{o;t}) \equiv$ (o-innovations, surplus p-innovations, recognizable actions). For $t \in t^*, c \in c^*$, the Markov-chain components, with explicitly marked recognizable actions,

$$f(\varDelta_t|u_{o;t}, d(t-1), c) f(u_{o;t}|d(t-1), c) \equiv \prod_{i=1}^{\mathring{\varDelta}} \varTheta_{\varDelta_{i;t}|\psi_{ic;t}} \prod_{i=\mathring{\varDelta}+1}^{\mathring{d}} \varTheta_{u_{o(i-\mathring{\varDelta});t}|\psi_{ic;t}} \tag{11.16}$$

and their weights $\{\alpha_c\}_{c\in c^*}$ are assumed to be known (well estimated).

The considered extension (see Section 5.1.5) of the true user's ideal pf is

$${}^{\lfloor U}f(d(\mathring{t})) \equiv \prod_{t\in t^*} \prod_{i=1}^{\mathring{\varDelta}_o} {}^{\lfloor U}\varTheta_{\varDelta_{i;t}|\psi_{i;t}} \prod_{i=\mathring{\varDelta}_o+1}^{\mathring{\varDelta}} {}^{\lfloor I}f(\varDelta_{i;t}|\psi_{i;t}) \prod_{i=\mathring{\varDelta}+1}^{\mathring{d}} {}^{\lfloor U}\varTheta_{u_{o(i-\mathring{\varDelta});t}|\psi_{i;t}}. \tag{11.17}$$

Unlike in normal case, Chapter 9, the KL divergence can be exactly optimized.

Proposition 11.9 (Optimal fully probabilistic industrial design) *Let the joint pf*

$$\begin{aligned}&{}^{\lfloor I}f(\varDelta(\mathring{t}), u_o(\mathring{t}))\\ &\equiv \prod_{t\in t^*} \frac{\sum_{c\in c^*} \alpha_c f(\varDelta_t|u_{o;t}, d(t-1), c) f(u_{o;t}|d(t-1), c)}{\sum_{c\in c^*} \alpha_c f(u_{o;t}|d(t-1), c)}\, {}^{\lfloor I}f(u_{o;t}|d(t-1))\end{aligned}$$

with components (11.16) be determined by the optional industrial advising strategy described by pfs $\left\{ {}^{\lfloor I}f(u_{o;t}|d(t-1)) \right\}_{t\in t^*}$. Then, the optimal strategy, minimizing the KL divergence $\mathcal{D}\left({}^{\lfloor I}f \middle|\middle| {}^{\lfloor U}f \right)$ to the user's ideal pf (11.17), is generated by the following algorithm, initialized by $\omega_\gamma(\phi_{\mathring{t}}) = 0, \quad \forall \phi_{\mathring{t}} \in \phi^*$.

```
For $t = \mathring{t}, \ldots, 1$
  For $\phi_{t-1} = 1, \ldots, \mathring{\phi}$
              Set $\psi_{\mathring{d};t} = \phi_{t-1}, \quad \omega_{\gamma}(\phi_{t-1}) = 0.$          (11.18)
    For $i = \mathring{d}, \ldots, 1$
      For $d_{i;t} = 1, \ldots, \mathring{d}_i$
              Set $\psi_{i-1;t} \equiv [d_{i;t}, \psi'_{i;t}], \quad \Theta_{d_{i\cdots\mathring{d};t}|\psi_{i;t}} = 0.$
        For $c_t = 1, \ldots, \mathring{c}$
              Set $\Theta_{d_{i\cdots\mathring{d}}|\psi_{i;t},c_t} = 1.$
          For $j = i, \ldots, \mathring{d}$
              $\Theta_{d_{i\cdots\mathring{d};t}|\psi_{i;t},c_t} = \Theta_{d_{i\cdots\mathring{d}}|\psi_{i;t},c_t}\Theta_{d_{j;t}|\psi_{j;t},c_t}$
          end of the cycle over $j$
              $\Theta_{d_{i\cdots\mathring{d};t}|\psi_{i;t}} = \Theta_{d_{i\cdots\mathring{d};t}|\psi_{i;t}} + \alpha_{c_t}\Theta_{d_{i\cdots\mathring{d};t}|\psi_{i;t},c_t}.$
        end of the cycle over $c_t$
      end of the cycle over $d_{i;t}$
    end of the cycle over $i$
  For $u_{o;t} = 1, \ldots, \mathring{u}_o$
    For $\Delta_t = 1, \ldots, \mathring{\Delta}$
          Set $\omega_0(\psi_{0;t}) \equiv \omega_{\gamma}(\phi_t)$ for any $d_{t-\partial}$ extending $\phi_t \to \psi_{0;t} = \Psi_t.$
    end of the cycle over $\Delta_t$
    For $i = 1, \ldots, \mathring{\Delta}$
              $\omega_i(\psi_{i;t}) = {}^{\lfloor\psi}\omega_{i-1}(\psi_{i;t})$
      For $\Delta_{i;t} = 1, \ldots, \mathring{\Delta}_i$
              $\omega_i(\psi_{i;t}) = \omega_i(\psi_{i;t}) + \frac{\Theta_{d_{i\cdots\mathring{d};t}|\psi_{i;t}}}{\Theta_{d_{(i+1)\cdots\mathring{d};t}|\psi_{i;t}}}$
    $\times \left[ {}^{\lfloor\Delta\Psi}\omega_{i-1}([\Delta_{i;t}, \psi_{i;t}]) + \chi\left(i \leq \mathring{\Delta}_o\right) \ln\left( \frac{\Theta_{d_{i\cdots\mathring{d};t}|\psi_{i;t}}}{\Theta_{d_{(i+1)\cdots\mathring{d};t}|\psi_{i;t}} \, {}^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}}} \right) \right].$
      end of the cycle over $\Delta_{i;t}$
       In the above recursion, we use
              ${}^{\lfloor\psi}\omega_{i-1}(\psi_{i;t})$ = the part of $\omega_{i-1}(\psi_{i-1;t})$
               independent of entries $\Delta_{i;t}$
              ${}^{\lfloor\Delta\Psi}\omega_{i-1}([\Delta_{i;t}, \psi_{i;t}])$ = the part of $\omega_{i-1}(c_t, \psi_{i-1;t})$
               that depends also on $\Delta_{i;t}$
    end of the cycle over $i$
            Set ${}^{\lfloor I}f(u_{o;t}|\phi_{t-1}) = {}^{\lfloor U}\Theta_{u_{o;t}|\phi_{t-1}} \exp[-\omega_{\mathring{\Delta}}(\psi_{\mathring{\Delta}})],$
            $\omega_{\gamma}(\phi_{t-1}) = \omega_{\gamma}(\phi_{t-1}) + {}^{\lfloor I}f(u_{o;t}|\phi_{t-1})$.
    end   of the cycle over $u_{o;t}$
      For   $u_{o;t} = 1, \ldots, \mathring{u}_o$
            ${}^{\lfloor I}f(u_{o;t}|\phi_{t-1}) = \frac{{}^{\lfloor I}f(u_{o;t}|\phi_{t-1})}{\omega_{\gamma}(\phi_{t-1})}$,
      end   of the cycle over $u_{o;t}$
            $\omega_{\gamma}(\phi_{t-1}) = -\ln(\omega_{\gamma}(\phi_{t-1}))$.
    end   of the cycle over $\phi_{t-1}$
  end   of the cycle over $t$
```

Proof. We apply Proposition 7.4 with the Bellman function in the form of the table $\omega_{\gamma}(\phi_t)$. Its initial value is zero. The optimal strategy has the form

$$\begin{aligned}
{}^{\lfloor I}f(u_{o;t}|\phi_{t-1}) &= {}^{\lfloor U}\Theta_{u_t|\phi_{t-1}} \frac{\exp[-\omega_{\gamma}(u_{o;t}, \phi_{t-1})]}{\gamma(\phi_{t-1})} \quad \text{with} \\
\gamma(\phi_{t-1}) &\equiv \sum_{u_{o;t} \in u_o^*} {}^{\lfloor U}\Theta_{u_t|\phi_{t-1}} \exp[-\omega_{\gamma}(u_{o;t}, \phi_{t-1})], \\
\omega_{\gamma}(\phi_{t-1}) &= -\ln(\gamma(\phi_{t-1})).
\end{aligned}$$

The decisive shifted conditional KL divergence has the form

$$\begin{aligned}
\omega_{\gamma}(u_{o;t}, \phi_{t-1}) = &\sum_{\Delta_t \in \Delta^*} \frac{\sum_{c_t \in c^*} \alpha_c \prod_{i=1}^{\mathring{d}} \Theta_{d_{i;t}|\psi_{i;t},c_t}}{\sum_{\tilde{c}_t \in c^*} \alpha_{\tilde{c}} \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \Theta_{u_{o(i-\mathring{\Delta});t}|\psi_{i;t},\tilde{c}_t}} \\
&\times \left[\ln \left(\frac{\sum_{c_t \in c^*} \alpha_c \prod_{i=1}^{\mathring{d}} \Theta_{d_{i;t}|\psi_{i;t},c_t}}{\sum_{\tilde{c}_t \in c^*} \left[\alpha_{\tilde{c}} \prod_{i=\mathring{\Delta}_o+1}^{\mathring{d}} \Theta_{d_{i;t}|\psi_{i;t},\tilde{c}_t} \right] \prod_{i=1}^{\mathring{\Delta}_o} {}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}} \right) + \omega_{\gamma}(\phi_t) \right].
\end{aligned}$$

Similarly as in Proposition 11.6, its evaluation is done recursively using the chain rule for expectations, Proposition 2.6. Conditional and marginal pfs are evaluated according to Propositions 11.2 and 11.3. For $i = 1, \ldots, \mathring{\Delta}$, it holds that

$$\begin{aligned}
\omega_i(c_t, \psi_{i;t}) = &\, {}^{\lfloor \psi}\omega_{i-1}(c_t, \psi_{i;t}) + \sum_{\Delta_{i;t} \in \Delta_i^*} \frac{\sum_{c_t \in c^*} \alpha_c \prod_{j=i}^{\mathring{d}} \Theta_{d_{j;t}|\psi_{j;t},c_t}}{\sum_{\tilde{c}_t \in c^*} \alpha_{\tilde{c}} \prod_{j=i+1}^{\mathring{d}} \Theta_{d_{j;t}|\psi_{j;t},\tilde{c}_t}} \\
&\times \left[{}^{\lfloor \Delta\Psi}\omega(c_t, [\Delta_{i;t}, \psi_{i;t}]) \right. \\
&\left. + \chi\left(i \leq \mathring{\Delta}_o\right) \ln \left(\frac{\sum_{c_t \in c^*} \alpha_c \prod_{j=i}^{\mathring{d}} \Theta_{d_{j;t}|\psi_{j;t},c_t}}{\sum_{\tilde{c}_t \in c^*} \alpha_{\tilde{c}} \prod_{j=i+1}^{\mathring{d}} \Theta_{d_{j;t}|\psi_{j;t},\tilde{c}_t} \, {}^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}}} \right) \right].
\end{aligned}$$

Initial value of this recursion is obtained by extending $\omega_\gamma(\phi_t)$ to $\omega_0(\psi_{0;t})$ by taking them as equal whenever ϕ_t is a part of $\psi_{0;t}$. □

The proved proposition, combined with the receding-horizon certainty-equivalence strategy, justifies the following algorithm that adds the estimation environment to the design algorithm.

Algorithm 11.3 (Optimal industrial advising)

Initial (offline) mode

- *Estimate the Markov-chain mixture describing the o-system with the state ϕ_t in the phase form; see Chapter 10.*
- *Specify the true user's ideal pf (11.17) on the response of the o-system.*
- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots$,*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Update estimates of the model parameters, Section 10.5, if you deal with the adaptive advisory system.*
3. *Initialize the iterative mode by setting $\tau = t + T$ and $\omega_\gamma(\phi_\tau) = 0$. Omit the zeroing of ω_γ if $t > 1$ and the IST strategy is used.*
   Iterative mode
   - *Apply algorithm given in Proposition 11.9 while replacing t by τ and stopping at $\tau = t + 1$.*
4. *Evaluate the pf ${}^{\lfloor I}f(u_{o;t+1}|\phi_t)$ resulting from this industrial design.*
5. *Present to the operator projections of the ideal pf*

$$
{}^{\lfloor I}f(d_{t+1}|\phi_t) = \frac{\sum_{c_{t+1}\in c^*} \alpha_{c_{t+1}} \prod_{i=1}^{\mathring{d}} \Theta_{d_{i;t+1}|\psi_{i;t+1},c_{t+1}}}{\sum_{c_{t+1}\in c^*} \alpha_{c_{t+1}} \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \Theta_{d_{i;t+1}|\psi_{i;t+1},c_{t+1}}} \, {}^{\lfloor I}f(u_{o;t+1}|\phi_t).
$$

6. *Go to the beginning of* Sequential mode.

Remark(s) 11.4

The grouped version of the industrial design is to be implemented; see Proposition 7.13 and Remark 5.

11.2.3 Simultaneous academic and industrial design

The simultaneous academic and industrial design provides the best problem formulation and solution. The academic actions of the simultaneous p-system $c_t \in c^* \equiv \{1, \ldots, \mathring{c}\}$ are generated by a causal strategy $d^*(t-1) \to c^*$. The industrial part generates the recommended recognizable actions $d^*(t-1) \to u^*_{o;t}$. The potential ideal pfs are

$$\begin{aligned}
{}^{\lfloor I}f(d_t,c_t|d(t-1)) &= f(\Delta_{o;t}|\Delta_{p+;t},u_{o;t},d(t-1),c_t) \\
&\times f(\Delta_{p+;t}|u_{o;t},d(t-1),c_t)\, {}^{\lfloor I}f(c_t|u_{o;t},d(t-1))\, {}^{\lfloor I}f(u_{o;t}|d(t-1)).
\end{aligned} \tag{11.19}$$

The user's ideal pf is

$$\begin{aligned}
&{}^{\lfloor U}f(d_t,c_t|d(t-1)) = {}^{\lfloor U}f(\Delta_{o;t}|u_{o;t},d_o(t-1)) \\
&\times f(\Delta_{p+;t}|u_{o;t},d(t-1),c_t)\, {}^{\lfloor U}f(c_t|u_{o;t},d(t-1))\, {}^{\lfloor U}f(u_{o;t}|d_o(t-1)) \\
&\propto \prod_{i=1}^{\mathring{\Delta}_o} \left[{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}} \right] f(\Delta_{p+;t}|u_{o;t},d(t-1),c_t) \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} \left[{}^{\lfloor U}\Theta_{u_{o(i-\mathring{\Delta});t}|\psi_{i;t}} \right] {}^{\lfloor U}\Theta_{c_t|\psi_t}
\end{aligned}$$

$f(\Delta_{o;t}|\Delta_{p+;t},u_{o;t},d(t-1),c_t) \equiv \prod_{i=1}^{\mathring{\Delta}_o} \Theta_{\Delta_{ic_t;t}|\psi_{ic_t;t}}$ is the pf derived from the c_tth learned component describing the o-innovations;

$f(\Delta_{p+;t}|u_{o;t},d(t-1),c_t) \equiv \prod_{i=\mathring{\Delta}_o+1}^{\mathring{\Delta}} \Theta_{\Delta_{i;t}|\psi_{ic_t;t}}$ is the pf also derived from the c_tth learned component and describing the surplus p-innovations;

$\left\{ {}^{\lfloor I}f(c_t,u_{o;t}|d(t-1)) \equiv {}^{\lfloor I}f(c_t|u_{o;t},d(t-1))\, {}^{\lfloor I}f(u_{o;t}|d(t-1)) \right\}_{t\in t^*}$ is the optimized simultaneous strategy;

${}^{\lfloor U}f(\Delta_{o;t}|u_{o;t},d_o(t-1)) = \prod_{i=1}^{\mathring{\Delta}_o} {}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}$ is the true user's ideal pf on the o-innovations; see Section 5.1.5;

${}^{\lfloor U}f(u_{o;t}|d_o(t-1)) = \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} {}^{\lfloor U}\Theta_{u_{o(i-\mathring{\Delta});t}|\psi_{i;t}}$ is the true user's ideal pf on the recognizable actions $u_{o;t}$; see Section 5.1.5;

${}^{\lfloor U}f(c_t|u_{o;t},d(t-1)) = {}^{\lfloor U}\Theta_{c_t|\psi_t}$ is the pf representing the optional knob of the p-system that can respect special requirements like stability of advices. Notice that the desirable c_t may depend on $u_{o;t}$.

Proposition 11.10 (Optimal simultaneous fully probabilistic design) *Let us consider the simultaneous academic and industrial design for the o-system described by the Markov-chain mixture with the state ϕ in the phase form. The data record d_t contains both innovations $\Delta_t = (\Delta_{o;t},\Delta_{p+;t}) =$ (innovations in d_o^*, innovations in d_{p+}^*) and recognizable actions $u_{o;t}$. The o-data are $d_{o;t} = (\Delta_{o;t},u_{o;t}) =$ (o-innovations, recognizable actions). The data record d_t is ordered as follows $d_t = (\Delta_{o;t},\Delta_{p+;t},u_{o;t})$.*

The assumed influence of advices and the user's ideal pf are described by (11.19). The parameters $\Theta_{d_{i;t}|\psi_{ic;t};c}$ of the Markov-chain mixture as well those determining the true user's ideal pf ${}^{\lfloor U}\Theta_{d_{i;t}|\psi_{ic;t},c}$ are known and fixed. They are extended as in Proposition 11.5 so that the corresponding factors of the learned mixture and of the user's ideal pf have the common regression vectors $\psi_{i;t} \equiv \left[d_{i+1;t},\psi'_{i+1;t}\right]' = \left[d'_{(i+1)\cdots\mathring{d};t},\phi'_{t-1}\right]'$, $i<\mathring{d}$, $\psi_{\mathring{d};t} \equiv \phi_{t-1}$.

Let us search for the advising strategy $\left\{ {}^{\lfloor I}f(c_t,u_{o;t}|d(t-1)) \right\}_{t\in t^*}$, selecting both the recommended pointers c_t and the recognizable actions $u_{o;t}$, that minimizes the KL divergence of

$$ {}^{\lfloor I}f(d(\mathring{t}),c(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor I}f(d_t,c_t|d(t-1)) \quad \textit{to the user's ideal pf} $$

$$^{\lfloor U}f(d(\mathring{t}), c(\mathring{t})) = \prod_{t\in t^*} {}^{\lfloor U}f(d_t, c_t|d(t-1)) \text{ *with factors given by (11.19).*}$$

Then, the optimal simultaneous advising strategy is described by pfs

$$^{\lfloor I}f(c_t, u_{o;t}|\phi_{t-1}) \propto {}^{\lfloor U}\Theta_{c_t|\psi_t}\, {}^{\lfloor U}\Theta_{u_t|\phi_{t-1}} \exp\left[-\omega_\gamma(c_t, \psi_t)\right],$$

where the array $\omega_\gamma(c_t, \psi_t)$ *is generated by the following formulas initialized by* $\omega_\gamma(\phi_{\mathring{t}}) = 0, \quad \forall \phi_{\mathring{t}} \in \phi^*$.

```
For   $t = \mathring{t}, \ldots, 1$
  For   $\phi_{t-1} = 1, \ldots, \mathring{\phi}$
          Set $\omega_\gamma(\phi_{t-1}) = 0.$                                   (11.20)
    For   $(c_t, u_{o;t}) = (1,1), \ldots, (\mathring{c}, \mathring{u}_o)$
      For   $\Delta_t = 1, \ldots, \mathring{\Delta}$
          Set  $\Psi_t \equiv \psi_{0;t} = \left[d_t', \phi_{t-1}'\right]' \equiv \left[\phi_t', d_{t-\partial}'\right]'$
          $\omega_0(c_t, \psi_{0;t}) \equiv \omega_0(\psi_{0;t}) = \omega_\gamma(\phi_t)$
           for any $d_{t-\partial}$  extending  $\phi_t \to \psi_{0;t} = \Psi_t$.
      end   of the cycle over $\Delta_t$
      For   $i = 1, \ldots, \mathring{\Delta}$
          $\omega_i(c_t, \psi_{i;t}) = {}^{\lfloor \psi}\omega_{i-1}(c_t, \psi_{i;t})$
        For   $\Delta_{i;t} = 1, \ldots, \mathring{\Delta}_i$
          $\omega_i(c_t, \psi_{i;t}) = \omega_i(c_t, \psi_{i;t}) + \Theta_{\Delta_{i;t}|\psi_{i;t}, c_t}$
          $\times \left[ {}^{\lfloor \Delta\Psi}\omega_{i-1}(c_t, [\Delta_{i;t}, \psi_{i;t}]) + \chi\left(i \le \mathring{\Delta}_o\right) \ln\left(\frac{\Theta_{\Delta_{i;t}|\psi_{i;t}, c_t}}{{}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}}}\right)\right].$
        end   of the cycle over $\Delta_{i;t}$
     In the above recursion, we use
          ${}^{\lfloor \psi}\omega_{i-1}(c_t, \psi_{i;t})$ = the part of $\omega_{i-1}(c_t, \psi_{i-1;t})$
           independent of entries $\Delta_{i;t}$,
          ${}^{\lfloor \Delta\Psi}\omega_{i-1}(c_t, [\Delta_{i;t}, \psi_{i;t}])$ = the part of $\omega_{i-1}(c_t, \psi_{i-1;t})$
          that depends also on $\Delta_{i;t}$.
      end   of the cycle over $i$
        Set  ${}^{\lfloor I}f(c_t, u_{o;t}|\phi_{t-1}) = {}^{\lfloor U}\Theta_{c_t|\psi_t}\, {}^{\lfloor U}\Theta_{u_{o;t}|\phi_{t-1}} \exp[-\omega_{\mathring{\Delta}}(c_t, \phi_{t-1})]$,
        $\omega_\gamma(\phi_{t-1}) = \omega_\gamma(\phi_{t-1}) + {}^{\lfloor I}f(c_t, u_{o;t}|\phi_{t-1})$,
    end   of the cycle over $(c_t, u_{o;t})$
    For   $(c_t, u_{o;t}) = (1,1), \ldots, (\mathring{c}, \mathring{u}_o)$
        ${}^{\lfloor I}f(c_t, u_{o;t}|\phi_{t-1}) = \frac{{}^{\lfloor I}f(c_t, u_{o;t}|\phi_{t-1})}{\omega_\gamma(\phi_{t-1})}$,
        end   of the cycle over $(c_t, u_{o;t})$
                  $\omega_\gamma(\phi_{t-1}) = -\ln(\omega_\gamma(\phi_{t-1}))$.
    end   of the cycle over $\phi_{t-1}$
end   of the cycle over t
```

Proof. Discrete nature of the recognizable actions implies that formally they play the same role as recommended pointers to components. Thus, the current proposition copies the results of Proposition 11.8 just with appropriate changes in notation. □

The proposition, combined with the certainty-equivalence strategy, justifies the following algorithm for the design of the fixed simultaneous advisory system.

Algorithm 11.4 (Optimal fixed simultaneous advising)

Initial (offline) mode

- *Estimate the Markov-chain mixture describing the o-system with the state ϕ_t in the phase form; see Chapter 10.*
- *Specify the true user's ideal pf on the response of the o-system*

$$^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) = \prod_{i=1}^{\mathring{\Delta}_o} {}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}} \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} {}^{\lfloor U}\Theta_{u_{o(i-\mathring{\Delta});t}|\psi_{i;t}}.$$

- *Specify <u>time invariant</u> user's ideal pf on the recommended pointers $^{\lfloor U}\Theta_{c_t|\psi_t}$.*
- *Select the length of the design horizon $\mathring{t} \geq 1$.*
- *Initialize the iterative mode by setting $\omega_\gamma(\phi_{\mathring{t}}) = 0$.*

Iterative (offline) mode

- *Correct the arrays $\omega_\gamma(c_t, \psi_t)$, for $t = \mathring{t}, \ldots, 1$, as given in (11.20).*
- *Take the final ω_γ as the array defining the optimal* steady-state strategy, *cf. Chapter 3.*

Sequential (online) mode, *running for $t = 1, 2, \ldots$,*

1. *Acquire the data record d_t and determine the state vector ϕ_t.*
2. *Evaluate the ideal pf*

$$^{\lfloor I}f(c_{t+1}, u_{o;t+1}|\phi_t) \propto {}^{\lfloor U}\Theta_{c_{t+1}|\psi_{t+1}} {}^{\lfloor U}\Theta_{u_{o;t+1}|\phi_t} \exp\left[-\omega(c_{t+1}, \psi_{t+1})\right].$$

3. *Present to the operator projections of the ideal pf*

$$^{\lfloor I}f(d_{t+1}|\phi_t) = \sum_{c_{t+1}\in c^*} {}^{\lfloor I}\Theta_{c_{t+1},u_{o;t+1}|\phi_t} \prod_{i=1}^{\mathring{\Delta}} \Theta_{\Delta_{i;t+1}|\psi_{t+1},c_{t+1}}.$$

4. *Go to the beginning of* Sequential mode.

Proposition 11.10, combined with the receding-horizon, certainty-equivalence strategy, justifies the following adaptive design algorithm.

Algorithm 11.5 (Optimal adaptive simultaneous advising)

Initial (offline) mode

- *Estimate the Markov-chain mixture describing the o-system with the state ϕ_t in the phase form; see Chapter 10.*
- *Specify the true user's ideal pf on the response of the o-system*

$$^{\lfloor U}f(d_{o;t}|d_o(t-1)) \equiv {}^{\lfloor U}f(d_{o;t}|\phi_{t-1}) = \prod_{i=1}^{\mathring{\Delta}_o} {}^{\lfloor U}\Theta_{\Delta_{i;t}|\psi_{i;t}} \prod_{i=\mathring{\Delta}+1}^{\mathring{d}} {}^{\lfloor U}\Theta_{u_{o(i-\mathring{\Delta});t}|\psi_{i;t}}.$$

- *Specify the user's ideal pf $^{\lfloor U}\Theta_{c_t|\psi_t}$ on the recommended pointers.*
- *Select the length of the receding horizon $T \geq 1$.*

Sequential (online) mode, *running for $t = 1, 2, \ldots,$*

1. *Acquire the data record d_t and determine the state vector ϕ_t as well as the data vector $\Psi_t = \left[d_t', \phi_{t-1}'\right]'$.*
2. *Update estimates of the model parameters; Section 10.5.*
3. *Initialize the iterative mode by setting $\tau = t + T$ and $\omega_\gamma(\phi_\tau) = 0$. Omit the initialization of ω_γ if $t > 1$ and the IST strategy is adopted.*
   Iterative mode
   - *Correct the arrays $\omega_\gamma(c_\tau, \psi_\tau)$ defining the optimal strategy using Proposition 11.10 with the horizon T and τ in the role of t.*
4. *Evaluate the ideal pf on recommended pointers c_{t+1} and recognizable actions $u_{o;t}$*

$$^{\lfloor I}\Theta_{c_{t+1},u_{o;t+1}|\phi_t} \propto {}^{\lfloor U}\Theta_{c_{t+1}|\psi_{t+1}} {}^{\lfloor U}\Theta_{u_{o;t+1}|\phi_t} \exp\left[-\omega_\gamma(c_{t+1}, \psi_{t+1})\right].$$

5. *Present to the operator projections of the ideal pf*

$$^{\lfloor I}f(d_{t+1}|\phi_t) = \sum_{c_{t+1}\in c^*} {}^{\lfloor I}\Theta_{c_{t+1},u_{o;t}|\phi_t} \prod_{i=1}^{\mathring{\Delta}} \Theta_{\Delta_{ic_{t+1};t+1}|\psi_{ic_{t+1}}}.$$

6. *Go to the beginning of* Sequential mode.

Remark(s) 11.5

1. *It is worth stressing that the user's ideal pf on c_t is conditioned on ψ_t. Thus, it can modify dependence of c_t and $u_{o;t}$.*
2. *Other design variants, like selection of the most probable advices or grouped version, can be constructed directly according to Chapter 7, using the above evaluations.*

11.3 Interaction with an operator

Here, we outline designs of strategies generating the presentation and signaling actions using the model of Section 7.1.3 in the case of Markov-chain mixtures.

11.3.1 Assigning priorities

Again, Markov-chain mixtures make this task formally much easier than in the normal case. Dimensionality is the main obstacle. Thus, similarly to Chapter 9, we restricts ourselves to the simplest case with $\mathring{z}_t = 1$. We also assume that there are no surplus p-innovations in order to simplify presentation of the results.

First we find explicitly the algorithm describing how to compute the influence of the $z_t \in \{1, \ldots, \mathring{d}_o\}$. It requires evaluation of marginal pfs $\Theta_{d_{z_t;t}|\phi_{t-1}}$, ${}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1}}$. It is reasonable to evaluate them jointly in order to decrease the computational load.

Algorithm 11.6 (Marginal pfs for fixed z_t, $d_{z_t;t}$, t, ϕ_{t-1}**)**

```
                Set $\Theta_{d_{z_t;t}|\phi_{t-1}} = 0,\ {}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1}} = 0$
    For $c_t \in c^*$
                Set $\Theta_{d_{z_t;t}|\phi_{t-1},c_t} = 0,\ {}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1},c_t} = 0$
      For $d_{1\cdots(z_t-1);t} \in d^*_{1\cdots(z_t-1)}$
                Create $\psi_{z_t;t} = [d_{1\cdots(z_t-1);t}, \phi'_{t-1}]'$
                Set $\Theta_{d_{1\cdots z_t;t}|\phi_{t-1},c_t} = 1,\ {}^{\lfloor I}\Theta_{d_{1\cdots z_t;t}|\phi_{t-1},c_t} = 1$
        For $i = 1, \cdots, z_t$
                $\Theta_{d_{1\cdots z_t;t}|\phi_{t-1},c_t} = \Theta_{d_{1\cdots z_t;t}|\phi_{t-1},c_t}\Theta_{d_{i;t}|\psi_{i;t},c_t}$
                ${}^{\lfloor I}\Theta_{d_{1\cdots z_t;t}|\phi_{t-1},c_t} = {}^{\lfloor I}\Theta_{d_{1\cdots z_t;t}|\phi_{t-1},c_t}\ {}^{\lfloor I}\Theta_{d_{i;t}|\psi_{i;t},c_t}$
        end of the cycle over $i$
                $\Theta_{d_{z_t;t}|\phi_{t-1},c_t} = \Theta_{d_{z_t;t}|\phi_{t-1},c_t} + \Theta_{d_{1\cdots(z_t-1);t}|\phi_{t-1},c_t}$
                ${}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1},c_t} = {}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1},c_t} + {}^{\lfloor I}\Theta_{d_{1\cdots(z_t-1);t}|\phi_{t-1},c_t}$
      end of the cycle over $d_{1\cdots(z_t-1);t}$
                $\Theta_{d_{z_t;t}|\phi_{t-1}} = \Theta_{d_{z_t;t}|\phi_{t-1}} + \alpha_{c_t}\Theta_{d_{z_t;t}|\phi_{t-1},c_t}$
                ${}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1}} = {}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1}} + {}^{\lfloor I}\Theta_{c_t|\phi_{t-1}}\ {}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1},c_t}$
    end of the cycles over $c_t$
```

The outputs of this algorithm provides the model relating presentation action to the data record

$$f(d_t|\phi_{t-1}, z_t) = \sum_{c_t \in c^*} \alpha_{c_t} \prod_{i=1}^{\mathring{d}} \Theta_{d_{i;t}|\psi_{i;t},c_t} \frac{{}^{\lfloor I}\Theta_{d_{z_t}|\phi_{t-1}}}{\Theta_{d_{z_t}|\phi_{t-1}}}. \qquad (11.21)$$

Proposition 11.11 (Optimal fully probabilistic presentation) *Let us consider that the academic, industrial or simultaneous design has provided for the optimum model*

$$^{\lfloor I}f(d_t|d(t-1)) = \sum_{c_t \in c^*} {}^{\lfloor I}\Theta_{c_t|\phi_{t-1}} \prod_{i=1}^{\mathring{d}} {}^{\lfloor I}\Theta_{d_{i;t}|\psi_{i;t},c_t}.$$

Let the signaling strategy make the operator fully alert, i.e., signaling actions $s(\mathring{t}) \equiv 1$*. Let us assume that* $\mathring{z}_t = 1$*. Let us specify the user* <u>*data invariant*</u> *ideal pf* $^{\lfloor U}f(z_t)$ *on the set of possible presentation actions* $z^* \equiv \{1, \ldots, \mathring{d}_o\}$*.*

The optimal presentation strategy assigns the higher priority to the entries of $d_{z_t;t}$*, the higher are values of the following pf*

$$f(z_t|\phi_{t-1}) \propto {}^{\lfloor U}f(z_t) \exp\left[-\omega_\gamma(z_t, \phi_{t-1})\right]. \tag{11.22}$$

The table $\omega_\gamma(z_t, \phi_{t-1})$ *is generated by the following algorithm initialized by* $\omega_\gamma(\phi_{\mathring{t}}) = 0,\ \forall \phi_{\mathring{t}} \in \phi^*$*,*

```
For   $t = \mathring{t}, \ldots, 1$
  For   $\phi_{t-1} = 1, \ldots, \mathring{\phi}$
                Set $\omega_\gamma(\phi_{t-1}) = 0.$                                    (11.23)
    For   $i = \mathring{d}, \ldots, 1$
      For   $d_{i;t} = 1, \ldots, \mathring{d}_i$
                Set  $\psi_{i-1;t} \equiv \left[d_{i;t}, \psi'_{i;t}\right]$,   $\Theta_{d_{i \cdots \mathring{d};t}|\psi_{i;t}} = 0.$
        For   $c_t = 1, \ldots, \mathring{c}$
                Set $\Theta_{d_{i \cdots \mathring{d}}|\psi_{i;t},c_t} = 1.$
          For   $j = i, \ldots, \mathring{d}$
                $\Theta_{d_{i \cdots \mathring{d};t}|\psi_{i;t},c_t} = \Theta_{d_{i \cdots \mathring{d}}|\psi_{i;t},c_t} \Theta_{d_{j;t}|\psi_{j;t},c_t}.$
          end   of the cycle over $j$
                $\Theta_{d_{i \cdots \mathring{d};t}|\psi_{i;t}} = \Theta_{d_{i \cdots \mathring{d};t}|\psi_{i;t}} + \alpha_{c_t} \Theta_{d_{i \cdots \mathring{d};t}|\psi_{i;t},c_t}.$
        end   of the cycle over $c_t$
      end   of the cycle over $d_{i;t}$
    end   of the cycle over $i$
    For   $d_t \equiv (d_{1;t}, \ldots, d_{\mathring{d};t}) = \mathbf{1}_{\mathring{d}}, \ldots, [\mathring{d}, \ldots, \mathring{d}]$
          Set $\omega_0(\psi_{0;t}) \equiv \omega_\gamma(\phi_t)$  for any $d_{t-\partial}$
          extending  $\phi_t = [d'_t, \phi'_{t-1}]' \rightarrow \psi_{0;t} = \Psi_t.$
    end   of the cycle over $d_t \equiv (d_{1;t}, \ldots, d_{\mathring{d};t})$
    For   $z_t = 1, \ldots, \mathring{d}_o$
          Evaluate $\Theta_{d_{z_t,t}|\phi_{t-1}}$,  ${}^{\lfloor I}\Theta_{d_{z_t,t}|\phi_{t-1}}$ using Algorithm 11.6.
```

```
         For  $i = 1, \ldots, \mathring{d}$
                  $\omega_i(z_t, \psi_{i;t}) \equiv {}^{\lfloor \psi}\omega_{i-1}(z_t, \psi_{i;t})$
                  $\omega_i(z_t, \psi_{i;t}) = \omega_i(z_t, \psi_{i;t}) + \Theta_{d_{i \cdots \mathring{d};t}|\psi_{i;t}} \frac{{}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1}}}{\Theta_{d_{z_t;t}|\phi_{t-1}}}$
         $\times \left[ {}^{\lfloor \Delta\Psi}\omega_{i-1}(z_t, [\Delta_{i;t}, \psi_{i;t}]) + \chi\left(i \le \mathring{\Delta}_o\right) \ln\left( \frac{\Theta_{d_{i \cdots \mathring{d};t}|\psi_{i;t}} \, {}^{\lfloor I}\Theta_{d_{z_t;t}|\phi_{t-1}}}{\Theta_{d_{z_t;t}|\phi_{t-1}} \, {}^{\lfloor U}\Theta_{d_{i;t}|\psi_{i;t}}} \right) \right].$
         end  of the cycle over $i$
           In the above recursion, we use
         ${}^{\lfloor \psi}\omega_{i-1}(z_t, \psi_{i;t})$ = the part of $\omega_{i-1}(z_t, \psi_{i-1;t})$ independent of entries $d_{i;t}$,
         ${}^{\lfloor \Delta\Psi}\omega_{i-1}(z_t, [d_{i;t}, \psi_{i;t}])$ = the part of $\omega_{i-1}(z_t, \psi_{i-1;t})$ that depends on $d_{i;t}$.
                  Set  ${}^{\lfloor I}f(z_t|\phi_{t-1}) = {}^{\lfloor U}\Theta_{z_t|\phi_{t-1}} \exp[-\omega_{\mathring{d}}(z_t, \psi_{\mathring{d};t})]$
                  $\omega_\gamma(\phi_{t-1}) = \omega_\gamma(\phi_{t-1}) + {}^{\lfloor I}f(z_t|\phi_{t-1})$
      end  of the cycle over $z_t$
      For  $z_t = 1, \ldots, \mathring{d}$
                  ${}^{\lfloor I}f(z_t|\phi_{t-1}) = \frac{{}^{\lfloor I}f(z_t|\phi_{t-1})}{\omega_\gamma(\phi_{t-1})}$,
      end  of the cycle over $z_t$
                  $\omega_\gamma(\phi_{t-1}) = -\ln(\omega_\gamma(\phi_{t-1}))$.
   end  of the cycle over $\phi_{t-1}$
end  of the cycle over $t$
```

Proof. With the prepared model a version of industrial design has arisen. Consequently, the corresponding algorithm can be copied with an appropriate change of notation. □

11.3.2 Stimulating the operator

The signaling strategy makes the operator alert. It asks him to follow the advised actions when the ideal pf resulting from an academic, industrial or simultaneous design gives significantly smaller KL divergence to the user's ideal pf than the KL divergence of the estimated model to it.

The model relating the signaling action $s_t \in s^* \equiv \{0, 1\}$ to the response of the optimized guided o-system is

$$ {}^{\lfloor I}f(d_t, s_t|d(t-1)) \equiv {}^{\lfloor I}f(d_t|s_t, d(t-1)) \, {}^{\lfloor I}f(s_t|d(t-1)) \qquad (11.24) $$

$$ {}^{\lfloor I}f(d_t|s_t = 0, d(t-1)) \equiv f(d_t|d(t-1)) \equiv \underbrace{\sum_{c \in c^*} \alpha_c \prod_{i=1}^{\mathring{d}} \Theta_{d_{i;t}|\psi_{i;t},c}}_{\text{learned mixture}} $$

$$
\begin{aligned}
{}^{\lfloor I}f(d_t|s_t = 1, d(t-1)) &\equiv {}^{\lfloor I}f(d_t|d(t-1)) \equiv \\
&\equiv \underbrace{\sum_{c_t \in c^*} {}^{\lfloor I}f(c_t|\phi_{t-1}) \prod_{i=1}^{\mathring{d}} {}^{\lfloor I}\Theta_{d_{i;t}|\psi_{i;t},c_t}}_{\text{designed mixture}} \\
{}^{\lfloor I}f(c_t|\phi_{t-1}) &\propto {}^{\lfloor U}f(c_t|\phi_{t-1}) \exp[-\omega_\gamma(c_t, \phi_{t-1})].
\end{aligned}
$$

The model (11.24) is a special version of the model (11.13) used in the academic design. Thus, the signalling design reduces completely to it.

Problem 11.3 (Completion of the Markov-chain design suite) *This chapter covers all basic steps of the design, taking advantage of the formal simplicity of the Markov-chain case. It is fair to underline that the described algorithms are applicable in low-dimensional cases only. All available art and possibly additional results will be needed in order to cope with high-dimensional arrays. Sure steps to be taken are*

- *exploitation of the sparse nature of the underlying transition matrices,*
- *use of the well-developed art of graphs describing relationships among non-trivially occurring state transitions,*
- *orientation on searches for approximate stationary solutions.*

12

Sandwich BMTB for mixture initiation

The research described in this work has been stimulated by successes and limitations of the mean tracking (MT) clustering algorithm [67]. We have designed factors of a specific type, called *MT factors* when unifying the overall approach to the learning and design. These factors have allowed us to interpret the MT clustering algorithm in a Bayesian way. It has opened up the possibility of using the tool set available within the Bayesian paradigm for these factors and to strengthen the original algorithm as well as its exploitation. It was found that built-in independence of entries of d_t prevents the full use of such factors for advising. This makes us use some excellent features of the MT algorithm only for supporting the decisive learning step, i.e., the learning initiation.

The considered *MT uniform factor* makes the static prediction of a real-valued data entry $d_t \in [-1, 1]$. The factor is described by the pdf

$$f(d_t|a_t, d(t-1), \Theta) = f(d_t|\mu) \equiv \mathcal{M}_{d_t}(\mu) \equiv \begin{cases} \frac{1-\varepsilon}{2b} & \text{if } |d_t - \mu| \le b \\ \frac{\varepsilon}{2(1-b)} & \text{otherwise} \end{cases}, \quad (12.1)$$

where the unknown parameter $\Theta \equiv \mu \in [-1+b, 1-b]$. The width $0 < b < 1$ and $0 < \varepsilon < 1,\ \varepsilon \approx 0$ are <u>fixed</u> optional parameters of this pdf.

The *sandwich initiation*, described here, uses Bayesian methodology for estimating the key fixed parameter, namely, the *box width* b. Then, it exploits the simplicity with which the MT algorithm searches for the box position at local stationary points in the data space. The local maxima found among them then serve as initial positions for the final Bayesian step described in detail in Chapter 8.

The MT algorithm moves windows through the data space without changing the width of the "inspecting" window. This important property can be exploited more generally by introducing *MT normal factors*. They are simply normal pdfs with a fixed noise variance. They allow us to make *dynamic clustering* in the vein of uniform MT factors. The normal MT factor is

$$f(d_t|a_t, d(t-1), \Theta) = \mathcal{N}_{d_t}(\theta' \psi_t, r),$$

where $\Theta \equiv \theta \equiv$ regression coefficients in $\theta^* \equiv \mathring{\psi}$-dimensional real space, ψ is the regression vector and r is a known positive noise variance. These factors serve the same purpose as MT factors. They are more computationally demanding but — unlike MT factors — they suit the modelling of dynamic systems. Their use in the BMTB sandwich is straightforward and therefore is skipped here. They are discussed in Chapter 13 as they serve for modelling the dependence of discrete-valued variables on continuous ones.

An important but restricted role of MT factors dictates that the layout of this chapter differs from that of Chapters 8 and 10. It focuses on facts related to the selection of the prior pdf $f(\Theta)$ for normal mixture estimation.

After preparing common tools, Section 12.1.1, the conceptual BMTB algorithm is described, Section 12.2. Then, the steps creating it, distinguished by prefixes B-, MT and -B, are in Sections 12.3, 12.4 and 12.5, respectively.

12.1 Common tools

12.1.1 Properties of the MT factor

The uniform MT factor is a nonstandard pdf. Thus, it makes sense to evaluate its moments.

Proposition 12.1 (Moments of the uniform MT factor) *Let $f(d|\mu) = \mathcal{M}_d(\mu)$; see (12.1). Then, for $i = 0, 1, \ldots$*

$$\begin{aligned}\mathcal{E}\left[d^i|\mu\right] &= \frac{\varepsilon}{2(i+1)(1-b)}\left[1-(-1)^{i+1}+(\mu-b)^{i+1}-(\mu+b)^{i+1}\right]\\ &+ \frac{1-\varepsilon}{2(i+1)b}\left[(\mu+b)^{i+1}-(\mu-b)^{i+1}\right] \qquad (12.2)\\ \mathcal{E}[d|\mu] &= \mu\left[1-\frac{b\varepsilon}{1-b}-\varepsilon\right] \to \mu, \quad \mathrm{cov}[d|\mu] = \frac{b^2}{3} \text{ both for } \varepsilon \to 0.\end{aligned}$$

Proof. It follows from the identity

$$m_i \equiv \mathcal{E}\left[d^i|\mu\right] \equiv \int d^i f(d|\mu)\, dd = \frac{\varepsilon}{2(1-b)}\left[\int_{-1}^{\mu-b} d^i\, dd + \int_{b+\mu}^{1} d^i\, dd\right]$$
$$+ \frac{1-\varepsilon}{2b}\int_{\mu-b}^{\mu+b} d^i\, dd \frac{\varepsilon}{2(i+1)(1-b)}\left[1-(-1)^{i+1}+(\mu-b)^{i+1}-(\mu+b)^{i+1}\right]$$
$$+ \frac{1-\varepsilon}{2(i+1)b}\left[(\mu+b)^{i+1}-(\mu-b)^{i+1}\right].$$

For $i = 1$, $\mathcal{E}[d|\mu] = \mu\left[1-\frac{b\varepsilon}{1-b}-\varepsilon\right]$. The variance is found in a similar way using the identity $\mathrm{cov}[d|\mu] = \mathcal{E}\left[d^2|\mu\right] - \mathcal{E}^2[d|\mu]$. □

12.1.2 KL divergence of MT factors

For judging the proximity of a pair of MT factors, we use their KL divergence.

Proposition 12.2 (KL divergence of parameterized MT factors) *Let d_t be a scalar quantity in $[-1,1]$ and $f(d(t)) = \prod_{t\in t^*} \mathcal{M}_{d_t}(\mu)$, $\tilde{f}(d(t)) = \prod_{t\in t^*} \mathcal{M}_{d_t}(\tilde{\mu})$; see (12.1). The parameters $\mu, \tilde{\mu}$ are assumed to be known. Their KL divergence is given by the formula*

$$\begin{aligned}\mathcal{D}(f\|\tilde{f}) &= \mathring{t}K\operatorname{abs}(\mu-\tilde{\mu}) \text{ with} \\ K &\equiv \frac{1-\varepsilon}{2b}\ln\left(\frac{(1-\varepsilon)(1-b)}{b\varepsilon}\right) + \frac{\varepsilon}{2(1-b)}\ln\left(\frac{b\varepsilon}{(1-b)(1-\varepsilon)}\right).\end{aligned} \tag{12.3}$$

Proof.

$$\begin{aligned}\mathcal{D}(f\|\tilde{f}) &\equiv \int f(d(t)|\mu)\ln\left(\frac{f(d(t)|\mu)}{f(d(t)|\tilde{\mu})}\right)dd(t) = \sum_{t\in t^*}\int \mathcal{M}_{d_t}(\mu)\ln\left(\frac{\mathcal{M}_{d_t}(\mu)}{\mathcal{M}_{d_t}(\tilde{\mu})}\right)dd_t \\ &= \sum_{t\in t^*}\left[\int_{\mu-b}^{\mu+b}\frac{1-\varepsilon}{2b}\ln\left(\frac{1-\varepsilon}{2b\mathcal{M}_{d_t}(\tilde{\mu})}\right)dd_t\right. \\ &+ \int_{-1}^{\mu-b}\frac{\varepsilon}{2(1-b)}\ln\left(\frac{\varepsilon}{2(1-b)\mathcal{M}_{d_t}(\tilde{\mu})}\right)dd_t \\ &+ \left.\int_{\mu+b}^{1}\frac{\varepsilon}{2(1-b)}\ln\left(\frac{\varepsilon}{2(1-b)\mathcal{M}_{d_t}(\tilde{\mu})}\right)dd_t\right].\end{aligned}$$

Let us evaluate a generic term assuming that $\tilde{\mu} \le \mu$

$$\begin{aligned}&\int_{\mu-b}^{\mu+b}\frac{1-\varepsilon}{2b}\ln\left(\frac{1-\varepsilon}{2b\mathcal{M}_{d_t}(\tilde{\mu})}\right)dd_t + \int_{-1}^{\mu-b}\frac{\varepsilon}{2(1-b)}\ln\left(\frac{\varepsilon}{2(1-b)\mathcal{M}_{d_t}(\tilde{\mu})}\right)dd_t \\ &+ \int_{\mu+b}^{1}\frac{\varepsilon}{2(1-b)}\ln\left(\frac{\varepsilon}{2(1-b)\mathcal{M}_{d_t}(\tilde{\mu})}\right)dd_t \\ &= \int_{\tilde{\mu}+b}^{\mu+b}\frac{1-\varepsilon}{2b}\ln\left(\frac{(1-\varepsilon)2(1-b)}{2b\varepsilon}\right)dd_t \\ &+ \int_{\tilde{\mu}-b}^{\mu-b}\frac{\varepsilon}{2(1-b)}\ln\left(\frac{2b\varepsilon}{2(1-b)(1-\varepsilon)}\right)dd_t \\ &= (\mu-\tilde{\mu})\underbrace{\left[\frac{1-\varepsilon}{2b}\ln\left(\frac{(1-\varepsilon)(1-b)}{b\varepsilon}\right) + \frac{\varepsilon}{2(1-b)}\ln\left(\frac{b\varepsilon}{(1-b)(1-\varepsilon)}\right)\right]}_{K}.\end{aligned}$$

For $\tilde{\mu} > \mu$, the role of these variables just exchanges. □

12.1.3 Estimation and prediction with MT factors

MT factors do not belong to the exponential family but their estimation is still relatively simple. They are similar to the uniform pdf with an unknown center μ and a fixed length $2b$. The fixed nonzero value $0.5\varepsilon/(1-b)$ considered out of the interval $[\mu-b,\mu+b]$ prevents the learning from breaking down due to outlying data.

Proposition 12.3 (Estimation and prediction of the MT factor) *Let natural conditions of decision making, Requirement 2.5, hold, and the MT factor (12.1) be estimated. Let the uniform prior pdf on $[-1+b,1-b]$ be used for μ. Then, the posterior pdf of μ has the form*

$$f(\mu|d(\mathring{t})) \propto \left[\frac{1-\varepsilon}{2b}\right]^{\nu(\mu,\mathring{t})} \left[\frac{\varepsilon}{2(1-b)}\right]^{\mathring{t}-\nu(\mu,\mathring{t})} \equiv \mathcal{M}_{\mu}(\nu(\mu,\mathring{t})), \text{ where} \tag{12.4}$$

$\nu(\mu,\mathring{t})$ denotes the number of data points among $d(\mathring{t})$ fulfilling $|d_t-\mu|\leq b$.

The MAP point estimate $\hat{\mu}(d(\mathring{t}))$ of μ maximizes $\nu(\mu,\mathring{t})$, i.e., $\nu\left(\hat{\mu}(d(\mathring{t})),\mathring{t}\right)\geq \nu(\mu,\mathring{t})$, $\forall \mu\in\mu^*\equiv[-1+b,1-b]$.

For a given MAP estimate of parameters the Bayesian prediction can be approximated by the "certainty-equivalence" predictor

$$f(d|d(t)) \approx \mathcal{M}_d(\hat{\mu}(d(\mathring{t}))). \tag{12.5}$$

Proof. By a direct evaluation. The prediction relies on the sharpness of the posterior pdf $f(\mu|d(\mathring{t}))$ around its modes. □

The fact that the MT algorithm [67] provides efficiently the MAP estimate $\hat{\mu}(d(\mathring{t}))$ of μ is the main reason for discussing this special factor. Let us recall the essence of the original MT algorithm [67].

Algorithm 12.1 (Basic MT algorithm)

Initial mode

- *Remove outliers from the raw data.*
- *Normalize data $d(\mathring{t})$ so that they belong to $[-1,1]$.*
- *Select the parameter $0<b<1$.*
- *Select the initial guess $\hat{\mu}_0$ in $\mu^*\equiv[b-1,1-b]$.*
- *Select the upper bound $\mathring{n}$ on the number n of iterations.*

Iterative mode

```
For   $n = 1,\ldots,\mathring{n}$
          Set $\hat{\mu}_n \equiv$ sample mean of data belonging to $[\hat{\mu}_{n-1}-b,\hat{\mu}_{n-1}+b]$.
          Break if $|\hat{\mu}_n-\hat{\mu}_{n-1}|<\varepsilon$.
end   of the cycle over $n$
```

Remark(s) 12.1

1. *The algorithm is described for a scalar d_t. It is applicable in high dimensions without a change and with a very low additional computational cost.*
2. *The estimation stops at the local maximum of the empirical pdf of data. This fits in well with our wish to find all significant modes since the local extremes are of direct interest.*
3. *The adequate choice for the fixed parameters of the MT factor decides upon the achieved quality. Their tuning is described below. It is based on practical similarity of uniform and normal factors. Essentially, a normal static mixture is fitted to the considered scalar normalized data and its (averaged) standard deviation is taken as b. The found mean value of the normal pdf may serve as an initial guess of the unknown μ (B-step). This connection is used in the opposite direction, too. The MT factors estimated in the MT step provide initial positions of factors in the final normal mixture (-B step).*
4. *The considered normalization is quite sensitive to outlying data. Their presence shrinks data points too much; consequently, some modes are easily "over-looked" during clustering.*
5. *Estimation of the mixture consisting of MT components is extremely simple and fast. It serves well as a good starting point for the estimation of normal mixtures that are able to describe correlation between entries of the multivariate record d. It is important in order to grasp dynamic properties of the modelled system: MT factors make static clustering of dynamic data; see Section 6.4.9.*
6. *MT factors are tightly related to histograms with a fixed box length distributed on a variable orthogonal grid.*
7. *Many windows describing MT factors can be shifted in parallel.*
8. *The original MT algorithm lacks*
   - *systematic but feasible initiation of the MT algorithm for large data sets, for large $\mathring{t}$,*
   - *recognition of whether the reached stationary point is a local minimum or maximum,*
   - *systematic merging and cancelling of components,*
   - *structure estimation whenever MT algorithm is applied to dynamic data as reflected in Proposition 6.17.*

   These aspects are also addressed here.

12.2 Conceptual BMTB algorithm

As said above, this chapter describes an attempt to use simplicity of the MT algorithm and its ability to find efficiently local stationary points in the data space. The initial univariate Bayesian estimation, *B-step* on respective data axes, serves well for the choice of critical parameters of the MT algorithm.

Then, the MT algorithm is applied. The obtained restricted results are extended to the initial mixture. This initial mixture resulting from *BMT part* of the discussed "sandwich" initializes the complete Bayesian estimation, which forms the final *-B step.*

Particular steps of the following conceptual BMTB algorithm are elaborated on subsequent sections. The presentation is predominantly made for multivariate static components, i.e., in terms of data records d_t. It is worth stressing that the dynamic case is addressed simply by treating data vectors Ψ_t instead of the data records.

Algorithm 12.2 (Conceptual BMTB algorithm)

B-step*, run for* $i = 1, \ldots, \mathring{d}$,

Initial mode

- *Remove outliers.*
- *Normalize the data so that the majority of them have entries in the range* $[-1, 1]$.
- *Specify the initial, univariate normal mixture covering safely and uniformly the whole range* $[-1, 1]$.

Learning mode

- *Use an approximate Bayesian estimation on axes* $d_i(\mathring{t})$ *(Section 8.5) for learning this normal mixture.*
- *Specify box widths* b_i *specific to the considered entry* $d_i(\mathring{t})$.

MT step

Initial mode

- *Select the structure of the data vectors* Ψ *to be described by the static normal mixture as justified by Proposition 6.17.*
- *Specify optional parameters of the MT step using estimation results obtained in the B-step.*

Learning mode

- *Apply the MT algorithm to get initial guesses of cluster positions in the space of data vectors* Ψ^*.
- *Use the BMT results for defining an initial guess of the mixture estimated in the -B step.*

-B step

Initial mode

- *Specify optional parameters of the Bayesian initialization and subsequent estimation using the BMT results.*

Learning mode

- *Apply an initiation algorithm for normal mixtures (see Section 8.4) using the initial mixture obtained in the BMT steps.*

- *Apply iterative learning to the mixture obtained; see Section 8.5.*
- *Validate the model; see Section 8.7. Stop if validation is successful; otherwise go to **MT step**.*

12.3 B-step: preparation of MT parameters

One-dimensional static mixtures are estimated in this step. For the considered purpose, it is sufficient to initialize the mixture estimation by selecting a sufficiently high number of components $\mathring{c}$, typically a few tens. The initial positions of component estimates are equidistantly distributed over the dominant range of scaled data, i.e., over the interval [-1,1]. Components can be just shifted copies of a normal pdf. Their common noise variance should make them distinguishable but overlapping, so that distance of centers should be about 4 to 6 standard deviations of noise and the parameter variance should be of the order of the noise variance. These values should be sufficiently fixed, so that the standard recommended value of degrees of freedom $\nu_{i;0} \approx 0.1\mathring{t}$ is applicable; see Remark 6.5. Universally, equal component weights respecting the recommendation $\kappa_{c;0} \approx 0.1\mathring{t}/\mathring{c}$ can be and should be chosen. Let us summarize formally the B-step.

Algorithm 12.3 (Initialization of B-step: apply entrywise!)

1. *Perform data preprocessing with outlier removal.*
2. *Scale the data to have zero mean and unit variance.*
3. *Select the number $\mathring{c} \approx 20$ of univariate components.*
4. *Define $\nu_{i;0} \approx 0.1\mathring{t}$, $i = 1, \ldots, \mathring{d}$, $\kappa_{c;0} \approx 0.1\mathring{t}/\mathring{c}$, $c = 1, \ldots, \mathring{c}$.*
5. *Define the distance between component centers $s = \frac{2}{\mathring{c}}$.*
6. *Complement the definition of the prior pdf by selecting $L'DL$ decomposition of respective extended information matrices as follows*

$$L_{c;0} = \begin{bmatrix} 1 & 0 \\ -1 + \left(c - \frac{1}{2}\right) s & 1 \end{bmatrix}, \quad D_{c;0} \equiv D_0 = \begin{bmatrix} \frac{(\nu_0 - 2)s^2}{16} & \\ & 1 \end{bmatrix}, \quad c \in c^*.$$

The learning mode consists of a direct application of the normal mixture estimation, presented in Section 8.5, to this static univariate mixture. Thus, it remains to use its results for selection of box widths and possibly of initial positions of boxes entering the MT algorithm. The discussion is formalized with the help of the following simple proposition.

Proposition 12.4 (Extension of single variate pdfs) *Let $f(x)$, $f(y)$ be a pair of pdfs and let us search for the joint pdf $f(x, y)$ that*

1. *has $f(x)$, $f(y)$ as its marginal pdfs, i.e., $f(x) = \int f(x, y)\, dy$, $f(y) = \int f(x, y)\, dx$,*

2. *is the most uncertain among such pdfs (in order to respect just the considered constraints); the KL divergence to a uniform, possibly improper, pdf is taken as the measure of uncertainty.*

Then, $f(x,y) = f(x)f(y)$.

Proof. We minimize a convex functional on a convex set so the we can formulate the problem in terms of the Lagrangian functional

$$\begin{aligned}
&\int f(x,y)\left[\ln\left(f(x,y)\right) + \lambda(x) + \lambda(y)\right]\, dxdy \\
&\quad = \int f(x,y) \ln\left(\frac{f(x,y)}{\left(\frac{\exp[-\lambda(x)-\lambda(y)]}{\int \exp[-\lambda(\tilde{x})-\lambda(\tilde{y})]\, d\tilde{x}d\tilde{y}}\right)}\right) dxdy \\
&\qquad - \ln\left(\int \exp[-\lambda(\tilde{x}) - \lambda(\tilde{y})]\, d\tilde{x}d\tilde{y}\right) \\
&\quad = \mathcal{D}\left(f(x,y)\left|\left|\frac{\exp(-\lambda(x))}{\int \exp[-\lambda(\tilde{x})]\, d\tilde{x}} \frac{\exp(-\lambda(y))}{\int \exp[-\lambda(\tilde{y})]\, d\tilde{y}}\right.\right.\right) \\
&\qquad - \ln\left(\int \exp[-\lambda(\tilde{x})]\, d\tilde{x}\right) - - \ln\left(\int \exp[-\lambda(\tilde{y})]\, d\tilde{y}\right).
\end{aligned}$$

The Lagrangian multipliers $\lambda(x), \lambda(y)$ have to be chosen so that the first requirement is met. The KL distance forming the first term in the last identity is minimized by the pdf in product form. The choice $\lambda(x) = -\ln(f(x))\ \lambda(y) = -\ln(f(y))$ meets the first constraint. The uniqueness of the minima implies that we have found the optimum. □

We would like to extend the mixtures on individual data entries to mixtures on the overall data space. The above proposition shows why it is not possible exactly for the target dimensions $\mathring{d}$ on the order of several tens. The number of multivariate components that have the specified marginal pdfs is extremely large and the majority of them are spurious. The spurious components are simply shadows of the true components. Their exclusion is called the *shadow cancelling problem.* The general solution of the above problem is computationally hard. A promising version is outlined in [173] and elaborated on in Section 12.3.2. Before it, a simpler version fitting the studied sandwich algorithm is proposed.

12.3.1 Simple choice of box width

The simple solution given here neglects information on the component centers and, for each individual data axis $d_{i;t}$, it specifies a single box width used further on.

The scalar normal mixture estimated for a given data record axis says that with probability α_c the noise variance is r_c. The posterior pdf of α is the Dirichlet pdf $Di_{\alpha}(\kappa_{\mathring{t}})$. It is determined by the statistic $\kappa_{\mathring{t}}$; see Proposition 10.1. The adopted approximate Bayesian estimation, Section 8.5 implies that the joint posterior pdf of $r_c,\ c \in c^*$, is the product of the inverse Wishart pdfs (8.23) $f(r_c|d(\mathring{t})) \equiv iW_{r_c}\left({}^{\lfloor d}D_{c;\mathring{t}}, \nu_{c;\mathring{t}}\right) \propto r_c^{-0.5(\nu_{c;\mathring{t}}+2)} \exp\left[-\frac{{}^{\lfloor d}D_{c;\mathring{t}}}{2r_c}\right]$. They are determined by the LS remainders ${}^{\lfloor d}D_{c;\mathring{t}}$ and the numbers of degrees of freedom $\nu_{c;\mathring{t}}$; see Proposition 8.7. We search for a single representative of these pdfs.

Proposition 12.5 (Representative of a group of iW_{r_c} pdfs) *The pdf $f(r|d(\mathring{t})) \equiv iW_r\left(\sum_{c\in c^*} \hat{\alpha}_{c;\mathring{t}}\, {}^{\lfloor d}D_{c;\mathring{t}}, \sum_{c\in c^*} \hat{\alpha}_{c;\mathring{t}}\nu_{c;\mathring{t}}\right)$ with $\hat{\alpha}_{c;\mathring{t}} = \frac{\kappa_{c;\mathring{t}}}{\sum_{c\in c^*}\kappa_{c;\mathring{t}}}$ minimizes the expected KL divergence*

$$\mathcal{E}\left[\sum_{c\in c^*} \alpha_c \int f(r|d(\mathring{t})) \ln\left(\frac{f(r|d(\mathring{t}))}{f(r=r_c|d(\mathring{t}))}\right) dr\right].$$

Proof. Proposition 2.7 implies that we have to minimize the conditional expectation $\mathcal{E}[\cdot|d(\mathring{t})]$ of the loss. It reads

$$\begin{aligned}
&\sum_{c\in c^*} \hat{\alpha}_{c;\mathring{t}} \int f(r|d(\mathring{t})) \ln\left(\frac{f(r|d(\mathring{t}))}{f(r_c = r|d(\mathring{t}))}\right) dr \\
&= \int f(r|d(\mathring{t})) \ln\left(\frac{f(r|d(\mathring{t}))}{\prod_{c=1}^{\mathring{c}}\left[f(r_c = r|d(\mathring{t}))\right]^{\hat{\alpha}_{c;\mathring{t}}}}\right) dr \\
&= \mathcal{D}\left(f(r|d(\mathring{t})) \middle|\middle| \frac{\prod_{c\in c^*}\left[f(r=r_c|d(\mathring{t}))\right]^{\hat{\alpha}_{c;\mathring{t}}}}{\int \prod_{c\in c^*}\left[f(r=r_c|d(\mathring{t}))\right]^{\hat{\alpha}_{c;\mathring{t}}} dr}\right) \\
&- \ln\left(\int \prod_{c\in c^*}\left[f(r=r_c|d(\mathring{t}))\right]^{\hat{\alpha}_{c;\mathring{t}}} dr\right).
\end{aligned}$$

Properties of the KL divergence imply that the minimizing pdf is the weighted geometric mean of pdfs corresponding to particular components. The weighted geometric mean of iW pdfs is iW pdf with the statistics being weighted arithmetic means of the combined statistics. □

The expected variance $\hat{r}$ corresponding to the found representative is

$$\hat{r} = \frac{\sum_{c\in c^*} \hat{\alpha}_{c;\mathring{t}}\, {}^{\lfloor d}D_{c;\mathring{t}}}{\sum_{c\in c^*} \hat{\alpha}_{c;\mathring{t}}\, {}^{\lfloor d}\nu_{c;\mathring{t}} - 2}.$$

This common estimate of the noise variance $\hat{r}$ is converted into the box width b by using probabilistic interpretation (12.1) of MT factors. For a negligible ε, they are simply uniform pdfs with half-width equal to b. The overall algorithm for determining the box width b we get by equating its variance $b^2/3$ with the adequate point estimate of r, which is the expected value of the pdf obtained in Proposition 12.5.

Algorithm 12.4 (Choice of box widths for respective data axes)

Initial phase

- *Initiate univariate normal mixture estimation using Algorithm 12.3.*
- *Estimate approximately the parameters of the normal mixture.*

Evaluation phase

Define box width

$$b = \sqrt{3 \frac{\sum_{c \in c^*} \hat{\alpha}_{c;\mathring{t}} \, {}^{\lfloor d}D_{c;\mathring{t}}}{\sum_{c \in c^*} \hat{\alpha}_{c;\mathring{t}} \, {}^{\lfloor d}\nu_{c;\mathring{t}} - 2}}. \tag{12.6}$$

12.3.2 Centers and box widths via shadow cancelling

The solution presented above neglects the information on positions of one-dimensional projections of cluster centers. Moreover, it works with an average variance estimate. The latter item was found to be critical when both very narrow and wide clusters are needed. This situation is often met, for instance, in the application on rolling mill; see Section 14.1. This makes us consider the shadow cancelling problem (see Chapter 6) and to apply the solution proposed in the paper [151] while relying on an algorithm described in [173].

The solution is built stepwise.

1. The prior and posterior pdfs on a multivariate component are constructed by using a fixed selection of marginal components as a building material. The selection is coded by an *index vector*. The corresponding component weight is also assigned to prior and posterior pdfs. Propositions 12.4 and 6.5 motivate the combination selected.
2. The v-likelihood corresponding to the inspected index vector is predicted using ideas developed in connection with the merging of components; see Section 6.6.4.
3. The optimization algorithm, proposed in [173], and selecting a given number of index vectors leading to the highest v-likelihood, is applied. The algorithm enables us to find a relatively large number of potential multivariate components that, after appropriate flattening, serve as initial positions and box widths for the MT step or can be directly used as initial estimates of the normal mixture.

Step 1 We assume that the modelled system is described by the normal static mixture written in the matrix form

$$f(d_t|\Theta) = \sum_{c \in c^*} \alpha_c \mathcal{N}_{d_t}(\theta_c, R_c)$$

with the $\mathring{d}$-vector expected values θ_c and diagonal covariance matrices R_c. The marginal pdf on the ith axis is the mixture

$$f(d_{i;t}|\Theta) = \sum_{c \in c^*} \alpha_c \mathcal{N}_{d_{i;t}}(\theta_{ic}, r_{ic}),$$

where the probabilistic weights α_c coincide with that of the multivariate mixture, θ_{ic} is the ith entry of θ_c and r_{ic} the ith diagonal entry of R_c. The posterior pdfs obtained through the estimation of "marginal" mixtures on respective axes are

$$f(\theta_{i1\cdots\mathring{c}_i}, r_{i1\cdots\mathring{c}_i}, \alpha|d_i(\tau)) = \prod_{c_i=1}^{\mathring{c}_i} GiW_{\theta_{ic_i},r_{ic_i}}(V_{ic_i;\tau}, \nu_{ic_i;\tau}) Di_{\alpha}(\kappa_{i1\cdots\mathring{c};\tau}),$$

where $\tau = \mathring{t}$. The result is valid if the conjugate prior of the same form but with $\tau = 0$ is used. For notation and derivation; see Chapter 8. The constructed approximate multivariate components are indexed by $m \in m^* \equiv \{1,\ldots,\mathring{m}\}$. The mth approximate component is identified by the *index vector* ${}^{\lfloor m}c \equiv \left[{}^{\lfloor m}c_1,\ldots,\ {}^{\lfloor m}c_{\mathring{d}}\right]'$. It is chosen so that it points to a single marginal component ${}^{\lfloor m}c_i$ on each axis $i = 1,\ldots,\mathring{d}$. It selects ${}^{\lfloor m}c_i$th univariate components

$$f(\theta_{i\,{}^{\lfloor m}c_i}, r_{i\,{}^{\lfloor m}c_i}|d_i(\tau)) = GiW_{\theta_{i\,{}^{\lfloor m}c_i}, r_{i\,{}^{\lfloor m}c_i}}(V_{i\,{}^{\lfloor m}c_i;\tau}, \nu_{i\,{}^{\lfloor m}c_i;\tau}). \tag{12.7}$$

We interpret the pdfs (12.7) as marginal pdfs of the constructed estimate of the multivariate parameters of the normal, static, parameterized component $\mathcal{N}_d\left(\theta_{{}^{\lfloor m}c}, R_{{}^{\lfloor m}c}\right)$. Using Proposition 12.4, the extension of these marginal pdfs to $\mathring{d}$-dimensional space, corresponding to the index vector ${}^{\lfloor m}c$, reads

$$f(\theta_{{}^{\lfloor m}c}, R_{{}^{\lfloor m}c}|d(\tau)) = \prod_{i=1}^{\mathring{d}} GiW_{\theta_{i\,{}^{\lfloor m}c_i}, r_{i\,{}^{\lfloor m}c_i}}(V_{i\,{}^{\lfloor m}c_i;\tau}, \nu_{i\,{}^{\lfloor m}c_i;\tau}),\ \ \tau = 0, \mathring{t}.$$

For the used univariate static mixtures, we deal with (2,2)-dimensional extended information matrices $V_{i\,{}^{\lfloor m}c_i;\tau}$. Their $L'DL$ decompositions are

$$L_{i\,{}^{\lfloor m}c_i;\tau} = \begin{bmatrix} 1 & 0 \\ \hat{\theta}_{i\,{}^{\lfloor m}c_i;\tau} & 1 \end{bmatrix},\ D_{i\,{}^{\lfloor m}c_i;\tau} = \mathrm{diag}\left[{}^{\lfloor d}D_{i\,{}^{\lfloor m}c_i;\tau},\ {}^{\lfloor \psi}D_{i\,{}^{\lfloor m}c_i;\tau}\right],$$

where $\hat{\theta}_{i\,{}^{\lfloor m}c_i;\tau}$ is the least-squares (LS) estimate of the offset, ${}^{\lfloor d}D_{i\,{}^{\lfloor m}c_i;\tau}$ is the LS remainder and ${}^{\lfloor \psi}D_{i\,{}^{\lfloor m}c_i;\tau}$ coincides with the LS-estimates covariance factor, respectively.

Furthermore, using Proposition 10.1, we get the marginal pdfs of weights of univariate mixtures, $i = 1,\ldots,\mathring{d}$, $\tau \in \{0, \mathring{t}\}$,

$$f_i \equiv f(\alpha_{{}^{\lfloor m}c_i}|d_i(\tau)) = Di_{\alpha_{{}^{\lfloor m}c_i}}\left(\kappa_{i\,{}^{\lfloor m}c_i;\tau}, \sum_{c=1}^{\mathring{c}} \kappa_{ic;\tau} - \kappa_{i\,{}^{\lfloor m}c_i;\tau}\right). \tag{12.8}$$

They provide the only available experience $\mathcal{P}$ for constructing the estimate $f \equiv f(\alpha_{{}^{\lfloor m}c}|\mathcal{P})$ of the objective pdf ${}^{\lfloor o}f \equiv {}^{\lfloor o}f(\alpha_{{}^{\lfloor m}c}|d(\tau))$. The estimate is chosen as the minimizer of the expected KL divergence. Proposition 2.7 implies that it minimizes

$$\mathcal{E}\left[\mathcal{D}\left(f\left|\right|{}^{\lfloor o}f\right)\right|\mathcal{P}\right].$$

The conditional expectation $\mathcal{E}[\cdot|\mathcal{P}]$ over the unknown pdf ${}^{\lfloor o}f$ is assumed to assign equal probabilities of the sets $\left\{{}^{\lfloor o}f \equiv {}^{\lfloor o}f(\alpha|d(\tau)) : {}^{\lfloor o}f \approx f_i\right\}$. Under this assumption, the optimal estimate f of ${}^{\lfloor o}f$ is the geometric mean of the pdfs forming $\mathcal{P}$; cf. Proposition 6.5. Thus, for a chosen ${}^{\lfloor m}c$, the Bayesian estimate of the weight $\alpha_{{}^{\lfloor m}c}$ assigned to the multivariate component in question, becomes

$$f(\alpha_{{}^{\lfloor m}c}|d(\tau)) \tag{12.9}$$
$$\equiv Di_{\alpha_{{}^{\lfloor m}c}}\left[\underbrace{\frac{1}{\mathring{d}}\sum_{i=1}^{\mathring{d}}\kappa_{i\,{}^{\lfloor m}c_i;\tau}}_{\kappa_{{}^{\lfloor m}c;\tau}},\underbrace{\frac{1}{\mathring{d}}\sum_{i=1}^{\mathring{d}}\left(\sum_{c\in\{1,\ldots,\mathring{c}\}}\kappa_{ic;\tau}-\kappa_{i\,{}^{\lfloor m}c_i;\tau}\right)}_{\rho_\tau-\kappa_{{}^{\lfloor m}c;\tau}}\right].$$

The found estimates $f(\alpha_{{}^{\lfloor m}c}|d(\tau))$, $m = 1,\ldots,\mathring{m}$, are marginal pdfs of the joint pdf $f(\alpha|d(\tau))$ describing the collection $\alpha = \left({}^{\lfloor 1}\alpha,\cdots,{}^{\lfloor \mathring{m}}\alpha\right)$ of the probabilistic weights of the multivariate mixture. Properties of the Dirichlet pdf (Proposition 10.1) imply that the pdf

$$f(\alpha|d(\mathring{t})) = Di_\alpha(\kappa_{1;\tau},\ldots,\kappa_{\mathring{m};\tau}) \equiv Di_\alpha(\kappa_\tau) \tag{12.10}$$

has the marginal pdfs (12.9). This motivates us to take the pdf (12.10), with the κ statistics having the entries $\kappa_{{}^{\lfloor m}c;\tau}$ defined in (12.9), as the estimate of mixture-components weights.

Step 2 For the chosen index vector ${}^{\lfloor m}c$, Step 1 provides the prior and posterior pdfs on parameters of the component ${}^{\lfloor m}c$ for $\tau = 0$ and $\tau = \mathring{t}$ respectively. The Bayes rule (2.8), applied to data belonging to the ${}^{\lfloor m}c$th component, implies that

$$f\left(\theta_{{}^{\lfloor m}c}, R_{{}^{\lfloor m}c}|d(\mathring{t}),\, {}^{\lfloor m}c\right) = \frac{\mathcal{L}\left(\theta_{{}^{\lfloor m}c}, R_{{}^{\lfloor m}c}, d(\mathring{t})\right) f\left(\theta_{{}^{\lfloor m}c}, R_{{}^{\lfloor m}c}|d(0),\, {}^{\lfloor m}c\right)}{f\left(d(\mathring{t})|\,{}^{\lfloor m}c\right)}.$$

This identity holds for arbitrary parameters and the likelihood function $\mathcal{L}\left(\theta_{{}^{\lfloor m}c}, R_{{}^{\lfloor m}c}, d(\mathring{t})\right)$ evaluated for $\theta_{{}^{\lfloor m}c} = 0$ and $R_{{}^{\lfloor m}c} I_{\mathring{d}}$ is independent of ${}^{\lfloor m}c$. It provides the following estimate of the v-likelihood related to the respective index vectors

$$f\left(d(\mathring{t})|\,{}^{\lfloor m}c\right)$$
$$= \frac{\mathcal{L}\left(\theta_{{}^{\lfloor m}c} = 0, R_{{}^{\lfloor m}c} = I_{\mathring{d}}, d(\mathring{t})\right) f\left(\theta_{{}^{\lfloor m}c} = 0, R_{{}^{\lfloor m}c} = I_{\mathring{d}}|d(0),\, {}^{\lfloor m}c\right)}{f\left(\theta_{{}^{\lfloor m}c} = 0, R_{{}^{\lfloor m}c} = I_{\mathring{d}}|d(\mathring{t}),\, {}^{\lfloor m}c\right)}.$$

This expression can be made more specific by omitting ${}^{\lfloor m}c$-independent factors; cf. formulas (8.22) and (10.3). Moreover, the identity, Proposition 8.7, is used

$$ {}^{\lfloor d}V_{i\,{}^{\lfloor m}c_i;\tau} = {}^{\lfloor d}D_{i\,{}^{\lfloor m}c_i;\tau} + \hat{\theta}^2_{i\,{}^{\lfloor m}c_i;\tau}\, {}^{\lfloor \psi}D_{\,{}^{\lfloor m}c_i;\tau}, \quad \tau \in \{0, \mathring{t}\}. $$

It results in the final expression

$$ f\left(d(\mathring{t}) \,\middle|\, {}^{\lfloor m}c\right) \propto \frac{\exp\left[-0.5\mathrm{tr}\left({}^{\lfloor d}V_{\lfloor m_{c;0}}\right)\right]}{\mathcal{I}(V_{\lfloor m_{c;0}}, \nu_{\lfloor m_{c;0}})} \frac{\mathcal{I}(V_{\lfloor m_{c;\mathring{t}}}, \nu_{\lfloor m_{c;\mathring{t}}})}{\exp\left[-0.5\mathrm{tr}\left({}^{\lfloor d}V_{\lfloor m_{c;\mathring{t}}}\right)\right]} \propto \exp $$

$$ \left[\frac{1}{2}\sum_{i=1}^{\mathring{d}}\left({}^{\lfloor d}D_{\,{}^{\lfloor m}c_i;\mathring{t}} - {}^{\lfloor d}D_{\,{}^{\lfloor m}c_i;0} + \hat{\theta}^2_{\,{}^{\lfloor m}c_i;\mathring{t}}\, {}^{\lfloor \psi}D_{\,{}^{\lfloor m}c_i;\mathring{t}} - \hat{\theta}^2_{\,{}^{\lfloor m}c_i;0}\, {}^{\lfloor \psi}D_{\,{}^{\lfloor m}c_i;0} \right)\right] $$

$$ \times \prod_{i=1}^{\mathring{d}} \frac{\mathit{\Gamma}(0.5\nu_{\,{}^{\lfloor m}c_i;\mathring{t}})\, {}^{\lfloor d}D_{\,{}^{\lfloor m}c_i;0}^{\,0.5\nu_{\,{}^{\lfloor m}c_i;0}}\, {}^{\lfloor \psi}D^{0.5}_{\,{}^{\lfloor m}c_i;0}}{\mathit{\Gamma}(0.5\nu_{\,{}^{\lfloor m}c_i;0})\, {}^{\lfloor d}D_{\,{}^{\lfloor m}c_i;\mathring{t}}^{\,0.5\nu_{\,{}^{\lfloor m}c_i;\mathring{t}}}\, {}^{\lfloor \psi}D^{0.5}_{\,{}^{\lfloor m}c_i;\mathring{t}}}. \tag{12.11} $$

Step 3 For any choice of the index vector ${}^{\lfloor m}c$, formula (12.11) provides the corresponding v-likelihood. Thus, we can search for the $\mathring{m}$ best extensions. The number of the index vectors is, however, extremely large so that an adequate search algorithm is needed. An algorithm solving a specific *index-based maximization* task is suitable. First, the task is specified and its solution is described. Then, it is connected with the addressed problem. Let us consider a set $(x^*, >)$ linearly ordered according to values of the function $l(x)$. We also consider its $\mathring{d}$-fold Cartesian product ordered according to the values of a scalar function $L\left(x_1(\mathring{c}_1), \ldots, x_{\mathring{d}}(\mathring{c}_{\mathring{d}})\right)$ defined on a priori given, nonempty finite sequences $x_i(\mathring{c}_i) \equiv (x_{i1}, x_{i2}, \ldots, x_{i\mathring{c}_i})\,,\ i = 1, \ldots, \mathring{d},\ x_{ic} \in x^*$. The lengths of respective sequences $\mathring{c}_i \in (0, \infty)$ may differ and the function $L(\cdot)$ is assumed to be isotonic, i.e.,

$$ l(x_{ic}) > l(\tilde{x}_{ic}) \ \Rightarrow L(x_{i1}, \ldots, x_{ic}, \ldots, x_{i\mathring{c}}) > L(x_{i1}, \ldots, \tilde{x}_{ic}, \ldots, x_{i\mathring{c}}). \tag{12.12} $$

We inspect values of

$$ L\left({}^{\lfloor m}c\right) \equiv L\left({}^{\lfloor m}x_{1\,{}^{\lfloor m}c_1}, {}^{\lfloor m}x_{2\,{}^{\lfloor m}c_2}, \ldots, {}^{\lfloor m}x_{\mathring{d}\,{}^{\lfloor m}c_{\mathring{d}}}\right) \tag{12.13} $$

while taking just one item ${}^{\lfloor m}x_{i\,{}^{\lfloor m}c_i}$ from each sequence $x_i(\mathring{c}_i)$, $i = 1, \ldots, \mathring{d}$.

We search for indexes forming *index vectors*

$$ {}^{\lfloor m}c \equiv \left[{}^{\lfloor m}c_1,\ {}^{\lfloor m}c_2,\ \ldots,\ {}^{\lfloor m}c_{\mathring{d}}\right] \tag{12.14} $$

giving $\mathring{m}$ largest values of $L(\cdot)$ with arguments of the form (12.13).

The set containing all potential index vectors (12.14) has the form

$$ {}^{\lfloor \bullet}c^* = \left\{\ {}^{\lfloor \bullet}c \equiv [c_1,\ c_2,\ \ldots,\ c_{\mathring{d}}] \mid 1 \le c_i \le \mathring{c}_i,\ \forall\ i = 1, 2, \ldots, \mathring{d}\ \right\}. $$

The addressed *index-based maximization task* is formalized as follows.

$$\boxed{\begin{array}{l} \text{Find different index vectors } {}^{\lfloor m}c \in {}^{\lfloor \bullet}c^*,\ m = 1,\ldots,\mathring{m} \le {}^{\lfloor \bullet}\mathring{c} \\[2ex] \min_{\ {}^{\lfloor \bullet}c \in \left\{ {}^{\lfloor 1}c,\, {}^{\lfloor 2}c, \ldots,\, {}^{\lfloor \mathring{m}}c \right\}} L\left({}^{\lfloor \bullet}c \right) \ge \max_{\ {}^{\lfloor \bullet}c \in\, {}^{\lfloor \bullet}c^* \setminus \left\{ {}^{\lfloor 1}c,\, {}^{\lfloor 2}c, \ldots,\, {}^{\lfloor \mathring{m}}c \right\}} L\left({}^{\lfloor \bullet}c \right). \end{array}}$$

Verbally, the worst index vector from the constructed set is better than the best index vector from its complement. Algorithm 12.5 given below solves the formulated task. The paper [173] proves it. The following symbols are used.

$\mathcal{C}$ denotes a sequence of index vectors. It contains the optimal index vectors when the algorithm stops.

${}^{\lfloor a}\mathcal{C}$ is an auxiliary sequence containing the candidates for the optimal index vectors.

$\mathcal{S}$ is the sequence of sons of the last index vector in $\mathcal{C}$. A son of the index vector ${}^{\lfloor \bullet}c$ has the form ${}^{\lfloor \bullet}c + e_i$, where e_i is $\mathring{d}$-vector having the only nonzero value equal 1 on ith position.

$\Sigma\mathcal{A}$ is the sequence of the values $L(\cdot)$ assigned to members of the sequence $\mathcal{A}$ of index vectors.

Algorithm 12.5 (Algorithm solving the index maximization)

Initial mode

- *Order the input sequences in the nonincreasing manner*

$$l(x_{i1}) \ge l(x_{i2}) \ge \cdots \ge l(x_{i\mathring{c}_i}),\ i = 1,\ldots,\mathring{d}.$$

- *Set* $\mathcal{C} = \mathbf{1}_{\mathring{d}} \equiv \underbrace{[1,1,\ldots,1]}_{\mathring{d}-times},\ {}^{\lfloor a}\mathcal{C} = \emptyset$ *and* $\sum {}^{\lfloor a}\mathcal{C} = \emptyset.$

Iterative mode

1. *Return* $\mathcal{C}$ *and stop if* $length(\mathcal{C}) = \mathring{m}$.
2. *Set* $\mathcal{S} = \emptyset$ *and* $\Sigma\mathcal{S} = \emptyset$.
3. *For* $i = 1,2,\ldots,\mathring{d}$ *If* $\mathcal{C}_{ilength(\mathcal{C})} < \mathring{c}_i$ *put* $\mathcal{C}_{length(\mathcal{C})} + e_i$ *as the last member of* $\mathcal{S}$ *and* $L(\mathcal{C}_{length(\mathcal{C})} + e_i)$ *as the last member of* $\Sigma\mathcal{S}$.
4. *Sort the sequence* $\Sigma\mathcal{S}$ *in a nondecreasing order and sort the sequence* $\mathcal{S}$ *using the same rule.*
5. *Merge the sequences* $\Sigma\, {}^{\lfloor a}\mathcal{C}$ *and* $\Sigma\mathcal{S}$ *while placing the equal items from* $\Sigma\, {}^{\lfloor a}\mathcal{C}$ *first and merge the sequences* ${}^{\lfloor a}\mathcal{C}$ *and* $\mathcal{S}$ *using the same rule.*
6. *Erase duplicates from the sequence* ${}^{\lfloor a}\mathcal{C}$ *as well as the corresponding members from* $\Sigma\, {}^{\lfloor a}\mathcal{C}$.
7. *Preserve the initial* $\min\left\{\mathring{m} - length(\mathcal{C}),\ length({}^{\lfloor a}\mathcal{C})\right\}$ *members of both* $\Sigma\, {}^{\lfloor a}\mathcal{C}$ *and* ${}^{\lfloor a}\mathcal{C}$*, erase all others.*
8. *Delete the first member from* $\Sigma\, {}^{\lfloor a}\mathcal{C}$ *while deleting the first member from* ${}^{\lfloor a}\mathcal{C}$ *and placing it as the last member of* $\mathcal{C}$.
9. *Go to the beginning of* Iterative mode.

Now we apply this algorithm in order to find index vectors ${}^{\lfloor \bullet}c$ that combine properly selected components on individual axes so that the each constructed multivariate component has a high v-likelihood.

The linearly ordered set x^* consists of statistics

$$x_{ic} \equiv \left(\nu_{ic;\mathring{t}}, \nu_{ic;0}, \hat{\theta}_{ic;\mathring{t}}, \hat{\theta}_{ic;0}, {}^{\lfloor d}D_{ic;\mathring{t}}, {}^{\lfloor d}D_{ic;0}, {}^{\lfloor \psi}D_{ic;\mathring{t}}, {}^{\lfloor \psi}D_{ic;0}\right).$$

The individual statistics are ordered according to the value of the v-log likelihood $x_{ic} > x_{\tilde{i}\tilde{c}} \Leftrightarrow l(x_{ic}) > l(x_{\tilde{i}\tilde{c}})$, where

$$\begin{aligned} l(x_{ic}) \equiv 0.5 &\left({}^{\lfloor d}D_{ic;\mathring{t}} - {}^{\lfloor d}D_{ic;0} + \hat{\theta}^2_{ic;\mathring{t}} \, {}^{\lfloor \psi}D_{ic;\mathring{t}} - \hat{\theta}^2_{ic;t} \, {}^{\lfloor \psi}D_{ic;0} \right) \\ &+ \ln \left(\frac{\Gamma(0.5\nu_{ic;\mathring{t}}) \, {}^{\lfloor d}D_{ic;0}^{0.5\nu_{ic;0}} \, {}^{\lfloor \psi}D_{ic;0}^{0.5}}{\Gamma(0.5\nu_{ic;0}) \, {}^{\lfloor d}D_{ic;\mathring{t}}^{0.5\nu_{ic;\mathring{t}}} \, {}^{\lfloor \psi}D_{ic;\mathring{t}}^{0.5}} \right). \end{aligned}$$

The sum of the marginal v-log likelihoods defines the function $L(\cdot)$ assigned to each index vector. Summing meets obviously the monotonicity property (12.12). It confirms the possibility to use Algorithm 12.5 to our problem.

This option finishes extension of the marginal posterior pdfs to the overall description of the multivariate mixture as the choice of index vectors determines also the estimate of component weights; see the discussion of the first step.

In order to avoid falsely low uncertainty, it is necessary to flatten the constructed estimate; see Section 6.4.3. It is wise to select separate flattening rates $\lambda_D \in (0,1), \lambda \in (0,1)$ so that the resulting sums of counters κ and of degrees of freedom ν are equal to the sums of prior values, respectively.

Let us describe the overall algorithm.

Algorithm 12.6 (Index-vector-based initiation of MT algorithm)

Initial mode

- *Remove outliers and normalize the data so that the majority of them has entries in the range* $[-1, 1]$.
- *Specify the prior and posterior pdfs on univariate components*

```
For  $i = 1, \ldots, \mathring{d}$
        Specify the number of components $\mathring{c}_i$ and shift in positions $s_i = \frac{2}{\mathring{c}_i}$
    For  $c = 1, \ldots, \mathring{c}_i$
            Select statistics $\kappa_{ic;0} \approx 0.1\mathring{t}/\mathring{c}$, $\nu_{ic;0} \approx 1$, $\hat{\theta}_{ic;0} = -1 + (c - 0.5)s_i$
            ${}^{\lfloor d}D_{ic;0} = \frac{(\nu_{ic;0} - 2)s_i^2}{16}$, ${}^{\lfloor \psi}D_{ic;0} = 1$
    end  of the cycle over $c$
end  of the cycle over $i$
```

```
For $i = 1, \ldots, \mathring{d}$
        Estimate univariate mixture using data $d_i(\mathring{t})$.
  For $c = 1, \ldots, \mathring{c}_i$
        Form the statistics describing both prior and posterior pdf
        $x_{ic} \equiv \left(\nu_{ic;\mathring{t}}, \nu_{ic;0}, \hat{\theta}_{ic;\mathring{t}}, \hat{\theta}_{ic;0}, {}^{\lfloor d}D_{ic;\mathring{t}}, {}^{\lfloor d}D_{ic;0}, {}^{\lfloor \psi}D_{ic;\mathring{t}}, {}^{\lfloor \psi}D_{ic;0}\right)$
        Evaluate the individual v-log-likelihood
        $l_{ic} = 0.5\left({}^{\lfloor d}D_{ic;\mathring{t}} - {}^{\lfloor d}D_{ic;0} + \hat{\theta}^2_{ic;\mathring{t}}\,{}^{\lfloor \psi}D_{ic;\mathring{t}} - \hat{\theta}^2_{ic;0}\,{}^{\lfloor \psi}D_{ic;0}\right)$
        $+ \ln\left(\frac{\Gamma(0.5\nu_{ic;\mathring{t}})\,{}^{\lfloor d}D_{ic;0}^{0.5\nu_{ic;0}}\,{}^{\lfloor \psi}D_{ic;0}^{0.5}}{\Gamma(0.5\nu_{ic;0})\,{}^{\lfloor d}D_{ic;\mathring{t}}^{0.5\nu_{ic;\mathring{t}}}\,{}^{\lfloor \psi}D_{ic;\mathring{t}}^{0.5}}\right)$
  end of the cycle over c
        Order $x_i = [x_{i1}, \ldots, x_{i\mathring{c}_i}]$ so that $l_{i1} \geq l_{i2} \geq \cdots \geq l_{i\mathring{c}_i}$
end of the cycle over i
```

- *Select the number $\mathring{m}$ of constructed multivariate components.*
- *Apply Algorithm 12.5 on the constructed sequences $x_i = [x_{i1}, \ldots, x_{i\mathring{c}_i}]$ to get index sequences ${}^{\lfloor m}c = \left[{}^{\lfloor m}c_1, \ldots, {}^{\lfloor m}c_{\mathring{d}}\right]$, $m = 1, \ldots, \mathring{m}$.*

Evaluation mode

```
For $m = 1, \ldots, \mathring{m}$
  Specify the posterior pdf on parameters of multivariate components
  $f\left(\theta_{{}^{\lfloor m}c}, R_{{}^{\lfloor m}c} \middle| d(\mathring{t}), {}^{\lfloor m}c\right) = \prod_{i=1}^{\mathring{d}} GiW_{\theta_{i\,{}^{\lfloor m}c_i}, r_{i\,{}^{\lfloor m}c_i}}(V_{i\,{}^{\lfloor m}c_i;\mathring{t}}, \nu_{i\,{}^{\lfloor m}c_i;\mathring{t}})$
  $\equiv \prod_{i=1}^{\mathring{d}} GiW_{\theta_{i\,{}^{\lfloor m}c_i}, r_{i\,{}^{\lfloor m}c_i}}\left(\hat{\theta}_{i\,{}^{\lfloor m}c_i;\mathring{t}}, {}^{\lfloor d}D_{i\,{}^{\lfloor m}c_i;\mathring{t}}, {}^{\lfloor \psi}D_{i\,{}^{\lfloor m}c_i;\mathring{t}}\nu_{i\,{}^{\lfloor m}c_i;\mathring{t}}\right).$
  Specify the statistics of the pdf on component weights
      $\kappa_{{}^{\lfloor m}c;\tau} = \frac{1}{\mathring{d}} \sum_{i=1}^{\mathring{d}} \kappa_{i\,{}^{\lfloor m}c_i;\tau}, \ \tau \in \{0, \mathring{t}\}.$
  One-shot flattening, Section 6.4.3, gives the constructed prior pdf $f(\Theta|\,{}^{\lfloor m}c)$.
  Use the centers $\hat{\theta}_{{}^{\lfloor m}c}$ as initial positions for the MT algorithm and recompute
  the corresponding marginal variances of data to the box widths
      $b_{{}^{\lfloor m}c_i} = \sqrt{3\frac{{}^{\lfloor d}D_{i\,{}^{\lfloor m}c_i}}{\nu_{i\,{}^{\lfloor m}c_i} - 2}\left[1 + {}^{\lfloor \psi}D^{-1}_{i\,{}^{\lfloor m}c_i}\right]}$,   cf. (8.31).
end of the cycle over m
```

12.4 MT step: make the MT algorithm feasible

The original MT algorithm proved to be unsuitable for the target extent of data sets $\mathring{t} \approx 10^5$. This stimulated a search for improvements. Algorithm 12.6 provides a rather promising solution. Here, alternative and complementary improvements are discussed.

12.4.1 Initialization

The original MT algorithm selects gradually each data record $d_t,\ t \in t^*$ as the initial center of the box for searching of stationary positions. This safest way is, however, too expensive for data files containing more than several tens of thousand records. Then, an incomplete search is necessary.

Random starts provide an alternative to an exhaustive search. It can be speeded up significantly by exploiting the deterministic nature of the MT algorithm. Essentially, the visited data records that would lead to the same trajectory are excluded from a further search. We want to preserve simplicity of the algorithm. For this reason, we want to recognize such points treated in connection with the box in question.

Considering a fixed axis, we can analyze univariate case. Let $\hat{d}$ be a center of the inspected box. Let us define

$$\begin{aligned}
\overline{D} &\equiv \min\{d_t : d_t > \hat{d} + b\}_{t\in t^*} && \text{the smallest data record right to the box}\\
\underline{D} &\equiv \max\{d_t : d_t < \hat{d} - b\}_{t\in t^*} && \text{the largest data record left to the box}\\
\overline{d} &\equiv \max\{d_t : d_t \le \hat{d} + b\}_{t\in t^*} && \text{the largest data record within the box}\\
\underline{d} &\equiv \min\{d_t : d_t \ge \hat{d} - b\}_{t\in t^*} && \text{the smallest data record within the box.}
\end{aligned} \tag{12.15}$$

These bounds specify ranges of data safely outside $(\underline{D}, \overline{D})$ and safely inside $(\underline{d}, \overline{d})$ the inspected box trajectory.

We arrive at the same intermediate and final box center if we replace $\hat{d}$ by any $\tilde{d}_t$ (in this box) such that the box centered in this data record contains the same data records. Formally:

$$\begin{aligned}
\tilde{d}_t &\le \underline{d} + b,\ \tilde{d}_t < \overline{D} - b\\
\tilde{d}_t &\ge \overline{d} - b,\ \tilde{d}_t > \underline{D} + b.
\end{aligned}$$

Thus, all such data points can be excluded from the set of potential starting points. Evaluation of $\underline{D}, \overline{D}, \underline{d}, \overline{d}$ can be done simultaneously when searching for the points belonging to the inspected box. The comparisons needed for determination of points $\tilde{d}$ can again be made cheaply on the data belonging to the box. It is sufficient to store them temporarily.

12.4.2 Merging and cancelling of centers

Merging and cancelling of boxes are now performed in the MT step using ad hoc rules for their closeness or (almost) emptiness. Within the "sandwich" context, it is possible to postpone this action to the beginning of the -B step; see Section 12.5. It is sufficient to apply cancelling and merging of Gaussian components (see Section 8.6.4) that arise from the MT factors in the -B step.

If need be, we can use finiteness of the KL divergence of a pair of MT factors (see Proposition 12.2) in order to decrease the number of components before entering the -B step; Section 12.5. For this, let us consider the set f^* of the MT components having the common box width and individual pdfs $f, \tilde{f}$ defined by centers $\mu, \tilde{\mu}$. Then,

$$\rho(\mu, \tilde{\mu}) \equiv \frac{\mathcal{D}\left(f \middle|\middle| \tilde{f}\right)}{\sup_{f \in f^*, \tilde{f} \in f^*} \mathcal{D}(f||\tilde{f})} = \sum_{i=1}^{\mathring{d}} \frac{|\mu_i - \tilde{\mu}_i|}{2}. \tag{12.16}$$

Thus, we can take $\mu, \tilde{\mu}$ as identical if $\rho(\mu, \tilde{\mu})$ is a small number, say of the order 10^{-3}.

12.4.3 Recognition of a false local maximum

Updating a box center $\hat{\mu}$ ends when it does not shift any more. It may, however, happen that the window sits between data clusters. We counteract this situation by formulating and testing the following hypothesis.

H_0 : the data $d_k,\ k \in k^* \equiv \{1, \ldots, \mathring{k}\}$ within the box $[\hat{\mu} - b, \hat{\mu} + b]$ have the uniform pdf $\mathcal{U}_d\left(\hat{\mu} - b, \hat{\mu} + b\right)$,

H_1 : the data $d_k,\ k \in \{1, \ldots, \mathring{k}_1\}$ fulfilling the inequality $d_k \leq \hat{\mu}$ have the uniform pdf $\mathcal{U}_d\left(\hat{\mu} - b, \hat{\mu} - b_1\right)$ and the data $d_k,\ k \in \{\mathring{k}_1{+}1, \ldots, \mathring{k}\}$ fulfilling the inequality $d_k > \hat{\mu}$ have the uniform pdf $\mathcal{U}_d\left(\hat{\mu} + b_2, \hat{\mu} + b\right)$.

The corresponding models relating observations to respective hypotheses are given by the following formulas.

$$\begin{aligned} f(d_k|H_0) &= \mathcal{U}_{d_k}\left(\hat{\mu} - b, \hat{\mu} + b\right) = \frac{1}{2b} \\ \underbrace{\Rightarrow}_{\text{static case}} \quad f(d(\mathring{k})|H_0) &= (2b)^{-\mathring{k}} \end{aligned}$$

$$\begin{aligned} f(d_k|b_1, b_2, H_1) &= \frac{\chi_{[\hat{\mu}-b,\hat{\mu}-b_1]}(d_k)}{2(b-b_1)} + \frac{\chi_{(\hat{\mu}+b_2,\hat{\mu}+b]}(d_k)}{2(b-b_2)} \\ \underbrace{\Rightarrow}_{\text{static case}} \quad f(d(\mathring{k})|b_1, b_2, H_1) &= \frac{\chi_{[\hat{\mu}-b,\hat{\mu}-b_1]}(\bar{d}_1) \times \chi_{(\hat{\mu}+b_2,\hat{\mu}+b]}(\underline{d}_2)}{2^{\mathring{k}}(b-b_1)^{\mathring{k}_1}(b-b_2)^{\mathring{k}-\mathring{k}_1}}, \end{aligned}$$

where $\mathring{k}$ is the number of data points within the inspected box. $\mathring{k}_1$ denotes the number of data points among them fulfilling the inequality $d_k \leq \hat{\mu}$ and

$\bar{d}_1$ their maximum. The minimum of the data d_k being within the box and fulfilling $d_k > \hat{\mu}$ is denoted $\underline{d}_2$.

Taking uniform prior pdfs $f(b_i) = \mathcal{U}_{b_i}(0, b),\ i = 1, 2$, we get

$$f(d(\mathring{k})|H_1) = \frac{1}{2^{\mathring{k}} b^2} \int_0^{\hat{\mu}-\overline{d}_1} (b - b_1)^{-\mathring{k}_1}\, db_1 \int_0^{\underline{d}_2-\hat{\mu}} (b - b_2)^{-\mathring{k}+\mathring{k}_1}\, db_2$$

$$= \frac{1}{2^{\mathring{k}} b^2} \frac{(b - \hat{\mu} + \overline{d}_1)^{-\mathring{k}_1+1} - b^{-\mathring{k}_1+1}}{(\mathring{k}_1 - 1)} \frac{(b + \hat{\mu} - \underline{d}_2)^{-\mathring{k}+\mathring{k}_1+1} - b^{-\mathring{k}+\mathring{k}_1+1}}{\mathring{k} - \mathring{k}_1 - 1}.$$

Using the Bayes rule without prejudice, we get

$$f(H_0|d(\mathring{k})) = \frac{f(d(\mathring{k})|H_0)}{f(d(\mathring{k})|H_0) + f(d(\mathring{k})|H_1)}. \tag{12.17}$$

When dealing with the multivariate data records, the entry-related quantities are conditionally independent and, for $\iota \in \{0, 1\}$, $f(d(\mathring{k})|H_\iota) = \prod_{i=1}^{\mathring{d}} f(d_i(\mathring{k})|H_{i\iota})$, where $H_{i\iota}$ refers to the hypothesis ι on the ith axis.

With these probabilities, it is straightforward to cancel those centers for which the posterior probability $f(H_0|d(\mathring{k}))$ is too small.

12.4.4 Improved MT algorithm

The above ideas lead to the following improved version of the MT algorithm.

Algorithm 12.7 (Improved MT algorithm)

Initial mode

- *Remove outliers from data sample $d(\mathring{t})$; see Section 6.2.2.*
- *Normalize resulting data to the range $[-1, 1]^{\mathring{d}}$.*
- *Attach to each data record d_t of the resulting data the flag $g_t = 1 \equiv$ ready_to_serve_as_starting_point.*
- *Select widths of box b_i, $i = 1, \ldots, \mathring{d}$, for individual data channels $d_{i;t}$, using the B-step; see Section 12.3.*
- *Select the upper bound $\mathring{n}$ on the number n of iterations.*
- *Select the significance level $\varepsilon \in \left[10^{-3}, 10^{-5}\right]$, used for cancelling of local minima and "merging" level, i.e., define small values of (12.16).*
- *Initialize the list $\mathcal{C}$ of found centers, $\mathcal{C} =$empty list, and the numbers $\kappa =$ of points in them, $\kappa =$empty list.*

Iterative mode

```
For   $n = 1, \ldots, \mathring{n}$
            Select a starting center $\hat{d} =$ a data record with flag $g_t = 1$.
            Set $\underline{d} = \underline{D} = +\infty \times \mathbf{1}_{\mathring{d}}$, $\overline{d} = \overline{D} = -\infty \times \mathbf{1}_{\mathring{d}}$.
            Set $e = +\infty$.
```

```
REPEAT while the data center move, while $e > \varepsilon$.
          Initialize list $\Upsilon$ of data in the box, $\Upsilon =$ empty list and
          the number $\mathring{k}$ of points within the inspected box, $\mathring{k} = 0$.
  For   $t = 1, \ldots, \mathring{t}$
        Set the $flag\_in = 1$(point is in box) and find all data in
        the current box and evaluate the bounds $\overline{D}_i, \overline{d}_i, \underline{D}_i, \underline{d}_i$; see (12.15),
    For   $i = 1, \ldots, \mathring{d}$
          IF $d_{i;t} > \hat{d}_i + b_i$
            $\overline{D}_i = \min\left(d_{i;t}, \overline{D}_i\right), \quad flag\_in = 0.$
          ELSE
            If $d_{i;t} < \hat{d}_i - b_i$
              $\underline{D}_i = \max\left(d_{i;t}, \underline{D}_i\right), \quad flag\_in = 0$
          end of the ELSE
    end   of the cycle over $i$
          If $flag\_in = 1$
              For $i = 1, \ldots, \mathring{d}$
                $\underline{d}_i = \min\left(d_{i;t}, \underline{d}_i\right), \quad \overline{d}_i = \max\left(d_{i;t}, \overline{d}_i\right)$
              end of the cycle over $i$.
             Set $\Upsilon = [\Upsilon, (d_t, t)], \quad \mathring{k} = \mathring{k} + 1,$
          end of the If
  end   of the cycle over $t$
 Mark the data suitable for initialization and compute a new box center.
          Set $s = 0$
  For   $k = 1, \ldots, \mathring{k}$
       Set the flag $g_t = 0$ for $(d_t, t)$ on the $k$th position of $\Upsilon$. Set $s = s + d_t$.
    For   $i = 1, \ldots, \mathring{d}$
          If $d_{i;t} > \underline{d}_i + b$ or $d_{i;t} \geq \overline{D}_i - b$ or $d_{i;t} \leq \underline{D}_i + b$ or $d_{i;t} < \overline{d}_i - b$
            Set the flag $g_t = 1$.
          end of the If
    end   of the cycle over $i$
  end   of the cycle over $k$
          Set $e = \rho\left(\hat{d}, \frac{s}{\mathring{k}}\right)$, see (12.16), and set $\hat{d} = \frac{s}{\mathring{k}}$.
end of the cycle REPEAT
Extend the list of centers and numbers of points in them
$\mathcal{C} = [\mathcal{C}, \hat{d}], \ \kappa = [\kappa, \mathring{k}]$.
```

```
Cancel the local minima as follows.
  For  $i = 1, \dots, \overset{\circ}{d}$
            Set $\overset{\circ}{k}_{i1} = 0,\ \overline{d}_{i1} = -\infty,\ \underline{d}_{i2} = +\infty$
  end  of the cycle over i
  For  $k = 1, \dots, \overset{\circ}{k}$
            For $(d_t, t)$ on the kth position of $\Upsilon$
    For  $i = 1, \dots, \overset{\circ}{d}$
            If $d_{i;t} < \hat{d}_i$
              Set $\overset{\circ}{k}_{i1} = \overset{\circ}{k}_{i1} + 1$ and $\overline{d}_{i1} = \max(\overline{d}_{i1}, d_{i;t})$
            Else
              Set $\underline{d}_{i2} = \min(\underline{d}_{i2}, d_{i;t})$.
            end Else
    end  of the cycle over i
  end  of the cycle over k
   Evaluate the probability (12.17), in the multivariate version,
   using $\overset{\circ}{k}$ and vectors $\overset{\circ}{k}_1$, $\overline{d}_1$, $\underline{d}_2$ found.
   Cancel the center found from the list $\mathcal{C}$ and
   the corresponding counter from $\kappa$ if this probability is small.
end  of the cycle over n
Merge centers from the list $\mathcal{C}$ with mutual distance (12.16) smaller than $\varepsilon$.
```

Remark(s) 12.2

1. *The conditions for the identity of the final boxes can be relaxed by setting for the flag $g_t = 0$ even if the conditions for $g_t = 0$ are violated slightly only.*
2. *The choice of initial points is not fixed here. Either a random choice from the data list with flag $g_t = 1$ can be made, or Algorithm 12.5 is employed.*
3. *The information on repetitive visiting of the same local extreme can be exploited for predicting the overall number of maxima and thus used for an efficient and reliable stopping [174].*
4. *The use of a shadow-cancelling-based choice of initial boxes avoids the need for labelling of suitable starting points.*
5. *The algorithm can be substantially accelerated by temporary sorting of each data channel and storing the sorted indexes. Then, much fewer data records have to be inspected when searching for the points belonging to the given box and evaluating $\underline{D}, \overline{D}$.*

12.5 -B step: MT results as initial mixture

Important but partial information obtained from the MT step has to be extended into the statistics determining the prior GiW pdf of the final normal mixture. The extension is discussed here.

12.5.1 Position and noise covariance

The center of a multivariate box on the data-vector space $\varPsi^*$ is a "natural" prior LS point estimate $\hat{\theta}_0$ of the offset in the static model modelling $\varPsi_t$. Similarly, the diagonal LS estimate of the matrix noise covariance is obtained by equating respective scalar variances to the square of the corresponding box width divided by 3; cf. Algorithm 12.4.

12.5.2 Remaining statistics

For the static model, fixed component c and factor i, the covariance factor C_{ic} of the LS estimate is scalar. It is just a shifted version of the corresponding $\nu_{ic} = \nu_c$, $\forall\ i$. The scalar ν_c equals the corresponding κ_c, again potentially up to some shift. Having no reasons for selecting different shifts, we set $C_{ic;0} = \nu_{ic;0} = \kappa_{c;0}$. The reasonable value $\kappa_{c;0}$ is proportional to the number $\mathring{k}_{c;\mathring{t}}$ of data vectors captured within the inspected box. The proportion factor is chosen so that the general recommendation for initial values of $\sum_{c \in c^*} \kappa_{c;0} \approx \gamma \mathring{t},\ \gamma \approx 0.1$, is met. It gives

$$C_{ic;0} = \nu_{ic;0} = \kappa_{c;0} = \mathring{k}_{c;\mathring{t}} \frac{\gamma \mathring{t}}{\sum_{\tilde{c} \in \mathring{c}} \mathring{k}_{\tilde{c};\mathring{t}}}. \tag{12.18}$$

This complements the definition of the prior GiW pdf for the normal mixture.

12.5.3 Conversion of static components to dynamic factors

The MT algorithm provides a static model that is extended to the dynamic case by applying it to the data vector $\varPsi_t$, which has unity as its last entry. The extension is justified by Proposition 6.17. In this way, we get statistics $\nu_{c;0}, V_{c;0}$ of the static matrix components $c \in c^*$. We need to convert them into statistics of related dynamic factors. This is done here using a special structure of the extended information matrix $V_c \equiv V_{c;0}$.

The matrix V_c contains some nontrivial entries resulting from the MT algorithm. The remaining ones, expressing unobserved correlations, are set to zero using the standard argument of insufficient reasons. The wish to preserve sample moments, while respecting relationships between uniform to normal pdfs, leads to the following form of the extended information matrix for the ith factor

$$V_{ic} = k_c \begin{bmatrix} \frac{b_{ic}^2}{3} + \hat{\mu}_{ic}^2 & & & \hat{\mu}_{ic} \\ & \frac{b_{(i+1)c}^2}{3} + \hat{\mu}_{(i+1)c}^2 & & \hat{\mu}_{(i+1)c} \\ & & \ddots & \vdots \\ & & & \hat{\mu}_{\mathring{\Psi}c} \\ \hat{\mu}_{ic} & \hat{\mu}_{(i+1)c} & \ldots \hat{\mu}_{\mathring{\Psi}c} & \mathring{\Psi} \end{bmatrix}, \qquad (12.19)$$

where k_c is the number of data in the cth final box found by MT algorithm. The symbol b_{ic} is used for the box width on the ith axis and $\hat{\mu}_{ic}$ is the corresponding entry of the cth box center.

We have to verify that V_{ic} is well defined, to verify its positive definiteness.

The submatrices A_{ic} arising by omission of the last row and column are diagonal with positive diagonal entries, and thus they are positive definite. Thus, it remains to show that $|V_{ic}| > 0$. This property is implied by the identity

$$\begin{aligned} |V_{ic}| &= |A_{ic}|(\mathring{\Psi} - [\hat{\mu}_{ic}, \hat{\mu}_{(i+1)c}, \ldots, \hat{\mu}_{\mathring{\Psi}c}]A_{ic}^{-1}[\hat{\mu}_{ic}, \hat{\mu}_{(i+1)c}, \ldots, \hat{\mu}_{\mathring{\Psi}c}]' \\ &= |A_{ic}| \left(\mathring{\Psi} - \sum_{j=i}^{\mathring{\Psi}} \frac{\hat{\mu}_{jc}^2}{b_{jc}^2 + \hat{\mu}_{jc}^2} \right) > |A_{ic}| \left(\mathring{\Psi} - (\mathring{\Psi} - i) \right) > 0. \end{aligned}$$

We need LDL' decompositions of these information matrices. The simple form of the matrices (12.19) allows us to generate them relatively cheaply by a special version of the algorithm presented in Proposition 8.2. The following proposition describes this version. In it, the fixed component subscript is suppressed.

Proposition 12.6 (**$L'DL$ form of MT-based V_i, $i = \mathring{\Psi}, \ldots, 1$**) *For $i = \mathring{\Psi}$, set $L_{\mathring{\Psi}} = 1$, $D_{\mathring{\Psi}} = \mathring{\Psi}$; see (12.19).*

Let us have the $L'DL$ decomposition of $V_{i+1} = L_{i+1}D_{i+1}L_{i+1}'$ and write the definition (12.19) recursively

$$V_i = \begin{pmatrix} g_i & [0, \ldots, 0, \hat{\mu}_i] \\ [0, \ldots, 0, \hat{\mu}_i]' & L_{i+1}'D_{i+1}L_{i+1} \end{pmatrix}.$$

Then, the decomposition $V_i = L_i'D_iL_i$ is obtained as follows

$$L_i = \begin{pmatrix} 1 & 0 \\ \beta_i & L_{i+1} \end{pmatrix} \quad D_i = \begin{pmatrix} a_i & 0 \\ 0 & D_{i+1} \end{pmatrix}, \quad \textit{where}$$
$$\beta_i = (L_{i+1}'D_{i+1})^{-1} \underbrace{[0, \ldots, 0, \hat{\mu}_i]'}_{b_i'}, \quad a_i = g_i - \beta_i'D_{i+1}\beta_i.$$

Proof. Multiplication of the matrices forming the $L'DL$ decompositions imply

$$V_i = \begin{bmatrix} a_i + \beta_i'D_{i+1}\beta_i & \beta_i'D_{i+1}L_{i+1} \\ L_{i+1}'D_{i+1}\beta_i & L_{i+1}'D_{i+1}L_{i+1} \end{bmatrix} \equiv \begin{bmatrix} g_i & b_i' \\ b_i & V_{i+1} \end{bmatrix}.$$

The comparison of respective entries implies the proved formulas. □

In this way, the transition from MT statistics, gained for the static components, to normal dynamic factors is completed. The following algorithm summarizes this transition.

Algorithm 12.8 (Static component → dynamic factors)

Initial mode

- *Perform initialization by the MT algorithm so that you will get* $\mathring{c}$ *of* $\mathring{\Psi}$*-dimensional boxes characterized by the number of data* $\mathring{k}_c$ *captured within the box, and* $\mathring{\Psi}$*-vectors of the box center* $\hat{\mu}_c$ *and widths* b_c*.*
- *Set* $L_{\mathring{\Psi}+1,c;0} = 1$*,* $D_{\mathring{\Psi}+1,c;0} = \mathring{\Psi}$*,* $c \in c^*$*.*
- *Select the ratio* $\gamma \approx 0.1$ *and evaluate* $s = \sum_{c \in c^*} \mathring{k}_c$*.*

Recursive mode

```
For   $c = 1, \ldots, \mathring{c}$
          Set $\kappa_{c;0} = \mathring{k}_c \frac{\rho}{s}$
  For   $i = \mathring{\Psi}, \ldots, 2$
        $\nu_{ic;0} = \kappa_{c;0}$
        Solve equation for $\beta$ with upper triangular matrix $L'_{ic;0} D_{ic;0}$
          $L'_{ic;0} D_{ic;0} \beta = [0, \ldots, 0, k_c \hat{\mu}_{ic}]'$
          Set $a = k_c \left( \frac{b_{ic}^2}{3} + \hat{\mu}_{ic}^2 \right) - \beta' D_{ic;0} \beta$
        Define
        $L_{(i-1)c;0} \equiv \begin{bmatrix} 1 & 0 \\ \beta & L_{ic;0} \end{bmatrix}, \quad D_{(i-1)c;0} \equiv diag\left[a, diag\left(D_{ic;0}\right)\right]$
  end   of the cycle over i
end   of the cycle over c
```

Problem 12.1 (BMTB algorithm with normal MT factors) *MT normal factors can be used instead of MT uniform factors to create an alternative initialization.*

It makes no sense to replace MT factors by a normal ones in a one-to-one fashion since it would just increase the computational complexity without bringing any obvious advantage. On the other hand, it is possible to use an approximate mixture estimation on a multivariate mixture made of MT-normal factors. The centers, diagonal covariances and component weights, found when solving the shadow cancelling problem, may serve as the needed start.

Moreover, the possibility of using the dynamic nature of MT normal components is worth inspection.

Obviously, there are a lot of aspects related to normal mixtures that can be exploited while preserving the mean-tracking ability of MT normal factors. This direction is worth elaborating.

13

Mixed mixtures

Learning described at the general level in Chapter 6 deals with the factorized version of the probabilistic component description. It allows us to use different models for describing different entries of data records and to model jointly continuous and discrete data, to deal with *mixed mixtures*. Specific aspects related to mixtures of this type form the content of the current chapter.

Continuous-valued quantities are modelled by normal factors (see Chapter 8) and discrete-valued ones by Markov-chain factors; see Chapter 10. While the former one can be made easily dependent on later ones, the opposite case has not been supported. This major and substantial restriction on the presented treatment is weakened here by exploiting a special version of normal factors suitable to this purpose. These *MT normal factors* provide a practical bridge between the world of discrete and continuous quantities. They are simply normal factors with a small *fixed variance*. Consequently, their learning just tracks positions of the modelled quantities, similarly as the original MT factors. Their mixture approximates the pf of a discrete-valued quantity. In this way, the difficult area of logistic regression, e.g., [175], is avoided.

Learning of the mixed mixtures is addressed in Section 13.1. Section 13.2 presents a promising attempt to estimate the ratio of mixtures. It allows us to model dependence of discrete quantities on continuous ones. More generally and more importantly it allows the use of data-dependent component weights.

The design part is omitted as it has not been elaborated in detail. The learning part and Chapters 9 and 11 imply, however, that it will lead to a feasible blend of manipulations with quadratic forms and tables. Moreover, the mild — and often inevitable — restriction on modelling of discrete quantities by the MT normal factors makes the design part identical with that of Chapter 9. It is obvious as the referred design uses constant point estimates of all parameters.

13.1 Learning with factors on mixed-type quantities

The joint presence of discrete and continuous valued quantities in respective factors is the key problem of the mixed mixtures. Both possible variants, i.e., the dependencies of the continuous factor output on discrete and continuous quantities and the discrete factor output on continuous and discrete quantities, are discussed in this section.

13.1.1 Factors in EF with a discrete regressor

We consider the case that the ith regression vector contains one or more discrete entries (regressors). To simplify notation, we map them on a scalar discrete regressor, say $\iota_t \in \iota^* \equiv \{1, \ldots, \mathring{\iota}\}$. Then, it holds.

Proposition 13.1 (Estimation in EF with a discrete regressor)
Let us consider the ith factor in the exponential family in the form

$$f(d_{i;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1), \Theta_i) = \prod_{\iota=1}^{\mathring{\iota}} \left[A(\Theta_{\iota i}) \exp\left[\langle B_\iota(\Theta_{\iota i}), C_\iota(\Psi_{\iota i;t})\rangle\right]\right]^{\delta_{\iota\iota_t}}$$

where $\Theta_i \equiv \{\Theta_{\iota i}\}_{\iota=1}^{\mathring{\iota}}$. Kronecker symbol δ (5.31) and the functions $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ occurring in the exponential family, Agreement 3.1, are employed. The individual factors in this product called parameterized subfactors*. Then, the conjugate prior pdf $f(\Theta_i|d(0))$ has the form*

$$f(\Theta_i|d(0)) = \prod_{\iota=1}^{\mathring{\iota}} A^{\nu_{\iota i;0}}(\Theta_{\iota i}) \exp\left[\langle V_{\iota i;0}, C_\iota(\Theta_{\iota i})\rangle\right].$$

The corresponding posterior pdfs $f(\Theta_i|d(t))$ preserve this functional form and their sufficient statistics evolve as follows.

$$V_{\iota i;t} = V_{\iota i;t-1} + \delta_{\iota_t,\iota} B_\iota(\Psi_{\iota i;t}),\ \nu_{\iota i;t} = \nu_{\iota i;t-1} + \delta_{\iota_t,\iota},\ \iota = 1, \ldots, \mathring{\iota},\ t \in t^*.$$

Proof. It is implied directly by the Bayes rule. □

The proposition says that the observed discrete-valued entry ι_t serves as a known pointer to the single pair of statistics $V_{\iota_t i;t-1}$, $\nu_{\iota_t i;t-1}$ updated at the time t.

The individual subfactors with indexes ιi are only updated on the subselection of data for which the observed discrete pointers $\iota_t,\ t \in t^*$, have the value ι. Thus, the observed discrete regressor ι_t effectively segments the processed data into $\mathring{\iota}$ parts. Let us formalize this conclusion.

Proposition 13.2 (Processing of factors with discrete regressor)
Let us consider learning of a mixed mixture with some factors, say $\{i_1, \ldots, i_{\mathring{k}}\}$, each depending on a discrete-valued observable quantity $\iota_{i_k;t}$, $k \in k^* \equiv$

$\{1,\ldots,\mathring{k}\}$. *Then, at time t, the parameter estimates of subfactors with indexes $\iota_{i_k;t}$, $k \in k^*$, are updated by the current data and possibly forgotten. They serve for computing component weights used in the approximate mixture estimation, Section 6.5. Other subfactors are untouched and unused.*

Proof. Omitted. □

Remark(s) 13.1

Often, the discrete-valued quantity in the condition keeps its fixed value for a long period. Then, it is potentially possible to presegment the data accordingly and process these segments independently. Before doing it, it is reasonable to consider explicitly such a quantity in the condition of the parameterized model because it may well happen that its different values lead to identical system responses. Then, the presegmentation can be done with respect to aggregated values of this quantity. Consequently, a smaller amount of segments have to be learned and overall learning efficiency is improved. This situation may also have significant interpretation consequences.

13.1.2 MT normal factors

The opposite case, when the *discrete-valued factor output* $d_{i;t}$ depends on continuous-valued, and possibly discrete, quantities is much more difficult. It is implied by the fact that in a nontrivial dynamic case a model out of the exponential family is needed. Essentially, Markov-chains are the only dynamic models of discrete outputs staying within EF.

Let us indicate how to arrive at this conclusion. For this, let it us assume that the scalar discrete-valued innovation Δ_t has a model in the dynamic EF with vector-valued functions $B(\Psi_t)$, $C(\Theta)$, i.e.,

$$\begin{aligned} f(\Delta_t|\psi_t,\Theta) &= A(\Theta)\exp\langle B(\Delta_t,\psi_t), C(\Theta)\rangle \equiv \exp\langle \beta_t, \gamma_{\Delta_t}(\Theta)\rangle \equiv f(\Delta_t|\beta_t,\Theta)\\ \beta'(\psi_t) &\equiv \left[B'(1,\psi_t),\ldots,B'(\mathring{\Delta},\psi_t),1\right]\\ \gamma'_{\Delta_t}(\Theta) &\equiv \left[0_{1,(\Delta_t-1)\mathring{C}}, C'(\Theta),0,\ldots,0,\ln(A(\Theta))\right]\end{aligned}$$

This identity holds on support of the pf $f(\Delta_t|\psi_t,\Theta) \leq 1$. For a fixed Θ, we can always select $\gamma_{\Delta_t} \geq 0$. For a nontrivial dynamic EF, some entries $\gamma_{i\Delta_t} > 0$ and $f(\Delta_t|\beta_t,\Theta)$ depends nontrivially on β_t. Taking the derivative of the normalizing condition $1 = \sum_{\Delta_t=1}^{\mathring{\Delta}} f(\Delta_t|\beta_t,\Theta)$ with respect to the ith entry of b_t, we get the contradiction

$$0 = \sum_{\Delta_t=1}^{\mathring{\Delta}} \gamma_{i\Delta_t}(\Theta) f(\Delta_t|\beta_t,\Theta) > 0.$$

Thus, no such model exists within EF and we either have to employ approximate techniques developed in the logistic regression, e.g., [176], or find another way out. The fixed box widths of the MT uniform factors, Chapter 12, inspire us to approximate the desired parameterized factor by a mixture of parameterized MT normal factors, introduced in Chapter 12, which model well the differences in data positions.

Thus, even if we know that the output $d_{i;t}$ of the ith factor is a discrete-valued one, we model it by the normal (scalar) component

$$f(d_{i;t}|d_{(i+1)\cdots\mathring{d};t}, d(t-1), \Theta_{ic}, c) = \mathcal{N}_{d_{i;t}}(\theta_{ic}'\psi_{ic;t}, r_{ic}), \qquad (13.1)$$

around the positions $\theta_{ic}'\psi_{ic;t}$. Here, $\Theta_{ic} \equiv \theta_i \equiv$ regression coefficients in $\theta_{ic}^* \equiv \mathring{\psi}_{ic}$-dimensional real space. $\psi_{ic;t}$ is the regression vector. The noise variance r_{ic} is a known and small positive number. Behavior around other positions has to be modelled by other MT factors and thus the mixture model is needed.

Let us assume that the predicted factor output $d_{i;t} \in \{1, 2, \ldots\}$. To distinguish sharply these values by normal pdf it is sufficient to take its standard deviation smaller than 0.05. It motivates the recommendation to fix the variance r_{ic} at the order 10^{-3} and smaller.

Using the results of Chapter 8, it is straightforward to describe the estimation and prediction of the MT normal factors. The notation of Chapter 8 is adopted and the subscript ic is temporarily suppressed. Recall also that the information matrices are split as follows.

$$V = \begin{bmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}V' \\ {}^{\lfloor d\psi}V & {}^{\lfloor \psi}V \end{bmatrix}, \ {}^{\lfloor d}V \text{ is scalar. The } L'DL \text{ decomposition is employed:}$$

$$V = L'DL, \ L \text{ is a lower triangular matrix with unit diagonal and}$$
$$D \text{ is diagonal matrix with nonnegative entries while}$$
$$L = \begin{bmatrix} 1 & 0 \\ {}^{\lfloor d\psi}L & {}^{\lfloor \psi}L \end{bmatrix}, \quad D = \begin{bmatrix} {}^{\lfloor d}D & 0 \\ 0 & {}^{\lfloor \psi}D \end{bmatrix}, \ {}^{\lfloor d}D \text{ is scalar.}$$

Proposition 13.3 (Estimation and prediction of the MT normal factors) *Let natural conditions of decision making, Requirement 2.5, hold. The MT normal factor (13.1) is treated, and a conjugate prior (3.13) as well as the conjugate alternative (see Section 3.1) in stabilized forgetting are used. Then, the prior, posterior and alternative pdfs are normal*

$$f(\theta|V) = \mathcal{N}_\theta\left(\hat{\theta}, r\, {}^{\lfloor \psi}V^{-1}\right) \ \textit{with} \ \hat{\theta} \equiv \mathcal{E}[\theta|L, D, \nu] = {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor d\psi}L \ \textit{and}$$

$$r \equiv {}^{\lfloor d}D = {}^{\lfloor d}V - {}^{\lfloor d\psi}V'\hat{\theta} = \textit{ a priori fixed value.}$$

The predictive pdf is also normal (see Proposition 3.2)

$$f(d|\psi, V) = \mathcal{N}_d\left(\hat{\theta}'\psi, r\left(1 + \psi'\, {}^{\lfloor \psi}V^{-1}\psi\right)\right) \ \textit{or alternatively}$$

$$f(d|\psi, V) = \frac{\exp\left\{-\frac{h_d^2}{2r(1+\zeta)}\right\}}{\sqrt{2\pi r(1+\zeta)}}, \ \textit{with} \qquad (13.2)$$

$$h = (L')^{-1}\Psi, \ h' = [h_d, h_\psi'], \ h_d \textit{ is scalar}, \ \zeta \equiv \sum_{i=2}^{\mathring{\Psi}} h_i^2/D_{ii}. \qquad (13.3)$$

Note that h_d coincides with the prediction error $\hat{e}$ (8.31). The involved sufficient statistic evolves according to the recursion with resetting of the entry $^{\lfloor d}V_t$

$$V_t = \lambda(V_{t-1} + \Psi_t\Psi_t') + (1-\lambda)\, {}^{\lfloor A}V_t, \quad \textit{initial and alternative } V_0,\ {}^{\lfloor A}V_t \textit{ given}$$
$${}^{\lfloor d}V_t = r + {}^{\lfloor d\psi}V_t'\hat{\theta}_t \ \Leftrightarrow\ {}^{\lfloor d}D_t = r.$$

The normalization is finite for all time moments iff $^{\lfloor \psi}D_0 > 0$, and $r > 0$.

Proof. The majority of evaluations can be found in [69] or can be derived as a special case of the general normal factor with the known r; Chapter 8. □

Remark(s) 13.2
The algorithm for updating these factors is identical to that for normal factors; see Chapter 8. We just have to change the value of the least-squares remainder $^{\lfloor d}D$ (8.14) after each data updating. The main difference is in evaluation of predictive pdfs; cf. (8.31) and (13.2).

The derived estimation is an important but particular step in the whole learning. Here, we review the remaining steps by referring to relevant sections, where their solutions are presented. We also point to the problems that remain to be solved.

- Data preprocessing is mostly oriented on individual signals in question. Generally, it is addressed in Section 6.2. Section 8.2 specializes in normal factors and thus also in MT normal factors. Markov chain oriented preprocessing is discussed in Section 10.2.

  Problem 13.1 (Mixed counterpart of PCA) *Reduction of dimensionality is the only considered preprocessing step dealing jointly with several signals. It is based on continuous-signal oriented PCA. Thus, a specific solution is needed for the mixed case. Surely, strongly unequal variances of respective data entries have to be taken into account.*

- Incorporation of physical knowledge is strictly signal-oriented within this text. Under this restriction, no additional problems are induced by the mixed case. General, normal, and thus MT normal, and Markov cases are addressed in Section 6.3, 8.3 and 10.3, respectively.
- Construction of the prior pdf is solved predominantly factorwise. Thus, adequate solutions are ready; see Sections 6.4, 8.4 and 10.4.
- Structure estimation is again solved factorwise. For particular solutions; see Sections 6.6, 8.6, 10.6.

  Problem 13.2 (Unsolved structure estimation problems) *The majority of the unsolved problems is not induced by the mixed case. It is, however, worthwhile to list them.*
  - *Structure estimation of MT normal factors is not described here in detail but its solution is a special and simplified case of the normal factors; Section 8.6.1 and references [93, 95, 162].*

  - *Structure estimation of Markov chains is not practically solved. The solution given in Proposition 10.11 has to be elaborated further on, for instance, in the direction outlined in [170] and definitely in the directions related to Bayesian networks.*
  - *Efficient estimation of the component structure is the open problem for all types factors and an adequate solution has to be searched for.*
  - *The estimation of the mixture structure is conceptually solved through the branching by factor splitting, factor merging and component cancelling. All problems mentioned in relevant Sections of Chapters 6, 8 and 10 persist for the mixed case, too. It brings, however, no new problems.*

- Model validation, presented in previous chapters, is applicable to the mixed case mostly without change. Algorithm 8.18, forgetting based validation, Section 6.7.3, and judging of models as multi-step predictors, Section 7.1.2 seem to be the most promising common techniques.

Problem 13.3 (Completion of the mixed-mixtures suite) *Learning as well as design algorithm is feasible for the mixed mixtures. This is especially true when normal MT factors are employed for modelling of discrete data. In this case, the learning is obtained through simplification of the learning of normal mixtures. The design is completely unchanged as the fixed point estimates of parameters are used.*

13.2 An approximate estimation of mixture ratios

The normal MT factors help us to cope with the modelling of dependence of discrete quantities on continuous ones. The solution is based on interpreting discrete quantities as continuous ones with a very narrow variation range. In this way, the normalization does not spoil the mixture character of the model. Estimation of a *ratio of mixtures* is another way how to approach the problem. A promising attempt in this respect is outlined in this section. It extends the descriptive power of all mixtures as it allows us to deal with dynamic mixtures having data-dependent component weights. The significance of this extension cannot be exaggerated. At the same time, it is fair to forewarn that the proposed solution is far from being the final one.

13.2.1 Mixture modelling of stationary pdf

We start with a reparameterization of the joint pdf of the observed data.

The the chain rule applied to the joint pdf $f(d(\mathring{t}))$ of the data sequence $d(\mathring{t})$ gives $f(d(\mathring{t})) = \prod_{t \in t^*} f(d_t|d(t-1))$. Finite memory of the majority of real systems justifies the assumption that there is a finite fixed-dimensional state $\phi_{t-1} = \phi(d(t-1))$ such that $f(d_t|d(t-1)) = f(d_t|\phi_{t-1})$. We assume moreover that the conditional pdf $f(d_t|\phi_{t-1})$ is a time invariant function of the data vector (cf. Agreement 9.1)

$$\Psi_t = [d_t', \phi_{t-1}']' \equiv [\Psi_{1;t}, \ldots, \Psi_{\mathring{\Psi};t}]', \; \Psi_{i;t} \text{ are scalars.} \tag{13.4}$$

The considered modelling specifies a parametric class of stationary pdfs on Ψ_t that defines indirectly the parameterized model used in estimation. Specifically, we assume that the stationary pdf $f(\Psi_t|\Theta)$ of the data vector Ψ_t, parameterized by an unknown parameter Θ, is at disposal. It defines the parameterized model of the system in question

$$f(d_t|d(t-1), \Theta) = f(d_t|\phi_{t-1}, \Theta) = \frac{f(\Psi_t|\Theta)}{f(\phi_{t-1}|\Theta)}, \tag{13.5}$$

where $f(\phi_{t-1}|\Theta) = \int f(\Psi_t|\Theta)\, dd_t$.

The possibility to parameterize time and data invariant function of Ψ_t is the main advantage of the indirect construction (13.5). Under rather general conditions, finite mixtures have a "universal approximation property" [39], i.e., almost any time-invariant pdf $f(\Psi_t)$ can be approximated by a finite mixture

$$f(\Psi_t|\Theta) = \sum_{c\in c^*} \alpha_c f(\Psi_t|\Theta_c, c), \tag{13.6}$$

$$\Theta \in \Theta^* = \left\{\Theta_c,\, c \in c^*,\, \alpha \in \left\{\alpha_c \geq 0,\, \sum_{c\in c^*} \alpha_c = 1\right\}\right\}.$$

The mixture form (13.6), the conditioning (13.5) and the chain rule give the parameterized model needed for the Bayesian parameter estimation

$$\begin{aligned} f(d_t|\phi_{t-1}, \Theta) &= \frac{\sum_{c\in c^*} \alpha_c f(\Psi_t|\Theta_c, c)}{\sum_{\tilde{c}\in c^*} \alpha_{\tilde{c}} f(\phi_{t-1}|\Theta_{\tilde{c}}, \tilde{c})} \\ &= \sum_{c\in c^*} \underbrace{\frac{\alpha_c f(\phi_{t-1}|\Theta_c, c)}{\sum_{\tilde{c}\in c^*} \alpha_{\tilde{c}} f(\phi_{t-1}|\Theta_{\tilde{c}}, \tilde{c})}}_{\alpha_c(\phi_{t-1})} f(d_t|\phi_{t-1}, \Theta_c, c). \end{aligned} \tag{13.7}$$

The resulting *dynamic mixture* (13.7) has time-invariant but *state-dependent component weights* $\{\alpha_c(\phi_{t-1})\}_{c\in c^*}$.

We use factorized versions of the permuted components $f(\Psi_{c;t}|\Theta_c, c)$ in the mixture (13.6). It means that we factorize them using the chain rule

$$f(\Psi_{c;t}|\Theta_c, c) = \prod_{i=1}^{\mathring{\Psi}} f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c). \tag{13.8}$$

The data record d_t is the obligatory part of all component-specific data vectors $\Psi_{c;t}$. The subsequent evaluations are simplified by keeping the data record d_t at leading positions of all data vectors $\Psi_{c;t}$.

The factor $f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{ic},c)$ predicts the ith entry of cth data vector $\Psi_{c;t}$, i.e., it predicts the factor output $\Psi_{ic;t}$, using the regression vector $\psi_{ic;t}$, formed by an appropriate subselection of entries $(\Psi_{(i+1)c;t},\ldots,\Psi_{\mathring{\Psi}c;t})$. It is parameterized by an unknown parameter Θ_{ic} formed by a subselection of Θ_c.

Using (13.8), the conditional pdf (13.7) gets the form

$$f(d_t|\phi_{t-1},\Theta)=\frac{\sum_{c\in c^*}\alpha_c\prod_{i=1}^{\mathring{\Psi}}f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{ic},c)}{\sum_{c\in c^*}\alpha_c\prod_{i=\mathring{d}+1}^{\mathring{\Psi}}f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{ic},c)},$$

i.e., the marginal pdfs in the denominator are obtained by omitting the initial factors in the products defining the components. This property stems from the assumed leading position of the predicted data record d_t in $\Psi_{c;t}\equiv[d_t',\phi_{c;t-1}']'$.

13.2.2 Extended quasi-Bayes estimation

The mixture estimation is a well-established art, [48, 49]. A specific estimation of the rational model (13.7) is, however, missing. A promising solution is outlined here.

The quasi-Bayes estimation, Section 6.5.1, takes the mixture as the marginal pdf of the pdf

$$f(\Psi_t,c_t|\Theta)=\alpha_{c_t}f(\Psi_{c_t;t}|\Theta_{c_t},c_t)=\prod_{c\in c^*}\left[\alpha_c f(\Psi_{c;t}|\Theta_c,c)\right]^{\delta_{c,c_t}}. \tag{13.9}$$

The quantity c_t is interpreted as an unobserved random pointer to the active component. The Kronecker symbol $\delta_{c,c_t}\equiv 1$ if $c=c_t$ and $\delta_{c,c_t}\equiv 0$ otherwise.

Recall that assuming the prior pdf in the form, cf. Agreement 6.1,

$$f(\Theta|d(t-1))\propto\prod_{c\in c^*}\alpha_c^{\kappa_{c;t-1}-1}\prod_{i=1}^{\mathring{d}}f(\Theta_{ic}|d(t-1)), \tag{13.10}$$

the quasi-Bayes estimation uses the weight $w_{c;t}\equiv f(c_t=c|d(t))$ as an approximation of the unknown δ_{c,c_t}. This choice used in the data updating $f(\Theta|d(t-1))\rightarrow f(\Theta|d(t))$ preserves the product form (13.10) of the parameter estimate.

In the inspected case, the dynamic model is the ratio of two finite mixtures with constant weights. The numerator is the mixture describing the data vector Ψ_t and the denominator is the mixture describing the state ϕ_{t-1}. Thus, it seems to be reasonable to

- assume that the prior pdf is also in the form (13.10), but with $\underline{\mathring{\Psi}\text{ factors}}$,
- approximate the inverse of the denominator by a finite mixture,
- compute the weights $w_{c_t;t}$ for the product version of the numerator and the weights $\tilde{w}_{\tilde{c}_t;t}$ related to the product version of the approximated denominator,

- update the prior pdf by the product of these product forms with the unknown Kronecker symbols replaced by respective weights $w_{c_t;t}$, $\tilde{w}_{\tilde{c}_t;t}$.

Let us develop this plan starting from a simple motivating proposition.

Proposition 13.4 (One-sided approximation of pdfs) *Let $f(\Theta|d(t))$ be a posterior pdf and $\hat{f}(\Theta|d(t))$ its approximation such that the following implication hold $\hat{f}(\Theta|d(t)) = 0 \Rightarrow f(\Theta|d(t)) = 0$ for all $t \in t^*$ and any $\Theta \in \Theta^*$. Let the asymptotic support of the pdf $\hat{f}(\Theta|d(t))$ consist of a single-point set $\left\{ {}^{\lfloor o}\Theta \right\} \subset \Theta^*$. Then, the asymptotic support of $f(\Theta|d(t))$ consists of the same point.*

Proof. The assumption implies that $\mathrm{supp}\left[f(\Theta|d(t))\right] \subset \mathrm{supp}\left[\hat{f}(\Theta|d(t))\right]$ for all $t \in t^*$. Thus, the inclusion holds asymptotically and the nonempty subset of the single-point set $\left\{ {}^{\lfloor o}\Theta \right\} \subset \Theta^*$ coincides with this set. □

Proposition 2.15, that describes asymptotic properties of the Bayesian estimation, shows that the support of the pdf $f(\Theta|d(t))$ concentrates almost surely on parameters that point to parameterized models closest to the objective model of reality. Consequently, Proposition 13.4 says that the pdf $\hat{f}(\Theta|d(t))$ approximating $f(\Theta|d(t))$ from above will provide the correct estimate if it concentrates on a single point.

The following proposition provides a specific upper bound $\hat{f}(\Theta|d(t))$ that can be constructed recursively.

Recall that $Di_\alpha(\kappa)$ is Dirichlet pdf determined by the vector statistics κ and describing the component weights α.

Proposition 13.5 (One sided approximation: ratio of mixtures) *Let us consider the model parameterized by $\Theta \equiv \left\{ \{\Theta_{ic}\}_{i=1}^{\mathring{\Psi}}, \alpha_c \right\}_{c\in c^*}$*

$$f(d_t|d(t-1),\Theta) \equiv \frac{\sum_{c\in c^*} \alpha_c \prod_{i=1}^{\mathring{\Psi}} f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{ic})}{\sum_{\tilde{c}\in c^*} \alpha_{\tilde{c}} \prod_{i=\mathring{d}+1}^{\mathring{\Psi}} f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{i\tilde{c}})}$$

and the prior upper bound

$$\hat{f}(\Theta|d(t-1)) = Di_\alpha(\hat{\kappa}_{t-1}) \prod_{c\in c^*} \prod_{i=1}^{\mathring{\Psi}} \hat{f}(\Theta_{ic}|d(t-1))$$

on the prior pdf $f(\Theta|d(t-1))$, i.e., fulfilling the inequality $f(\Theta|d(t-1)) \leq C(d(t-1))\hat{f}(\Theta|d(t-1))$ with a finite $C(d(t-1))$. This guarantees the inclusion of their supports $\mathrm{supp}\left[f(\Theta|d(t-1))\right] \subset \mathrm{supp}\left[\hat{f}(\Theta|d(t-1))\right]$.

After observing the new data vector Ψ_t, the following pdf forms the upper bound on the posterior pdf $f(\Theta|d(t))$.

$$\hat{f}(\Theta|d(t)) \propto \hat{f}(\Theta|d(t-1)) \tag{13.11}$$
$$\times \begin{cases} \frac{1}{\left(\sum_{\bar{c}\in\tilde{c}^*} \alpha_{\bar{c}}\right)^2} \sum_{c\in c^*} \alpha_c \sum_{\tilde{c}\in\tilde{c}^*} \alpha_{\tilde{c}} \frac{\prod_{i=1}^{\mathring{\Psi}} f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{ic})}{\prod_{\tilde{i}=\mathring{d}+1}^{\mathring{\Psi}} f(\Psi_{\tilde{i}\tilde{c};t}|\psi_{\tilde{i}\tilde{c};t},\Theta_{\tilde{i}\tilde{c}})} & \textit{if } \tilde{c}^* \neq \emptyset \\ \sum_{c\in c^*} \prod_{i=1}^{\mathring{d}} f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{ic}) & \textit{if } \tilde{c}^* = \emptyset. \end{cases}$$

The set $\tilde{c}^* \equiv \tilde{c}^*(\Psi_t)$ contains pointers $\tilde{c} \in c^*$ to such components for which

$$\int \frac{\hat{f}(\Theta_{i\tilde{c}}|d(t-1))}{f(\Psi_{i\tilde{c};t}|\psi_{i\tilde{c};t},\Theta_{i\tilde{c}})}\, d\Theta_{i\tilde{c}} < \infty \textit{ for all } i = \mathring{d}+1,\ldots,\mathring{\Psi}. \tag{13.12}$$

Proof. The Bayes rule implies that supp $[\, f(\Theta|d(t))]$ is contained in the support of the pdf proportional to

$$\frac{\sum_{c\in c^*} \alpha_c \prod_{i=1}^{\mathring{\Psi}} f(\Psi_{ic;t}|\psi_{ic;t},\Theta_{ic})}{\sum_{\tilde{c}\in c^*} \alpha_{\tilde{c}} \prod_{\tilde{i}=\mathring{d}+1}^{\mathring{\Psi}} f(\Psi_{\tilde{i}\tilde{c};t}|\psi_{\tilde{i}\tilde{c};t},\Theta_{\tilde{i}\tilde{c}})} \hat{f}(\Theta|d(t-1)). \tag{13.13}$$

Let us consider the case $\tilde{c}^* \neq \emptyset$. By omitting all components in denominator out of $\tilde{c}^*$, we get an upper bound on the expression (13.13). It has the same form as (13.13) but summation in the denominator is taken over $\tilde{c}^*$ only. The Jensen inequality implies the following bound on the new denominator

$$\frac{1}{\sum_{\tilde{c}\in\tilde{c}^*} \alpha_{\tilde{c}} \prod_{i=\mathring{d}+1}^{\mathring{\Psi}} f(\Psi_{i\tilde{c};t}|\psi_{i\tilde{c};t},\Theta_{i\tilde{c}})}$$
$$= \frac{1}{\left(\sum_{\bar{c}\in\tilde{c}^*} \alpha_{\bar{c}}\right) \sum_{\tilde{c}\in\tilde{c}^*} \frac{\alpha_{\tilde{c}}}{\sum_{\bar{c}\in\tilde{c}^*} \alpha_{\bar{c}}} \prod_{i=\mathring{d}+1}^{\mathring{\Psi}} f(\Psi_{i\tilde{c};t}|\psi_{i\tilde{c};t},\Theta_{i\tilde{c}})}$$
$$\underbrace{\leq}_{(2.14)} \frac{1}{\left(\sum_{\bar{c}\in\tilde{c}^*} \alpha_{\bar{c}}\right)^2} \sum_{\tilde{c}\in\tilde{c}^*} \frac{\alpha_{\tilde{c}}}{\prod_{i=\mathring{d}+1}^{\mathring{\Psi}} f(\Psi_{i\tilde{c};t}|\psi_{i\tilde{c};t},\Theta_{i\tilde{c}})}.$$

It proves the first part of the claim.

Let us consider the case with $\tilde{c}^* = \emptyset$. In each cth term of the sum over $c \in c^*$ occurring in (13.13), we simply omit all terms in denominator except the cth one. It leads to cancelling both α_c and parameterized factors with indexes $i = \mathring{d}+1,\ldots,\mathring{\Psi}$ and proves this degenerate case. □

Problem 13.4 (Refinement of bounds to factor structure) *The component is shifted to $c^* \setminus \tilde{c}^*$ even if just its single factor makes the integral (13.12) infinite. This condition can be probably refined. It should done at least for normal mixtures.*

For the case $\tilde{c}^* \neq \emptyset$, $\tilde{c}^* \neq c^*$ the factor $\left(\sum_{\bar{c}\in\tilde{c}^*} \alpha_{\bar{c}}\right)^{-2}$ "spoils" the resemblance of the upper bound to a mixture. The following simple proposition helps us to get rid of it.

Proposition 13.6 (One-sided approximation: ratio of weights) *Let $\tilde{c}^*$ be nonempty set and the statistic $\hat{\kappa}_{t-1}$ determining the Dirichlet pdf has entries $\hat{\kappa}_{\tilde{c};t-1} > \frac{2}{\mathring{\tilde{c}}}$ for all $\tilde{c} \in \tilde{c}^*$. Then,*

$$\frac{\prod_{c\in c^*} \alpha_c^{\hat{\kappa}_{c;t-1}-1}}{\left(\sum_{\tilde{c}\in\tilde{c}^*} \alpha_{\tilde{c}}\right)^2} \leq \prod_{c\in c^*\setminus\tilde{c}^*} \alpha_c^{\hat{\kappa}_{c;t-1}-1} \prod_{\tilde{c}\in\tilde{c}^*} \alpha_{\tilde{c}}^{\hat{\kappa}_{\tilde{c};t-1}-\eta-1}, \quad \eta \equiv \frac{2}{\mathring{\tilde{c}}}. \tag{13.14}$$

The first product in (13.14) is set to 1 if $c^* = \tilde{c}^*$.

Proof. Diving the inequality to be proved by its positive right-hand side, we get the equivalent to be proved

$$\frac{\prod_{\tilde{c}\in\tilde{c}^*} \alpha_{\tilde{c}}^{\eta}}{\left(\sum_{\tilde{c}\in\tilde{c}^*} \alpha_{\tilde{c}}\right)^2} \leq 1, \quad \text{for } \alpha_{\tilde{c}} \geq 0 \text{ and } \sum_{\tilde{c}\in\tilde{c}^*} \alpha_{\tilde{c}} \leq 1.$$

It can be simply shown that its maximum is reached for identical entries $\alpha_{\tilde{c}} = A \geq 0, A \leq 1,\ \tilde{c} \in \tilde{c}^*$. For them, we get the value $\frac{A^{\eta\mathring{\tilde{c}}-2}}{\mathring{\tilde{c}}^2}$. It is clear that $A^{\eta\mathring{\tilde{c}}-2} \leq 1$ for $\eta \geq \frac{2}{\mathring{\tilde{c}}}$. Thus, the claim holds as $\mathring{\tilde{c}} \geq 1$. □

Remark(s) 13.3 *The approximation is not needed for the case $\tilde{c}^* = c^*$. We can effectively deal with both cases at once by setting $\eta = 0$ for $\tilde{c}^* = c^*$.*

The combination of Propositions 13.5, 13.6 updates the upper bound in the form of the mixture. Thus, the heuristic leading to quasi-Bayes estimation; see Section 6.5.1, can almost be copied. To make it, the upper bound is viewed formally as the marginal pdf of the following pdf (the sign ˆ is dropped)

For $\tilde{c}^* \neq \emptyset$ (13.15)

$$f(\Theta, c_t, \tilde{c}_t|d(t)) \propto \prod_{c\in c^*} \left[\alpha_c \prod_{i=1}^{\mathring{\Psi}} f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic})\right]^{\delta_{c,c_t}}$$

$$\times \prod_{\tilde{c}\in\tilde{c}^*} \left[\alpha_{\tilde{c}}^{1-\eta} \prod_{\tilde{i}=\mathring{d}+1}^{\mathring{\Psi}} \frac{1}{f(\Psi_{\tilde{i}\tilde{c};t}|\psi_{\tilde{i}\tilde{c};t}, \Theta_{\tilde{i}\tilde{c}})}\right]^{\delta_{\tilde{c},\tilde{c}_t}} Di_{\alpha}(\kappa_{t-1}) \prod_{\tilde{i}=1,\bar{c}=1}^{\mathring{\Psi},\mathring{c}} f(\Theta_{\tilde{i}\bar{c}}|d(t-1)).$$

For $\tilde{c}^* = \emptyset$

$$f(\Theta, c_t|d(t)) \propto \prod_{c\in c^*} \left[\prod_{i=1}^{\mathring{d}} f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic})\right]^{\delta_{c,c_t}}$$

$$\times Di_{\alpha}(\kappa_{t-1}) \prod_{\tilde{i}=1,\bar{c}=1}^{\mathring{\Psi},\mathring{c}} f(\Theta_{\tilde{i}\bar{c}}|d(t-1)).$$

Comparing to the quasi-Bayes estimation, the recognition whether the component belongs to $\tilde{c}^*$ is an additional logical operation needed. Moreover, the expectations $\mathcal{E}[\alpha_c \alpha_{\tilde{c}}^{1-\eta}|d(t-1)]$ have to be evaluated. It seems to be sufficient to use the crudest approximation

$$\mathcal{E}[\alpha_c\, \alpha_{\tilde{c}}^{1-\eta}|d(t-1)] = \tag{13.16}$$
$$= \mathcal{E}[\alpha_c|d(t-1)]\, [\mathcal{E}[\alpha_{\tilde{c}}|d(t-1)]]^{1-\eta} \propto \frac{\kappa_{c;t-1}}{\sum_{c\in c^*} \kappa_{c;t-1}} \left[\frac{\kappa_{\tilde{c};t-1}}{\sum_{\tilde{c}\in\tilde{c}^*} \kappa_{\tilde{c};t-1}}\right]^{1-\eta}.$$

The overall *extended quasi-Bayes estimation* algorithm we put together for mixtures with parameterized factors in exponential family; see Section 3.2,

$$f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic}) = A(\Theta_{ic}) \exp\left\langle B([\Psi_{ic;t}, \psi'_{ic;t}]'), C(\Theta_{ic})\right\rangle$$

and the corresponding conjugate (upper bounds on) prior pdfs

$$f(\Theta_{ic}|d(t-1)) = \frac{A^{\nu_{ic;t-1}}(\Theta_{ic}) \exp\langle V_{ic;t-1}, C(\Theta_{ic})\rangle}{\mathcal{I}(V_{ic;t-1}, \nu_{ic;t-1})}.$$

Here, $V_{ic;t-1}$ and $\nu_{ic;t-1}$ form the (approximate) sufficient statistic and $\mathcal{I}(V_{ic;t-1}, \nu_{ic;t-1})$ denotes the corresponding normalizing integral.

The evaluation of the posterior pf $f(c_t, \tilde{c}_t|d(t))$ — decisive in the design of the learning algorithm — is straightforward for the degenerated case with $\tilde{c}^* = \emptyset$. Integration of the joint pdf (13.15) over Θ^* implies that

$$\text{For } \tilde{c}^* = \emptyset \tag{13.17}$$
$$f(c_t = c|d(t)) \propto \prod_{i=1}^{\mathring{d}} \frac{\mathcal{I}\left(V_{ic;t-1} + B\left([\Psi_{ic;t}, \psi'_{ic;t}]'\right), \nu_{ic;t-1} + 1\right)}{\mathcal{I}(V_{ic;t-1}, \nu_{ic;t-1})}.$$

The normalization of the pf $f(c_t|d(t))$ provide the weights

$$w_{c;t} = \frac{f(c_t = c|d(t))}{\sum_{c_t\in c^*} f(c_t|d(t))} \approx \delta_{c_t,c}. \tag{13.18}$$

Use of this approximation in the formula (13.15) provides the updating rule of the statistics determining the approximate posterior pdf in the degenerate case $\tilde{c}^* = \emptyset$.

The case $\tilde{c}^* \neq \emptyset$ is treated separately for $c_t = \tilde{c}_t = c$ and $c_t = c \neq \tilde{c} = \tilde{c}_t$. The expression (13.15) reads

$$\text{For } \tilde{c}^* \neq \emptyset, \text{ and } c_t = \tilde{c}_t = c \tag{13.19}$$
$$f(\Theta, c_t = c, \tilde{c}_t = c|d(t)) \propto Di_\alpha(\kappa_{t-1})$$
$$\times \alpha_c^{2-\eta} \prod_{i=1}^{\mathring{d}} A(\Theta_{ic})^{\nu_{ic;t-1}+1} \exp\left\langle V_{ic;t-1} + B\left([\Psi_{ic;t}, \psi_{ic;t}]\right), C(\Theta_{ic})\right\rangle.$$

Thus, with the approximation (13.16), it holds

$$\text{For } \tilde{c}^* \neq \emptyset, \quad \text{and } c_t = \tilde{c}_t = c \tag{13.20}$$
$$f(c_t = c, \tilde{c}_t = c|d(t))$$
$$\propto \kappa_{c;t-1}^{2-\eta} \prod_{i=1}^{\mathring{d}} \frac{\mathcal{I}\left(V_{ic;t-1} + B\left(\left[\Psi_{ic;t}, \psi_{ic;t}\right], \nu_{ic;t-1} + 1\right)\right)}{\mathcal{I}\left(V_{ic;t-1}, \nu_{ic;t-1}\right)}.$$

The remaining case $c_t = c \neq \tilde{c} = \tilde{c}_t$ specializes the expression (13.15) to

$$\text{For } \tilde{c}^* \neq \emptyset, \text{ and } c \neq \tilde{c}, \; f(\Theta, c_t = c, \tilde{c}_t = \tilde{c}|d(t)) \propto Di_{\alpha}(\kappa_{t-1}) \tag{13.21}$$
$$\times \alpha_c \alpha_{\tilde{c}}^{1-\eta} \prod_{i=1}^{\mathring{\Psi}} A(\Theta_{ic})^{\nu_{ic;t-1}+1} \exp\left\langle V_{ic;t-1} + B\left(\left[\Psi_{ic;t}, \psi_{ic;t}\right]\right), C(\Theta_{ic})\right\rangle$$
$$\times \prod_{\tilde{i}=\mathring{d}+1}^{\mathring{\Psi}} A(\Theta_{\tilde{i}\tilde{c}})^{\nu_{\tilde{i}\tilde{c};t-1}-1} \exp\left\langle V_{\tilde{i}\tilde{c};t-1} - B\left(\left[\Psi_{\tilde{i}\tilde{c};t}, \psi_{\tilde{i}\tilde{c};t}\right]\right), C(\Theta_{\tilde{i}\tilde{c}})\right\rangle.$$

Thus, with the approximation (13.16), it holds

$$\text{For } \tilde{c}^* \neq \emptyset, \quad \text{and } c \neq \tilde{c}, \quad f(c_t = c, \tilde{c}_t = \tilde{c}|d(t)) \tag{13.22}$$
$$\propto \kappa_{c;t-1} \kappa_{\tilde{c};t-1}^{1-\eta} \prod_{i=1}^{\mathring{\Psi}} \frac{\mathcal{I}\left(V_{ic;t-1} + B\left(\left[\Psi_{ic;t}, \psi_{ic;t}\right]\right), \nu_{ic;t-1} + 1\right)}{\mathcal{I}\left(V_{ic;t-1}, \nu_{ic;t-1}\right)}$$
$$\times \prod_{\tilde{i}=\mathring{d}+1}^{\mathring{\Psi}} \frac{\mathcal{I}\left(V_{\tilde{i}\tilde{c};t-1} - B\left(\left[\Psi_{\tilde{i}\tilde{c};t}, \psi_{\tilde{i}\tilde{c};t}\right]\right), \nu_{\tilde{i}\tilde{c};t-1} - 1\right)}{\mathcal{I}\left(V_{\tilde{i}\tilde{c};t-1}, \nu_{\tilde{i}\tilde{c};t-1}\right)}.$$

The marginalization and normalization of the pf $f(c_t, \tilde{c}_t|d(t))$ provide the weights

$$w_{c;t} = \frac{\sum_{\tilde{c}_t \in \tilde{c}^*} f(c_t = c, \tilde{c}_t|d(t))}{\sum_{\tilde{c}_t \in \tilde{c}^*} \sum_{c_t \in c^*} f(c_t, \tilde{c}_t|d(t))} \approx \delta_{c_t,c} \tag{13.23}$$
$$\tilde{w}_{\tilde{c};t} = \frac{\sum_{c_t \in c^*} f(c_t, \tilde{c}_t = \tilde{c}|d(t))}{\sum_{\tilde{c}_t \in \tilde{c}^*} \sum_{c_t \in c^*} f(c_t, \tilde{c}_t|d(t))} \approx \delta_{\tilde{c}_t,\tilde{c}}.$$

To simplify the notation, let us define $\tilde{w}_c \equiv 0$, for $c \in c^* \setminus \tilde{c}^*$. This is the last preparatory step in completing the updating rule for all statistics involved. The resulting extended quasi-Bayes estimation preserves the functional form (6.3) with Dirichlet pdf $Di_{\alpha}(\kappa_t)$ assigned to the component weights and with conjugate pdfs $f(\Theta_{ic}|d(t)) \propto A^{\nu_{ic;t}}(\Theta_{ic}) \exp\langle V_{ic;t}, C(\Theta_{ic})\rangle$ for the parameterized factors from the exponential family.

For $\tilde{c}^* \neq \emptyset$, the updating $Di_{\alpha}(\kappa_t) \propto \prod_{c \in c^*} \alpha_c^{w_c} \prod_{\tilde{c} \in \tilde{c}^*} \alpha_{\tilde{c}}^{(1-\eta)\tilde{w}_{\tilde{c}}} Di_{\alpha}(\kappa_{t-1})$ is performed. It gives $\kappa_{c;t} \equiv \kappa_{c;t-1} + w_c + (1-\eta)\tilde{w}_c$.

The different updating is obtained for the factor outputs and for individual predicted entries of the state vector. Specifically, for each $c \in c^*$, For $i = 1, \ldots, \mathring{d}$, $f(\Theta_{ic}|d(t)) \propto [f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic})]^{w_c} f(\Theta_{ic}|d(t-1))$ giving

$$V_{ic;t} = V_{ic;t-1} + w_{c;t} B\left(\left[\Psi_{ic;t}, \psi'_{ic;t}\right]'\right), \; \nu_{ic;t} = \nu_{ic;t-1} + w_{c;t}.$$

For $i = \mathring{d}, \ldots, \mathring{\Psi}$, $f(\Theta_{ic}|d(t)) \propto [f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic})]^{w_c - \tilde{w}_c} f(\Theta_{ic}|d(t-1))$ giving

$$V_{ic;t} = V_{ic;t-1} + (w_{c;t} - \tilde{w}_{c;t}) B\left(\left[\Psi_{ic;t}, \psi'_{ic;t}\right]'\right)$$

$$\nu_{ic;t} = \nu_{ic;t-1} + w_{c;t} - \tilde{w}_{c;t}.$$

For $\tilde{c}^* = \emptyset$, no updating is performed on component weights $Di_\alpha(\kappa_t) = Di_\alpha(\kappa_{t-1})$. It gives $\kappa_{c;t} \equiv \kappa_{c;t-1}$.

The updating of the factors differs for the factor outputs and for individual predicted entries of the state vector. Specifically, for each $c \in c^*$

For $i = 1, \ldots, \mathring{d}$, $f(\Theta_{ic}|d(t)) \propto [f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic})]^{w_c} f(\Theta_{ic}|d(t))$ giving

$$V_{ic;t} = V_{ic;t-1} + w_{c;t} B\left(\left[\Psi_{ic;t}, \psi'_{ic;t}\right]'\right), \; \nu_{ic;t} = \nu_{ic;t-1} + w_{c;t}.$$

For $i = \mathring{d}, \ldots, \mathring{\Psi}$, $f(\Theta_{ic}|d(t)) = f(\Theta_{ic}|d(t-1))$ giving

$$V_{ic;t} = V_{ic;t-1}, \; \nu_{ic;t} = \nu_{ic;t-1}.$$

Let us summarize these results into a complete algorithm.

Algorithm 13.1 (Extended Quasi-Bayes estimation in EF)

Initial (offline) mode

- *Select the complete structure of the mixture and set time $t = 0$.*
- *Select prior pdfs $f(\Theta_{ic})$ of the individual factors in the conjugate form (3.13) with respect to the parameterized factors $f(\Psi_{ic;t}|\psi_{ic;t}, \Theta_{ic}, c)$.*
  In other words, specify the statistics $V_{ic;t}$, $\nu_{ic;t}$, $c \in c^*$, $i = 1, \ldots, \mathring{\Psi}$!
- *Select initial values $\kappa_{c;0} > 2$, say, about $0.1\mathring{t}/\mathring{c}$, describing the prior pdf of the component weights α.*

Sequential (online) mode,

```
For  $t = 1, \ldots, \ldots$
            Acquire the data $d_t$ and create the data vector $\Psi_t$
            Determine the set $\tilde{c}^*$, i.e., set $\tilde{c}^* = c^*$, $\mathring{\tilde{c}} = \mathring{c}$ and
  For  $c = 1, \ldots, \mathring{c}$
    For  $i = \mathring{d} + 1, \ldots, \mathring{\Psi}$
            Evaluate $\mathcal{I}_{-ic} \equiv \dfrac{\mathcal{I}\left(V_{ic;t-1} - B\left(\left[\Psi_{ic;t}, \psi'_{ic;t}\right]'\right), \nu_{ic;t-1} - 1\right)}{\mathcal{I}\left(V_{ic;t-1}, \nu_{ic;t-1}\right)}$
            Set $\tilde{c}^* = \tilde{c}^* \setminus \{c\}$, $\mathring{\tilde{c}} = \mathring{\tilde{c}} - 1$, and break i-cycle if $\mathcal{I}_{-ic} = \infty$
    end  of the cycle over i


    For  $i = 1, \ldots, \mathring{\Psi}$
```

```
            Evaluate $\mathcal{I}_{+ic} \equiv \frac{\mathcal{I}\left(V_{ic;t-1}+B\left(\left[\Psi_{ic;t},\psi'_{ic;t}\right]'\right),\nu_{ic;t-1}+1\right)}{\mathcal{I}\left(V_{ic;t-1},\nu_{ic;t-1}\right)}$
      end   of the cycle over i
    end   of the cycle over c

Treat the degenerate case if $\tilde{c}^* = \emptyset$, i.e., set $s = 0$ and
    For   $c = 1,\ldots,\mathring{c}$
            $f_{c;t} \equiv \prod_{i=1}^{\mathring{d}} \mathcal{I}_{+ic}$
            $s = s + f_{c;t}$
    end   of the cycle over c
    Update the statistics in the degenerate case, i.e.,
    For   $c = 1,\ldots,\mathring{c}$
            Set $w_{c;t} \equiv \frac{f_{c;t}}{s}$
      For   $i = 1,\ldots,\mathring{d}$
            Update $V_{ic;t} = V_{ic;t-1} + w_{c;t}B\left(\left[\Psi_{ic;t},\psi'_{ic;t}\right]'\right)$
                   $\nu_{ic;t} = \nu_{ic;t-1} + w_{c;t}$
      end   of the cycle over i
      For   $i = \mathring{d}+1,\ldots,\mathring{\Psi}$
            $V_{ic;t} = V_{ic;t-1}$
            $\nu_{ic;t} = \nu_{ic;t-1}$
      end   of the cycle over i
            $\kappa_{c;t} = \kappa_{c;t-1}$
    end   of the cycle over c
End the degenerate branch, i.e., go to the end of the cycle over time t.

Treat the nondegenerate case, $\tilde{c}^* \neq \emptyset$, i.e.
                  Set $s = 0$ and $\eta = \begin{cases} 0 & \text{if } \tilde{c}^* = c^* \\ \frac{2}{\mathring{\tilde{c}}} & \text{otherwise} \end{cases}$
        For   $c = 1,\ldots,\mathring{c}$
          For   $\tilde{c} = 1,\ldots,\mathring{\tilde{c}}$
                  if $\tilde{c} = c$,
                      $f_{cc} = \kappa_{c;t-1}^{2-\eta} \prod_{i=1}^{\mathring{d}} \mathcal{I}_{+ic}$
            else,
                $f_{c\tilde{c}} = \kappa_{c;t-1}\kappa_{\tilde{c};t-1}^{1-\eta} \prod_{i=1}^{\mathring{\Psi}} \mathcal{I}_{+ic} \prod_{i=\mathring{d}+1}^{\mathring{\Psi}} \mathcal{I}_{-i\tilde{c}}$
            end if-else,
            $s = s + f_{c\tilde{c}}$
        end    of the cycle over $\tilde{c}$
      end    of the cycle over $c$


    For    $c = 1, \ldots, \mathring{c}$
           $w_c = \sum_{\tilde{c} \in \tilde{c}^*} \frac{f_{c\tilde{c}}}{s}$
           if $c \in \tilde{c}^*$    $\tilde{w}_c = \sum_{\tilde{c} \in c^*} \frac{f_{\tilde{c}c}}{s}$
           else $\tilde{w}_c = 0$
    end    of the cycle over $c$
               Update statistics in the non-degenerate case, i.e.,
  For    $c = 1, \ldots, \mathring{c}$
    For    $i = 1, \ldots, \mathring{d}$
           $V_{ic;t} = V_{ic;t-1} + w_c B\left(\left[\Psi_{ic;t}, \psi'_{ic;t}\right]'\right),$
           $\nu_{ic;t} = \nu_{ic;t-1} + w_c$
    end    of the cycle over $i$
    For    $i = \mathring{d} + 1, \ldots, \mathring{\Psi}$
           $V_{ic;t} = V_{ic;t-1} + (w_c - \tilde{w}_c)\, B\left(\left[\Psi_{ic;t}, \psi'_{ic;t}\right]'\right),$
           $\nu_{ic;t} = \nu_{ic;t-1} + w_c - \tilde{w}_c$
    end    of the cycle over $i$
           $\kappa_{c;t} = \kappa_{c;t-1} + w_c + (1-\eta)\tilde{w}_c$
  end    of the cycle over $c$
   Evaluate the estimates of $\Theta$ using $f(\Theta|V_t, \nu_t, \kappa_t)$.
end    of the cycle over $t$
```

Remark(s) 13.4

1. *The derived algorithm allows negative weights to new data. In this way, it is closer to the reinforcement learning and as such it is expected to have better transient properties than the quasi-Bayes estimation applied to $f(\Psi|\Theta)$.*

2. *The algorithm may fail when some* $\kappa_{c;t-1}$ *falls below* $2/\mathring{\tilde{c}}$*. Then, this component has to be shifted to* $\tilde{c}^*$*, too.*
3. *Using Jensen inequality (2.14), it can be shown that the approximation of the unknown* δ_{c,c_t} *by its expectation provides a lower bound on the approximating function. Thus, the approximated upper bound on the correct pdf is not guaranteed to be upper bound anymore. This made us to take Proposition 13.4 as motivating one only.*
4. *Normalization of a dynamic factor predicting discrete quantity makes the treated pdf rational: thus the proposed solution can be directly used to this case vital for mixed mixtures.*

Problem 13.5 (Complete extension of approximate estimation)
Preliminary limited experience with the extended quasi-Bayes estimation is mixed: the expected significant improvement with respect to a plain application of the Bayes rule to data vectors has not been achieved. At the same time, the illogical frozen updating of factors in complement of $\tilde{c}^*$ *indicates that the proposed algorithm can be improved. We expect that the same methodology can be followed but the optimized upper bound has to be refined.*

Moreover, the extension of the quasi-Bayes estimation concerns only manipulation of weights of data. Thus, it is obvious that similar extensions are possible for all remaining versions of approximate estimation, Section 6.5.

14

Applications of the advisory system

The applications described here confirm usefulness of the developed theory and algorithms. At the same time, the applications proved to be significant for developing both the theory and algorithms. They helped us to discover errors, inconsistencies, drawbacks and bugs. They stimulated solutions of many subproblems that arose in attempts to implement the theoretical results. The process is still unfinished but the discussed interaction can be unanimously claimed to be useful.

Three rather different applications are outlined here. Advising at cold rolling mill is described in Section 14.1. Nuclear medicine application related to treatment of thyroid gland cancer is reflected in Section 14.2. Prediction problems related to urban traffic control are in Section 14.3. Practical conclusions are made in Section 14.4; see also [177].

The presented experience with *applications* we are working with reflects the research stage that corresponds with the last arrow in the verification chain $\boxed{\text{theory} \to \text{software} \to \text{offline experiments} \to \text{implementation.}}$

14.1 Operation of a rolling mill

Support of rolling mill operators in their decisions how to adjust key operating parameters of the machine was the main application output of the project that stimulated the results presented in this text. In this application, the fixed advisory system with periodically refreshed advisory mixture is used.

14.1.1 Problem description

Nowadays rolling mills are usually equipped with a control system enabling high quality production for properly adjusted manual settings. It is, however, difficult to find optimal settings for all possible working conditions and every type of the material being processed.

The rolling mill

A cold rolling mill is a complex machine used for reduction of metal strip thickness. The thickness is reduced between working rolls of the mill. It is affected by the rolling force in conjunction with input and output strip tensions being applied. For reversing mills, the strip is moved forward and backward in several passes until the desired thickness is achieved.

Two types of cold rolling mills were selected for full-scale experiments. They differ in the arrangement of rolls. The 4-high rolls arrangement consists of two working rolls each being supported by a single back-up roll. Data from two mills of this type were used for offline experiments.

A fine reversing cold rolling mill with 20-high arrangement of rolls, Fig. 14.1, was employed for both offline test and final online advisory system implementation. About 12 material types — alloys of iron, copper, nickel, zinc, etc. — are processed on the machine. Contact meters on both sides of the rolling mill provide measurements of strip thickness with $\pm 1\,\mu$m accuracy. For illustration, rolling forces are of order of 10^6 N, electric currents of drives about 10^2 A, strip tensions about 10^4 N and rolling speed in orders of 0.1–$1\,\mathrm{ms}^{-1}$.

The machine has been already equipped with the two-level control and information system including the adaptive thickness controller [178]. For properly adjusted settings, the controller is capable to keep deviations of the output strip thickness on the level of the measurement accuracy.

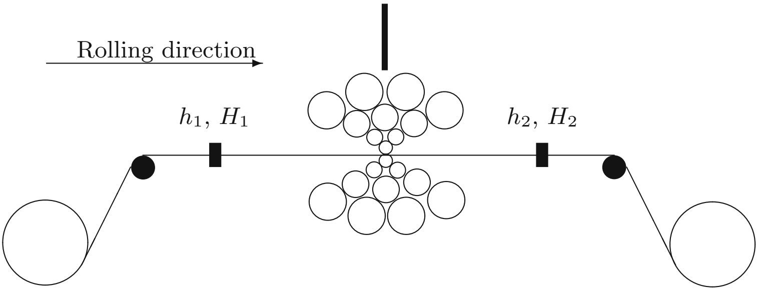


Fig. 14.1. Schematic diagram of the 20-high reversing rolling mill. For the selected rolling direction, H_1 and h_1 denote the input strip thickness and deviation from its nominal value, respectively; similarly, H_2 and h_2 for the output thickness.

Data collection

The modern distributed control system of rolling mills enables to archive every sample of process data. The collection is triggered by the strip movement.

Several tens of signals are collected on each particular mill. Thus, for each pass, $(\mathring{d}, \mathring{t})$ data matrix is produced where $\mathring{d}$ is the number of channels specific for a particular mill and the number of samples $\mathring{t}$ varies from 3 000 to 30 000 depending on a particular pass. Data samples are recorded each 4 cm of the strip, i.e., sampling period varies according to the strip speed. For instance, for the speed 1 m/s, the sampling period is 40 ms. The rolling mill experts selected $\mathring{d} = 10$ most important data channels as adequate for advising purposes. The specific selection depends on a particular rolling mill as described below.

14.1.2 Problem and its solution

The advisory system is to help adjust the main process quantities in order to maintain the best possible product quality. Moreover, for the 20-high mill, the goal is a better utilization of the rolling speed range in order to increase production potential without a loss of the high product quality. The practical implementation aspects described here relate to this problem.

Offline phase

The offline phase of the solution consists of data preprocessing, Section 6.2, estimation of the prior mixture information, Section 6.4, structure and parameter estimation, Sections 6.6 and 6.5, and the design of the advisory system; see Chapters 7 and 9.

Data

Data from two 4-high rolling mills were utilized for offline experiments. 58 quantities are recorded. The number of considered quantities was reduced to 10.

Most important selected data channels for mill 1 (rolling copper and its alloys) and mill 2 (rolling iron and its alloys) differ. The difference is caused by differences of control systems at respective mills. In contrast with the 20-high rolling mill, they contain explicitly rolling forces and hydraulic pressures on rolling cylinders.

Additional offline tests and a final online implementation concern the 20-high rolling mill. The available 44 data channels were also reduced to 10 relevant ones shown in Table 14.1.

For reversing rolling mill input/output pairs of signals refer to right/left sides of the mill, or vice versa according to the rolling direction. The front/rear pairs of signals refer to operator/drive sides of a rolling mill.

Particular sets of important data channels for the given rolling mills were selected by combining experience, correlation analysis, availability and reliability of particular data channels. The quantities measured with insufficient precision, heavily correlated or irrelevant for the addressed problem were omitted. Formally, they were treated as surplus data d_{o+} of the o-system.

Table 14.1. Offline and online processed quantities for the 20-high rolling mill

#	Symbol	Description	Typical range	Unit	Type	Filter
1	T_1	Input strip tension	$\langle 0; 50\rangle$	kN	u_o	Outliers 3
2	T_2	Output strip tension	$\langle 0; 50\rangle$	kN	u_o	Outliers 3
3	v_R	Ratio of input and output strip speeds	$\langle 0.5; 0.99\rangle$		u_o	Outliers 5
4	v_1	Input strip speed	1^{st} pass: $\langle 0.1; 0.3\rangle$ another: $\langle 0.5; 0.7\rangle$	m/s m/s	u_o	Outliers 5
5	v_2	output strip speed	1^{st} pass: $\langle 0.1; 0.3\rangle$ another: $\langle 0.5; 0.7\rangle$	m/s m/s	u_o	Outliers 5
6	I_1	Electric current of the input coiler	$\langle 50; 200\rangle$	A	Δ_{p+}	Smooth
7	I_2	Electric current of the output coiler	$\langle 50; 200\rangle$	A	Δ_{p+}	Smooth
8	I	Electric current of the mill main drive	$\langle 20; 180\rangle$	A	Δ_{p+}	Smooth
9	h_1	Deviation of input thickness from nominal value	$\langle -50; 50\rangle$	μm	Δ_{p+}	—
10	h_2	Deviation of output thickness from nominal value	$\langle -10; 10\rangle$	μm	Δ_o	Limits $\langle -10; 10\rangle$

Quality markers

The concept of quality markers is introduced in Section 5.1.4. The deviation of the output strip thickness from its nominal value is the key quality marker, which must be kept as low as possible, practically in units of μm. To evaluate the output quality, three statistical performance indicators (capability coefficients), usual in statistical process control [179], are taken as quality markers:

1. Statistical coefficient C_p defined

$$C_p = \frac{tol^+_{h_2} + |tol^-_{h_2}|}{6\,\sigma_{H_2}}\,, \tag{14.1}$$

   where H_2 denotes output thickness, h_2 is its deviation from the nominal value H_{2nom}, $tol^+_{h_2}$, $tol^-_{h_2}$ are boundaries of tolerance range of h_2 and $\bar{H}_2$, σ_{H_2} are mean and standard deviation of the output thickness H_2, respectively. The coefficient C_p describes variability of output thickness H_2 despite its magnitude, e.g., bias.
2. Statistical coefficient C_{pk} defined

$$C_{pk} = \frac{\min(\bar{h}_2 - tol^-_{h_2},\ tol^+_{h_2} - \bar{h}_2)}{3\,\sigma_{H_2}}\,, \tag{14.2}$$

where $\bar{h}_2$ denotes the mean of h_2. The coefficient C_{pk} describes the relative difference of the output thickness deviation h_2 from the mean of the tolerance range related to the variability of H_2.

3. The coefficient C_{per} representing the percentage of h_2 being within the tolerance range $\langle tol_{h_2}^{-},\, tol_{h_2}^{+} \rangle$.

The aim of the quality control is to keep values of the coefficients C_p, C_{pk} and C_{per} as high as possible.

These markers were used according to the customer's wish to compare the product quality before and after implementation of the advisory system.

Preprocessing of data files

Data files corresponding to particular passes were grouped according to the material type. Within each group, three subgroups were created for the first, and next even and odd passes through the mill. The corresponding data files within a subgroup were merged into a file having 5×10^5 samples on average. The subsequent mixture estimation provided acceptable results but it was time consuming. Therefore only representative shorter parts from particular data files were merged, preserving mutual ratios of data lengths.

Consequently, a typical data file within each subgroup contains roughly 30,000 samples of 10 selected data channels. As a result, n files have been available for each rolling mill, where $n = 3 \times n_{\mathrm{mat}}$ and n_{mat} is the number of material types processed on the given rolling mill. The selection and merging procedures were automated to allow repetition for new sets of process data.

The selected quantities and the way of their filtering for the 20-high rolling mill are in Table 14.1. There, the number given for outliers removal means a multiple of standard deviation as a half-width of an acceptance interval around the estimated mean; see 6.2.2 and 6.2.3.

Before further analysis, data in each channel were scaled to zero mean and unit variance.

Static mixture estimation

The static mixtures can be considered as a smooth approximation of multidimensional histograms representing the number of occurrences of data points in the 10-dimensional (10-D) space. As an input, the data gained during the operation of an experienced operator, yielding a high-quality product, are used. The adjustment of the actual working point into a location where data points occurred most often during the good rolling should lead to a high-quality product.

Preprocessed data files were used for an iterative estimation in the offline mode using the Mixtools function mixinit [180] that provides initialization by hierarchical factor splitting and implements the normal version of Algorithm 6.8. Typically, ten iterations per data file were used. The runtime of estimation for a single file with 30,000 records on a 1 GHz machine was about

10 minutes for default options, 160 minutes when batch quasi-Bayes estimation was applied. On average, nine components were estimated.

Comparison of the estimated mixture with the empirical pdf constructed from multidimensional histograms served as the main indicator of the estimation success. The estimated mixture seems to be quite fair for some selected 2-D projections, while for another projection with crumbled data clusters it is problematic. Improvements have been achieved by utilizing nondefault options for the identification procedure. Essentially, the changes have to reflect high differences of noise-level in processed data-record entries. At the end, an acceptable compromise was achieved considering different importance of data channels from the application point of view. A two-dimensional projection of a typical static mixture is shown in the Fig. 14.2. For the 20-high rolling mill, estimated static mixtures are used as inputs for online operation according to the type of material and pass number.

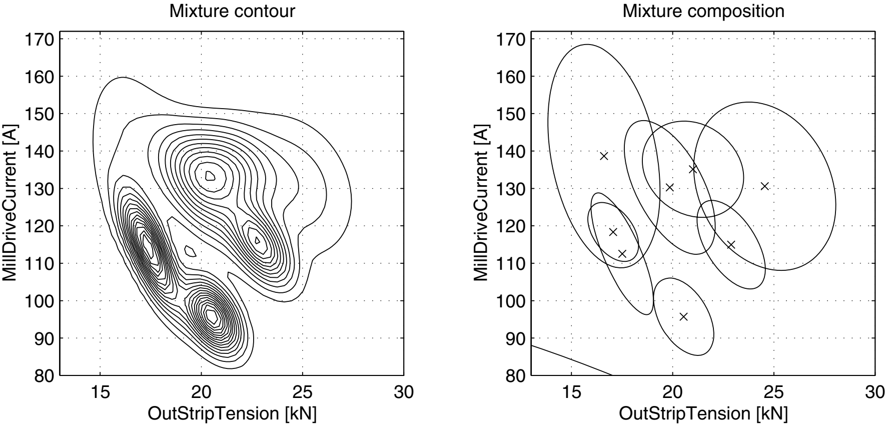


Fig. 14.2. Left plot: 2-D marginal projection of a static mixture for two selected data channels. Right plot: composition of the mixture from particular components.

Dynamic mixture estimation

Despite its clear interpretation, the static approach does not respect dependence of a current state of the system on its recent history, i.e., does not model the system evolution. The first-order auto-regression was found sufficient to describe estimated dynamic components. The order of the model was chosen with respect to sampling rate, human reaction times and closed-loop behavior. The estimates were periodically corrected (after each pass) and used in the fixed advisory system. The fully adaptive advisory system was not used yet.

Zero- and second-order models were also tried The quality of these alternatives was compared according to values of logarithmic v-likelihood and by inspecting prediction errors. Based on these tests, the first-order model was chosen as optimal. The same comparison was used for a finer tuning of optional parameters of the estimation procedure.

Typical processing time of the dynamic estimation on the same machine as above is approximately 5 minutes for default options and 120 minutes for the batch quasi-Bayes option with three detected components in average.

As an alternative to a dual static/dynamic estimation, a pseudo-dynamic approach has been investigated; cf. Section 6.4.9. The static mixture describing data vector, containing also one delayed data record was estimated. The processing combines intuitively appealing histogram interpretation while respecting the dynamic nature of dependencies among data records. Dynamic models are obtained by appropriate conditioning; see Proposition 7.2.

Simultaneous design

The *simultaneous design*, Section 5.4.6, was applied for advising at the 20-high rolling mill.

The user's ideal pdf ${}^{\lfloor U}f$ for the 20-high rolling mill was constructed as a static normal pdf that respects the treated regulated problem and aims of the control.

Specifically, the following correspondence to general notions applies

$$\begin{aligned} u_o &= (T_1, T_2, v_R, v_1, v_2) \ \ \text{recognizable actions} \\ \varDelta_o &= (h_2) \ \ \text{o-innovations} \\ \varDelta_{p+} &= (I_1, I_2, I, h_1) \ \ \text{surplus p-data.} \end{aligned}$$

The user's ideal pdf is expressed as (see Section 5.4.6)

$$\begin{aligned} {}^{\lfloor U}f(d(\mathring{t})) &= \prod_{t=1}^{\mathring{t}} {}^{\lfloor U}f(d_t|d(t-1)) \\ &= \prod_{t=1}^{\mathring{t}} {}^{\lfloor U}f(\varDelta_{o;t})\, {}^{\lfloor U}f(u_{o;t})\, {}^{\lfloor U}f(c_t)\, {}^{\lfloor I}f(\varDelta_{p+;t}|d(t-1)). \qquad (14.3) \end{aligned}$$

The user's ideal pf ${}^{\lfloor U}f(c_t)$ served only for exclusion of dangerous components. It was set as uniform one on nondangerous components as the user cannot make any statement about them.

The term ${}^{\lfloor I}f(\varDelta_{p+;t}|d(t-1))$ concerns quantities that cannot be influenced by the operator and therefore they are "left to their fate", i.e., no requirements are put on them.

The term ${}^{\lfloor U}f(\varDelta_{o;t})$ is the key factor in the true user's ideal. It expresses the main management aim. As the innovation h_2 is continuous, ${}^{\lfloor U}f(\varDelta_{o;t})$ has a form of normal pdf $\mathcal{N}_{h_2}(0, \sigma_{h_2})$. The zero expectation expresses the wish to get zero deviations of the output thickness from the technologically prescribed value. The value of σ_{h_2} expresses acceptable spread of these deviations. It is related to the desired coefficient C_p (14.1). For instance, if C_p is required to be 4/3, then $\sigma_{h_2} \equiv \sigma_{H_2}$ must be 8-times contained in the interval $\langle tol_{h_2}^-, tol_{h_2}^+ \rangle$. Then $\sigma_{h_2} = (tol_{h_2}^+ + |tol_{h_2}^-|)/8$.

The pdf ${}^{\lfloor U}f(u_{o;t})$ concerning recognizable actions was chosen as a product of one-dimensional normal pdfs. Their means and variances express the technologically desirable ranges of these quantities; see Table 14.1. Their choice was made similarly as for the output thickness.

The parameters of time-invariant pdfs ${}^{\lfloor U}f(c_t)$, ${}^{\lfloor U}f(\mathit{\Delta}_{o;t})$ and ${}^{\lfloor U}f(u_{o;t})$ together with a list of corresponding data channels are passed to the design procedures that generate the optimized ideal mixture ${}^{\lfloor I}f(d_t|d(t-1))$.

The advisory mixture is obtained by projection of ${}^{\lfloor I}f$ to the space $(u_o, \mathit{\Delta}_o)$, i.e., the advisory mixture is

$$ {}^{\lfloor I}f(u_{o;t}, \mathit{\Delta}_{o;t}|\mathit{\Delta}_{p+;t}, d(t-1)). \tag{14.4}$$

The overall procedure of simultaneous design is described in Algorithm 7.9.

Online phase

The online phase was fully implemented for the 20-high rolling mill.

Generating advices

Based on *offline* results of mixture estimations, the advisory part of the system uses the newest available data for generating recommendations to operators. The shapes and weights of advisory mixture components are recalculated for each new data record and assign the probability of the best possible location of the working point. Marginal pdfs of the advisory mixture (14.4) are evaluated for all o-data channels $d_{i;t}$, $i = 1, \ldots, \mathring{d}_o$. The recommended values correspond with the modes of these one-dimensional projections.

The limited extent of the o-data allowed us to avoid a presentation problem.

A heuristic solution of signaling problem is temporarily implemented. A simple distance between the actual working point and its best recommended location is evaluated permanently and set up as an “alarm” when it is found that the setting should be changed.

Mixture updating and adaptivity

The adaptive advisory system updates online the mixture estimate, repeats the design with each new data record and uses the updated advisory mixture. Historical (slower machines) and practical (separation of advising and updating) reasons make us use a fixed advisory system. To allow learning, the estimated mixtures are updated and redesigned online by a parallel process during each pass, for each material and each pass subgroup (first, odd and even; see paragraph concerning data preprocessing). For the new pass of the given subgroup, the updated advisory mixtures are used.

14.1.3 Implementation

The advisory system has been implemented into the distributed control system by Compureg Plzeň, s.r.o. The implementation was realized in the form of a dedicated server and a visualization node equipped with a special graphical user interface (*GUI*). A specific effort was made to enable the bulk of computation to be executed either under MS Windows or Linux operating systems.

The server

The server executes the following two main applications.

Adviser loads a proper mixture file at the beginning of the pass and then, when appropriate, uses new data record for evaluating the advisory mixture and its marginal projections to be displayed for operators;

Updater uses data samples corresponding to a good rolling for updating the mixture. Thus, the system can refine the estimates and learn new possible locations of "good" working points.

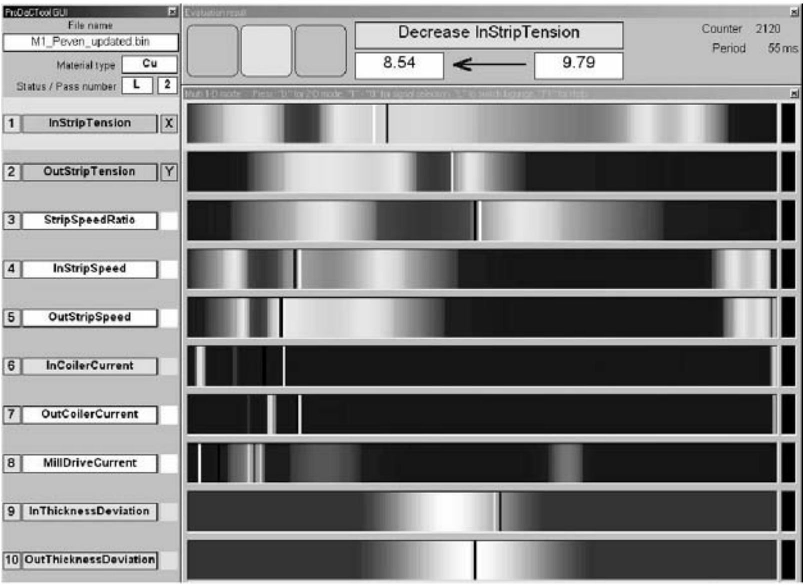
Both applications use a multithreading technique that allows us to optimize the system performance. Programs are coded in ANSI C, and the Tcl language/interpreter is used for interface to a human supervisor. This together with the appropriate version of the Mixtools library [180] makes porting to another computer platform easy.

Graphical user interface

A visualization node was introduced for the control system on which the dedicated GUI for operators is executed. The application can be used in a one-dimensional (1-D; Fig. 14.3) and two-dimensional (2-D; Fig. 14.4) modes, which project selected quantities of (14.4), and compares them with their actual values. Values of the projected pdfs are transformed into a color scale and actual working points are marked by a black line. Its basic properties and appearance can be easily modified through a configuration file. Whatever mode is selected, the GUI displays the overall status of the system in the form of traffic lights where "green" status means "no recommendations". The quantity to be changed primarily is emphasized, together with its recommended value. For 2-D mode, the operator can select data channels whose marginal mixture projection should be plotted. OpenGL functions are used for smooth online animations.

14.1.4 Results

Offline processing of data from two 4-high rolling mills was made mainly to confirm functionality of developed algorithms and to verify generality of the



File name: M1_Peven_updated.bin

Material type: Cu

Status / Pass number: L 2

Decrease InStripTension

8.54 ← 9.79

Counter 2120

Period 55 ms

Fig. 14.3. Screenshot of the Graphical User Interface for 1-D mode.

approach. Qualitatively, the results were satisfactory. As the advisory loop was not closed, no overall quantitative evaluation was made.

The case of the 20-high rolling mill was brought to online use enabling us to compare statistically the production before and after the implementation.

The tolerance range was set to $tol^{+}_{h_2} = 10$ μm, $tol^{-}_{h_2} = -10$ μm for all cases. Data collected within 20 months prior to implementation were taken as the basis for comparisons. More than two months of the new system operation were taken into account for the final evaluation. Comparisons were made for several sorts of materials, which were processed in both the periods.

Numbers for comparisons were obtained by SQL queries to production databases. The following values were evaluated:

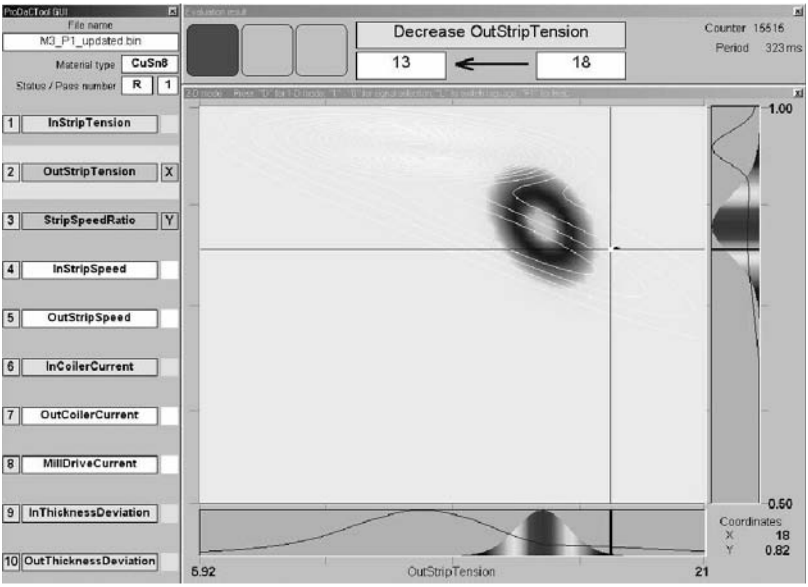
- Averages of the rolling speed,
- Averages of coefficients C_p, C_{pk} and C_{per}.

Quality markers for results evaluation are represented by relative differences of these values (in percent) between the production before and after installation of the advisory system. For example, $\varDelta\bar{C}_p$ is defined as

$$\varDelta\bar{C}_p = \frac{(\bar{C}_p)_{\text{after}} - (\bar{C}_p)_{\text{before}}}{(\bar{C}_p)_{\text{before}}} \cdot 100\%.$$

The values $\varDelta\bar{C}_{pk}$, $\varDelta\bar{C}_{per}$ and $\varDelta\bar{v}_2$ are defined similarly.

Six common material types, marked by capital letters, were tested in both periods. Percentage improvements of monitored values are summarized in Table 14.2. First pass and further passes are distinguished in comparisons since


ProDaCTool GUI

File name: M3_P1_updated.bin

Material type: CuSn8

Status / Pass number: R 1

1 InStripTension

2 OutStripTension X

3 StripSpeedRatio Y

4 InStripSpeed

5 OutStripSpeed

6 InCoilerCurrent

7 OutCoilerCurrent

8 MillDriveCurrent

9 InThicknessDeviation

10 OutThicknessDeviation

Evaluation result

Decrease OutStripTension

13 ← 18

Counter 15516

Period 323ms

Coordinates

X 18

Y 0.82

Fig. 14.4. Screenshot of the Graphical User Interface for 2-D mode.

conditions for both cases differ. The C_{per} coefficient was evaluated for last passes only.

Table 14.2. Percentage improvements of quality markers for the 20-high mill.

Material		A	B	C	D	E	F	Weighted
Quality Marker								mean
Pass 1	$\Delta\bar{C}_p$	0.98	0.50	-0.69	-0.14	0.66	0.72	0.45
	$\Delta\bar{C}_{pk}$	43.36	15.45	28.57	16.13	39.39	0.00	4.84
	$\Delta\bar{v}_2$	13.33	29.41	58.33	17.65	47.37	10.53	17.08
Pass 2+	$\Delta\bar{C}_p$	31.87	24.68	83.33	-2.94	89.29	-1.08	9.51
	$\Delta\bar{C}_{pk}$	25.30	6.25	105.41	5.00	85.45	1.43	12.36
	$\Delta\bar{v}_2$	16.67	42.86	62.96	16.67	26.83	29.41	33.41
Last pass	$\Delta\bar{C}_{per}$	42.76	16.22	33.33	25.81	45.95	0.68	5.50

14.1.5 Conclusions

The system turned out to stabilize the highest product quality and delimit secure machine settings for the whole variety of working conditions. Obtained results, Table 14.2, validate prior expectations that the rolling speed can be increased by up to 40% (depending on the pass number and material type) while preserving the highest product quality.

The results are encouraging in spite of the fact that the improvements are limited by the amount of information contained in the available data — relatively small and fine mill and low rolling speed. At the same time, we are aware that utilization for high-speed steel mills will undoubtedly need more computing power and probably a further development as well.

14.2 Treatment of thyroid gland cancer

Treatment of thyroid gland tumor using ^{131}I is another important application area tried.

The task is to help physicians to decide on administration of an individual amount of radioactive ^{131}I for a specific patient using information hidden in retrospective data. The decisions in this application were supported by a fixed advisory system. Its advisory mixture can be modified periodically after collecting enough new patient data records.

14.2.1 Problem description

Treatment of thyroid gland carcinoma is a complex process involving endocrinology, histology, surgery, radiology and other branches of medicine. The current section pertains to support of the radiological treatment [181].

Thyroid gland and iodine

The thyroid gland accumulates anorganic iodine from blood and incorporates it into thyroid hormones. A stable isotope of iodine is $^{127}_{53}$I, simply denoted as ^{127}I. Radioactive isotope is chemically identical to the stable one but it has a different number of neutrons in the nucleus. Such a nucleus is unstable, i.e., it can spontaneously change and emit one or more ionizing particles. The nucleus of ^{131}I produces ionizing particles of β- and γ-radiation due to nuclear decays. Thyroid accumulating ^{131}I becomes a source of ionizing radiation. The thyroid tissue is almost unaffected by γ-particles (high-energy photons). The majority of them are not absorbed in the tissue and they can be detected outside the body. Almost all β-particles (electrons) are absorbed in the tissue and their energy is passed to the tissue. The half-life ${}^{\lfloor \mathrm{p}}T$ of ^{131}I — the time in which one half of the radioactive nuclei takes a change — is approximately 8 days.

Radiotherapy of thyroid cancer

A thyroid tumor is usually removed by surgery. What remains or reappears is destroyed by radiotherapy. The radiotherapy has a diagnostic and a therapeutic stage.

In the diagnostic stage, ^{131}I in the form of sodium or potassium iodide of a low diagnostic (tracer) activity is administered orally to a patient. This activity is about $70\,\mathrm{MBq} = 70\times 10^6$ changes per second in mean. Then, several measurements are performed, detecting γ-radiation from the accumulating tissues, and some quantities describing the kinetics of ^{131}I are evaluated. This information is used for the decision about the next therapeutic stage.

In the therapeutic stage, the patient is administered the solution with ^{131}I of a high activity in the range 3–10 GBq. The therapeutic dose, representing the absorbed energy of β-radiation caused by ^{131}I accumulated in the thyroid, destroys the thyroid tissue.

About 3–6 months later, the diagnostic stage is repeated to examine the therapy result.

Notations, terms and quantities

The notation adopted in this section follows the conventions introduced in this book. Hence, it differs from a notation usual in other thyroid-related papers.

Continuous time will be denoted as ρ, discrete time index as t. Discrete (indexed) time is denoted as ρ_t. A value of a quantity X in a time instant t is denoted as $X_t \equiv X_{\rho_t}$. A quantity X as a function of time ρ is denoted as X_ρ.

After the administration, ^{131}I is distributed over the body and accumulated in thyroid gland and, possibly, other tissues, called *lesions*, and partially over a patient's whole body. Lesions are usually metastases of the primary tumor. Here, we include thyroid gland among them.

The thyroid activity rapidly increases, reaches its maximum within hours and then slowly decreases within days or weeks. The activity of lesions and the whole body can be evaluated and sequences $\{A_t, \rho_t\}$ of activity A_t sampled in time ρ_t. However, activity is not directly measurable and just a few, rather noisy, indirect measurements are available. A nontrivial, theoretical and Bayesian algorithmic solution of its estimation is summarized in [182].

Administered activity A_0 is activity of ^{131}I salt solution drunk by the patient. The diagnostic administration is denoted as $^{\lfloor \mathrm{d}}A_0$ and the therapeutic one as $^{\lfloor \mathrm{t}}A_0$.

Dosimetric data are sequences of activity in time $\{(A_t, \rho_t)\}$ concerning the given lesion or the whole body.

The *effective half-life* $^{\lfloor \mathrm{ef}}T$ is a parameter of a mono-exponential model describing decrease of activity of a lesion in time. The mono-exponential model uses the maximum A_1 of the activity course in the time ρ_1 and predicts the activity A_ρ in time ρ as follows

$$A_\rho = A_1 \; \exp\left(- \frac{\rho - \rho_1}{^{\lfloor \mathrm{ef}}T} \ln 2 \right) . \tag{14.5}$$

The model describes the activity course only for $\rho > \rho_1$. Effective half-life $^{\lfloor \mathrm{ef}}T$ combines physical and biological mechanisms of activity decrease and quantifies its rate.

Residence time τ is a time for which all the administered activity hypothetically "resided" in a selected tissue to cause the same irradiation effect, i.e.,

$$\tau = \frac{1}{A_0} \int_0^{+\infty} A_\rho \, \mathrm{d}\rho. \tag{14.6}$$

The residence time is proportional to radiation dose — energy per mass unit — absorbed in the tissue. Values concerning the diagnostic or therapeutic administration are denoted ${}^{\lfloor \mathrm{d}}\tau$ or ${}^{\lfloor \mathrm{t}}\tau$, respectively.

Relative activity (uptake) ${}^{\lfloor \mathrm{r}}A$ is a percentage of the absolute lesion activity A from the administered activity corrected to the physical decay, i.e.,

$${}^{\lfloor \mathrm{r}}A = \frac{A}{A_0 \ \exp\left(-\frac{\rho}{{}^{\lfloor \mathrm{p}}T} \ln 2\right)} \ 100\,\%. \tag{14.7}$$

Time ρ is set to zero at the administration moment. Uptake informs about the ability of the lesion to accumulate ^{131}I. Usual values of the maximum thyroid uptake are up to 5 %.

Excretion of the activity ${}^{\lfloor \mathrm{r}}E$ is a relative activity diluted from the body within some time interval. Typically, the intervals used for its description are 0–24 hours, ${}^{\lfloor \mathrm{r}}E_2$, and 24–48 hours, ${}^{\lfloor \mathrm{r}}E_3$, after the administration. These values are estimated from measurements of the whole-body activity and carry additional information about ^{131}I kinetics in the organism. They are evaluated after the diagnostic administration only.

14.2.2 Problem and its solution

The aim is to recommend such a value of administered activity ${}^{\lfloor \mathrm{t}}A_0$ for therapy so that 3–6 months later the maximum uptake ${}^{\lfloor \mathrm{rmn}}A$ is less than 0.18 %. The value ${}^{\lfloor \mathrm{t}}A_0$ should be chosen as low as possible.

The solution consists of offline estimation of a static mixture on the historical patient data and online generation of a recommendation on ${}^{\lfloor \mathrm{t}}A_0$.

Therapeutic activity ${}^{\lfloor \mathrm{t}}A_0$ is a key quantity influencing the treatment success, because the dose absorbed in the thyroid tissue is proportional to it. A decision on the optimal value of ${}^{\lfloor \mathrm{t}}A_0$ is the responsibility of the physician. The value of ${}^{\lfloor \mathrm{t}}A_0$ must be high enough so that the affected tissue is destroyed. This is quantified by the negligible accumulation ability after the therapy. However, the administered activity must be low enough to meet prescribed irradiation limits and to minimize secondary radiation risks.

Offline phase

Data

The data used for the mixture analysis were collected from 1992 on more than 6500 patients. After reorganizing and preprocessing, the final data file contains about 1200 complete records of 12 biophysical quantities directly measured or estimated using Bayesian methodology [183, 184].

The quantities chosen for each record are described in the Table 14.3.

Table 14.3. Quantities used in the advisory system for nuclear medicine

#	Symbol	Description	Typical range	Unit	Type
1	g	Patient's gender: 0=female, 1=male	$\{0, 1\}$		$\mathit{\Delta}_{p+}$
2	a	Patient's age	$\langle 7; 85\rangle$	year	$\mathit{\Delta}_{p+}$
3	$\lfloor^{\mathrm{d}}A_0$	Diagnostic administered activity before therapy	$\langle 10; 130\rangle$	MBq	$\mathit{\Delta}_{p+}$
4	$\lfloor^{\mathrm{l}}n$	Number of lesions	$\langle 1; 5\rangle$		$\mathit{\Delta}_{p+}$
5	$\lfloor^{\mathrm{efd}}T$	Diagnostic thyroidal effective half-life	$\langle 0.1; 8\rangle$	day	$\mathit{\Delta}_{p+}$
6	$\lfloor^{\mathrm{d}}\tau$	Diagnostic thyroidal residence time	$\langle 0.000\,4; 1\rangle$	day	$\mathit{\Delta}_{p+}$
7	$\lfloor^{\mathrm{rmd}}A$	Maximum diagnostic thyroid uptake before therapy	$\langle 0.01; 25\rangle$	%	$\mathit{\Delta}_{p+}$
8	$\lfloor^{\mathrm{r}}E_2$	Excretions 0–24 hours	$\langle 15; 88\rangle$	%	$\mathit{\Delta}_{p+}$
9	$\lfloor^{\mathrm{r}}E_3$	Excretions 24–48 hours	$\langle 0.8; 34\rangle$	%	$\mathit{\Delta}_{p+}$
10	$\lfloor^{\mathrm{t}}n$	Number of previous therapies	$\langle 0; 5\rangle$		$\mathit{\Delta}_{p+}$
11	$\lfloor^{\mathrm{t}}A_0$	Therapeutic administered activity	$\langle 1\,700; 8\,300\rangle$	MBq	u_o
12	$\lfloor^{\mathrm{rmn}}A$	Maximum diagnostic thyroid uptake after therapy	$\langle 0.01; 1.22\rangle$	%	$\mathit{\Delta}_o$

One data record contains information about one patient concerning his single therapeutic and two diagnostic administrations: 2–3 days before and 3–6 months after the therapy. Data records of different patients are undoubtedly mutually independent. One patient can have more records if more therapies have been absolved. Although an individual patient's history is causal, time interval between two therapies is in the order of months to years. This allows us to assume weak mutual dependence of records even for one patient. Therefore a static model of the whole patient population provides "natural" data description. With the "delayed" diagnostic uptake after therapy in the record, the model is pseudo-dynamic.

Quality marker

The less thyroid tissue remained after therapy, the less activity it accumulates. The therapy is taken as successful if the accumulation activity is negligible.

The marker of the therapy success is a maximum relative activity $\lfloor^{\mathrm{rmn}}A$ (uptake) reached by the thyroid gland in the diagnostic administration of ^{131}I after the therapy. Ideally, after a successful therapy, ${}^{\lfloor\mathrm{rmn}}A = 0$. Practically, ${}^{\lfloor\mathrm{rmn}}A < 0.18\,\%$ of the administered activity represents a successful therapy.

Preprocessing of the data files

Records with missing data were excluded. For the complete preserved records, the data were tested for physically and medically meaningful ranges to avoid cases with mistyped values or outliers of another origin. Due to this operation, other records of the original number were lost.

Because of numerical reasons, each quantity was scaled to zero mean and unit variance. All the estimations and computations were performed with the scaled data. The results are presented in their original scales.

Static mixture estimation

The addressed problem and its technical conditions lead to the following correspondence of data entries to the general notions presented in Chapter 5

$$\begin{aligned} u_o &= {}^{\lfloor\mathrm{t}}A_0 \ \ \text{recognizable action} \\ \mathit{\Delta}_o &= {}^{\lfloor\mathrm{rmn}}A \ \text{o-innovation} \\ \mathit{\Delta}_{p+} &= (g, a,\ {}^{\lfloor\mathrm{d}}A_0,\ {}^{\lfloor\mathrm{l}}n,\ {}^{\lfloor\mathrm{efd}}T,\ {}^{\lfloor\mathrm{d}}\tau,\ {}^{\lfloor\mathrm{rmd}}A,\ {}^{\lfloor\mathrm{r}}E_2,\ {}^{\lfloor\mathrm{r}}E_3,\ {}^{\lfloor\mathrm{t}}n) \ \ \text{surplus p-data.} \end{aligned} \tag{14.8}$$

The meaning of respective quantities is summarized in Table 14.3.

The 12-dimensional data space $(\,{}^{\lfloor\mathrm{t}}A_0,\ {}^{\lfloor\mathrm{rmn}}A, \mathit{\Delta}_{p+})$ was examined for the occurrence of clusters. The parameters and structure of the normal static mixture $f(\,{}^{\lfloor\mathrm{t}}A_0,\ {}^{\lfloor\mathrm{rmn}}A, \mathit{\Delta}_{p+}|\mathit{\Theta})$ (see 5.9) were estimated. Three processing ways were tried.

BMTB algorithm: An older, less elaborated variant of the sandwich algorithm providing Bayesian extension of the mean-tracking algorithm [154] was tried; see Chapter 12. It is computationally very cheap, but, unfortunately, it failed in the sparsely populated data space available.

AutoClass software: This freely available software [42] detects clusters using the EM algorithm (see Chapter 8) with random starts. Repetitive searches increase the probability of finding a better description of clusters but the computational cost increases, too. Over 700 000 random starts have been tried to get stable results. Low dimensions of the data set have allowed it. The best result was used as partial prior information for processing corresponding fully to the theory and algorithm described in Chapter 8.

Hierarchical splitting, quasi-Bayes algorithm and AutoClass: An older version of initialization [44] and quasi-Bayes algorithm as implemented in Mixtools [180] were used. The results were enhanced by exploiting partially results obtained by AutoClass.

Simultaneous design

The simultaneous design, described in Section 5.4.6 and elaborated in Proposition 9.15, was applied. It means that the component weights were interpreted more as elements of the mixture approximating the objective distribution of data records than objective values reflecting characteristics of the patient population.

As the parameters Θ are assumed to be known in the design, they will be omitted in the notation below. Using the notation introduced in (14.8), the user ideal pdf ${}^{\lfloor U}f(d(\mathring{t}))$ was defined in the same way like in (14.3). Uniform pf ${}^{\lfloor U}f(c_t)$ was chosen. The true user's ideal for the recognizable action ${}^{\lfloor U}f(u_{o;t})$ was set to $\mathcal{N}_{{}^{\lfloor \mathrm{t}}A_0}(3\,500, 800)$. This pdf has practical support within the usual range of administered activities. The true user ideal pdf of $\Delta_{o;t}$ was chosen conditioned by the value of $u_{o;t}$ in the following way:

$$
{}^{\lfloor U}f(\Delta_{o;t}|u_{o;t}) = \mathcal{N}_{{}^{\lfloor \mathrm{rmn}}A}(g({}^{\lfloor \mathrm{t}}A_0), 0.03),
$$

where $g({}^{\lfloor \mathrm{t}}A_0) = a\,{}^{\lfloor \mathrm{t}}A_0 + b$, $a < 0$. The values of a and b were chosen so that the range of ${}^{\lfloor \mathrm{t}}A_0$ shown in the Table 14.3 is linearly mapped to the required range of ${}^{\lfloor \mathrm{rmn}}A \in (0; 0.18)$ with a negative slope. Using this formulation, the intention to minimize therapeutic administered activity ${}^{\lfloor \mathrm{t}}A_0$ was expressed.

After the design made by Algorithm 7.9, the constructed ideal pdf was projected into $d_o^* \equiv {}^{\lfloor \mathrm{rmn}}A,\ {}^{\lfloor \mathrm{t}}A_0{}^*$ giving the advisory mixture ${}^{\lfloor I}f(u_o, \Delta_o|\Delta_{p+})$.

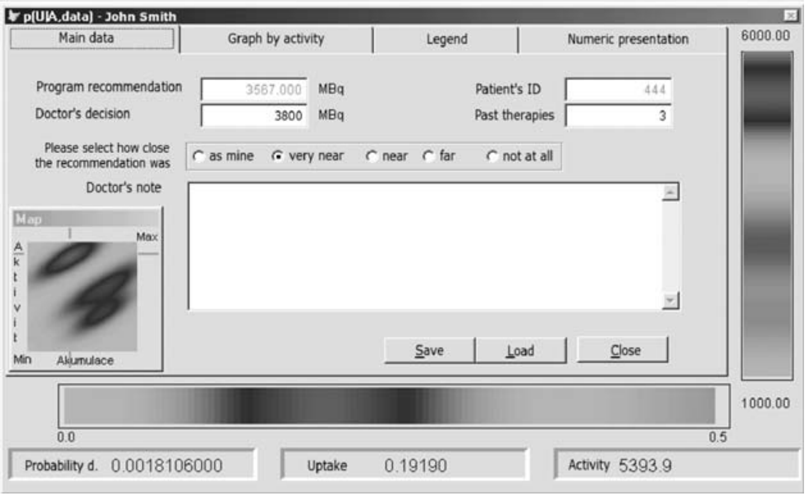
Online phase

The statistics of the ideal pdf ${}^{\lfloor I}f({}^{\lfloor \mathrm{rmn}}A,\ {}^{\lfloor \mathrm{t}}A_0|\Delta_{p+})$ conditioned by the actual patient's data Δ_{p+} were evaluated. The maximum of the pdf indicates the recommended value of ${}^{\lfloor \mathrm{t}}A_0$ with predicted value of corresponding ${}^{\lfloor \mathrm{rmn}}A$. The example with a two-dimensional pdf map and one-dimensional marginal pdfs of both quantities is shown in the Figure 14.5.

14.2.3 Implementation

The initial versions of advisory modules were programmed in C++. An original API capable of transferring MATLAB mex-functions into stand-alone applications was used to take advantage of ready-made Mixtools algorithms. The application accepting data Δ_{p+} and computing a grid of the corresponding ${}^{\lfloor I}f({}^{\lfloor \mathrm{rmn}}A,\ {}^{\lfloor \mathrm{t}}A_0|\Delta_{p+})$ was integrated into the system *Iodine III* [185]. This system, written in MS Visual FoxPro, cares about data management and evaluation of a range of Bayesian estimates of various biophysical quantities. The presented GUI is a part of this system.

Graphical presentations of advices are shown in Figure 14.5. The interactive color pdf map allows physicians to try, for the given patient, various planned administrations by clicking on the map and examining the consequences, i.e., to predict values of ${}^{\lfloor \mathrm{rmn}}A$. The GUI allows to classify the advices for testing purposes, save the data, restore data of previously processed patients and adjust GUI.



p(U|A,data) - John Smith

Program recommendation 3567.000 MBq

Doctor's decision 3800 MBq

Patient's ID 444

Past therapies 3

Please select how close the recommendation was ○ as mine ● very near ○ near ○ far ○ not at all

Doctor's note

Probability d. 0.0018106000

Uptake 0.19190

Activity 5393.9

Fig. 14.5. Screenshot of *Iodine III* with interactive probability density map on the left and bars with marginal pdfs at the bottom and on the right

14.2.4 Results

With the cooperation of physicians from the Clinic of Nuclear Medicine and Endocrinology, Motol Hospital, Prague, the advices were tested for 101 patients. The recommended values of ${}^{\lfloor \mathrm{t}}A_0$ were compared to those decided by the physicians.

It was observed that the best coincidence was reached in the cases when a patient first visited the clinic for a radio-destruction of thyroid remnants after thyroid removal during surgery. These patients usually suffer from a primary tumor, i.e., only the thyroid is affected and no metastases are usually developed.

In these cases, the treatment procedure usually does not depend on other quantities like histology of the tumor, biochemical analyzes, etc. The differences between the recommended and decided value of ${}^{\lfloor \mathrm{t}}A_0$ in these cases are below 15 %. These cases represent 46 % of the tested patients. A difference below 10 % was observed in 31 % of the tested patients.

In cases of therapy repetition due to the tumor reappearance, the coincidence was much lower. A difference above 50 % was observed in 15 % of patients. Generally, the advisory system recommends lower values than the physicians. Decision in this category depends also on other examinations than the used dosimetric ones. Unfortunately, these vital data have not been available in the clinic in a form suitable for automatic processing.

14.2.5 Conclusions

The described advisory system is an original application of the mixture theory in this field. However, the task is specific to a small amount of the data available; therefore the results of the learning stage depend heavily on prior information. The AutoClass initialization has improved performance of the tested version of the Mixtools. Even with the prior obtained on the basis of 700 000 random starts, the model designed for the advising has acceptable physical interpretation only in about one-half of the cases. The advices are sensitive to mixture estimation. The key assumption, that the requested information $\underline{\text{is}}$ contained in the available data set, seems to be violated by using only a subset of data that is only accessible in a systematic way. The *GIGO principle* (garbage in, garbage out) is manifested here.

Anyway, the subset of cases relevant to the data set and the way of its processing, 46 % in this application, exhibits satisfactory results. Moreover, the quantitatively demonstrated conclusion that dosimetric data are insufficient for controlling the therapy has initiated the reorganization of data acquisition. Consequently, the advising based also on additional informative data will be possible in the near future. The nature of the considered data implies that the use of mixed data mixtures, Chapter 13, will be inevitable.

14.3 Prediction of traffic quantities

With an increasing number of cars, various significant transportation problems emerge, especially those concerning urban traffic, [186], [187]. Feedback control via existing traffic lights is a viable way to decrease them.

14.3.1 Problem description

The control design for complex traffic systems has to be build by solving partial, mutually harmonized, subtasks. Among them, the prediction of the traffic state is of primary importance. Here, we demonstrate appropriateness of predicting traffic quantities based on a mixture model.

Prediction of urban traffic state

Solution of urban transportation problems via reconstruction of the street network is expensive and very limited as it has to respect the existing urban conditions. Thus, the capacity of the network has to be efficiently exploited. This makes feedback traffic control utilizing the available traffic lights extremely important. Its design has to be able to predict a future traffic state for a given regime of traffic lights. Only with such a model, the regime can be varied to increase the permeability of city crossroads network. For it, mixture modelling seems to be suitable. Changes of daily and seasonal traffic indicate it clearly.

Notations, terms and quantities

Controlled networks are split into microregions. They are logically self-contained transportation areas of several cross-roads with their adjoining roads. Their modelling and feedback control exploit data measured by detectors based on inductive electric coils placed under the road surface. The presence of a huge metallic object above the coil changes its magnetic properties and thus individual cars are detected. Each detector signals a presence or absence of a car above it. From this signal, basic transportation quantities are evaluated:

- *Occupancy* o, which is defined as the portion (in %) of the time when the inspected place is occupied by cars.
- *Intensity* q expressing the number of cars per hour.
- *Density* ρ, which is defined as the number of cars per kilometer of the traffic flow. Specifying an average length of cars, it is computed as a ratio of occupancy and average car length.

Intensity and density describe the traffic state at the detector position.

14.3.2 Problem and its solution

Crossroads in cities are controlled by setting a proper green proportion and cycle length of the signal lights. Mostly, these signals are set according to a working plan whose changes during the day are given by a fixed schedule. Automatic feedback in selection of these plans can improve quality of the traffic significantly.

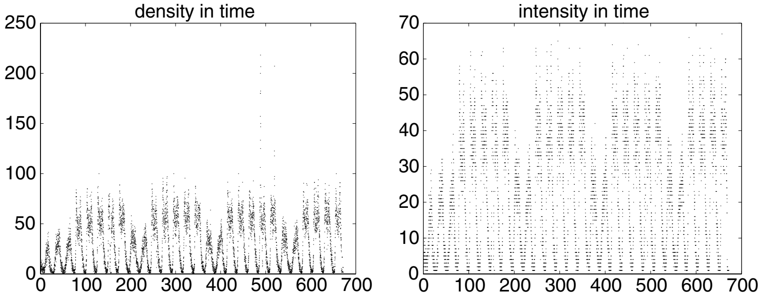
For the control, knowledge of traffic flow state incoming into the microregion in the near future is necessary. Reliable and possibly multistep prediction of the traffic flow can decide about practical success of such a feedback control.

Offline phase

Data

For experiments, 20-dimensional data records measured along two traffic lanes of the Strahov tunnel in Prague were used. The length of the tunnel is 2 km with two lanes in each direction. The detectors are placed after 500 m under each lane. Traffic in the tunnel is relatively fluent and almost no traffic congestions are observed.

Time course of the measured quantities on detectors reflects natural transportation periodicity. It is clearly seen on the Fig. 14.6, showing a typical time course of intensity and density of traffic flow for about 4 weeks.



Time course of density ρ

Time course of intensity q

Fig. 14.6. Traffic quantities from the middle of the tunnel

The daily periodicity is visible there. At night, the traffic intensity is very low. In the morning, the demands rapidly rise due to the cars of commuters and increased commercial transport. The intensity of the traffic reaches soon its maximum and a slight fall is observed around noon. Then, it rises again and the saturation lasts until the evening when the intensity starts to fall gradually to zero. The weekly periodicity connected with alternating work days and weekends is also strongly reflected in the data.

The available detectors provide 5-minutes samples of the traffic intensity and density. Data from 10 detectors are recorded giving 20-dimensional data record each 5 minutes, 288 vectors per day.

About 8 500 data records reflecting approximately 4 weeks period were used for learning and 300 data records corresponding to 1 day served for testing.

Quality marker

Quality of models used as predictors is judged using the relative standard deviation R_i of the prediction error $\hat{e}_{i;t} \equiv d_{i;t} - \hat{d}_{i;t}$ of the ith data entry $d_{i;t}$,

ith channel,

$$R_i \equiv \sqrt{\frac{\sum_{t \in t^*} \hat{e}_{i;t}^2}{\sum_{t \in t^*} (d_{i;t} - \bar{d}_i)^2}}, \quad i = 1, \ldots, \mathring{d} = 20, \; \bar{d}_i \equiv \text{sample mean of } d_{i;t} \; t \in t^*. \tag{14.9}$$

In (14.9), the point prediction $\hat{d}_{i;t}$ is an approximate conditional expectation of $d_{i;t}$.

Preprocessing of data

Experience shows that detectors are fault-prone and their reparation is far from being easy. Therefore the filtration of their data is necessary. Here, outlier filtration and normalization to zero mean and unit standard deviation were applied on the raw data.

Model for prediction

The traffic system switches among several very different states. A mixture model has been chosen as a proper tool for modelling of this effect.

The model is built in the learning offline phase and is used for prediction with another part of the data. During the learning, the joint pdf of the modelled data record is estimated. Then, the pdf predicting selected channel(s) conditioned on other measured data is created. It is used for point or interval predictions of the channels of interest.

Estimation of mixture models

Both static and dynamic normal mixture models were used.

Experiments with the static case inspected the contribution of multivariate modelling and of the switching among various components.

Auto-regressions of the first-order used in the dynamic case were found adequate for modelling and predicting 20-dimensional data records. Mixtures with components having higher order brought no improvements.

The initialization based on hierarchical factor splitting and implementing the normal version of Algorithm 6.8, [44], was used. It determined a finer structure of components and their number. Up to six components were found. In all cases, at least three of these components had non-negligible weights. Some components had almost zero weights but it is interesting that their omission decreased the prediction quality visibly. The rarely populated components described outlying data and consequently improved parameter estimates of the remaining components.

Online phase

The traffic application has not reached the implementation phase yet mainly because of economical reasons. Thus, testing of predictive capabilities of the learnt models on validation data was used as a substitution of their online use. It imitates well application of the models as fixed predictors.

The test were performed in the experimental and development environment provided by the Mixtools toolbox [180].

14.3.3 Experiments

The experiments verified the ability of the estimated mixture model to predict well the state of the traffic flow at the northern exit of the tunnel using the data along the whole tunnel.

Prediction quality of the inspected mixture model (MIX) was compared with the one-component mixture, i.e., with the auto-regression (AR) model. The comparison of results is used both for the static and dynamic case.

The choice of the best model variant was based on the corresponding v-likelihood. The function of the obtained predictors was quantified by the performance index R_i (14.9). Its dependence on the number of modelled channels $\mathring{d}$ and the number of prediction steps, are presented in Tables 14.4 and 14.5.

Static mixture estimation

Static normal components predict the future data by their offsets only. The prediction made with them shows the need for dynamic modelling and superiority of the mixture model in comparison to the corresponding AR model. This superiority stems from the data-dependent switching between components, i.e., between different offsets.

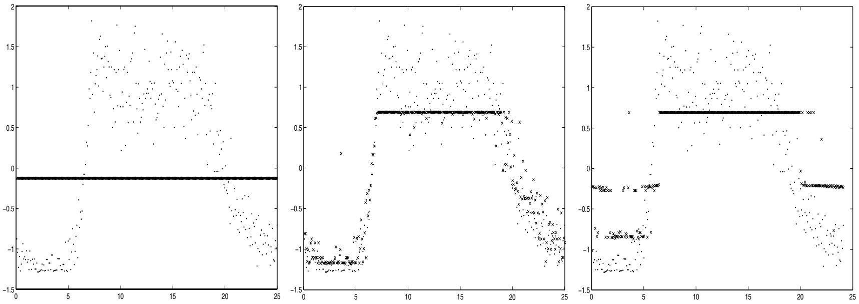
The two-dimensional model of intensity and density at the tunnel exit, 1st and 2nd channel, $\mathring{d} = 2$, was estimated only.

Table 14.4 shows prediction quality expressed by the marker $R_i, i = 1, 2$ (14.9) for two modelled channels and the prediction horizon ranging from 1 to 18 steps, which corresponds to the range from 5 to 90 minutes.

Table 14.4. Multistep ahead predictions with static two-dimensional model. R_i is the marker (14.9) for channels $i = 1, 2$ ($\mathring{d} = 2$). Results related to the mixture and auto-regression models are marked "MIX" and "AR", respectively.

	R_1		R_2	
Steps ahead	MIX	AR	MIX	AR
1 (5 min)	0.418	1.021	0.422	1.020
6 (30 min)	0.512	1.025	0.516	1.024
12 (60 min)	0.634	1.031	0.638	1.030
18 (90 min)	0.756	1.028	0.760	1.037

Figure 14.7 shows typical behavior of the compared autoregression "AR" and static mixture "MIX" models on one-step and 18-step ahead predictions.



1-step-ahead “AR” 1-step-ahead “MIX” 18-step-ahead “MIX”

Fig. 14.7. Prediction with auto-regression “AR” and static mixture “MIX” models. Dots denote measured data, × predictions. Data are normalized to zero mean and unit variance.

Dynamic mixture estimation

First-order normal components have been used. Mixtures with $\mathring{d} = 2$ and $\mathring{d} = 20$ were estimated, the latter was projected into the space of the channels 1 and 2.

Table 14.5 shows a comparison of the marker R_i, (14.9), for different number of modelled channels used for multiple-step predictor. The length of predictions covers the range from 5 to 90 minutes (one step means 5 minutes).

Table 14.5. Multistep predictions with dynamic models. $\mathring{d}$ — the number of modelled channels, R_i the prediction marker (14.9) for channels $i = 1, 2$. The results are related to the automatically initialized mixture model with four components found.

	R_1				R_2			
Steps ahead	1	6	12	18	1	6	12	18
$\mathring{d} = 2$	0.34	0.45	0.56	0.68	0.48	1.05	1.09	1.12
$\mathring{d} = 20$	0.32	0.41	0.54	0.67	0.86	0.91	0.93	0.96

14.3.4 Results

Comparison of auto-regression and mixture models

It can be seen, that the mixture models predict better, even though the signal lights are preset into the mode preventing congestions inside the Strahov tunnel and thus limiting other traffic modes. The mixture model also utilize better the information from the additional entries of data records.

Static models

The good property of mixtures is clearly seen on Fig. 14.7. The static AR model is able to predict only the mean value of the whole data sample. The projected mixture, due to the data-dependent weights of individual components, is able to follow even the ascending and descending trends of the data; see Section 7.1.2. With the increasing length of prediction the ability to follow data declines but the switching among components is clearly visible.

Dynamic models

They describe and thus predict data of this kind much better. For short-term predictions, projected high dimensional model does not dominate over the low-dimensional one. For longer-term predictions, advantage of the high-dimensional model is clearly visible as it includes more information about the traffic state.

14.3.5 Conclusions

In summary, mixtures are worth being considered for modelling and prediction of high-dimensional transportation data. Their ability to make reasonable multistep ahead predictions indicates their ability to grasp a significant portion of physical properties of the modelled systems. This is a necessary precondition for applicability of a model in a control design.

14.4 Conclusions on the applications

14.4.1 Lessons learned

The applicability of the advisory system based on mixture models has been tested on three real processes: rolling of metal strips, treatment of thyroid gland tumor and prediction of transportation data. All of them have a multimodal nature that makes the use of ordinary linear models inefficient.

The rolling mill application confirmed that the adopted methodological basis is sound and that the resulting algorithms can be applied with a measurable increase of efficiency of the managed system. At the same time, it has provided invaluable feedback to the reported research and clarified priorities for continuing research.

A specific trait seen on a nuclear medicine application is insufficiency of data given by the technical conditions of the treatment and data measurement. It makes the most visible the key application problem: the amount of information on the system carried by the data has to be sufficient. According to the *GIGO principle* (garbage in, garbage out), the quality of advices depends on how much information in the data represents the system behavior in its entirety. It is clear that lack of information in the data is crucial for successful estimation and advising.

This application has also confirmed the real need for a full coverage of mixed-mixtures modelling.

All cases, and especially the experiments with traffic data, confirmed that mixture models, in spite of necessary approximations in their estimation, are an excellent tool for description of multimodal systems.

14.4.2 Other application areas

The described applications are just samples from an extreme application domain of the described theory and algorithms. During the development of the advisory system we met, for instance, the following tasks that can use efficiently the research outcomes

- prediction of nonlinear phenomena like load in energy networks (electricity, gas, heat),
- modelling, prediction and evaluation of life-saving signals of prematurely born children,
- use of mixtures for speculative short-term monetary operations,
- support of operators of large furnaces for producing glass or metal,
- call for creating operation model to be used for training of operators of complex processes,
- attempt to use mixture models for fault detection and isolation [188],
- effort to use mixture models in complex social phenomena studied, for instance, in connection with research in electronic democracy [129],

15

Concluding remarks

This work has its roots in decades of research within the Department of Adaptive Systems where the authors are working. The preliminary version of this text arisen during solution of project IST-1999-12058 supported by European Commission as a theoretical and algorithmic basis of the project solution. It was substantially modified, improved, completed and refined in subsequent projects running within the department.

The IST project itself has been quite ambitious both with respect to its content, planned short period between elaboration of the theoretical solutions and the full scale use. The result is necessarily marked by the fact that we needed to collect or design theoretical solutions that can be converted in implementable algorithms. As such, it is unbalanced in its level, it is counter-pedagogical, contains many repetitions and surely some inconsistencies, various solutions have an ad hoc character, etc.

These problems may hide achievements we are proud of. First of all, our formulation of the design of the advisory systems is novel and represents an extension of the classical decision-making theory for a dynamic uncertain system. Consequently, it has the generic nature and an extreme applicability width. The following list underlines some points within this framework we feel as crucial achievements.

1. Academic, industrial and simultaneous advising were formulated and solved using approximate probabilistic optimization of advising.
2. Dynamic mixture modelling and prediction provide a novel and powerful decision-supporting tool.
3. Problem of the mixture estimation with mixed dynamic components has not been solved before to the extent presented here.
4. The designed estimation algorithms admit adaptive online clustering with an extreme potential for a permanent improvements of managing.
5. The basic initialization of the mixture estimation seems to be a novel tool applicable in high-dimensional data spaces.

It also solves – to a substantial extent – the structure estimation problem that is known to be crucial in the mixture estimation.

6. Both estimation and design for the basic normal factors, use efficient, numerically robust, factorized algorithms.
7. Claims above are supported by the diversity of the successful applications we have managed to implement within the project span; see Chapter 14.

We are aware that a lot of things are unfinished and some of them are missing. The crucial steps that should be done are as follows.

1. Offline prediction of closed loop performance of the advisory system should be done, in the vein of the project DESIGNER [115, 148, 189].
2. Filtering counterpart of the learning should be made. This would allow us to exploit process knowledge more deeply and to express management aims more directly.
3. The set of employed dynamic factors can be and should be significantly extended.
4. The models of the "rational" form – ratios of a mixture pair – should be addressed in depth in order to cope with the dynamic case more rigorously.
5. The number of ad hoc solutions should be decreased.
6. Improvements of various aspects of the proposed solutions should be considered.
7. Experience with real applications should be accumulated further on and reflected in the tool set we described here.

The above points can be summarized into the optimistic statement: **there are many interesting problems and an extensive unexploited application potential on the definitely promising way described in this text.**

On the margin of the proposed advisory system, it is worth stressing:

- The advisory system is coming in a proper time as users are gradually willing to work with sophisticated multivariate data processing.
- The described concept of the advisory system has no viable competitor that would be able to give optimized advices using dynamic, black-box, high-dimensional models. Its generic nature makes the resulting product adaptable to a surprisingly wide range of problems.
- The advisory system is expected to serve as a "clever" node in a more complex hierarchical support operator or their group. It cannot cover all tasks taken traditionally as operator support. It has to rely on an information system available as well as on the signals generated by other specialized modules.
- The advisory system has a specific niche on the "market" by providing relatively quickly an advanced addition to existing information systems. It will benefit by the expected speed of its implementation and its adaptive abilities.

- Abilities of the advisory system to reduce information overflow and focus operator attention on significant quantities are more important than originally expected.
- The theoretical and algorithmic outcomes of the research may well complement tool set for building specialized modules like fault-detectors [188].

References

1. M. Krstić, I. Kannellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*, Wiley-Interscience, New York, 1995.
2. B.D.O. Anderson and J.B. Moore, *Optimal Control : Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, New Jersey, 1989.
3. D.P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena Scientific, Nashua, US, 2001, 2nd edition.
4. B. Tamer (ed.), *Control Theory*, IEEE Press, New York, 2001.
5. T.H.Lee, C.C.Hang, and K.K.Tan, "Special issue in intelligent control for industrial application", *Journal of Adaptive Control and Signal Processing*, vol. 15, no. 8, 2001.
6. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.
7. G.F. Luger and W.A. Stubblefield, *Artificial Intelligence and the Design of Expert Systems*, chapter 8, pp. 296–297, Benjamin/Cummings, 1989.
8. B. S. Everitt and D. J. Hand, *Applied Multivariate Data Analysis*, Edward Arnold, London, 1991.
9. J. Girst, "The practical application of rule and knowledge-based system techniques to industrial process", in *Rule-Based Systems for Real-Time Planning and Control, IEE Colloquium*, October 1991.
10. S.R. Schmidt and R.G. Launsby, *Understanding Industrial Designed Experiments*, Air Academy Press, Colorado, 1992.
11. J.T.Kim, K.C. Kwon, I.K. Hwang, D.Y. Lee, W.M. Park, J.S. Kim, and S. Lee, "Development of advanced I & C in nuclear power plants: ADIOS and ASICS", *Nuc. Eng. Des.*, vol. 207, no. 1, pp. 105–119, 2001.
12. M. Setnes and R. Babuška, "Fuzzy decision support for the control of deteregent production", *Int. J. of Adaptive Control and Signal Processing*, vol. 15, no. 8, pp. 769–785, 2001.
13. C. Lindheim and K.M. Lien, "Operator support systems for new kinds of process operation work", *Comput. Chem. Eng.*, vol. 21, pp. S113–S118, 1997.
14. I. Ivanov, "Multivariate techniques for process monitoring and optimization at Tembec", *Pulp Pap. — Can.*, vol. 102, no. 7, pp. 23–25, 2001.
15. A. Mjaavatten and B.A. Foss, "A modular system for estimation and diagnosis", *Comput. Chem. Eng.*, vol. 21, no. 11, pp. 1203–1218, 1997.

16. Y.D. Pan, S.W. Sung, and J.H. Lee, "Data-based construction of feedback-corrected nonlinear prediction model using feedback neural networks", *Control Engineering Practice*, vol. 9, no. 8, pp. 859–867, 2001.
17. I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
18. A. Alessandri, T. Parisini, and R. Zoppoli, "Sliding-window neural state estimation in a power plant heater line", *International Journal of Adaptive Control and Signal Processing*, vol. 15, no. 8, pp. 815–836, 2001.
19. M. Malek, M.P. Toitgans J.L. Wybo, and M. Vincent, "An operator support system based on case-based reasoning for the plastic moulding injection process", in *Lecture Note in Artificial Intelligence*, vol. 1488, pp. 402–413. Springer-Verlag, New York, 2001.
20. A. Bonastre, R. Ors, and M. Peris, "Distributed expert systems as a new tool in analytical chemistry", *Trac–Trends Anal. Chem.*, vol. 20, no. 5, pp. 263–271, 2001.
21. G.A. Sundstrom, "Designing support contexts: Helping operators to generate and use knowledge", *Control Eng. Practice*, vol. 5, no. 3, pp. 375–381, 1997.
22. M.B. Perry, J.K. Spoerre, and T. Velasco, "Control chart pattern recognition using back propagation artificial neural networks", *Int. J. Prod. Res.*, vol. 39, no. 15, pp. 3399–3418, 2001.
23. C.W. Lu and M.R. Reynolds, "Cusum charts for monitoring an autocorrelated process", *J. Qual. Technol.*, vol. 33, no. 3, pp. 316–334, 2001.
24. F. Doymaz, J. Chen, J.A. Romagnoli, and A. Palazoglu, "A robust strategy for real-time process monitoring", *J. Process Control*, vol. 11, no. 4, pp. 343–359, 2001.
25. T. Bohlin, *Interactive System Identification: Prospects and Pitfalls*, Springer-Verlag, New York, 1991.
26. L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, London, 1987.
27. A. Fink, O. Nelles, M. Fisher, and R. Iserman, "Nonlinear adaptive control of a heater exchanger", *International Journal of Adaptive Control and Signal Processing*, vol. 15, no. 8, pp. 883–906, 2001.
28. B. Kuijpers and K. Dockx, "An intelligent man-machine dialogue system based on AI planning", *J. Appl. Intell.*, vol. 8, no. 3, pp. 235–245, 1998.
29. T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley, 1958.
30. K.A. Hoo, K.J. Tvarlapati, M.J. Piovoso, and R. Hajare, "A method of robust multivariate outlier replacement", *Comput. Chem. Eng.*, vol. 26, no. 1, pp. 17–39, 2002.
31. J.H. Chen and J.L. Liu, "Derivation of function space analysis based PCA control charts for batch process", *Chemical Engineering Science*, vol. 56, no. 10, pp. 3289–3304, May 2001.
32. M. Kano, S. Hasebe, I Hashimoto, and H. Ohno, "A new multivariate process monitoring method using principal component analysis", *Computers and Chemical Engineering*, vol. 25, no. 7-8, pp. 1103–1113, August 2001.
33. C. Rosen and J.A. Lennox, "Multivariate and multiscale monitoring of wastewater treatment operation", *Water Res.*, vol. 35, no. 14, pp. 3402–3410, 2001.
34. J.H. Chen and K.C. Liu, "Online batch process monitoring using dynamic PCA and dynamic PLS models", *Chem. Eng. Sci.*, vol. 57, no. 1, pp. 63–75, 2002.

35. S.J. Qin, S. Valle, and M.J. Piovoso, "On unifying multiblock analysis with application to decentralized process monitoring", *J. Cheometrics*, vol. 15, no. 9, pp. 715–742, 2001.
36. M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis", *Journal of the Royal Society Series B — Statistical Methodology*, vol. 61, pp. 611–622, 1999.
37. S. Kullback and R. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.
38. M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analyzers", *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.
39. S. Haykin, *"Neural Networks: A Comprehensive Foundation*, Macmillan, New York, 1994.
40. B. Lennox, G.A. Montague, A.M. Frith, C. Gent, and V. Bevan, "Industrial application of neural networks – an investigation", *J. of Process Control*, vol. 11, no. 5, pp. 497–507, 2001.
41. P.V. Varde, S. Sankar, and A.K. Verma, "An operator support system for research reactor operations and fault diagnosis through a framework of PSA knowledge based systems", *Reliabl. Eng. Syst. Safety*, vol. 60, no. 1, pp. 53–69, 1998.
42. J. Stutz and P. Cheeseman, "AutoClass - a Bayesian approach to classification", in *Maximum Entropy and Bayesian Methods*, J. Skilling and S. Sibisi, Eds. Kluwer, Dordrecht, 1995.
43. P. Paclík, J. Novovičová, P. Pudil, and P. Somol, "Road sign classification using Laplace kernel classifier", *Pattern Recognition Letters*, vol. 21, no. 13/14, pp. 1165–1174, 2000.
44. M. Kárný, P. Nedoma, I. Nagy, and M. Valečková, "Initial description of multimodal dynamic models", in *Artificial Neural Nets and Genetic Algorithms. Proceedings*, V. Kůrková, R. Neruda, M. Kárný, and N. C. Steele, Eds., Vienna, April 2001, pp. 398–401, Springer-Verlag.
45. S. Gourvenec, D.L. Massart, and D.N. Rutledge, "Determination of the number of components during mixture analysis using the Durbin-Watson criterion in the orthogonal projection approach and in the simple-to-use interactive self-modelling mixture analysis approach", *Chemometr Intell. Lab. Syst.*, vol. 61, no. 1–2, pp. 51–61, 2002.
46. A.M. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.
47. R.H. Elliot, L. Assoun, and J.B. Moore, *Hidden Markov Models*, Springer-Verlag, New York, 1995.
48. B.S. Everitt and D.J. Hand, *Finite Mixture Distributions*, Chapman and Hall, 1981.
49. D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixtures*, John Wiley, New York, 1985.
50. S. Richardson and P.J. Green, "On Bayesian analysis of mixtures with an unknown number of components, with discussion", *Journal of the Royal Statistical Society, Series B*, vol. 59, no. 4, pp. 731–792, 1997.
51. D.W. Scott, *Multivariate Density Estimation*, John Wiley, New York, 1992.
52. B.W. Silverman, *Density Estimation*, Chapman and Hall, 1991.
53. J. Coste, A. Spira, P. Ducimetiere, and J.B. Paolaggi, "Clinical and psychological diversity of non specific low-back pain. a new approach towards the

classification of clinical subgroups", *Journal of Clinical Epidemiol*, vol. 44, pp. 1233–1245, 1991.

54. R. Huth, I. Nemesova, and N. Klimperova, "Weather categorization based on the average linkage clustering technique: An application to European midlatitudes", *International Journal of Climatology*, vol. 13, 1993.
55. N. Jardine and R. Sibson, "The construction of hierarchic and nonhierarchic classification", *Computer Journal*, vol. 11, pp. 177–184, 1968.
56. D.W. Ginsberg and W.J. White, "Expert system developement using ellipsoidal-based clustering", *Minerals Engineering*, vol. 6, no. 1, pp. 31–40, 1993.
57. E.L. Sutanto, "Use of cluster analysis for the study of machine processes", in *Colloquium on Intelligent Manufacturing Systems*, London, December 1995, pp. 9/1–9/5.
58. E.L. Sutanto, "Intelligent reasoning for complex processes using multivariate cluster analysis", in *Proceedings of Computer Intensive Methods in Control and Data Processing CMP98*, Prague, September 1998, pp. 69–76.
59. R. Kulhavý and P.I. Ivanova, "Memory-based prediction in control and optimization", in *Proceedings of IFAC World Congress*, pp. 469–485. Pergamon, Oxford, 1999.
60. G. Bontempi, M. Birattari, and H. Bersini, "Lazy learning for local modelling and control design", *International Journal of Control*, vol. 72, no. 7-8, pp. 643–658, 1999.
61. M.H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.
62. J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, John Wiley, Chichester, 1997, 2nd ed.
63. M. Kárný, "Towards fully probabilistic control design", *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.
64. M. Kárný, J. Böhm, T.V. Guy, and P. Nedoma, "Mixture-based adaptive probabilistic control", *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 2, pp. 119–132, 2003.
65. M. Kárný, P. Nedoma, N. Khailova, and L. Pavelková, "Prior information in structure estimation", *IEE Proceedings — Control Theory and Applications*, vol. 150, no. 6, pp. 643–653, 2003.
66. P. Nedoma, M. Kárný, T.V. Guy, I. Nagy, and J. Böhm, *Mixtools (Program)*, ÚTIA AV ČR, Prague, 2003.
67. E.L. Sutanto, *Mean-tracking algorithm for multivariable cluster analysis in manufacturing processes*, PhD thesis, University of Reading, Reading, UK, 1996.
68. M. Kárný, A. Halousková, J. Böhm, R. Kulhavý, and P. Nedoma, "Design of linear quadratic adaptive control: Theory and algorithms for practice", *Kybernetika*, vol. 21, 1985, Supplement to Nos. 3, 4 ,5, 6.
69. V. Peterka, "Bayesian system identification", in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.
70. M. Kárný, I. Nagy, and J. Novovičová, "Mixed-data multimodelling for fault detection and isolation", *Adaptive control and signal processing*, , no. 1, pp. 61–83, 2002.
71. R.L. Keeny and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, J. Wiley, New York, 1978.
72. M.M. Rao, *Measure Theory and Integration*, John Wiley, New York, 1987.

73. V. Jarník, *Integral Calculus II*, Academia, Prague, 1984, (in Czech).
74. I. Vajda, *Information Theory and Statistical Decision Making*, Alfa, Bratislava, 1982, (in Slovak).
75. A.A. Feldbaum, "Theory of dual control", *Autom. Remote Control*, vol. 21, no. 9, 1960.
76. A.A. Feldbaum, "Theory of dual control", *Autom. Remote Control*, vol. 22, no. 2, 1961.
77. R. Bellman, *Introduction to the Mathematical Theory of Control Processes*, Academic Press, New York, 1967.
78. J. Šindelář and M. Kárný, "Dynamic decision making under uncertainty allows explicit solution", Tech. Rep. 1916, ÚTIA AVČR, POB 18, 18208 Prague 8, CR, 1997.
79. H. Kushner, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, New York, 1971.
80. I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Akadémiai Kiadó, Budapest, 1986.
81. M. Loeve, *Probability Theory*, van Nostrand, Princeton, New Jersey, 1962, Russian translation, Moscow 1962.
82. L. Berec and M. Kárný, "Identification of reality in Bayesian context", in *Computer-Intensive Methods in Control and Signal Processing: Curse of Dimensionality*, K. Warwick and M. Kárný, Eds., pp. 181–193. Birkhäuser, Boston, 1997.
83. I.N. Sanov, "On probability of large deviations of random variables", *Matematičeskij Sbornik*, vol. 42, pp. 11–44, 1957, (in Russian), translation in Selected Translations Mathematical Statistics and Probability, I, 1961, 213–244.
84. R. Kulhavý, *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*, Springer-Verlag, London, 1996.
85. R. Kulhavý and M. Kárný, "Tracking of slowly varying parameters by directional forgetting", in *Preprints of the 9th IFAC World Congress*, vol. X, pp. 178–183. IFAC, Budapest, 1984.
86. J.J. Milek and F.J. Kraus, "Time-varying stabilized forgetting for recursive least squares identification", in *IFAC Symposium ACASP'95*, Cs. Bányász, Ed., pp. 539–544. IFAC, Budapest, 1995.
87. L.Y. Cao and H. Schwartz, "A directional forgetting algorithm based on the decomposition of the information matrix", *Automatica*, vol. 36, no. 11, pp. 1725–1731, November 2000.
88. R. Kulhavý and M.B. Zarrop, "On the general concept of forgetting", *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.
89. E. Mosca, *Optimal, Predictive, and Adaptive Control*, Prentice Hall, 1994.
90. R.K. Mehra and D.G. Lainiotis (Eds.), *System Identification – Advances and Case Studies*, Pergamon Press, New York, 1976.
91. R. Koopman, "On distributions admitting a sufficient statistic", *Transactions of American Mathematical Society*, vol. 39, pp. 399, 1936.
92. E.F. Daum, "New exact nonlinear filters", in *Bayesian Analysis of Time Series and Dynamic Models*, J.C. Spall, Ed. Marcel Dekker, New York, 1988.
93. L. Berec, *Model Structure Identification: Global and Local Views. Bayesian Solution*, Ph.D. Thesis, Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Prague, 1998.

94. J. Rojíček and M. Kárný, “A sequential stopping rule for extensive simulations”, in *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*, J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, Eds., Praha, September 1998, pp. 145–150, ÚTIA AV ČR.
95. M. Kárný and R. Kulhavý, “Structure determination of regression-type models for adaptive prediction and control”, in *Bayesian Analysis of Time Series and Dynamic Models*, J.C. Spall, Ed. Marcel Dekker, New York, 1988, Chapter 12.
96. R. Kulhavý, “A Bayes-closed approximation of recursive nonlinear estimation”, *International Journal Adaptive Control and Signal Processing*, vol. 4, pp. 271–285, 1990.
97. R. Kulhavý, “Recursive Bayesian estimation under memory limitations”, *Kybernetika*, vol. 26, pp. 1–20, 1990.
98. R. Taylor, *Introduction to Functional Analysis*, Academia, Prague, 1973, (Czech translation).
99. V.S. Vladimirov, *Generalized Functions in Mathematical Physics*, Mir Publishers, Moscow, 1979.
100. K.J. Astrom and B. Wittenmark, *Adaptive Control*, Addison-Wesley, Reading, Massachusetts, 1989.
101. M. Kárný, “Adaptive systems: Local approximators?”, in *Preprints of the IFAC Workshop on Adaptive Systems in Control and Signal Processing*, Glasgow, August 1998, pp. 129–134, IFAC.
102. V. Peterka, “Adaptive digital regulation of noisy systems”, in *Preprints of the 2nd IFAC Symposium on Identification and Process Parameter Estimation.* ÚTIA ČSAV, Prague, 1970, paper 6.2.
103. K.J Astrom, *Introduction to Stochastic Control*, Academic Press, New York, 1970.
104. O.L.R. Jacobs and J.W. Patchell, “Caution and probing in stochastic control”, *International Journal of Control*, vol. 16, pp. 189–199, 1972.
105. V. Peterka, “Predictor-based self-tuning control”, *Automatica*, vol. 20, no. 1, pp. 39–50, 1984, reprinted in: *Adaptive Methods for Control System Design*, M.M. Gupta, Ed., IEEE Press, New York, 1986.
106. D.W. Clarke, C. Mohtadi, and P.S. Tuffs, “Generalized predictive control”, *Automatica*, vol. 23, no. 2, pp. 137–160, 1987.
107. D.W. Clarke, *Advances in Model-Based Predictive Control*, Oxford University Press, Oxford, 1994.
108. M. Cannon, B. Kouvaritakis, A. Brooms, and Y. Lee, “Efficient nonlinear model predictive control”, in *Proceedings of American Control Conference, June 28-30*, Chicago, 2000.
109. Y. Sheng, M. Tomizuka, and M. Ozaki, “Dynamic modelling and adaptive predictive control of drilling of composite materials”, in *Proceedings of American Control Conference, June 28-30*, Chicago, 2000.
110. P.R. Kumar, “A survey on some results in stochastic adaptive control”, *SIAM J. Control and Applications*, vol. 23, pp. 399–409, 1985.
111. N.M. Filatov and H.B. Unbehauen (Ed.), *Adaptive Dual Control - Theory and Applications*, Springer, Berlin, 2004.
112. R. Murray-Smith and T.A. Johansen, *Multiple Model Approaches to Modelling and Control*, Taylor & Francis, London, 1997.

113. M. Kárný and A. Halousková, "Preliminary tuning of self-tuners", in *Lecture Notes: Advanced Methods in Adaptive Control for Industrial Application (Joint UK-CS seminar)*, K. Warwick, M. Kárný, and A. Halousková, Eds., vol. 158. Springer-Verlag, 1991, held in May 1990, Prague.
114. M. Kárný and A. Halousková, "Pretuning of self-tuners", in *Advances in Model-Based Predictive Control*, D. Clarke, Ed., pp. 333–343. Oxford University Press, Oxford, 1994.
115. J. Bůcha, M. Kárný, P. Nedoma, J. Böhm, and J. Rojíček, "Designer 2000 project", in *International Conference on Control '98*, London, September 1998, pp. 1450–1455, IEE.
116. G. Belforte and P. Gay, "Optimal experiment design for regression polynomial models identification", *International Journal of Control*, vol. 75, no. 15, pp. 1178–1189, 2002.
117. M. Zarrop, *Experiment Design for Dynamic System Identification. Lecture Notes in Control and Information Sciences 21*, Springer, New York, 1979.
118. A.V. Oppenheim and A.S. Wilsky, *Signals and systems*, Englewood Clifts, Jersye, 1983.
119. R. Kruse and C. Borgelt, "Data mining with graphical models", *Lecture Notes In Computer Science*, vol. 2534, pp. 2–11, 2002.
120. M. Kárný, N. Khailova, P. Nedoma, and J. Böhm, "Quantification of prior information revised", *International Journal of Adaptive Control and Signal Processing*, vol. 15, no. 1, pp. 65–84, 2001.
121. M. Ishikawa and T. Moriyama, "Prediction of time series by a structural learning of neural networks", *Fuzzy Sets and Systems*, vol. 82, no. 2, pp. 167–176, September 1996.
122. M. Kárný, "Estimation of control period for selftuners", *Automatica*, vol. 27, no. 2, pp. 339–348, 1991, extended version of the paper presented at 11th IFAC World Congress, Tallinn.
123. L. Berec and J. Rojíček, "Control Period Selection: Verification on Coupled Tanks", in *Preprints of European Control Conference ECC'97* (on CD-ROM), G. Bastin and M. Gevers, Eds. ECC, Brussels, 1997.
124. M.E.P. Plutowski, "Survey: Cross-validation in theory and practice", Research report, Department of Computational Science Research, David Sarnoff Research Center, Princeton, New Jersey, 1996.
125. M. Novák and J. Böhm, "Adaptive LQG controller tuning", in *Proceedings of the 22nd IASTED International Conference on Modelling, Identification and Control*, M. H. Hamza, Ed., Calgary, February 2003, Acta Press.
126. M. Novák, J. Böhm, P. Nedoma, and L. Tesař, "Adaptive LQG controller tuning", *IEE Proceedings — Control Theory and Applications*, vol. 150, no. 6, pp. 655–665, 2003.
127. A. Wald, *Statistical Decision Functions*, John Wiley, New York, London, 1950.
128. M. Kárný, J. Kracík, I. Nagy, and P. Nedoma, "When has estimation reached a steady state? The Bayesian sequential test", *International Journal of Adaptive Control and Signal Processing*, vol. 19, no. 1, pp. 41–60, 2005.
129. M. Kárný and J. Kracík, "A normative probabilistic design of a fair governmental decision strategy", *Journal of Multi-Criteria Decision Analysis*, vol. 12, no. 2-3, pp. 1–15, 2004.
130. T.V. Guy, J. Böhm, and M. Kárný, "Probabilistic mixture control with multimodal target", in *Multiple Participant Decision Making*, J. Andrýsek,

M. Kárný, and J. Kracík, Eds., Adelaide, May 2004, pp. 89–98, Advanced Knowledge International.

131. J. Andrýsek, “Approximate recursive Bayesian estimation of dynamic probabilistic mixtures”, in *Multiple Participant Decision Making*, J. Andrýsek, M. Kárný, and J. Kracík, Eds., pp. 39–54. Advanced Knowledge International, Adelaide, 2004.
132. R. Horst and H. Tuy, *Global Optimization*, Springer, 1996, 727 pp.
133. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, 1996.
134. Tao Chen, Kai-Kuang Ma, and Li-Hui Chen, “Tristate median filter for image denoising”, *IEEE Transactions on Image Processing*, vol. 8, no. 12, pp. 1834–1838, December 1999.
135. T. Cipra, “Dynamic credibility with outliers”, *Applications of Mathematics*, vol. 41, no. 2, pp. 149–159, 1996.
136. M. Tanaka and T. Katayama, “A robust identification of a linear system with missing observations and outliers by the EM algorithm”, *Transactions of the Institute of Systems, Control and Information Engineering*, vol. 1, no. 4, pp. 117–129, September 1988.
137. L. Tesař and A. Quinn, “Detection and removal of outliers from multidimensional ar processes”, in *Proceedings of The Irish Signal and Systems Conference*, Maynooth, Ireland, August 2001, pp. 117–122, National University of Ireland, Maynooth College.
138. I. Nagy and M. Kárný, “A view on filtering of continuous data signals”, *Kybernetika*, vol. 28, no. 6, pp. 494–505, 1992.
139. T.V. Guy and M. Kárný, “On structure of local models for hybrid controllers”, in *Artificial Neural Nets and Genetic Algorithms. Proceedings*, V. Kůrková, R. Neruda, M. Kárný, and N. C. Steele, Eds., Vienna, April 2001, pp. 340–344, Springer-Verlag.
140. S.J. Godsill, *The Restoration of Degraded Audio Signals*, PhD thesis, University of Cambridge, Department of Engineering, December 1993.
141. L. Tesař and A. Quinn, “Method for artefact detection and supression using alpha-stable distributions”, in *Proceedings of ICANNGA Conference*, Prague, March 2001.
142. Kárný M., “Local filter robust with respect to outlying data”, Tech. Rep. 1644, ÚTIA ČSAV, Prague, 1990.
143. R.M. Rao and A.S. Bopardikar, *Wavelet Transforms: Introduction to Theory and Applications*, Addison,Wesley,Longman, July 1998.
144. L. He and M. Kárný, “Estimation and prediction with ARMMAX model: a mixture of ARMAX models with common ARX part”, *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 4, pp. 265–283, 2003.
145. V. Peterka, “Real-time parameter estimation and output prediction for ARMA-type system models”, *Kybernetika*, vol. 17, pp. 526–533, 1981.
146. V. Peterka, “Control of uncertain processes: applied theory and algorithms”, *Kybernetika*, vol. 22, 1986, Supplement to No. 3, 4 ,5, 6.
147. M. Kárný, “Quantification of prior knowledge about global characteristics of linear normal model”, *Kybernetika*, vol. 20, no. 5, pp. 376–385, 1984.
148. N. Khaylova, *Exploitation of Prior Knowledge in Adaptive Control Design*, PhD thesis, University of West Bohemia, Pilsen, Czech Republic, 2001.
149. B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, London, 1997.

150. B. Harris and G. Heinel, “The relation between statistical decision theory and approximation theory”, in *Optimizing Methods in Statistics*, J.G. Rustagi, Ed., pp. 263–272. Academic Press, New York, 1979.
151. J. Šindelář, P. Nedoma, and M. Kárný, “Algorithm for selecting the best variants in the shadow cancelling problem”, *Kybernetika*, 2004, submitted.
152. E.L. Sutanto and K. Warwick, “Cluster analysis: An intelligent system for the process industries”, in *Cybernetics and Systems '94*, Robert Trappl, Ed., Vienna, 1994, vol. 1, pp. 327–344, World Scientific.
153. E.L. Sutanto and K. Warwick, “Cluster analysis for multivariable process control”, in *Proceedings of the American Control Conference 1995*, Seattle, June 1995, vol. 1, pp. 749–750.
154. E.L. Sutanto, J.D. Mason, and K. Warwick, “Mean-tracking clustering algorithm for radial basis function centre selection”, *International journal of Control*, vol. 67, no. 6, pp. 961–977, August 1997.
155. M. Kárný, J. Kadlec, and E. L. Sutanto, “Quasi-Bayes estimation applied to normal mixture”, in *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*, J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, Eds., Prague, September 1998, pp. 77–82, ÚTIA AV ČR.
156. M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1972.
157. M. Valečková, M. Kárný, and E. L. Sutanto, “Bayesian M-T clustering for reduced parametrisation of Markov chains used for nonlinear adaptive elements”, in *Preprints of the IFAC Workshop on Adaptive Systems in Control and Signal Processing*, Glasgow, August 1998, pp. 381–386, IFAC.
158. V. Peterka, “Adaptation for LQG control design to engineering needs”, in *Lecture Notes: Advanced Methods in Adaptive Control for Industrial Application; Joint UK-CS seminar*, K. Warwick, M. Kárný, and A. Halousková, Eds., vol. 158. Springer-Verlag, New York, 1991, held in May 1990, Prague.
159. G.J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, Academic Press, New York, 1977.
160. M. Kárný, “Parametrization of multi-output multi-input autoregressive-regressive models for self-tuning control”, *Kybernetika*, vol. 28, no. 5, pp. 402–412, 1992.
161. V. Šmídl, A. Quinn, M. Kárný, and T.V. Guy, “Robust estimation of autoregressive processes using a mixture based filter bank”, *System & Control Letters*, , no. 4, pp. 315–324, 2005.
162. M. Kárný, “Algorithms for determining the model structure of a controlled system”, *Kybernetika*, vol. 19, no. 2, pp. 164–178, 1983.
163. T.V. Guy, M. Kárný, and J. Böhm, “Linear adaptive controller based on smoothing noisy data algorithm”, in *European Control Conference. ECC '99.* (CD-ROM), Karlsruhe, August 1999, VDI/VDE GMA.
164. A. Renyi, *Probability theory*, Academia, Prague, 1972, in Czech.
165. M. Valečková and M. Kárný, “Estimation of Markov chains with reduced parametrisation”, in *Preprints of the 2nd European IEEE Workshop on Computer-Intensive Methods in Control and Signal Processing, CMP'96*, L. Berec, J. Rojíček, M. Kárný, and K. Warwick, Eds., pp. 135–140. ÚTIA AV ČR, Prague, 1996.
166. T.S. Fergusson, “A Bayesian analysis of some nonparametric problems”, *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

167. R.M. Gray and D.I. Neuhoff, “Quantization”, *IEEE Transaction On Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.
168. R. Jiroušek, “Introduction into Bayesian network theory”, Tech. Rep. LISp-94-04, LISp, Department of Information and Knowledge Engineering, Prague University of Economics, 1994, in Czech.
169. F.V. Jensen, *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York, 2001.
170. M. Valečková, M. Kárný, and E.L. Sutanto, “Bayesian M-T clustering for reduced parameterisation of Markov chains used for nonlinear adaptive elements”, *Automatica*, vol. 37, no. 6, pp. 1071–1078, 2001.
171. H. Mine and S. Osaki, *Markovian Decision Processes*, Elsevier, New York, 1970.
172. M. Hendrikx, J. van Nunen, and J. Wessels, “On iterative optimization of structured Markov decision processes with discounted rewards”, *Math. Operationsforsch. und Statistics, ser. Optimization*, vol. 15, no. 3, pp. 439–459, 1984.
173. J. Šindelář, M. Kárný, and P. Nedoma, “Algorithms selecting several best “diagonal” variants”, Tech. Rep., ÚTIA AV ČR, 1998.
174. L. Tesař, M. Kárný, and J. Šindelář, “Bayesian stopping rule for a global maxima search in discrete parameter space”, *Kybernetika*, 2004, submitted.
175. S. Rabe-Hesketh and A. Skrondal, “Parameterization of multivariate effects models for categorical data”, *Biometrics*, vol. 57, no. 4, pp. 1256–1263, 2001.
176. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics. Springer, New York, 2001.
177. A. Quinn, P. Ettler, L. Jirsa, I. Nagy, and P. Nedoma, “Probabilistic advisory systems for data-intensive applications”, *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 2, pp. 133–148, 2003.
178. P. Ettler and F. Jirkovský, “Digital controllers for škoda rolling mills”, in *Lecture Notes: Advanced Methods in Adaptive Control for Industrial Application (Joint UK-CS seminar)*, K. Warwick, M. Kárný, and A. Halousková, Eds., vol. 158, pp. 31–35. Springer Verlag, 1991.
179. Samuel Kotz and Norman L. Johnson, *Process Capability Indices*, Chapman & Hall, London, 1993.
180. P. Nedoma, M. Kárný, I. Nagy, and M. Valečková, “Mixtools. MATLAB Toolbox for Mixtures”, Tech. Rep. 1995, ÚTIA AV ČR, Prague, 2000.
181. J. Harbert, W. Eckelman, and R. Neumann, *Nuclear Medicine. Diagnosis and Therapy*, Thieme, New York, 1996.
182. L. Jirsa, *Advanced Bayesian Processing of Clinical Data in Nuclear Medicine*, FJFI ČVUT, Prague, 1999, 95 pages, Ph.D. Thesis.
183. J. Heřmanská, K. Vošmiková, L. Jirsa, M. Kárný, and M. Šámal, “Biophysical inputs into the software MIRDose”, *Sborník lékařský*, vol. 99, no. 4, pp. 521–527, 1998.
184. J. Heřmanská, M. Kárný, J. Zimák, L. Jirsa, and M. Šámal, “Improved prediction of therapeutic absorbed doses of radioiodine in the treatment of thyroid carcinoma”, *The Journal of Nuclear Medicine*, vol. 42, no. 7, pp. 1084–1090, 2001.
185. J. Zimák, J. Heřmanská, N. Medić, L. Jirsa, M. Kárný, M. Šámal, and L. Kárná, “Guide to *Iodine III*”, Tech. Rep., Clinic of Nuclear Medicine, Motol Hospital, Prague, 2001, 28 pp.

186. D. Helbing, “Traffic data and their implications for consistent traffic flow modelling”, in *Transportation Systems*. 1997, pp. 809–814, Chania.

187. F. McLeod, N. Hounsell, and Saeed Ishtia, “Traffic data and their implications for consistent traffic flow modelling”, in *Transportation Systems*. 1997, pp. 911–916, Chania.

188. A. Rakar, T.V. Guy, P. Nedoma, M Kárný, and D. Juričić, “Advisory system prodactool: Case study on gas conditioning unit”, *Journal of Adaptive Control and Signal Processing*, 2004, submitted.

189. M. Kárný, “Tools for computer-aided design of adaptive controllers”, *IEE Proceedings — Control Theory and Applications*, vol. 150, no. 6, pp. 642, 2003.

Index